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

Ph.D. thesis

Marta Szachniuk

Promotor: prof. dr hab. inż. Jacek Błażewicz

Poznan University of Technology
Institute of Computing Science

Poznań, 2005

I would like to thank

prof. Jacek Błażewicz for brilliant ideas and hours of discussion,

dr Mariusz Popenda and Łukasz Popenda for excellent cooperation, time and biophysical support,

Igor for love, forbearance and being my private encyclopedia of everything,

my family for acceptance of all my life choices.

TABLE OF CONTENTS

I. Introduction 6

I.1 RNA structural analysis as bioinformatics problem 6

I.2 Scope of the thesis 9

II. Notions and definitions 11

II.1 RNA roles and structure 11

II.2 Structural analysis of biopolymers 16

II.3 Nuclear magnetic resonance (NMR) spectroscopy 21

II.4 Computational complexity of combinatorial problems 28

III. Basic concepts of RNA structural analysis in solution with the use of 2D NMR techniques 38

III.1 Strategy of RNA structure determination with NMR 38

III.2 NOE assignment in structure determination 43

III.3 Existing methods of NOE assignment 50

IV. New model of the NOE path construction problem 52

V. Computational complexity analysis of the NOE assignment problem 56

V.1 Complexity of the problem in the theoretical model 56

V.2 Complexity of the problem in the experimental model 63

VI. Algorithms for NOE pathway construction 67

VI.1 Input data and its preprocessing 67

VI.2 Feasibility and optimization criteria 71

VI.3 Enumerative algorithm 74

VI.4 Tabu search algorithm 80

VI.5 Evolutionary algorithm 83

VII. Computational experiments 89

VII.1 Overview of experimental data set 89

VII.2 Experimental results 96

VIII. Conclusions 104

Index of basic notions 106

Bibliography 109

TABLE OF FIGURES

Figure 2.1.1. An exemplary nucleotide structure. 12
Figure 2.1.2. Sugars: deoxyribose (a) and ribose (b). 12
Figure 2.1.3. Adenine (a), Guanine (b), Cytosine (c), Thymine (d) and Uracil (e). 13
Figure 2.1.4. DNA double helix. 14
Figure 2.2.1. Standard numbering convention in RNA nucleotides. 17
Figure 2.2.2. Selected motifs of RNA secondary structure: duplex (a), hairpin loop (b), internal loop (c), bulge loops (d), (e), junction (f) and pseudoknot (g). 18
Figure 2.2.3. Examples of tertiary structures: RNA duplex with mismatches r(GGACUUCGGUCC) (a), A-RNA duplex (b), hammerhead ribozyme (c) and tRNA (d) (Bourne and Weissig, 2003). 20
Figure 2.3.1. NMR spectrometer schema (Perkel, 2004). 24
Figure 2.3.2. One- (a) and two-dimensional (b) 1H NMR spectrum for 2'-OMe$(CGCGCG)_2$ (Popenda, 1998). 25
Figure 2.3.3. NMR spectrum of r$(CGCGCG)_2$: line charts (a), (b), contour diagram (c), spatial chart (d) (Popenda, 1998) 26
Figure 3.2.1. An exemplary 2D−NOESY spectrum r$(CGCGCG)_2$ in D_2O with bounded regions (Popenda, 1998). 44
Figure 3.2.2. Main NOE interactions in r(CGUA) (Popenda, 1998). 46
Figure 3.2.3. NOE pathway drafted in region [5−6]×[7−8] of 2D−NOESY spectrum of r$(CGCGCG)_2$ (Popenda et al., 1997). 47
Figure 4.1. The relationship between aromatic/anomeric region of the NOESY spectrum of r$(CGCGCG)_2$ (a) and the corresponding NOESY graph (b). 53
Figure 4.2. Relationship between NOE pathway (a) in the spectrum of r$(CGCGCG)_2$ and the corresponding NOE path (b). 55
Figure 5.1.1. A square subgraph. 59
Figure 5.1.2. Transformation procedure: (a) input graph, (b)-(e) succeeding steps of a transformation to a NOESY graph (e) 60
Figure 6.1.1. An exemplary input file rcgcgcg.list with spectral data. 68
Figure 6.1.2. An exemplary input file rcgcgcg.inf with supplemental data. 70
Figure 6.3.1. NOE pathway found in the spectrum of r$(CGCGCG)_2$. 79
Figure 6.3.2. Assignment file for r$(CGCGCG)_2$. 80

Figure 7.1.1. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for r$(CGCGCG)_2$ and the NOE path (b). 91
Figure 7.1.2. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for 2'−OMe$(CGCGCG)_2$ and the NOE path (b). 91
Figure 7.1.3. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for d$(GACTAGTC)_2$ and the NOE path (b). 92
Figure 7.1.4. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for r$(GGCAGGCC)_2$ and the NOE path (b). 92
Figure 7.1.5. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for r$(GAGGUCUC)_2$ and the NOE path (b). 93
Figure 7.1.6. [7−8]×[5−6] region of 2D−NOESY spectrum (a) for r$(GGCGAGCC)_2$ and he NOE path (b)..... 93
Figure 7.1.7. [5−6]×[7−8] region of 2D−NOESY spectrum (a) for r$(GGAGUUCC)_2$ and the NOE path (b). 94
Figure 7.2.1. Precisions of optimal solutions found by tabu search. 101
Figure 7.2.2. Precisions of optimal solutions found by evolutionary algorithm for p=250 (a), p=500 (b), p=750 (c), p=1000 (d). 101
Figure 7.2.3. Precisions of optimal solutions found by TS and EA for test T1. 102
Figure 7.2.4. Precisions of optimal solutions found by TS and EA for test T2. 102

I. INTRODUCTION

I.1 RNA STRUCTURAL ANALYSIS AS BIOINFORMATICS PROBLEM

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.

Inventing the first computers in late 30-ties of the 20th century started the process of introducing calculation into every field of life. About twenty years later, the first biochemical theories, based on computation, started to appear: Margaret Dayhoff's atlas of protein sequences, Needleman-Wunsch algorithm searching for similarities in protein sequences, DNA sequencing software, Smith-Waterman algorithm for identification of common molecular subsequences etc. (Setubal and Meidanis, 1997; Kanehisa, 2000; Lesk, 2002). Thus, biology, computer science and information technology merged into a single discipline called bioinformatics. The ultimate goal of the field has been to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology could be discerned. There have been three important sub-disciplines within bioinformatics: the development of new algorithms and statistics for associating relationships among members of large data sets, the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures, and, finally, the development and implementation of tools that allow for an efficient access and management of different types of information (Lesk, 2002). Structural analysis, as one of bioinformatics main sub-disciplines, started to evolve those years. The development of new analytical methods rapidly sped up and the scope of experiments performed was extended. Owing to this, many molecular structures have been solved and modified, and the new sub-disciplines of medicine, biochemistry, genetics and biotechnology have emerged.

Structural analysis of proteins and nucleic acids contributes to clarifying their biological functions and structure related properties in materials, components, liquid crystals, molecular electronics, drugs, pesticides and polymers, aiding identification of new diseases, raising new specimens of plants and animals, aiding progress in drug design as well as other synthetic methods and mechanisms, cataloguing new compounds, controlling compound quality etc (Williams and Fleming, 1996). Initially, the research in that area

concentrated on proteins and deoxyribonucleic acid (DNA). However, knowledge gained when studying these molecules appeared insufficient to answer all the questions that have been posed 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. Recognition of the importance of RNA in many biological processes has increased dramatically in recent years. The discoveries of non-coding regulatory RNA and RNA interference have involved a broad line of disciplines. The discovery of catalytic activity by RNA (Cech, 1990; Cech, 1993) has stimulated speculation that life may have originated through the formation and evolution of RNA molecules. This hypothesis, dubbed the "RNA World" (Gilbert, 1986), has invalidated the apparent monopoly of proteins as biological catalysts and changed the view on their role in the early stages of evolution.

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 structure has been less dynamic than the spread of processing protein and DNA structures.

Following the most common view (Guntert, 1998; Neidle, 1999; Westhof and Auffinger, 2000; Neidle, 2002), the subjects of RNA structural analysis are: its primary structure determined by the sequence of nucleoside monophosphates in the chain, the secondary structure describing one- and two-strand fragments as well as the formation of loops or helices and the tertiary structure, which characterizes the three-dimensional shape of the entire chain. For proteins, the quaternary structure, which describes interactions between polymer subunits is also analyzed. *The thesis discusses and solves key problems of the tertiary structure analysis of RNA molecules.*

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. However, only X–ray crystallography and NMR spectroscopy give the complete information about the structure at this level of study. Recent years, yielded a quick spread of NMR spectroscopy, which has been now a well established method for structure determination of biomolecules in solution (Gunther, 1996; Hilbers and Wijmenga, 1996; Williams and Fleming, 1996). A suggestion of using NMR spectroscopy as a choice for probing the structure and dynamics of RNA molecules has resulted from two reasons: the

difficulty in crystallization of RNA molecules caused by their conformational flexibility, and the interest in the relationship between dynamic behavior of RNA and its biological function in solution.

Tertiary structure determination 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 (Cavanach et al., 1996) and nucleic acids (Varani and Tinoco, 1991; Wijmenga and van Buuren, 1998). 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. Extracting the structural information from the NMR spectrum is an art itself. The procedure assigning 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- (2D) and three-dimensional (3D) spectra resulting from NMR experiments performed for the determination of tertiary structures. For short DNA and RNA duplexes the assignment is performed manually in accordance with the experimenter's knowledge and intuition. However, the larger the molecules we deal with, the more difficult the task becomes. *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 elucidation of protein structures (Moseley and Montelione, 1999). Several programs exist which automatize the process of their assignment (Lukin et al., 1997; Zimmerman et al., 1997; Leutner et al., 1998; Atreya et al., 2000; Moseley et al., 2001; Linge et al., 2003). 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. Only one proposition of automatic assignment method dedicated exactly for RNA chains has appeared (Roggenbuck et al., 1990). It has been based on the Reduced Adjacency Matrix and Backtracking procedures and has been applied for one RNA octamer duplex. However, the method has not been developed and its application is very limited.

I.2 SCOPE OF THE THESIS

Following the above description (Section 1.1) of the state of the art in the considered field, 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 generate 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* (Szachniuk et al., 2003; Szachniuk et al., 2004). 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.

An organization of the thesis is as follows. Chapter 2 defines basic notions of the disciplines involved in the thesis subject. It outlines the biological basis of RNA structure and functions, describes the main aspects of structural analysis of biopolymers, introduces to a nuclear magnetic resonance spectroscopy being a tool of structural analysis, and discusses main ideas of computational complexity theory, which is a starting point of the proposed combinatorial model. RNA acts in close relation to proteins and DNA, thus, these polymers roles are also referred. Chapter 3 characterizes a concept of RNA structural analysis in solution used in the context of the two-dimensional NOESY technique.

In Chapter 4, the new model of NOE path construction problem is proposed. The combinatorial complexity of the problem in question is discussed in Chapter 5. Chapter 6 contains the detailed specification of three new algorithms based on enumerative, tabu and genetic procedures, which automatically generate NOE paths on the basis of spectral data obtained from NMR experiments for RNA molecules. Experimental data as well as the results of computational experiments performed with the use of all proposed algorithms are presented in Chapter 7. Chapter 8 sums up the results of application of the enumerative, tabu and evolutionary approaches to the problem of NOE paths assignment and points out the directions for further research.

II.NOTIONS AND DEFINITIONS

The problem of automatic generation of NOE pathways belongs to interdisciplinary problems, which demand a lot of knowledge arising from different fields of science. To understand the nature and sense of it one should plumb into arcana of ribonucleic acid biology, cognize the principles of structural analysis of biopolymer molecules, sense the phenomenon of nuclear magnetic resonance, and be familiar with the theory of computational complexity as well as mathematics and combinatorial problems. The following sections explain the basis of all the mentioned branches of science and lead the reader to the appropriate starting point of the NOE pathway problem.

II.1 RNA ROLES AND STRUCTURE

In this section, the basics of biomolecule structure and roles are described. For years, biomolecules functions as well as their composition have been in the center of interest of the researchers representing different fields of science. Recognition of these functions is one of the main goals of structural study. The thesis' focus is set to some aspects of ribonucleic acid structure, however it cannot be discussed separately from other biomolecules, which play crucial roles in all the processes in flora and fauna world. Thus, some paragraphs introduce the theme of proteins and deoxyribonucleic acid, while the others magnify the subject of RNA functioning.

Life at its simplest can be pictured with the use of three basic notions: proteins, DNA and RNA, which build the molecular life model. All of these molecules belong to the family of biopolymers and compose the living cells of organisms as well as viruses. *Biopolymer* is defined as a *polymer* found in the nature, i.e. a long, repeating chain of atoms, formed through the linkage of many identical small molecules called *monomers*. *Amino acids* are natural monomers, which polymerize to form proteins, while *nucleic acids* compose deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) (Stryer, 2000; Bourne and Weissig, 2003). A nucleic acid chain, called also a *strand*, appears most often in single or double form. Typically, DNA molecule contains two strands whereas RNA molecules can be single- or double-stranded.

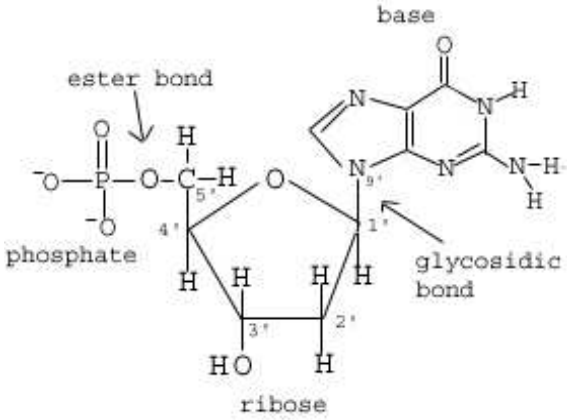


Figure 2.1.1. An exemplary nucleotide structure.

The strand is composed of covalently-bond *nucleotides* (Figure 2.1.1), i.e. organic molecules consisting of a heterocyclic base (a purine or a pyrimidine), a sugar (deoxyribose in DNA or ribose in RNA), and a phosphate group (Stryer, 2000). Figure 2.1.2 shows structures of the sugars.

Figure 2.1.2. Sugars: deoxyribose (a) and ribose (b).

The length of a nucleic chain is represented by the number of monomer units, i.e. building blocks of DNA and RNA. There are five different monomers: adenosine, guanosine, cytidine, thymidine and uridine, which contain the following *bases* (respectively): adenine and guanine (which are purines), cytosine, thymine, and uracil (which are pyrimidines). Uracil is found in RNA only, while thymine replaces it in DNA. Commonly, the nucleotides in nucleic acid strand are represented as letters related to bases – an adenosine nucleotide is abbreviated as A, thymidine as T, guanosine as G, cytidine as C, and uridine as U. Figure 2.1.3 shows structural models of the bases.

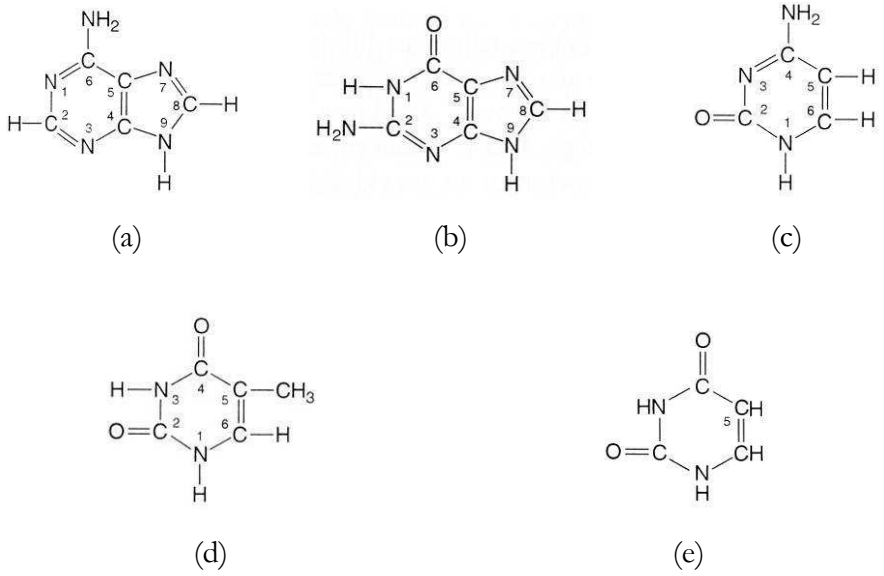


Figure 2.1.3. Adenine (a), Guanine (b), Cytosine (c), Thymine (d) and Uracil (e).

In case of a double-stranded nucleic acid, bases are paired between two strands. Therefore, its length is given by the number of base pairs. *Oligonucleotides* refer to short nucleic acid chains (< 50 bases or base pairs) and *polynucleotides* have longer chains (Neidle, 1999; Stryer, 2000).

The following paragraphs present the basic information about proteins, DNA and RNA roles in the essential processes of the living organisms.

Proteins play a crucial role in almost all biological processes and can be viewed as the working horses of living cells. They are chemicals that make up cell and organ structure, carry out reactions throughout the body and catalyze reactions (Kanehisa, 2000; Stryer, 2000).

DNA acts as an information repository of all the cells in organism. It carries genetic instructions for the biological development of cellular forms of life. It is sometimes referred to as the molecule of heredity as it is inherited and used to propagate traits. During reproduction, DNA is replicated and transmitted to the offspring. DNA molecule is very stable. It is composed of a long linear strand of millions of nucleotides, and is most often found paired with a partner strand. In a DNA molecule proposed by Watson and Crick, adenine is paired with thymine and they form two hydrogen bonds, while guanine is paired with cytosine and they form three hydrogen bonds (Watson and Crick, 1953). In less typical structures, other base pairs (e.g. (G:T) and (C:T)) may also form hydrogen bonds, although their strengths are not as strong as in case of (C:G) and (A:T) pairing (Stryer, 2000; Bourne and Weissig, 2003). Due to this specific base pairing, DNA's two

strands are complementary to each other. These strands wrap around each other in the double helix (Figure 2.1.4).

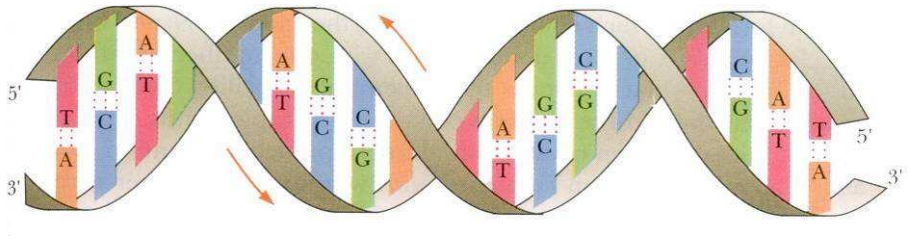


Figure 2.1.4. DNA double helix.

In the family of genetic material, RNA has long been the poor cousin of DNA. DNA makes up the genes, the master instructions of life, while RNA merely conveys those instructions to other parts of the cell. But surprising new discoveries (Gilbert, 1986; Cech, 1990; Cech, 1993; Stryer, 2000) have shown that cells contain an army of RNA snippets that do much more that just act as DNA's messenger. The discoveries have turned scientists' attention to more detailed examination of RNA roles and structure.

Structurally, *RNA* is similar to DNA, except for the critical presence of an additional hydroxyl group attached to each pentose ring at 2' position (i.e. replacing deoxyribose with ribose), as well as by the use of uracil instead of thymine. The other major difference between RNA and DNA is that RNA often posseses single-stranded fragments (Neidle, 1999; Stryer, 2000). However, RNA molecules often fold into more complex structures by making use of complementary internal sequences so that if one part of a single strand is complementary to another part of the same strand, these two parts bind together. This allows for the formation of hairpin loops, coils, etc., which then direct the formation of higher-order structures.

RNA can assume a wide variety of conformations in the execution of its biological roles. Moreover, its importance as a therapeutic target is widely recognized. Biological roles of RNA respond to its three types (Popenda, 1998; Neidle, 1999; Stryer, 2000; Neidle, 2002):

- **rRNA** – ribosomal RNA is a component of the ribosomes, the protein synthetic factories in the cell, and makes about 80% of all RNA in the cell;
- **mRNA** – messenger RNA is transcribed directly from DNA of a gene, next exported from the nucleus, carrying instructions to the cytoplasm, finally, bound to ribosomes and translated into the encoded protein;

- **tRNA** – transfer RNA is a short chain (74–93 nucleotides), that attaches the correct amino acid to the protein chain being synthesized at the ribosome of the cell, according to directions coded in the mRNA;

Ribosomal and transfer RNAs are examples of *RNA genes* (sometimes referred to as *non-coding RNA*), which encode functional RNA molecules and, in contrast to messenger RNA, do not code for proteins. Both forms participate in the process of *translation*, called also *protein biosynthesis* (Stryer, 2000). Since the late 1990s, many new RNA genes have been found, which raised the speculation of their much more significant role than previously thought.

Transcription, i.e. process of copying DNA to RNA by an enzyme called RNA polymerase, as well as translation participate in the processes of *gene expression* – the central dogma of molecular biology. In the process, enzymes scan down a messenger RNA copied from the DNA, and use ribosomes to build proteins based on the code that is read. In brief, DNA is transcribed into mRNA which is translated into proteins (Stryer, 2000).

One of the most significant and promising discoveries in modern molecular biology has been gene silencing mechanism. Double-stranded RNA (*dsRNA*), an origin of the discovery, has become a subject of the interest. dsRNA is composed of two complementary strands, similar to the DNA. It forms the genetic material of some kinds of viruses (*retroviruses*, i.e. viruses, which has a genome consisting of RNA) and is involved in some cellular processes (Caplen et al., 2001). Scientists have found that tiny fragments of RNA with two strands instead of the usual one can be used to shut off specific genes. The technique, known as RNA interference (*RNAi*), is based on the ability of dsRNA to suppress the expression of a gene corresponding to its own sequence (Caplen et al., 2001). Over the last few years, RNAi techniques have been widely used to discover the functions of genes by turning them off and seeing what happens to the plant or animal.

In the past two decades, the apparent monopoly of proteins as biological catalysts has been invalidated by the discovery that several RNA molecules can form active sites that catalyze chemical reactions. Since RNA also serves as a carrier of information, it seemed reasonable to suggest that ancient RNA molecules might have acted as starting point for the origin of life.

For many years there has been a general dissatisfaction with the protein hypothesis of the origin of life. Proteins cannot replicate themselves, which makes them unsuitable as a starting point for the development of life. However, there seemed to be no alternative

available until the discovery of RNA catalytic capability (Cech, 1993). The newer hypothesis, stating that RNA was, before the emergence of the first cell, the dominant, and probably the only form of life, has been dubbed the *RNA World* (Gilbert 1986). The general idea that evolution based on replication preceded protein synthesis was first proposed in late 1960s, following the elucidation of the genetic code. Next, it was reinitiated after the discovery of ribozymes and culminated after finding RNA catalytic capability, but yet it remains just the hypothesis.

The above paragraphs describe the variety of RNA roles and their importance in many biological processes. These processes can be traced on the basis of the data obtained from structural analysis, one of the most crucial sources of such knowledge acquisition. The following section is devoted to an introduction of the main currents of structural analysis of RNA, as well as DNA and proteins, as the general aspects are the same for the whole group of biopolymers.

II.2 STRUCTURAL ANALYSIS OF BIOPOLYMERS

This section contains an introduction to an extensive theme of structural analysis of RNA, DNA and proteins. Structural studies of proteins and nucleic acids are critical for understanding biological processes at the molecular level. They are a subject of structural biology, which aims to determine the structure of biological macromolecules, in particular proteins and nucleic acids, and explore what causes them to assume the specific conformations. Macromolecules carry out most of the functions of a cell. Moreover, it is the structure of molecules, especially their folding into a specific three-dimensional shape, which determines the molecule ability to perform most of its functions. This spatial shape, called the tertiary structure of a molecule, depends in a complicated way on the molecule's basic composition, i.e. its primary structure. Some aspects of the tertiary structure recognition are considered in the thesis. However, knowing the primary structure of a polymer often does not help either to deduce its spatial shape or to predict localized structuring, such as, for example, the formation of loops or helices in nucleic acids, or alpha helices and beta sheets in proteins, called secondary structure. In general, the subject of the structural analysis of biopolymers are primary, secondary, tertiary and quaternary structures and these four levels of structural organization are explained in the following paragraphs.

The exact chemical composition of a polymer and the sequence in which its units are arranged are called the *primary structure*, which is the first level of a polymer architecture.

Specification of the primary structure of a single-stranded biopolymer, such as a molecule of DNA, RNA or protein, means naming the species of every subunit in order from the beginning to the end of the strand. Thus, the convention for a protein is to list its constituent amino acid residues as they occur from the amino terminus (N-end) to the carboxylic acid terminus (C-end). The three-letter or single abbreviations code the amino acid residues: Ala (A), Cys (C), Asp (D), Glu (E), Phe (F), Gly (G), His (H), Ile (I), Lys (K), Leu (L), Met (M), Asn (N), Pro (P), Gln (Q), Arg (R), Ser (S), Thr (T), Val (V), Trp (W), Tyr (Y) (Stryer, 2000). The convention for a nucleic acid sequence is to list the nucleotides as they occur from the 5' end to the 3' end of the polymer chain. The 5' and 3' refer to the numbering of carbons around the ribose, which have the terminal hydroxyl groups attached. The three-letters or single abbreviations are used to code nucleotide residues: ADE (A), CYT (C), GUA (G), URI (U), THY (T) (Kanehisa, 2000; Stryer, 2000).

Figure 2.2.1 shows RNA nucleotides (A, C, G, U) with standard numbering convention. All of the nucleotides are shown in the *anti* conformation, which is typical for nucleotides with standard Watson-Crick base pairing. In more unusual cases, the nucleotides occur in the *syn* conformation, which is energetically less privileged. Conformations *syn* and *anti* determine a position of heterocyclic base with a relation to the sugar ring (Hilbers and Wijmenga, 1996; Popenda, 1998).

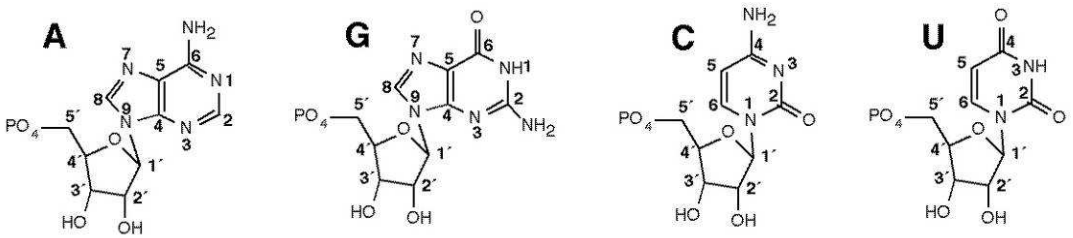


Figure 2.2.1. Standard numbering convention in RNA nucleotides.

Primary structure of a single-stranded polymer is determined in the process of *sequencing* (Kanehisa, 2000; Lesk, 2002). There are many various biophysical techniques for determining sequence information (Williams and Fleming, 1996; Blazewicz et al., 1997; Blazewicz et al., 1999; Kanehisa, 2000; Blazewicz et al., 2004c). Sequencing procedure results in a symbolic linear depiction known as a sequence which succinctly summarizes much of the atomic-level structure of the sequenced molecule.

The *secondary structure* generally reflects how individual residues in a biopolymer are connected to each other. It does not, however, refer to their current position in three-dimensional space, as the current positions are considered to be tertiary structure.

Secondary structure of a protein includes such structural motifs as alpha helices, beta sheets, and turns, or segments that completely lack secondary structure called random coils (Stryer, 2000). RNA secondary structure is generally divided into helices and various kinds of loops, and defines which nucleotides form hydrogen bonds. RNA secondary structure can also include pseudoknots and base triples (Tabaska et al., 1998; Ieong et al., 2003; Ruan et al., 2004). For many RNA molecules, the secondary structure is highly important to the correct functioning of the RNA, often, more important than the current sequence, since it defines active places within the chain (Dandekar, 2002). RNA secondary structure can be predicted with some accuracy by computer procedures. Various methods of secondary structure analysis exist and are applied to protein and nucleic acids structure detection and prediction (Jacobson and Zuker, 1993; Gulko and Haussler, 1996; Juan and Wilson, 1999; Luck et al., 1999; Akmaev et al., 2000; Ieong et al., 2003; Ruan et al., 2004). Figure 2.2.2 illustrates some motifs appearing as the RNA secondary structures.

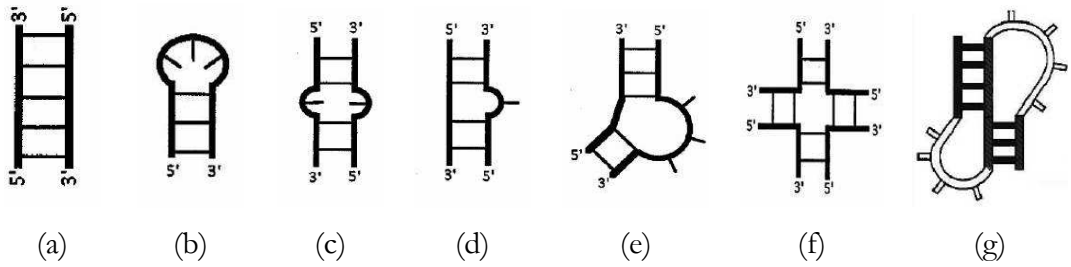


Figure 2.2.2. Selected motifs of RNA secondary structure: duplex (a), hairpin loop (b), internal loop (c), bulge loops (d), (e), junction (f) and pseudoknot (g).

Many biopolymers spontaneously fold into characteristic shapes, which determine their biological functions and depend in a complicated way on their primary structures. These three-dimensional shapes molecules assume, are called *tertiary structures*. By folding into a specific three-dimensional shape, molecules are able to carry out their physiological roles inside a cell. Folding is a spontaneous process, mainly guided by Van der Waals forces and entropic contributions to the Gibbs free energy (Westhof and Auffinger, 2000; Neidle, 2002). The determination of the folded structure is a lengthy and complicated process. The methods include X−ray crystallography, as well as spectral methods: nuclear magnetic resonance (NMR), Raman spectroscopy, fluorescence, ultraviolet/visible luminescence (UV−VIS), infrared (IR), electron paramagnetic resonance (EPR), mass spectrometry (MS) (Williams and Fleming, 1996). One of the major areas of interest in protein tertiary structure elucidation is the prediction of the *native structure*, i.e. operative or functional

form, from the sequences. The other approach, strictly bioinformatic, is looking for patterns among the diverse sequences that are known to give rise to particular shapes.

Tertiary structure is stabilized by hydrogen bonding, hydrophobic interactions, ionic bonds, as well as primary and secondary elements (Stryer, 2000; Westhof and Auffinger, 2000). However, tertiary structure determination is usually complicated because of structure dynamics and flexibility. Especially RNA changes its conformations often and quickly, thus, its structure is often hard to be determined (Westhof and Auffinger, 2000).

Tertiary structure can be represented by Cartesian coordinates given for all molecule atoms (algebraic representation), inter-atomic and inter-residue distances (geometric representation), dihedral angles (trigonometric representation) and electron density distribution (probabilistic representation) (Williams and Fleming, 1996; Westhof and Auffinger, 2000).

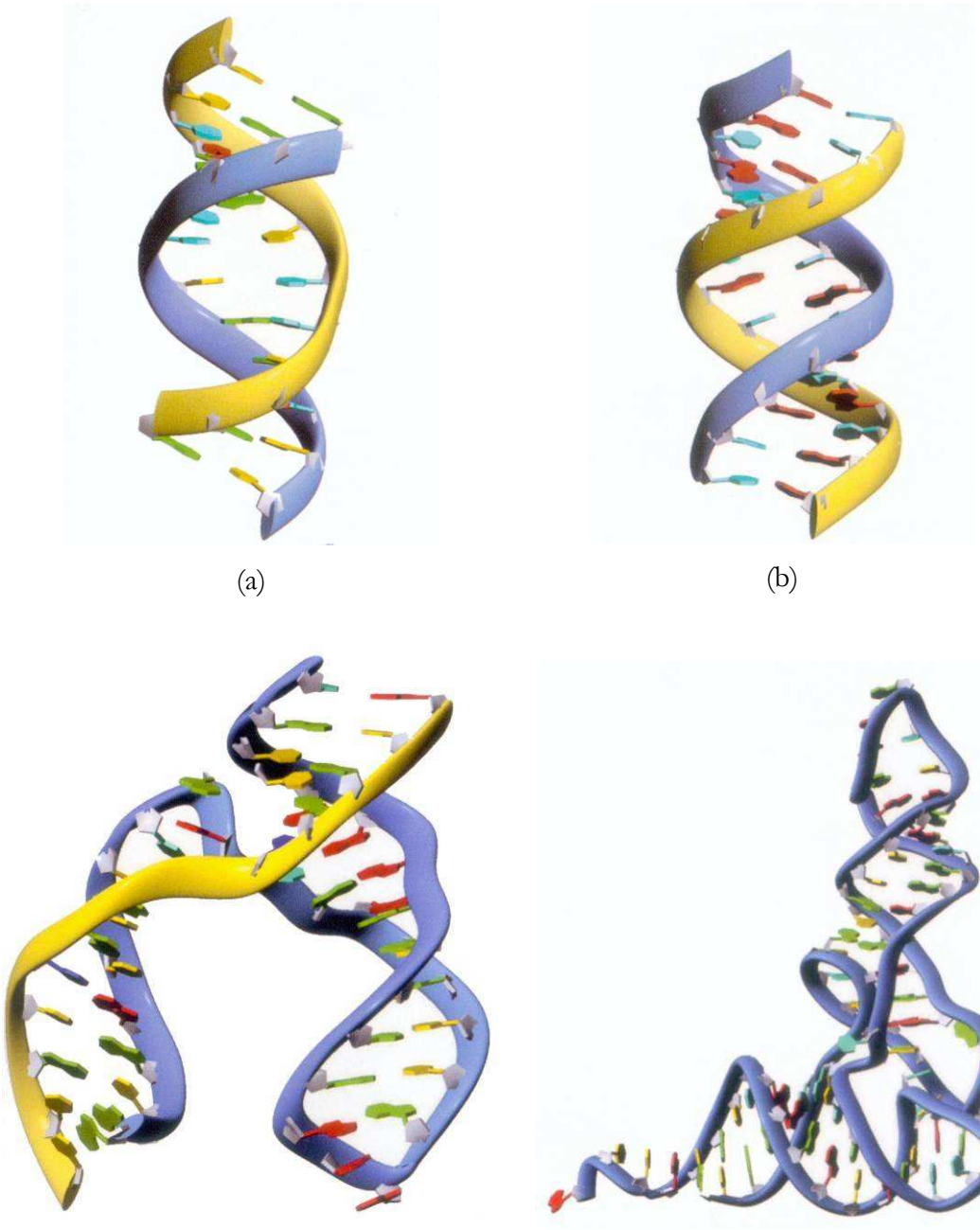


Figure 2.2.3. Examples of tertiary structures: RNA duplex with mismatches r(GGACUUCGGUCC) (a), A-RNA duplex (b), hammerhead ribozyme (c) and tRNA (d) (Bourne and Weissig, 2003).

Quaternary structure is the last level of structural organization being considered during proteins analysis. The study of nucleic acids typically stops after tertiary structure determination (Garrett and Grisham, 1995; Stryer, 2000). Quaternary structure refers to the spatial arrangement of identical or non-identical subunits within the polymer (Garrett and Grisham, 1995; Neidle, 1999). The subunits form complexes held together by non-covalent forces. Single subunit is called a monomer, two subunits – a dimmer, three –

a trimer etc. The forces that stabilize a quaternary structure are much the same as those that stabilize the secondary and tertiary structure, i.e. the non-covalent interactions (Garrett and Grisham, 1995), but they occur between molecules, not in their interiors. The number of interactions plays an important role in stabilization, i.e. the more contacts the more stable a structure is. A considerable range of quaternary structure is found in proteins: from dimeric creatine kinase to octomeric tryptophanase and ribulose diphosphate carboxylase, which assemble sixteen subunits.

In the section, the main aspects of structural analysis of biopolymers have been discussed. Analytical problems considered in the thesis are connected with a determination of RNA tertiary structure on the basis of spectral data. The considered method of structural data acquisition is nuclear magnetic resonance (NMR) spectroscopy, which will be described in the next section.

II.3 NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY

A set of information given in this section aims to explain some basic facts of nuclear magnetic resonance spectroscopy, which is considered now to be one of the best methods of the biomolecule structure determination in solution. This is a method supplying structural information processed by the algorithms presented in the thesis. The subject of NMR spectroscopy is very complex and only the general information is included: a short historical introduction, NMR phenomenon discussion, NMR roles and applications, spectrometer's architecture and performance, types of NMR experiments and spectra. A detailed description of NMR structural analysis of RNA molecules will be given in Chapter 3.

Nuclear magnetic resonance spectroscopy (or *NMR spectroscopy*) is a powerful technique for obtaining structural and dynamic information on molecules at the atomic level. NMR phenomenon itself serves observing physical, chemical, and electronic properties of molecules, allows to determine spatial structure of chemical compounds, and it is also the underlying principle of magnetic resonance imaging and a technique used to build quantum computers.

Phenomenon of nuclear magnetic resonance has been discovered by F. Bloch and E. Purcell, in 1945, who were awarded a Nobel prize seven years later. At the beginning of 1960s, the first protein NMR spectra were acquired, and so NMR spectroscopy was involved in biological sciences. Ten years later, a true revolution in medicine was induced by the discovery of NMR tomography possibility to obtain a view of the living organism

interior. The method of creating images, known as *magnetic resonance imaging* (MRI), has started to develop (Nobel prize for P.C. Lauterbur and Sir P. Mansfield in 2003). With an invention of two– and *d*–dimensional ($d>2$) NMR pulse techniques (Nobel prize for R.R. Ernst in 1991) structural study of biomolecules has started to develop rapidly (Nobel prize for K. Wuthrich in 2002). At present, NMR spectroscopy is widely used in determination of biomolecular spatial structure, analysis of response mechanisms and their speed, describing biopolymer conformational changes, analysis of intermolecular effects, preparing three-dimensional representations of DNA and RNA molecules, modeling their interactions with other biologically active substances etc. (Hausser and Kalbitzer, 1993; Gunther, 1996; Williams and Fleming, 1996). In the past ten years, nuclear magnetic resonance spectroscopy has proved itself as a potentially powerful alternative to X–ray crystallography for the determination of macromolecular three-dimensional structure. NMR has the advantage over crystallographic techniques in the fact that experiments are performed in aqueous solution as opposed to a crystal lattice. The physical principles that make NMR possible, limit the application of this technique to macromolecules of less than 35–40 kD[1]. At present however, computer analysis and extracting structural information from NMR spectra, as well as three-dimensional computer modeling of analyzed molecules belong to the main directions of molecular biology.

NMR spectroscopy uses a common property of most atomic nuclei, i.e., the fact that they are magnetic and therefore their magnetic moments tend to align either in a parallel or in an anti-parallel manner with respect to an external magnetic field. From quantum mechanics, each particle has a spin value of ½. The combination of multiple particles in the nucleus results in an overall spin quantum number. Some isotopes posses zero magnetic spin (e.g. ^{12}C, ^{16}O), thus they do not exist in spin states and are transparent to NMR spectroscopy. However, each of the four most abundant elements in biological material, i.e. H (hydrogen), C (carbon), N (nitrogen) and O (oxygen), has at least one naturally occurring isotope with non-zero nuclear spin and is in principle observable in an NMR experiment (Gunther, 1996). The most interesting in NMR study are isotopes ^{1}H, ^{15}N, ^{19}F and ^{13}C giving narrow resonance signals in the spectra. These isotopes have an odd number of protons and even number of neutrons, which results in spin quantum number equal to ½.

---

[1] kiloDalton (kD) is a unit for expressing a size of a polymer. One kD is equal to the weight of one thousand hydrogen atoms.

In the presence of an external magnetic field of the NMR instrument, the spin angular momentum of nuclei with isotopes of overall non-zero spin undergo a cone-shaped rotation motion called *precession* tending to align itself with two energy states: with or against the magnetic field. Electromagnetic radiation with an energy that matches exactly the energy difference of the two states is absorbed. The corresponding frequency (i.e. rate of precession) is determined by the nucleus under investigation as well as by the chemical environment. Each atom in a given molecule provides a signal with a slightly different frequency, dependent on the strength of the external field and is unique for the isotope. The exact frequency of the spin flips identifies the type of an atom that is involved and is due to the atoms to which the proton is bound (e.g. C, N, O, S) and the local chemical environment. Thus, each proton should, in principle, be characterized by a unique frequency value.

NMR experiment starts with a preparation of the sample. Most NMR spectra are recorded for compounds dissolved in a solvent. Therefore, signals can be observed for the solvent as well as for the compound and this must be accounted for in solving spectral problems. To avoid spectra dominated by the solvent signal, most spectra are recorded in a deuterated solvent like chloroform (CDCl3), acetone-d6, D_2O, or benzene-d6. Most spectra shown in the following chapters of the thesis have been recorded for RNA molecules dissolved in deuterated water (D_2O).

Prepared sample is held by an NMR probe. This device is placed into the bore of the magnet. The probe contains also the coil for irradiating the sample with radio frequency energy and for receiving the very weak radio frequency resonance back from the sample.

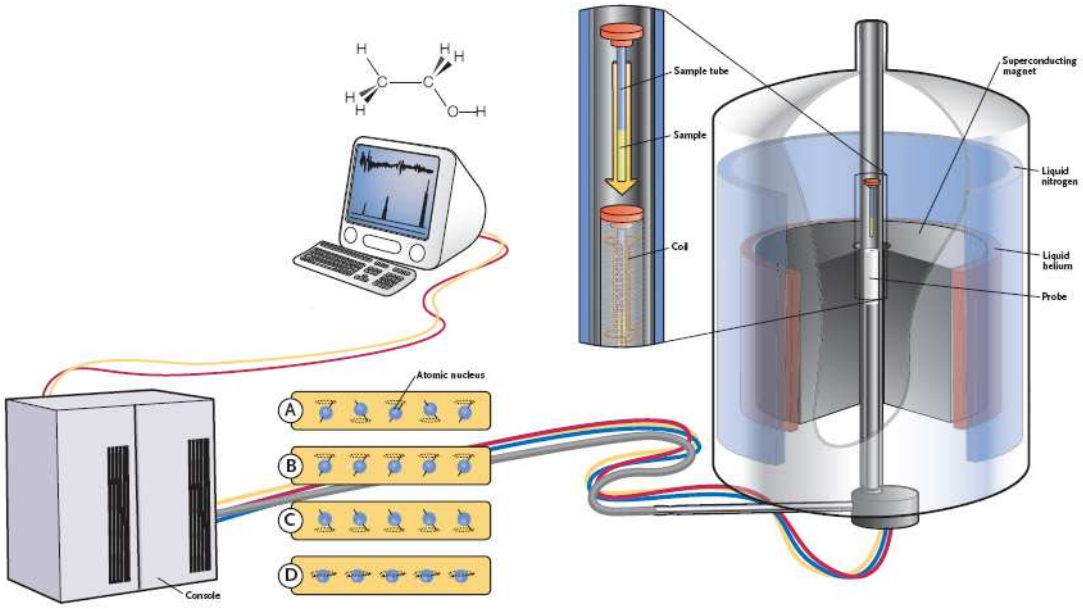


Figure 2.3.1. NMR spectrometer schema (Perkel, 2004).

In general, an NMR spectrometer (Figure 2.3.1) consists of a superconducting magnet, NMR console and a computer workstation with the control software. Thus, the spectrometer provides a uniform, stable magnetic field generated by the magnet (commonly a superconducting solenoid), creates radiofrequency (RF) pulses at the proper frequency (NMR console generates them), collects and processes the data (computer's part of the job).

The NMR magnet provides magnetic field with the strength depending on the magnet type. Typically, it is specified in terms of the resonance frequency for the hydrogen atom expressed in MHz. Standard spectrometers used today produce magnetic field with strength 4,7 T to 21,1 T corresponding to 200-900 MHz resonance frequency of hydrogen nuclei. The magnetic field is steady. NMR console generates and controls short bursts or pulses of high-power RF energy used to excite the sample in the probe. The NMR console also receives and detects the very weak signals coming back from the probe. Powerful radio frequency pulse is applied to the sample. Because this sharp-edged pulse can be looked at as the combination of a number of different frequencies, it can excite all of the nuclei in a sample at once (Ejchart and Kozerski, 1981). The data from all of these nuclei is collected simultaneously in the form of superimposed, exponentially decreasing sine waves whose frequencies reflect the position of the resonance lines relative to the transmitter. When the experiment is complete the free induction decay (FID) data can be Fourier transformed to provide the more familiar frequency-domain NMR spectrum. For organic structure determination, the two most important types of NMR spectra are proton

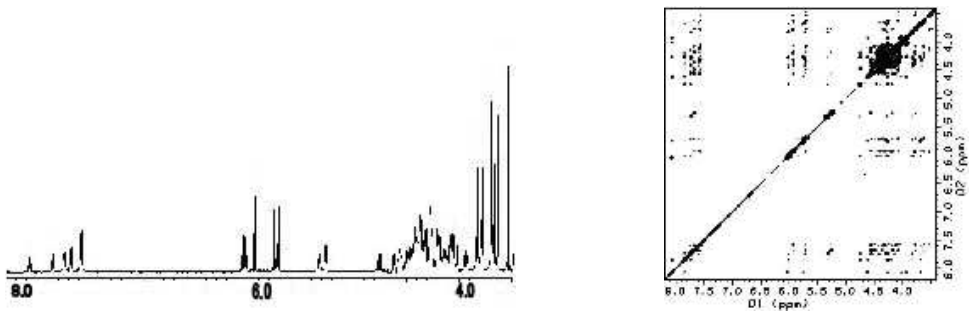
(^{1}H NMR) and carbon (^{13}C NMR) spectra (Gunther, 1996; Jardetzky, 1996b). They give information about the number of hydrogens and carbons in a molecule, their connections, and the information about functional groups. Generally, structure can be determined on the basis of structural constraints, like distances between protons, dihedral angles, mutual ties orientation, inter-nuclei vectors orientation, obtained from NMR experiments (Jardetzky, 1996b). All of the constraints can be measured or calculated out of the spectral data.

An NMR spectrum is typically presented as intensity against applied radio frequency chart. A signal in the spectrum is referred to as a resonance. The relative frequency of a signal is known as its *chemical shift*. Signal frequency is dependent on the strength of the magnetic field. Every magnet's field is a little different, thus, using an instrument-independent way to measure the position of resonance lines in the spectrum seems a natural procedure. The chemical shift (δ) is defined by the frequency of the resonance (ν_{sig}) expressed with reference to a standard compound (ν_{ref}) which is defined to be at 0 ppm. The scale is made more manageable by expressing it in parts per million (*ppm*) and is independent of the spectrometer frequency (ν_{sp}):

$$\delta = \frac{\nu_{sig} - \nu_{ref}}{\nu_{sp}} \times 10^6 \; ppm \,. \tag{2.3.1}$$

As the reference compound, usually TMS, i.e. tetramethysilane ($Si(CH_3)_4$) is used (Jardetzky, 1996a).

It is often convenient to describe the relative positions of the resonances in an NMR spectrum. For example, a peak at a chemical shift of 10 ppm is said to be downfield or deshielded with respect to peak at 5 ppm, while the peak at 5 ppm is upfield or shielded with respect to the peak at 10 ppm.



(a) (b)

Figure 2.3.2. One- (a) and two-dimensional (b) 1H NMR spectrum for 2'-OMe(CGCGCG)$_2$ (Popenda, 1998).

Figure 2.3.2 shows an example of an NMR spectrum. When NMR was first thought about, it was assumed that every hydrogen atom in a sample would have the same resonance frequency. However, NMR experiments showed multiple lines in the spectrum of the same sample. It has been eventually determined that lines appear with different frequencies (ppm) depending on the electronic environment around the nucleus (Ejchart and Kozerski, 1981). The electrons around the nucleus shield it from the magnetic field, so each type of nucleus seems to be in a different field, and therefore it has a different resonance frequency. If the electron density around one nucleus is higher than the other, the electrons provide a different amount of shielding from the full magnetic field. This results in the two nuclei having different resonance frequencies.

The same spectrum can be presented in a couple of different ways (Figure 2.3.3): linear charts (one per each dimension), a contour diagram, or a spatial chart (after optimization with Lorentz function).

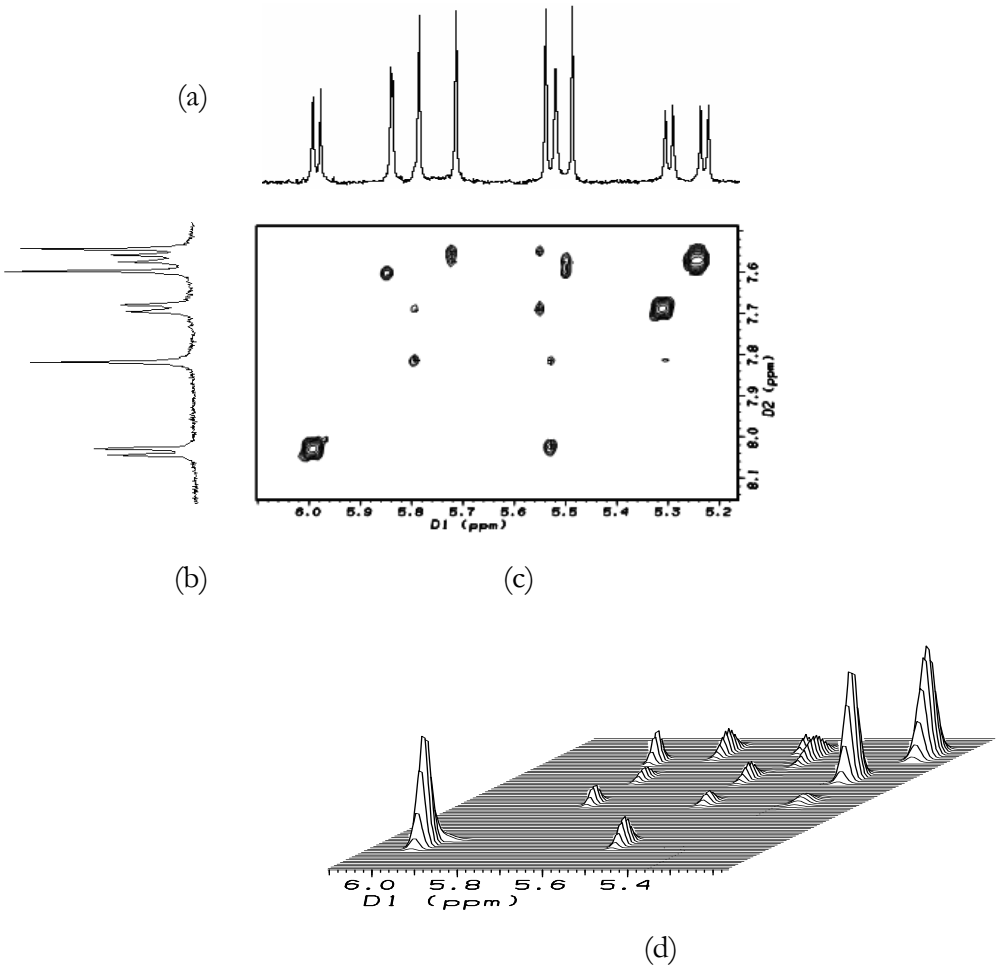


Figure 2.3.3. NMR spectrum of r(CGCGCG)$_2$: line charts (a), (b), contour diagram (c), spatial chart (d) (Popenda, 1998).

NMR spectroscopy provides possibility of trying one– and *d*–dimensional techniques. The *d*–dimensional experiment (for $d>1$) is simply a set of (d–1)–dimensional experiments performed for the same sample with one NMR parameter being variable. The more dimensions an experimenter tries, the more detailed information about the sample he acquires. These experiments, however, are expensive and time consuming. Thus, at present, one– (1D) and two–dimensional (2D) techniques are most commonly used, three– (3D) and four–dimensional (4D) experiments are performed as well.

All NMR experiments are carried out using pulse sequences. The single pulse, which is the basic 1D NMR experiment, is the simplest form of pulse sequence. It is convenient to visualize the effect of the pulse with reference to the rotating frame; i.e., the *x* and *y*-axes rotate about the *z*-axis at the observation frequency. This way, the radio frequency pulse becomes a stationary electric field vector in the *xy* plane. The direction of the vector is governed by the RF phase. The main types of one-dimensional experiments are: basic 1D NMR, decoupled 1D NMR, gated decoupled 1D NMR, inverse gated decoupled 1D NMR, Nuclear Overhauser Effect 1D NMR (NOE), distortionless enhancement of NMR signals by polarization transfer (DEPT) and insensitive nuclei enhanced by polarization transfer (INEPT) (Gunther, 1996; Hilbers and Wijmenga, 1996). They differ by a pulse sequence, which determines the experiment and gives different information about the sample.

The construction of NMR two-dimensional experiment, in addition to preparation and detection steps, which define 1D experiments, is expanded by an indirect evolution time t_1 and a mixing sequence (Jardetzky, 1996b; Ejchart and Kozerski, 1981). Thus, the basic 2D NMR experiment consists of exciting the nuclei with two pulses or groups of pulses, then receiving the free induction decay. The acquisition is carried out many times, incrementing the delay (i.e. evolution time t_1) between the two pulse groups. The FID is then Fourier transformed in both directions to yield the spectrum. The spectrum is conventionally displayed as a contour diagram. 2D NMR spectroscopy includes homonuclear and heteronuclear correlation experiments. In homonuclear experiments, signals generated by the same isotope (usually ^{1}H) are detected. These signals can be produced by the atoms being in close relation through bond or through space. The first group of signals (through bond) can be detected in the following experiments: COSY (Correlated Spectroscopy – correlates scalarly coupled protons), TOCSY (Total Correlation Spectroscopy – identifies protons belonging to the same scalar coupling network), INADEQUATE (Incredible Natural Abundance Double Quantum Transfer

Experiment), ISIPID (Inadequate Sensivity Improvement by Proton Indirect Detection). The interactions occurring between atoms lying closely in space or in chemical exchange are pictured in the spectra resulting from NOESY (Nuclear Overhauser Effect Spectroscopy) and ROESY (Rotational Nuclear Overhauser Effect Spectroscopy) experiments (Hilbers and Wijmenga, 1996). Heteronuclear experiments serve the detection of signals generated by different isotopes. Standard heteronuclear experiments are: HETCOR (Heteronuclear Correlation), HSQC (Heteronuclear Single Quantum Correlation), HMQC (Heteronuclear Multiple Quantum Correlation) and HMBC (Heteronuclear Multiple Bond Correlation) (Wuthrich, 1986; Hilbers and Wijmenga, 1996).

The two-dimensional spectra obtained from the above experiments map out the atoms, which interact. The diagonal as well as the projection on each axis of the spectrum is the 1D spectrum. The off-diagonal peaks indicate the presence of coupling between pairs of atoms. 2D experiments differ mainly by a pulse sequence, which determines the experiment and results in giving different information about the analyzed sample.

The main subject of the thesis is concerned with the two-dimensional NOESY spectrum, which is a 2D version of NOE experiment and yields a display of atoms that are close in space. In 1D NOE experiment, irradiation at specific frequencies before signal acquisition enhances the intensities of nearby nuclei. These nuclei induce relaxation, which leads to signal enhancement through dipole-dipole interactions. The presence of Overhauser enhancements between pairs of protons is indicated by off-diagonal peaks in the spectrum. The cross-peaks, lying in aromatic/anomeric region of the 2D–NOESY spectrum, can be sequenced due to a specific order and form the path, called NOE pathway. Existing methods of NOE pathway construction have been manual and very labor-consuming. A great necessity of proposing automatic methods has been, thus, a contribution and a main goal of the thesis. Therefore, the problem of NOE pathway analysis has been formulated in the category of combinatorial problems, which has facilitated a construction of the appropriate algorithms. The main aspects of combinatorial problems and computational complexity theory are, thus, introduced in the next section.

II.4 COMPUTATIONAL COMPLEXITY OF COMBINATORIAL PROBLEMS

Since the basic problem of finding NOE pathway will be modeled as a combinatorial problem, several basic issues from the latter field, including complexity issues and algorithm design, are recalled in this section. The combinatorial model as well as a proof of

problem NP–hardness will be based on graph theory notions. Thus, the chapter begins with the recollection of graph theory fundamentals. Next, definitions of combinatorial problems are given, followed by the description of complexity classes of the problems and application of the exact as well as approximation algorithms.

An *undirected graph* G is a finite nonempty set of objects called *vertices* together with a, possibly empty, set of unordered pairs of distinct vertices of G called *edges*. The vertex set of G is denoted by V, while the edge set is denoted by E. Thus, graph G may be also defined as a pair of objects $G=(V,E)$ (Chartrand and Lesniak, 1986; Gross and Yellen, 2004).

A *directed graph* or *digraph* D is a finite nonempty set of objects called vertices together with a, possibly empty, set of ordered pairs of distinct vertices of D called *arcs* or directed edges. As with graphs, the vertex set of D is denoted by V and the arc set of D is denoted by E. Digraphs are more natural than graphs for representing situations in which order or direction is involved in the relationships between pairs of objects. Thus, one may also define digraph D as a set of vertices V and a relation E defined on $V\times V$ (Chartrand and Lesniak, 1986).

The problems considered in the thesis, are more suitable to be represented as undirected graphs. Thus, definitions recalled in the following paragraphs are connected with graphs (the terminology used in discussing digraphs is quite similar to that used for graphs, although not always the same).

The edge $e=\{u,v\}$ is said to join the vertices u and v. If edge $e=\{u,u\}$ of graph G connects vertex u to itself, then e is called a *loop*. If $e=\{u,v\}$ is an edge of graph G, then u and v are *adjacent vertices*, while u and e are *incident*, as are v and e. Furthermore, if e_1 and e_2 are distinct edges of G incident with a common vertex, then e_1 and e_2 are *adjacent edges*.

The cardinality of the vertex set of graph G is called the *order* of G and is denoted by p, while the cardinality of its edge set is the *size* of G and is denoted by q. A graph having $p<2$ is a *trivial graph*, while a *nontrivial graph* has $p\geq 2$.

A number of edges of G incident with vertex v in graph G is a *degree* of vertex v. The degree of a vertex v is denoted $\deg(v)$. A vertex of degree 0 in G is called an *isolated vertex*. A graph is *connected* if it does not contain isolated vertices.

Many graph problems are devoted to traversing vertices and edges of a given graph. Any traversal of consecutive elements of graph G can be written as a finite alternating sequence of its vertices and edges, e.g. $u=u_0, e_1, u_1, e_2, \ldots, u_{n-1}, e_n, u_n=v$, where u is an initial vertex, v is a terminal vertex, and $e_i=\{u_{i-1},u_i\}$ for $i=1, 2, \ldots, n$. Often only the

vertices of a walk are indicated since the edges presence is then evident, e.g. $u=u_0,u_1,\ldots,u_{n-1}$, $u_n=v$. Such a sequence is called a *walk* of G, and number n (i.e. number of occurrences of edges) is called the *length of the walk* (Chartrand and Lesniak, 1986). It is important to note that there may be repetition of vertices and edges in a walk. A walk can be *closed* or *open* depending on whether $u=v$ or $u\neq v$. A *trail* is a walk in which no edge is repeated, while a *path* is walk in which no vertex is repeated. By definition, every path is a walk and it is also true that every path is a trail. However, the converse of this statement is not true in general.

A nontrivial (i.e. of length>1) closed trail of graph G is referred to as a *circuit* of G, and a circuit v_1, v_2, …, v_n, v_1 (for $n\geq 3$) whose n vertices v_i are distinct, is called a *cycle*. An *acyclic graph* has no cycles.

Historically most famous trail and circuit are Eulerian trail and Eulerian circuit, which have originated from the problem of Koenigsberg bridges considered by L. Euler in the XVIII century. An *Eulerian trail* of a connected graph G is an open trail of G containing all the edges of G. An *Eulerian circuit* of G is a circuit containing all the edges of G. A graph possessing an Eulerian circuit is called an *Eulerian graph*.

Another important property defines a Hamiltonian graph. A path in graph G traversing all the vertices of G is called a *Hamiltonian path* of G. A cycle of graph G containing every vertex of G is a *Hamiltonian cycle*. Thus, a *Hamiltonian graph* is one that posseses a Hamiltonian cycle.

Looking for Eulerian trail or Hamiltonian path in a given graph is a good example of a search problem – one of basic notions of computational complexity theory which will be discussed here. Let us begin with the notion of a *problem* Π which can be defined as a general question to be answered, usually possessing several parameters, or free variables, whose values are left unspecified. A problem is described by giving a general description of all its parameters and a statement of what properties the answer or solution is required to satisfy.

An *instance* I of a problem Π is obtained by specifying particular values for all the problem parameters (Garey and Johnson, 1979; Blazewicz, 1988; Drozdowski, 1997; Janiak, 1999). Each instance is associated with a fixed encoding scheme, which maps particular input values into the string describing them. Number of symbols in the string is an *input length* for an instance I of a problem Π and is usually used as a measure of *instance size*.

Problem Π can be solved by applying a step-by-step procedure, called an *algorithm* which produces a solution to any instance of Π. In algorithms development the focus is set to efficiency, which is usually determined by time requirements of an algorithm to generate a solution. These requirements are expressed by a *time complexity function*, which gives, for each possible input length, the largest amount of time needed by the algorithm to solve a problem instance of a given size. Depending on the time complexity, algorithm can belong to a group of polynomial, pseudo-polynomial or exponential time algorithms (Garey and Johnson, 1979; Blazewicz et al., 1983).

Let us say that a function $f(x)$ is $O(g(x))$ whenever there exists a constant c such that $|f(x)| \leq c \cdot |g(x)|$ for all values of $x \geq 0$ (Garey and Johnson, 1979). A *polynomial time algorithm* has its time complexity function estimated by $O(p(x))$, where p is some polynomial function and x denotes the instance size. Any algorithm whose time complexity function cannot be so bounded is called an *exponential time algorithm*. Finally, an algorithm is *pseudo-polynomial* if time of its execution can be bounded by a polynomial function dependent on instance size and the maximum value of any problem parameter. A problem which cannot be possibly solved by any polynomial time algorithm is referred to as *intractable problem* and is hard to be solved. Thus, it is often important to determine whether a given problem is intractable or not and then develop an appropriate (e.g. exact or approximate) algorithm to solve it. However, the time of algorithm execution cannot be reliably measured without establishing a model of the computer system. Two models will be used in the following paragraphs: Deterministic Turing Machine (DTM), being an example of realistic computer system model, and Nondeterministic Turing Machine (NDTM) – an unrealistic model. Any algorithm performed in a polynomial time on DTM is also polynomial when executed on any other realistic machine. Whereas, NDTM is capable of performing computations nondeterministically, which can be interpreted as ability of executing unbounded number of computations in time unit (Drozdowski, 1999).

At the beginning of the discussion on problem different types, let us recall that only combinatorial problems are considered in the thesis. A *combinatorial problem* is the one that concerns finite (or at least enumerable) collection of objects and parameters having integer values. In general, combinatorial problems belong to the domain of discrete mathematics. Three main kinds of combinatorial problems can be distinguished: the decision problems, the search problems and the optimization problems. A notion of *decision problem* refers to the problems stated as a question having only two possible solutions, either the answer "yes" or the answer "no" (e.g. Does a given graph posses

a Hamiltonian cycle?). A set of all the instances of a decision problem Π is denoted by D_{Π}, where yes-instances (i.e. instances with the answer being "yes") form subset $Y_{\Pi} \subseteq D_{\Pi}$ and a subset $N_{\Pi} \subseteq D_{\Pi}$ contains all no-instances (i.e. instances with the answer being "no") of Π. Defining a *search problem* one states an instance of a problem and asks an algorithm to find any feasible solution for it (e.g. Find a Hamiltonian cycle for a given graph.). An algorithm, designed to solve a given search problem, results in either giving a feasible solution or the statement that no feasible solution exists. Finally, an *optimization problem* is stated as a command to find the best solution for a given problem (e.g. Find a minimum cost Hamiltonian cycle for a given graph.). The final result of an algorithm, which solves an optimization problem is either the optimal solution or the statement that no feasible solution exists. In many combinatorial problems the search and optimization versions are equivalent, since the only one feasible solution, being an optimum automatically, exists for them.

Each optimization problem Π has its decision version as well as its search version (but not vice versa). In the sense of computational complexity both, the decision and the search versions of problem Π, are not computationally harder than the optimization version of the same problem. Moreover, the optimization version of problem Π is at least as complex as the search version of Π, while the search version of Π is at least as complex as its decision version (Garey and Johnson, 1979; Blazewicz et al., 1983; Blazewicz, 1988; Drozdowski, 1999). Hence, it is possible to analyze optimization problems computational hardness by considering their decision counterparts. By showing that the decision version of problem Π is hard, one automatically proves that the optimization and search versions of the same problem must be hard as well. Note, however, that an "easiness" of the decision version does not imply an "easiness" of the search and optimization versions of the same problem.

Decision problems fall into sets of comparable complexity, called complexity classes. By definition, the complexity class *NP* (Non-deterministic Polynomial) is the set of all decision problems that can be solved in polynomial time by NDTM (a non-deterministic Turing machine), which is an unrealistic model of computer-like machine. Solution correctness of every problem in NP can be verified effectively in time bounded by a polynomial dependent only on the instance size. NP class contains both – "easy" and "hard" problems. The first subset ("easy") of NP is *P* (Polynomial) complexity class, being the set of all decision problems that can be solved in polynomial time by DTM (deterministic Turing machine), which is equivalent to all realistic computing models.

Problems in P are easy also in common sense, i.e. they can be solved by polynomial time algorithms on any computer. The relationship between the classes P and NP, P⊆NP, is fundamental for the theory of NP–completeness. There is a widespread belief that P≠NP, although no proof of this conjecture has been done. The hardest problems in NP are counted into *NP–complete* complexity class, NP–complete⊆NP. These problems are solved by pseudo-polynomial or exponential algorithms, although a possibility of applying polynomial time algorithms for them is not excluded. A construction of a polynomial algorithm for NP–complete problem would prove that P=NP, and vice versa. NP–complete complexity class includes *strongly NP–complete* problems, for which no pseudo-polynomial time algorithms exist.

The notion of NP–completeness is crucial in the combinatorial model of NOE pathways problem considered in the thesis. Thus, the following paragraphs provide more details on this notion.

By definition, decision problem Π is NP–complete if and only if it is in NP and every other problem from NP class polynomially transforms to Π. A *polynomial transformation* of problem Π_1 to problem Π_2 is a transformation performed in a polynomial time with preserving the equivalence of both problems. The equivalence means, that the solution for an instance of problem Π_1 belongs to Y_{Π_1} if and only if the solution for an instance of Π_2 (being the polynomial transformation of Π_1) belongs to Y_{Π_2}. A polynomial transformation is denoted by $\propto$. Thus, $\Pi_1 \propto \Pi_2$ means that decision problem Π_1 is polynomially transformed to decision problem Π_2.

Proving that given problem Π_1 is NP–complete involves the following steps. First, it should be shown that Π_1 belongs to the NP class of problems. Next, known NP–complete problem Π_2 must be selected and polynomially transformed to Π_1 (Garey and Johnson, 1979; Blazewicz, 1988). To prove that problem Π_1 is in NP one should generate an algorithm solving Π_1 on a NDTM in polynomial time. *NDTM* is a computer-like machine consisting of guessing and checking modules. The first module guesses a solution (or – in fact – all finite solutions at the same time), while the second – checks whether the solution is correct. Thus, a program for NDTM should describe, in terms of NDTM symbols and states, what does each module exactly do to solve a given problem.

To each strongly NP–complete problem any other problem of this complexity class can be pseudo-polynomially transformed. Again, proving that given problem Π_1 is strongly NP–complete one should show that Π_1 is in NP and construct a pseudo-polynomial transformation of a known strongly NP–complete problem Π_2 to Π_1. *Pseudo-*

polynomial transformation is a pseudo-polynomial procedure (i.e. time of its execution depends on instance size and maximum value of any problem parameter) preserving an equivalence of problem before and after transformation. Additionally, it is said, that the maximum value cannot exponentially grow after the transformation and instance size cannot exponentially shrink.

Search problem Π is said to be *NP–hard* if any other NP–hard (or NP–complete) problem is *T–transformable* to Π. *T–transformation*, called also a *polynomial Turing transformation* of problem Π_2 to Π_1, is an algorithm solving problem Π_2 on DTM in polynomial time by the use of some hypothetical polynomial-time procedure solving problem Π_1 (on DTM). T–transformation is denoted by $\propto_T$, thus, $\Pi_2 \propto_T \Pi_1$ means that decision problem Π_2 is polynomially transformed to search problem Π_1. When some search or optimization problem appears NP–hard, it means that no polynomial time algorithm exists (unless P=NP) for the problem and heuristic algorithms must be used to solve it.

It is important to know, that there exist NP–hard search problems, which have their decision versions computationally easy (c.f. Johnson, 1985). In the thesis a problem of this kind is considered. Each instance of such a problem in its decision version can be answered "yes", i.e. a solution exists, but a construction of the solution is hard, what makes a search version of the problem computationally intractable. A good example of such a problem has been analyzed in (Johnson, 1985), where it has been proved, that finding a Hamiltonian cycle in given graph G in a polynomial time is hard, even though we know, G is a Hamiltonian graph. Unless P=NP, there simply cannot be constructed an algorithm, that, given a graph containing a Hamiltonian cycle, is guaranteed to find the one in polynomial time. The reasoning is as follows. If we had such an algorithm A, we could use it to tell in polynomial time whether an arbitrary graph G has a Hamiltonian cycle. Let p be the polynomial that bounds A's running time on graphs with Hamiltonian cycles. Apply A to G. If G has a Hamiltonian cycle, A will find one in time $p(|G|)$. If G does not have such a cycle, then after $p(|G|)$ steps A cannot have found one, and we will know that none exists.

The above problem is an example of a *promise problem* (Johnson, 1985). Promise problem consists of an instance domain (e.g. Given graph G.), a goal (e.g. Find a Hamiltonian cycle in G.) or question (e.g. Does Hamiltonian cycle exist in G?) and a promise concerning problem solution. Particularly, promise can determine that *at least* one solution exists and, in such case, the decision version of the promise problem is trivially easy (the answer is always "yes"). Search version of the latter can be easy or hard.

Let us consider promise problems having easy decision versions. We say that these problems belong to *PROMISE-P* complexity class. However, finding the solution to the promise problem can be hard, even for problems with a promise of solution existence (i.e. at least one solution is promised to exist). Thus, there exists a class of promise problems with easy decision versions and computationally intractable search versions. Such a problem is considered in the thesis. On the other hand, in order to prove that a given decision promise problem is hard, one should demonstrate a parsimonious transformation. A *parsimonious transformation* from problem Π_2 to problem Π_1 is a polynomial transformation that keeps a number of solutions for instances of Π_1 equal to a number of solutions for equivalent instances of Π_2 (Garey and Johnson, 1979). Proving that a decision promise problem Π_1 is hard one should execute a parsimonious transformation of decision promise problem Π_2 not belonging to PROMISE–P, to problem Π_1.

One more group can be distinguished within the class of promise problems. *Uniquely promised problem* consists of an instance domain, a goal or question and a promise that *at most* one solution to the problem exists. Solution uniqueness assumption does not necessarily influence problem computational complexity. Thus, the decision version of uniquely promised problem can be easy (polynomial) or hard (NP–complete) and its search version can be easy or hard (NP–hard) as well. A good example of NP–hard uniquely promised problem, i.e. uniquely promised Hamiltonian cycle, has been discussed in details in (Johnson, 1985).

Finally, let us discuss NP–hard optimization problems. Any problem from this class can be solved by an exponential algorithm giving an exact solution or by a heuristic algorithm generating a suboptimal solution in a polynomial time. The most common exact method applied for NP–hard optimization problems is branch and bound algorithm, whereas, tabu search, genetic algorithm and simulated annealing algorithm belong to a set of most popular heuristics used for obtaining suboptimal solutions. Two of these heuristics are considered in the thesis, tabu search and genetic method.

Branch and Bound (B&B) algorithm is a general search method dealing with optimization problems over a search space that can be presented as a directed, rooted tree structure, called a *search tree* (Papadimitriou and Steiglitz, 1982). Each node of the tree responds to one step of the algorithm, while leaves represent potential solutions to the considered problem. The method starts by analyzing the original problem with the complete feasible region, which is called the root problem and becomes the root of the

search tree. In each step current node is scored by the bounding procedure applied to it. If there is a chance for solution enhancement, the current node is divided into subnodes, which together cover the whole of their ancestor (ideally they partition the parent node). Next, the procedure steps into one subnode and is applied recursively, generating a tree of subproblems. The score of the best leaf found so far is kept as a bound. Whenever a node is reached whose score is worse, the search tree is pruned at that node, i.e. its subtree will not be searched, since it is guaranteed not to contain a leaf with better score. The search proceeds until all nodes have been solved or pruned, or until some specified threshold is met between the best solution found and the lower bounds on all unsolved subproblems. Branch and Bound is guaranteed to find the optimal solution, but its complexity in the worst case is as high as that of exhaustive search (exponential).

The basic concept of *tabu search* is a meta-heuristic superimposed on another heuristic (Glover and Laguna, 1997). It executes heuristic local search but also steps outside the local optimum area. The search begins by marching to a local optimum in a search space X, which, usually, contains only acceptable solutions. Each solution is valued by a goal function, which is optimized during the search. Solution $x \in X$ has a neighborhood $N(x) \subset X$. Every neighboring solution $x' \in N(x)$ can be achieved from x by one move. Usually, the step improves goal function value by choosing better solution, however, if the improvement is not possible other moves are accepted. To avoid retracing the steps used, the method records recent moves in one or more tabu lists and verifies each move with the lists. The lists are historical in nature and form the tabu search memory. The role of the memory can change as the algorithm proceeds. At initialization the goal is to make a coarse examination of the solution space. This procedure is known as diversification strategy (algorithm jumps between different search areas). Once candidate locations are identified, the search is more focused to produce local optimum solutions in a process of intensification. Differences between the various implementations of the tabu method have to do with the size, variability and adaptability of the tabu memory to a particular domain.

Genetic (or *evolutionary*) *algorithm* (GA, EA) is adaptive heuristic search method premised on the evolutionary ideas of natural selection and genetics. The basic concept of genetic algorithm is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Darwin of survival of the fittest. As such it represents an intelligent exploitation of a random search within a defined search space to solve a problem. GA is modeled loosely on the principles of the

evolution via natural selection, employing a population of individuals (solutions) that undergo selection in the presence of variation-inducing operators such as mutation and recombination (crossover). A fitness function is used to evaluate individuals, and reproductive success varies with fitness. The algorithm begins with a random generation of an initial population $M(0)$. Next it computes and saves the fitness $u(x)$ for each individual $x \in M(t)$, where M(t) is the current population. Afterwards, GA defines selection probabilities $p(x)$ for each individual $x \in M(t)$ so that $p(x)$ is proportional to the fitness $u(x)$ of x. The next population $M(t+1)$ is generated by probabilistic selection of individuals from $M(t)$ and producing offspring via genetic operators. The whole process, starting from evaluation of individuals in current population and finishing in generation of the new population, is repeated until a satisfying solution is obtained (Holland, 1975).

III. BASIC CONCEPTS OF RNA STRUCTURAL ANALYSIS IN SOLUTION WITH THE USE OF 2D NMR TECHNIQUES

Liquid state NMR has emerged as a structure determination technique for macromolecules in the mid 1980s (Case, 1998). Previously, X–ray diffractometry had been the most powerful method of tertiary structure elucidation. However, the latter technique is effective for the compounds forming single crystals after crystallization. Since single crystals are rare in natural substances and RNA is rather insusceptible for crystallization, NMR has taken X–ray place in an analysis of RNA and other compounds. On the other hand, there is a long way to pass from NMR experiment, through data extraction to a presentation of a suggested structure. A collection of different NMR experiments should be run and a set of parameters, later used in the final structure modeling, gathered. A sequence of steps executed during NMR structure analysis is described in the first of the following sections. One of the first analytical steps, i.e. *resonance signal identification*, was not fully solved, and this one defines the thesis subject. Thus, the second section characterizes the background of this step. Finally, the other existing methods of solving the problem of resonance signals identification are summarized in the last section of this chapter.

III.1 STRATEGY OF RNA STRUCTURE DETERMINATION WITH NMR

In this section the process of structural analysis with NMR techniques is explained. It is similar for different biomolecules; nevertheless, the description is focused just on RNA, which is considered in the thesis.

Let us start with the general description of RNA tertiary structure elucidation with NMR, presented from a point of view of structural biology. The process is composed of several stages. At the beginning NOE (Nuclear Overhauser Effect) signals are identified and assigned to appropriate protons of the molecule being analyzed. Next, ribose possible conformation, i.e. its spatial structure, is examined. The following step is dedicated to a description of a phosphate group alignment in a relation to ribose. Afterwards, the whole polynucleotide chain conformation is suggested and duplexes geometry is described. The information content available from NMR measurements after these steps is not sufficient to completely determine the structure. It allows for a generation of a set of possible

structures, that respond to a collection of parameters acquired. The cardinality of this set can be reduced during structure refinement step. NMR refinement usually involves simulated annealing procedure for optimizing functions that consider additional constraints and describe the molecular structure of the compound.

In the detailed chemical view, NMR fundamental structure determination explores techniques that correlate carbon atoms to directly attached protons (CH, CH2, CH3), assign proton-proton ($^{1}H-^{1}H$) spin couplings and determine coupling networks, assign long-range proton-carbon ($^{1}H-^{13}C$) connectivity and determine 3D structure by through-space proton-proton interactions (Wuthrich, 1986; Varani and Tinoco Jr., 1991; Williams and Fleming, 1996; Wijmenga and van Buuren, 1998; Nilges, 1999; Zidek et al., 2001). An experimental plan involves:

1. preparation of RNA sample,
2. carrying out a range of one- and multidimensional NMR experiments for the sample,
3. identification of resonance signals,
4. calculation of structural restraints,
5. generating a family of structures and their analysis.

RNA sample is prepared by a chemical synthesis or enzymatic methods. For a liquid state NMR, the compound is dissolved in a solvent. To avoid spectra dominated by solvent signals the solvent is deuterated, which means that some hydrogen protons are replaced by deuter (D or $^{2}_{1}H$) – heavier hydrogen isotope, transparent to NMR. Most commonly used solvents are acetic acid, acetone, acetonitrile, benzene, chloroform, dimethyl sulfoxide, methanol, methylene chloride, pyridine or deuterated water (D_2O). Chemical solution prepared in this way is then placed in a probe and situated in NMR spectrometer. Next NMR experiments are executed. Their variety depends on the information a spectroscopist wants to acquire. The whole set of possible one- and two-dimensional spectra has been described in Chapter 2.

The next step in the process is a signal identification, which determines execution of the following calculations. On the basis of spectral data, number of protons is verified and information about proton-proton *resonance signals* is obtained. These signals, called *correlation signals* or *NOE signals* are detected during NOESY (Nuclear Overhauser Effect Spectroscopy) experiment and displayed in a 2D–NOESY spectrum. NOE correlation signal appears when two protons, close in space, resonate. Usually, the resonance appears

between hydrogen protons, for which mutual distance is at most 6 Å[2] (Nilges, 1999). Cross-peaks in the spectrum represent the signals. Each cross-peak is characterized by its ordinary number, two coordinates of its center, given in ppms (see Section 2.3 for ppm definition), widths of the cross-peak and a volume corresponding to signal intensity (see Section 3.2 for spectrum detailed description). However, the information about protons, that generated the signal is unknown and should be guessed. In other words, appropriate protons must be assigned to each signal and so the signal is identified. This identification proceeds in two steps. First a specific *NOE connectivity pathway* (see Section 3.2 for pathway detailed description) is found and drafted in the spectrum. Next, the pathway is mapped onto primary sequence of the molecule. For example, if RNA molecule r(CGUA)$_2$ is analyzed and NOE pathway found is s_4-s_1-s_2-s_8-s_7-s_6-s_5-s_9-s_3, where s_i denotes i-th cross-peak in the spectrum, then mapping the pathway onto the sequence results in the following assignment: s_4 is a signal generated by H1' and H6 protons of C1 (denoted C1:H1'–C1:H6), s_1 is generated by H1' proton of C1 and H8 proton of G2 (C1:H1'–G2:H8), s_2 is generated by H8 and H1' protons of G2 (G2:H8–G2:H1'), s_8 is generated by H1' proton of G2 and H6 proton of U3 (G2:H1'–U3:H6), etc. (see Section 3.2 for assignment detailed description). After an assignment an experimenter knows what signal (how intensive, how positioned etc.) is generated by any chosen pair of hydrogen protons and what chemical shift each proton has. This information is sufficient to generate a draft of a structure, although in order to obtain more precise view other parameters must be calculated. These parameters, calculated with the use of information derived from assignments, are called structural restraints (e.g. NOE distances, torsion angles etc.). Without signal identification no structural restraints calculation can be performed. Finding NOE pathway in the spectrum, however, is a hard problem. Thus, this step is a bottleneck of the whole process of biomolecule structure determination and, since the last decade of 20th century, it has been one of the top problems in structural biology and bioinformatics (Kraulis, 1989; Roggenbuck et al., 1990; Bartels et al., 1997; Croft et al., 1997; Mumenthaler et al., 1997; Zimmerman et al., 1997; Guntert, 1998; Moseley and Montelione, 1999; Atreya et al., 2000; Linge et al., 2003; Adamiak et al., 2004; Balley-Kellog et al., 2004). In the thesis, several algorithms for NOE pathway reconstruction are

[2] Angstrom (Å) is a unit of length, used commonly in measuring atoms sizes or intermolecular distances. One angstrom equals 10^{-10} meters, 0.1 nanometers or 100 picometers.

proposed (see Chapter 6), which eliminate the difficulties on this stage of structure elucidation.

Having resonance signals identified, one can calculate structural restraints, basing primarily on chemical shifts obtained through the sequence-specific assignments. Traditionally, structure determination relies on the measurement of a maximum possible number of local restraints. The more restraints are calculated the more precise a final structure is obtained. A set of possible restraints includes: NOE distances, homo- and heteronuclear scalar couplings, chemical shifts, torsion angles, dipolar couplings, etc. (Wuthrich, 1986; Nauhaus and Williamson, 1989; Hilbers and Wijmenga, 1996; Jardetzky and Schmitt, 1996a; Varani et al., 1996; Guntert, 1998; Mollova and Pardi, 2000). Their calculation is beyond a scope of the thesis, thus just general information is given in the following paragraphs.

The most important of local restraints is the $^{1}H–^{1}H$ *nuclear Overhauser effect*, which provides distance information for pairs of protons separated by less than 5 Å. The accuracy of the NOE-derived distance usually decreases with the actual value of the distance because the effect of indirect NOE magnetization transfer tends to be worse for protons further apart (Wuthrich, 1986; Bax et al., 2001). NOE distance *distNOE* between two hydrogen protons h_1, h_2 is calculated according to the following formula:

$$distNOE(h_1,h_2) = \sqrt[6]{\left(\frac{ins(h_1,h_2)}{ins(h_{ref})}\right)^{-1}} \times 1.78\,, \tag{3.1.1}$$

where $ins(h_1,h_2)$ is an intensity of the signal generated by protons h_1 and h_2, and $ins(h_{ref})$ is an intensity of a reference point.

The second commonly used restraint is three-bond *J* coupling (scalar coupling), either homonuclear $^{1}H–^{1}H$, $^{13}C–^{13}C$, or heteronuclear $^{13}C–^{1}H$, $^{13}C–^{15}N$, or $^{15}N–^{1}H$, which are related to the intervening dihedral angles via empirically parameterized Karplus relationships, which can be approximated in the form (Wuthrich, 1986; Jardetzky and Schmitt, 1996a; Bax et al., 2001):

$$J = A + B\cdot\cos\theta + C\cdot\cos^2\theta\,, \tag{3.1.2}$$

where *A*, *B*, *C* are coefficients that depend upon the electronegativity of the surrounding atoms and θ is a dihedral angle about the central bonds. Other scalar couplings, through one- or two-bond interactions are also measured. *J* values measure the influence of a magnetic nucleus on a neighboring magnetic nucleus. The influence is determined by the chemical structure of a compound.

Twelve torsion angles determining molecule tertiary structure, are measured for nucleic acids: α, β, γ, δ, ε, ζ, χ, ν_0, ν_1, ν_2, ν_3, ν_4. Each torsion angle is defined by four atoms. Atom sequences defining sugar-phosphate backbone torsion angles are as follows: α: O3'–P–O5'–C5', β: P–O5'–C5'–C4', γ: O5'–C5'–C4'–C3', δ: C5'–C4'–C3'–O3', ε: C4'–C3'–O3'–P, ζ: C3'–O3'–P–O5', χ (pyrimidines): O4'–C1'–N1–C2,

χ (purines): O4'–C1'–N9–C4.

Endocyclic sugar torsion angles are defined by sequences:

ν_0: C4'–O4'–C1'–C2', ν_1: O4'–C1'–C2'–C3', ν_2: C1'–C2'–C3'–C4', ν_3: C2'–C3'–C4'–O4', ν_4: C3'–C4'–O4'–C1' (Hilbers and Wijmenga, 1996). Torsion angles can be calculated from a formula:

$$\theta = sign\cos^{-1}\left(-\left(\left(e_{12}\times e_{23}\right)\cdot\left(e_{43}\times e_{32}\right)\right)\right) \tag{3.1.3}$$

for $-\pi\leq\theta\leq\pi$, where *sign* is a sign of the following formulation:

$$\left(-\left(e_{12}\times e_{23}\right)\times\left(e_{43}\times e_{32}\right)\right)\cdot e_{23} \tag{3.1.4}$$

and e_{ij} denote atoms electronegativity.

Another relatively recent addition to the arsenal of experimental restraints includes cross-correlated relaxation. In contrast to the other parameters, cross-correlated relaxation can, at least in principle, report on the relative orientation of dipolar on chemical shift anisotropy tensors anywhere in the molecule.

A family of structural restraints determines tertiary structure of the molecule. Obtaining structure of high accuracy and precision is clearly of great importance if maximal insight is to be obtained into the relationships between three-dimensional structure and biological function, including the identification of subtle structural differences between similar RNA chains or the conformational changes that accompany polymer folding. This in turn requires careful consideration of the way in which spectral simulations are carried out and of the sampling and optimization procedures used. The final step of structure determination is then connected to structure refinement, which basically uses optimization procedures. The first, initial set of structures, generated e.g. with the use of distance geometry methods (most popular), approximately satisfies the covalent and experimental constraints. These structures are refined against the experimental data in various ways. Coupling constants data are commonly used in the refinement process. They are often converted to dihedral angle constraints and used in a manner similar to the distance constraints derived from NOE information. Because of the complexity of the interactions that influence the chemical shift, this information is

more useful at the final structure refinement stage, where the overall structure has already been determined. There exist several structure refinement methods (Case, 1998; Clore and Gronenborn, 1998; Chen et al., 1999; Nilges, 1999).

III.2 NOE ASSIGNMENT IN STRUCTURE DETERMINATION

This section is devoted to an explanation of NOE assignment problem. The assignment, being the first computational step in elucidation of RNA tertiary structure (see Section 3.1), remains a bottleneck of the whole procedure. In brief, the assignment is based on the 2D–NOESY spectrum, where NOE connectivity pathway should be sketched. NOE pathway can be also called a magnetization transfer pathway or H6/H8–H1' connectivity pathway. Completing information about assigned protons can result in the calculation of NOE distances and other restraints derived from them. The following paragraphs contain a detailed description of the NOESY spectrum and characterize the properties of NOE pathway. Finally, the last part of the assignment, structure mapping onto the pathway, is explained.

Two-dimensional *NOESY spectrum* displays information obtained during NMR NOESY experiment, which detects interactions appearing between pairs of protons being close in space but not closely connected by chemical bonds. These interactions, called *NOE signals*, are through-space correlations via the Nuclear Overhauser Effect (NOE). NOE is very sensitive to the distance and may therefore be used to estimate it. Interaction strength is proportional to $(distNOE)^{-6}$, where *distNOE* is the distance between interacting protons. The spectrum contains the diagonal and the off-diagonal *cross-peaks*. The diagonal, as well as a projection of the spectrum on each axis, is the 1D spectrum. The cross-peaks indicate NOE signals between two protons. Since signal position is measured as a chemical shift, thus, both dimensions ($D1$, $D2$) of the spectrum are expressed in ppm units and range from 4 to 8 ppm. Each nucleotide contains about 10 hydrogen protons, but only some (H1', H2', H5, H6, H8) are analyzed during NOE pathway construction. Typical chemical shift ranges of the protons are displayed in Table 3.2.1.

Table 3.2.1. Typical chemical shifts of the selected protons.

Proton	Residues including proton	Chemical shift range
H1'	C, U, A, G	5 – 6 ppm
H2'	C, U, A, G	4 – 5 ppm
H3'	C, U, A, G	4 – 5 ppm
H4'	C, U, A, G	4 – 5 ppm
H5'	C, U, A, G	4 – 5 ppm
H5"	C, U, A, G	4 – 5 ppm
H5	C, U	5 – 6 ppm
H6	C, U	7 – 8 ppm
H8	A, G	7 – 8 ppm

A typical 2D–NOESY spectrum contains nine characteristic regions of the correlated signals (Figure 3.2.1). Since the spectrum is symmetric with respect to the diagonal, the appropriate regions can respond to each other. However, vertical spectrum resolution differs from the horizontal one; thus, the symmetric regions do not necessarily look the same.

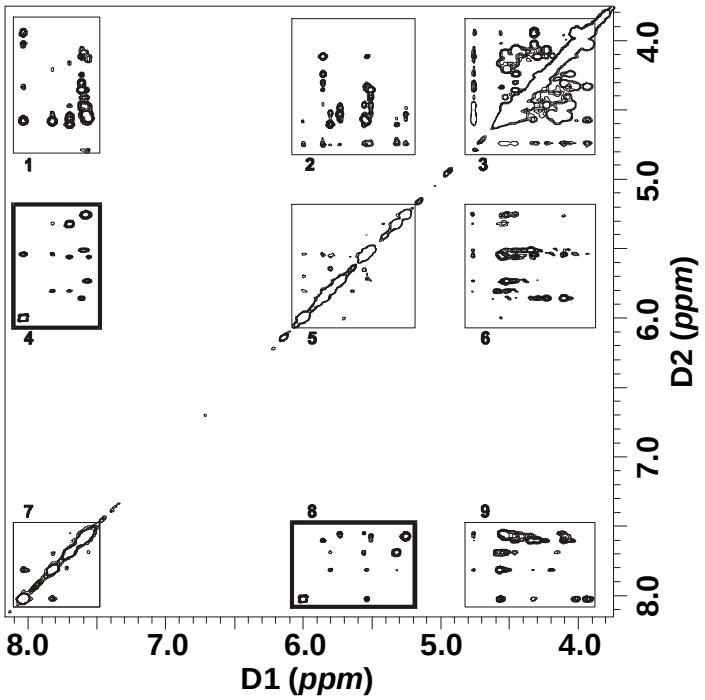


Figure 3.2.1. An exemplary 2D–NOESY spectrum r(CGCGCG)$_2$ in D_2O with bounded regions (Popenda, 1998).

Each spectral region displays interactions between other pairs of protons, due to the chemical shift range the region covers (Table 3.2.2).

Table 3.2.2. Interactions displayed in different spectral regions.

Region [ppm]	Interactions pictured in the region
[4–5]×[4–5]	inter- and intranucleotide correlations between H2', H3', H4', H5', H5"
[4–5]×[5–6] [5–6]×[4–5]	inter- and intranucleotide correlations between H2', H3', H4', H5', H5" interacting with H1' or H5
[4–5]×[7–8] [7–8]×[4–5]	inter- and intranucleotide correlations between H2' and H6, or H2' and H8
[5–6]×[5–6]	inter- and intranucleotide correlations between H1' and H5
[5–6]×[7–8] [7–8]×[5–6]	inter- and intranucleotide correlations between H1' and H6 or H1' and H8, intranucleotide correlations between H5 and H6 in pyrimidine nucleotides (these regions are distinguished by a bold frame in Figure 3.2.1)
[7–8]×[7–8]	internucleotide correlations between H6 and H6, H6 and H8, H8 and H8

Each hydrogen proton can interact with a proton belonging to the same nucleotide, and such an interaction is called the *intranucleotide* NOE signal, as well as with a proton positioned in the neighboring nucleotide generating *internucleotide* signal. Usually, intranucleotide signals are more intensive than internucleotide ones, what results from the distance between them. Thus, volumes of *intra* cross-peaks should be bigger than volumes of *inter* cross-peaks. However, this is not a principle.

Region [5–6]×[7–8] ppm (or symmetric [7–8]×[5–6] ppm), called an aromatic/anomeric region of the spectrum, is fundamental in signal assignment process (Wuthrich, 1986). This region should contain $P+2N-1$ cross-peaks, where N denotes the number of all nucleotides in analyzed RNA chain and P denotes the number of pyrimidine nucleotides in this chain. The number of interactions is given approximately, as unsuspected or noised interactions can occasionally appear. A major task of the assignment procedure is to find a sequence-specific connectivity pathway $H8/H6_{(i)}–H1'_{(i)}–H8/H6_{(i+1)}$ ($i \leq 2N-1$, where N denotes number of nucleotides), called *NOE pathway*. Formation of such a path is possible because each aromatic H6/H8 proton of nucleotide residue (H6 in C and U, H8 in A and G) is in close proximity to two anomeric protons: its own and the preceding (from 5' side) H1' proton. Figure 3.2.2 shows a short exemplary NOE pathway going through the four-nucleotide strand r(CGUA), where the main NOE interactions between protons of our interest are marked with arrows.

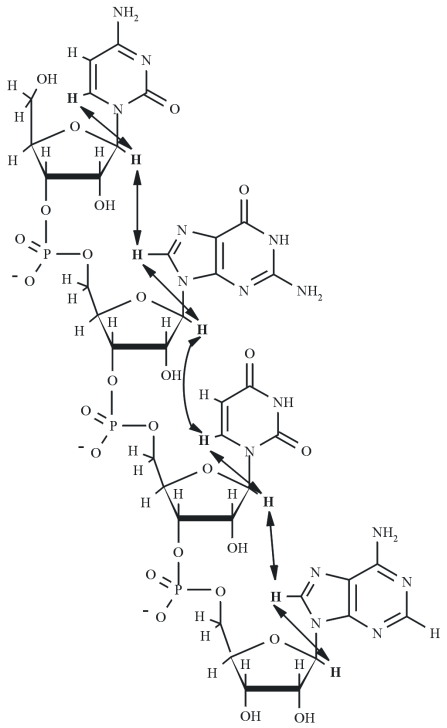


Figure 3.2.2. Main NOE interactions in r(CGUA) (Popenda, 1998).

The pathway is composed of intranucleotide and internucleotide interactions, which give rise to the alternately appearing cross-peaks. In case of ideal A-RNA duplexes, the NOE pathway starts with the intranucleotide interaction at 5' end of the strand and its length equals $2N-1$, where N denotes a number of residues (nucleotides) in the RNA chain. Each proton, except for the starting and terminal ones, 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 chemical shifts of the corresponding protons. In fact, every cross-peak is characterized by the two coordinates of its centre, widths in both dimensions and the value of signal intensity. All of these data, obtained from NMR experiment are taken into account during NOE pathway construction. Since each of the considered protons participates in (at most) two interactions forming connectivity in the NOE pathway, every two consecutive points in the pathway have exactly one coordinate in common (i.e. chemical shift of the common proton). This results in consecutive connections within the pathway laying vertically or horizontally. Moreover, for this reason every two neighboring connections are perpendicular. Summing up, the NOE pathway can be defined in the following way:

Definition 3.2.1 (*NOE pathway*)

Let $P_S=s_1, s_2, \ldots, s_l$ be a sequence of cross-peaks placed in the aromatic/anomeric region of 2D–NOESY spectrum S, representing NOE correlation signals occurring between protons of a RNA molecule. We will call P_S the *NOE pathway* in S, if the following conditions are satisfied (Szachniuk et al., 2003; Adamiak et al., 2004; Blazewicz et al., 2004a):

1. the path starts from a cross-peak representing an intranucleotide signal,
2. every cross-peak $s_i \in S$, occurs in path P_S at most once,
3. intra- and internucleotide signals appear alternately in P_S,
4. only cross-peaks having exactly one common coordinate are connected,
5. every two neighboring connections of P_S are perpendicular,
6. P_S does not contain collinear connections,
7. a length of P_S equals $l=2N-1$, where N denotes a number of nucleotides.

Figure 3.2.3 demonstrates an exemplary NOE pathway found in the aromatic/anomeric region of the 2D–NOESY spectrum generated for r(CGCGCG)$_2$.

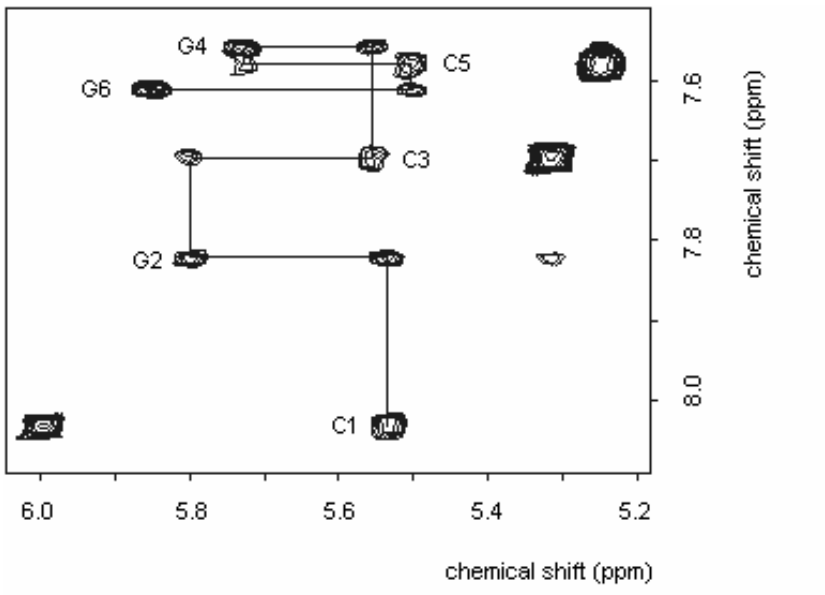


Figure 3.2.3. NOE pathway drafted in region [5–6]×[7–8] of 2D–NOESY spectrum of r(CGCGCG)$_2$ (Popenda et al., 1997).

Generally, NOE pathway can be constructed only on the basis of the information generated during NMR experiment for each registered NOE interaction: two coordinates of cross-peak centre given in ppm, widths in both dimensions and the value of signal intensity. Let us call them the *spectral data*. However, in practice, a collection of additional

information is used. In the thesis, they will be called the *supplemental data*. The supplemental data concern the analyzed structure and the obtained spectral information, some of them come from the other NMR experiments. The first group includes molecule sequence (i.e. its primary structure, which is known) and NOE pathway length (i.e. a number of cross-peaks in the pathway). Usually, the length is calculated on the basis of the sequence and equals $2N-1$, where N is a number of nucleotides in the chain (i.e. number of one-letter abbreviations used to code the sequence). However, RNA chain may be disjunctive and then the shorter pathway is looked for. The second group of supplemental information helps in an interpretation of the spectral data and concerns: volume intervals, spectrum resolution, overlapping signals, doublets, additional signal rejection, pathway potential starting points, known signal positions in the pathway, H5–H6 interactions. The meaning of several of these data is in some cases crucial for a correct NOE pathway construction, while sometimes it only reduces the search space and advances the construction process.

In many cases there is no difference between volume ranges of intra and inter cross-peaks or these ranges are not separable. However, if they do not overlap, it is useful to define them, which automatically differentiates the cross-peaks and divides their set into the subsets of inter and intranucleotide signals.

It may be also useful to define the divergence value, which depends on the digital resolution of a spectrum in both dimensions. This parameter is involved in the deviation of cross-peaks coordinates within the specified range, if there are problems with connecting the appropriate peaks because of some errors in the spectrum.

One of the hardest spectral problems is signal overlapping, i.e. positioning of many signals on the same line, while only two cross-peaks with the same coordinate are accepted to form a connectivity within the pathway. The problem appears when long RNA chains are studied (the longer is a chain the more crowded is a spectrum and more signals overlap). Thus, condition 6 from Definition 3.2.1 should be omitted during pathway construction. Usually, the fragment with overlapping signals can be specified by giving chemical shift range.

Nonequivalent nuclei with non-zero spins can sometimes generate signals, which split into doublet, triplet, quartet etc. This means, that one signal is represented by two (in case of a doublet), three (in case of a triplet) or more cross-peaks in the spectrum. These cross-peaks are positioned very close to one another and should be interpreted as one. Thus, in such cases it is necessary to define a distance between cross-peaks, which

should be treated as a multiple representation of one signal. Consequently, their coordinates, widths and volumes are used to calculate average values which are next assigned to a single representation of the signal.

In some cases, basing on the spectrum examination, one can distinguish cross-peaks representing signals discriminated during NOE pathway construction (i.e. not generated by H1', H6, H8 protons). If these cross-peaks can be determined they are next rejected during traversing the search space.

Sometimes, it is easy to point out a priori the expected positions of some cross-peaks within the NOE pathway. This concerns especially the starting and the ending point in the pathway arrangement.

The last supplemental information concerns H5–H6 signals, which are placed in the analyzed region of the NOESY spectrum. These interactions are not involved in forming the NOE pathway and must be discriminated by the pathway construction procedure. However, their placement can determine positioning of some other cross-peaks within the pathway. For the molecules including citidine and/or uridine, every cross-peak representing intranucleotide signal generated by citidine/uridine protons has one coordinate equal to one cross-peak representing H5–H6 interaction. Thus, on the basis of molecule sequence and the determined H5–H6 signals one can verify the correctness of a constructed NOE pathway.

For each RNA simplex and self-complementary RNA duplex one NOE pathway only exists in the spectrum. For noncomplementary duplexes two NOE pathways exist. They will be called the *original pathways*. *Constructing an original pathway is the main goal of the discussed problem*.

When the pathway is found, it should be compared with the sequence of an analyzed molecule. Upon this, each cross-peak of a pathway should be assigned to a pair of protons, that generated it. Thus, the exact chemical shifts of the protons involved in the pathway are recognized. This last step of sequence-specific assignment procedure proceeds in the following way. In an ideal case, each i–th odd cross-peak of the pathway sequence represents intranucleotide interaction between protons of i–th nucleotide (for i=1,2,…,N). Each i–th even cross-peak represents internucleotide interaction between a proton of i–th nucleotide and a proton of (i+1)–th nucleotide in the chain (for i=1,2,…,N). Thus, for example, having ACU chain analyzed and a pathway s_1-s_2-s_3-s_4-s_5 found we assign: s_1 is generated by A1:H1' (meaning H1' proton of A nucleotide, being first nucleotide in a sequence) and A1:H8, s_2 is generated by A1:H1' and C2:H6, s_3 is

generated by C2:H1' and C2:H6, s_4 is generated by C2:H1' and U3:H6, s_5 is generated by U3:H1' and U3:H6. Next, signal coordinates are taken and chemical shifts of the protons can be appointed. For example, if s_1 coordinates are (5.09,7.6) then it is obvious that H1' proton of A has a chemical shift of 5.09 ppm and H8 proton of A has a chemical shift of 7.6 ppm (see Table 3.2.1 for information of chemical shifts ranges for different hydrogen protons). In such a way, chemical shifts of all the analyzed protons are appointed and the first draft of molecule tertiary structure can be suggested (Case, 1998).

III.3 EXISTING METHODS OF NOE ASSIGNMENT

At present, automatization of NMR spectra analysis makes the strong impact on protein structures determination (Moseley and Montelione, 1999). This concerns also assignment step, which has been performed manually for a long time. Recent years have brought a development of different programs for automatic assignments of protein spectra (Bartels et al., 1997; Croft et al., 1997; Lukin et al., 1997; Mumenthaler et al., 1997; Zimmerman et al., 1997; Leutner et al., 1998; Atreya et al., 2000; Moseley et al., 2001; Linge et al., 2003; Balley-Kellog et al., 2004; Langmead et al., 2004). Some of these programs aim at nucleic acids assignment. However, in practice, they are not useful in RNA structural study at this stage. Thus, for short DNA and RNA duplexes the assignment is still performed manually in accordance with the experimenter's knowledge and intuition. Graphic interactive methods have slightly eased this manual procedure (Kraulis, 1989). However, no longer than ca 15-nucleotide chains can be analyzed and assigned in this way. For longer oligonucleotides, the spectra become overcrowded and cross-peaks often cover up the others. Thus, due to a considerable large number of signals and their overlapping, the manual assignment step becomes very troublesome.

To the author's knowledge one proposition of automatic procedure dedicated to RNA assignments has appeared (Roggenbuck et al., 1990). The proposed algorithm was based on backtracking (BT) and reduced adjacency matrix (RAM) procedures. However, no experimental results, except for one self-complementary octamer duplex, were reported. A number of alternative paths generated was high and the algorithm has not been practically developed.

Thus, since introducing automatic procedures for RNA assignments remains a crucial necessity, the new methods are proposed in the thesis. They are based on a graph model of an assignment problem that is described in Chapter 4. The first algorithm enumerates all feasible paths found in the graph that represents NOESY spectrum. The next heuristics optimize solution parameters and try to generate the most valuable solution. All the algorithms are presented in Chapter 6.

IV. NEW MODEL OF THE NOE PATH CONSTRUCTION PROBLEM

Respecting the biochemical description of the assignment problem, presented in Section 3.2, a graph model of the problem is proposed in this chapter. It reduces the NOE pathway (Definition 3.2.1) reconstruction to a variant of the Hamiltonian path problem. The model formulates a background for the complexity analysis and for the construction of algorithms for automatic assignment of resonance signals. It takes into account the specificity of the required connectivity between consecutive proton signals in the NMR spectrum. The model corresponds to the basic problem of NOE pathway defined by the spectral data (see Section 3.2). Since the complete set of the supplemental data is not always needed it will not be considered in model construction.

The process of sequential assignments of H6/H8–H1' corresponds to a construction of a path between vertices of a graph. Thus, converting 2D–NOESY spectrum to a certain graph structure seems to be an attractive idea. Since NOE pathway is drafted in aromatic/anomeric region of the spectrum, this region is considered during graph construction. Cross-peaks are obvious candidates for graph vertices. Possible connections, that can be suggested during NOE pathway reconstruction define edges of the graph. These connections must satisfy all the conditions described in Section 3.2. The following definition (Szachniuk et al., 2003; Adamiak et al., 2004; Blazewicz et al., 2004a; Blazewicz et al., 2005) characterizes a new type of a graph, which can be used to represent the selected region of NOESY spectra and NOE sequence properties (see Section 3.2 for NOESY spectrum and NOE pathway descriptions):

Definition 4.1 (*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 NOESY spectrum and hast the following properties of the corresponding cross-peak: the number, two coordinates and widths in two dimensions.
2. A number $|V|$ of vertices in graph G equals a number n of cross-peaks.

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 $V_0 = \{v_i: w_i = 0, i \in \{0,1,...,n\}\}$ and $V_1 = \{v_i: w_i = 1, i \in \{0,1,...,n\}\}$.
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 volumes having exactly one common coordinate) that can be drafted in the spectrum.

Let us stress that, in general, NOESY graph G may not correspond to any particular NOESY spectrum S, and thus, NOE path may not exist in it. We will call this case a *theoretical* one (theoretical model of the problem). It will be showed that both, decision and search versions of this problem are computationally hard. On the other hand, if the NOESY graph G corresponds to a given spectrum S (in this case we will talk about *experimental* or *real* model), then in the ideal case (no errors) its decision version is trivial, i.e. NOE path (or better NOE pathway) must always exist. However, finding one is still a computationally hard problem.

For a given 2D–NOESY spectrum, one can construct a NOESY graph G_S corresponding to the spectrum aromatic/anomeric region S. Figure 4.1 shows the relationship between the [5–6]×[7–8] region of a NOESY spectrum of r(CGCGCG)$_2$ (Figure 4.1.a) and the corresponding NOESY graph (Figure 4.1.b) obtained according to definition 4.1.

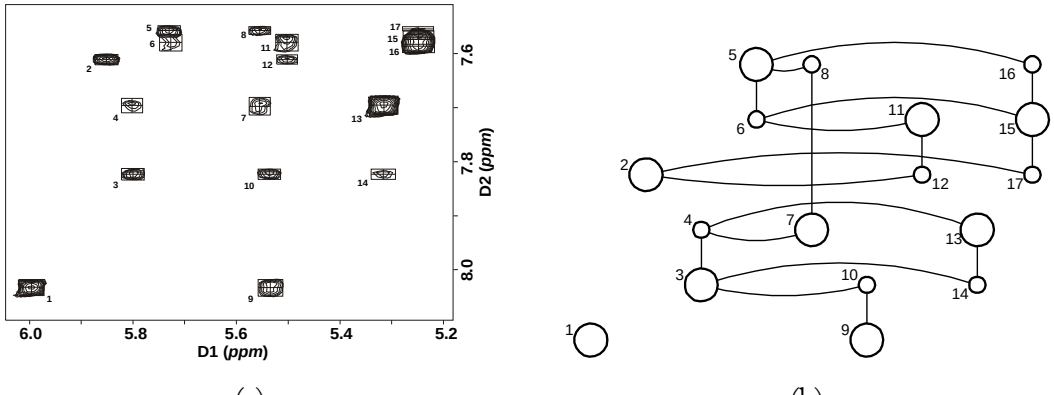


Figure 4.1. The relationship between aromatic/anomeric region of the NOESY spectrum of r(CGCGCG)$_2$ (a) and the corresponding NOESY graph (b).

In the example presented in Figure 4.1, spectrum S contains seventeen cross-peaks, and consequently, graph G_S contains the same number of vertices. Some cross-peaks lay so close to one another (see 5–6, 11–12, 15–16–17) that, for an inexperienced observer, they seem to be the single peaks. However, they are registered as different proton signals by peak-picking procedure of NMR software. Thus, there are nine intranucleotide resonances corresponding to nine vertices with weight 1 (displayed as big circles), and eight internucleotide resonances represented by eight vertices with weight 0 (displayed as small circles) in graph G_S. Edges of G_S correspond to all possible proper connections that can be drawn in spectrum S.

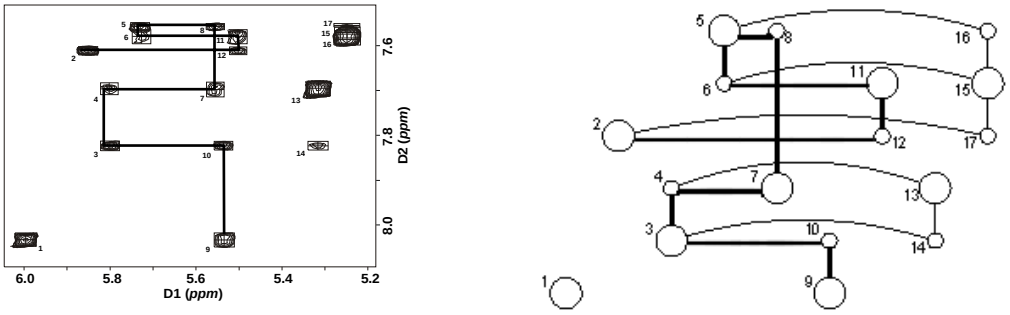
After converting spectrum S to graph G_S, an appropriate connectivity pathway can be looked for in G_S. This, however, requires a formulation of the NOE pathway problem in terms of graph theory. In order to differentiate an experimental and a theoretical problem, a notion of *NOE path* will be used with the reference to the graph theoretical model (NOESY graph which does not correspond to any specific spectrum and, thus, not necessarily contains the path), while the *NOE pathway* notion will still define H6/H8–H1' connectivity sequence in the spectrum, as well as in a graph representing the spectrum (i.e. an experimental model).

Definition 4.2 (*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 (Szachniuk et al., 2003; Adamiak et al., 2004; Blazewicz et al., 2004a):

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 are collinear.
6. A length of P_G equals $l=2|V_1|-1$.

NOE pathway P_S constructed in the 2D–NOESY spectrum S of RNA molecule is the solution of the assignment problem in the real model. NOE path P_G found in NOESY graph G_S corresponding to spectrum S, is the appropriate solution of the same problem. Figure 4.2 shows the relationship between the NOE pathway (Figure 4.2.a) found

in aromatic/anomeric region of the NOESY spectrum of r(CGCGCG)$_2$ and the corresponding path (Figure 4.2.b) in NOESY graph corresponding to this spectrum.



(a) (b)

Figure 4.2. Relationship between NOE pathway (a) in the spectrum of r(CGCGCG)$_2$ and the corresponding NOE path (b).

There is conformity between the problems of the NOE path and the Hamiltonian path in a graph. However, the first problem assumes additional constraints on the search space of the algorithms solving this problem: an edge can join only vertices having exactly one common coordinate (thus only horizontal and vertical edges are correct) and different weights, horizontal and vertical edges occur alternately in the path. These constraints have been considered in the proposed model of the problem, containing definitions of a NOESY graph and a NOE path. For some instances, additional information, that reduces the search space, is available and this is considered by algorithms during path correctness verification (Adamiak et al., 2004; Blazewicz et al., 2004a).

A problem of NOE path construction in the NOESY graph (a search problem), similarly to the corresponding spectral problem, is computationally hard. Its computational complexity is discussed in the next chapter.

V. COMPUTATIONAL COMPLEXITY ANALYSIS OF THE NOE ASSIGNMENT PROBLEM

This chapter is devoted to a presentation of a computational complexity of H6/H8–H1' (NOE) connectivity pathway construction in the theoretical as well as in the experimental model as discussed in Chapter 4. Classification of a problem into a proper complexity class helps in designing and verification of algorithms solving the problem. The basic question stated at the beginning of problem analysis is whether or not a problem is computationally intractable. This question will be answered in the following paragraphs of the chapter.

Let us remind that a notion of *NOE pathway* concerns the experimental model of the problem (searching the spectrum, searching a graph representation of the spectrum), while *NOE path* denotes the corresponding problem in the theoretical model (searching any NOESY graph). It will be 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. The appropriate proof is shown in Section 5.1. One can immediately try to classify decision and search versions of the corresponding real (experimental) problem of NOE pathway to the same complexity classes. However, a slight difference between theoretical and experimental models (in the latter NOESY graph G_S corresponds to a given spectrum S) appears crucial to negate such a classification general correctness. It makes a decision version of NOE pathway problem trivially easy. A search version, however, remains NP–hard. These proofs are given in Section 5.2.

V.1 COMPLEXITY OF THE PROBLEM IN THE THEORETICAL MODEL

In this section, the computational complexity of the NOE path problem is analyzed. The problem concerns a construction of the specific path, called NOE path (Definition 4.2), in a NOESY graph (Definition 4.1). NOE path problem corresponds to the real problem of NOE pathway reconstruction in the two-dimensional NOESY spectrum being the thesis subject, however, in the theoretical model, NOESY graph does not represent any specific spectrum and NOE path may not exist.

The following theorem establishes the hardness of both decision and search versions of finding the NOE path in the theoretical model.

Theorem 5.1.1 (NOE path complexity)
The problem of finding NOE path in a NOESY graph is NP–hard.

Proof
Since each search problem has its decision version, which is not computationally harder, the decision version of the problem will be used in the proof of Theorem 5.1.1.

Let us denote the search problem of finding NOE path by Π_{search}, and its decision version – by Π_{dec}. Formulations of both versions of the problem will be based on Definition 4.1 and Definition 4.2, respectively:

Definition 5.1.1 (NOE path construction – search version: Π_{search})
Instance: NOESY graph $G=(V,E)$.
Goal: Find (construct) NOE path in G.

Definition 5.1.2 (NOE path construction – decision version: Π_{dec})
Instance: NOESY graph $G=(V,E)$.
Question: Does G contain a NOE path?

The proof of problem Π_{search} strong NP–hardness will be based on demonstrating that its decision version Π_{dec} is strongly NP–complete. A sketch of the standard proving procedure is composed of the following steps:

1. By showing an algorithm on NDTM solving Π_{dec} in a polynomial time one proves that $\Pi_{dec} \in$ NP.
2. Select a known NP–complete problem Π'_{dec} .
3. Transform the selected problem Π'_{dec} to problem Π_{dec} .
4. Prove the transformation is a polynomial transformation.

Let us start with the first step of the above sketch. A nondeterministic algorithm is simple. It only needs to guess an ordering of the vertices and check in polynomial time whether all the conditions from Definition 4.2 are satisfied. The guessing module of NDTM generates random sequences of the vertices, which are next verified (in parallel) in polynomial time

by a deterministic module. The latter module provides an answer "yes" or "no" depending on the sequence correctness.

Selecting a problem Π'_{dec} transformable to Π_{dec} seems obvious. In what follows, Π'_{dec} will denote the Hamiltonian path problem, which is known to be strongly NP–complete (Karp, 1972; Garey and Johnson, 1979). It can be defined as follows.

Definition 5.1.3 (Hamiltonian path construction – decision version: Π'_{dec})
Instance: Graph $G_H=(V_H,E_H)$, $|V_H| = n$.
Question: Does G_H contain a Hamiltonian path, that is an ordering $<w_1,w_2, ..., w_n>$ of the vertices of G_H, such that $\{w_i,w_{i+1}\}\in E_H$ for all $i=1, 2, \ldots, n$?

It may be assumed that graph $G_H=(V_H,E_H)$ has no self-loops and no vertex with degree exceeding three and the problem remains strongly NP–complete (Garey and Johnson, 1979).

An arbitrary graph $G_H=(V_H,E_H)$, being an instance of the Hamiltonian path problem (restricted in the above sense), can be polynomially transformed to a NOESY graph $G=(V,E)$. 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 5.1.1 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}.

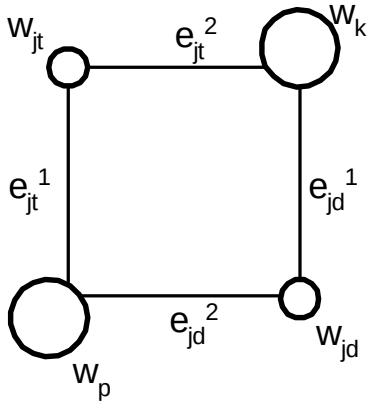


Figure 5.1.1. A square subgraph.

As a result of the proposed reduction of graph $G_H=(V_H,E_H)$, a NOESY graph $G=(V,E)$, consistent with Definition 4.1 is obtained. A set of vertices and a set of edges of the obtained graph G contain the following elements:

$$V = \bigcup_{i=1..|V_H|} v_i \cup \bigcup_{j=1..|E_H|} \{v_{jt}, v_{jd}\}, \tag{5.1.1}$$

$$E = \bigcup_{j=1..|E_H|} \{e_{jt}^{\ 1}, e_{jt}^{\ 2}, e_{jd}^{\ 1}, e_{jd}^{\ 2}\} \tag{5.1.2}$$

A cardinality of the vertex set and a cardinality of the edge set of G equal:

$$|V| = |V_H| + 2|E_H|, \tag{5.1.3}$$

$$|E| = 4|E_H|. \tag{5.1.4}$$

Figure 5.1.2 illustrates the following steps of the procedure transforming a simple graph G_H (Figure 5.1.2.a) into a corresponding NOESY graph G (Figure 5.1.2.e).

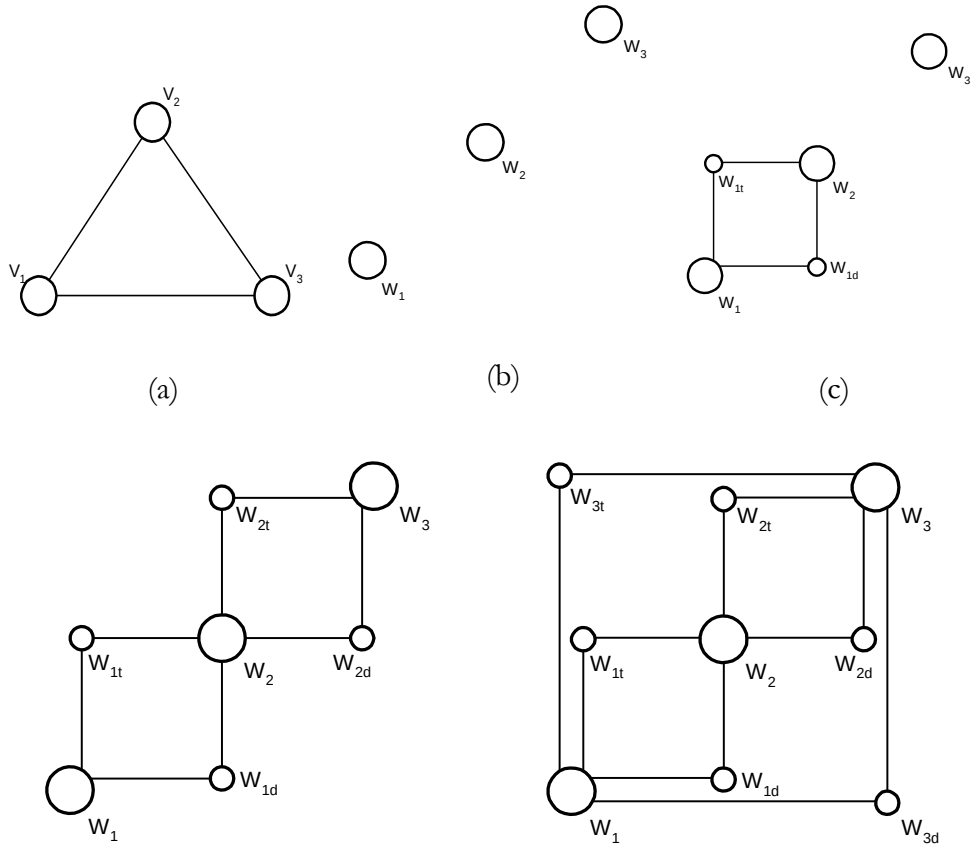


Figure 5.1.2. Transformation procedure: (a) input graph, (b)-(e) succeeding steps of a transformation to a NOESY graph (e).

To complete the proof of Theorem 5.1.1, it must be revealed that the proposed transformation procedure *R* is a polynomial time transformation, i.e. it is performed in a polynomial time and preserves the equivalence of the problem solutions before (Π'_{dec}) and after (Π_{dec}) the transformation.

Since the time used for the transformation procedure is bounded from the above by the input length of problem Π'_{dec}, it can be performed in time bounded by a polynomial function of the size of an input (i.e. number of vertices and edges of the transformed graph) and equals $O(n^2)$: for all n input vertices the same number of new vertices is created (complexity $O(n)$), and for every edge pair of vertices and edges is created (complexity $O(n^2)$ since in a complete graph number of edges equals $\frac{1}{2}(n-1)n$).

To prove problem solutions equivalence, it should be demonstrated that the solution for an instance of problem Π'_{dec} belongs to $Y_{\Pi dec'}$ (the answer is "yes") if and only if the solution for an instance of Π_{dec} also belongs to $Y_{\Pi dec}$. Thus, the following proposition truthfulness must be revealed:

Proposition 5.1.2 (*Hamiltonian and NOE paths equivalence*)
Graph $G_H=(V_H,E_H)$ contains a Hamiltonian path if and only if the corresponding NOESY graph $G=(V,E)$ contains a NOE path.

At the beginning, let us assume that graph $G_H=(V_H,E_H)$ contains a Hamiltonian path $P_H=w_{[1]},w_{[2]},\ldots,w_{[n]}$, $n=|V_H|$. For this path, we construct the corresponding path $P=v_{[1]},v_{[2]},\ldots,v_{[m]}$ ($m=|V|$) in graph $G=(V,E)$ resulted from the transformation procedure $R:G_H\rightarrow G$. Let us prove, that path P satisfies conditions 1–6 from Definition 4.2. From the transformation it has been known that for every vertex $w_i\in V_H$ (for $i\in\langle 1,n\rangle$) in graph G_H, there exists exactly one vertex $v_i\in V_1$ in graph G. Consequently, the first vertex in P is the one with weight equal to 1: $v_{[1]}\in V_1$ and condition 1 is satisfied. By definition, the Hamiltonian path P_H satisfies condition 2 (i.e. it is a simple path), thus, the corresponding NOE path P in graph G also satisfies this condition. Let us now discuss condition 3. It can be observed that each vertex $v_i\in \mathrm{V}_1$, $i\in\langle 1,m\rangle$ in graph G is adjacent to vertices from V_0 subset only, and each vertex $v_i\in \mathrm{V}_0$, $i\in\langle 1,m\rangle$ in graph G is adjacent to vertices from V_1 subset only. Therefore, vertices with weight 1 (belonging to V_1 subset) and vertices with weight 0 (belonging to V_0 subset) appear alternately in P and condition 3 is satisfied. Since graph G construction allows to choose one of two possible traversals between every two vertices of G and G structure contains only horizontal and vertical edges, it is always possible to construct NOE path P taking succeeding edges perpendicular. This makes condition 4 satisfied. From the procedure of transformation of graph G_H to graph G it is evident that NOE path P with properties 2 and 4 will also satisfy condition 5. Finally, let us now consider condition 6. Assume that length of a path in a graph is measured as a number of graph vertices occurring in the path. Consequently, a length of the Hamiltonian path P_H in graph $G_H=(V_H,E_H)$ equals $|V_H|$. From the transformation procedure it is evident, that $|V_1|=|V_H|$, where V_1 is a subset of vertices of G with weight equal to 1. It is also known, that each edge of Hamiltonian path P_H corresponds to a structure containing two edges and one vertex with weight equal to 0 in an appropriate NOE path P. Hamiltonian path includes $|V_H|-1$ edges and so, the corresponding NOE path will contain additional number $|V_1|-1$ of vertices with weight 0. Thus, the length of NOE path P in graph G corresponding to Hamiltonian path P_H in graph G_H equals $l=|V_1|+|V_1|-1=2|V_1|-1$. This way, satisfiability of the last condition (6) has been

proved and the first part of the proof has been completed.

In the above paragraph, it has been proved that if graph G_H contains Hamiltonian path P_H, then corresponding graph G contains NOE path P obeying properties 1–6 from Definition 4.2. Now, let us show, that the reversed statement is true as well.

At this point, assume that graph $G=(V,E)$ contains a NOE path satisfying conditions 1–6. For every vertex $v_i \in V_1$ in graph G, there exists exactly one vertex $w_i \in V_H$ in graph G_H. Additionally $|V_H|=|V_1|$. Thus, if graph G contains NOE path P which includes all the vertices $v_i \in V_1$, then in graph G_H there exists a path covering all the vertices $w_i \in V_H$. Moreover, if NOE path in G satisfies condition 2 (the path is simple), then – following a procedure of G construction – one may say that the corresponding path in graph G_H also satisfies this condition. Summing up, it may be claimed that the corresponding path in graph G_H is a Hamiltonian path.

In the above paragraph, it has been proved that if graph G contains NOE path P obeying properties 1–6 from Definition 4.2, then the corresponding graph G_H contains Hamiltonian path P_H. Consequently, we observe that NOESY graph G contains a NOE path *if and only if* the corresponding graph G_H contains a Hamiltonian path. In this way, it has been proved that Proposition 5.1.2 of Hamiltonian and NOE paths equivalence is true.

Since it has been proved, that problem of finding NOE path Π_{dec} is in NP and $\Pi'_{dec} \propto \Pi_{dec}$, where Π'_{dec} is an NP–complete Hamiltonian path problem, it is obvious that Π_{dec} is also NP–complete. In the sense of computational complexity, the decision version of any problem is known to be not computationally harder than the search (and optimization) version of the same problem. Thus, the search version of NOE path construction Π_{search} is at least as hard as the decision problem of NOE path construction Π_{dec} and it can be classified as an NP–hard problem. By proving this, the truthfulness of Theorem 5.1.1 has been shown. ■

We see that computational intractability of NOE path problem in the theoretical model has been proved. Usually, this finishes the discussion on computational complexity of the analyzed real problem, as the theoretical and experimental models have similar properties and complexity. However, it is not true in case of NOE path problem. A slight difference between the real and theoretical problems makes quite a big difference between computational complexity of their decision versions. Thus, the following section continues the discussion on computational complexity concentrating on the experimental

model.

V.2 COMPLEXITY OF THE PROBLEM IN THE EXPERIMENTAL MODEL

This section is devoted to complexity analysis of the NOE pathway construction being one of the first steps of RNA tertiary structure determination and the thesis subject. The problem concerns a construction of the specific path, called NOE pathway (Definition 3.2.1), in a NOESY spectrum (see Section 3.2 for a description). Complexity discussion is typically based on some theoretical model of the problem in question. The appropriate model for NOE pathway problem has been designed and introduced in Chapter 4. Section 5.1 has been centered on an examination of computational complexity of finding NOE path in a general NOESY graph, not necessarily corresponding to a spectrum. However, apart from the fact that both, theoretical and experimental problem, correspond to each other, they are not completely the same in the sense of computational complexity. The following paragraphs concentrate on the real problem of finding NOE pathway in a spectrum (or a NOESY graph corresponding to the spectrum).

Atoms of every RNA molecule placed in the strong magnetic field provided by NMR spectrometer, participate in inter- and intranucleotide interactions called NOE correlation signals. These signals are registered and pictured in a NOESY spectrum. Having such a spectrum one knows that it contains cross-peaks which –when identified and connected in a specific way – form a path, called NOE pathway. In non-ideal cases several cross-peaks may be missing for some reason and then a construction of the whole NOE pathway is not possible. Since such situations are quite rare and do not play crucial role in problem analysis, only ideal cases will be considered in this section. The question "whether or not NOESY spectrum *S* contains a NOE pathway?" will be always answered "yes". The decision problem of NOE pathway is then trivial, while – as one can remember – its version in the theoretical model is NP–complete (see Section 5.1).

Let us denote the experimental NOE pathway construction problem by Π_E, its search version – by $\Pi_{Esearch}$, and its decision version – by Π_{Edec}. A formulation of the latter version of the problem is based on the Definition 3.2.1 of NOE pathway.

Definition 5.2.1 (NOE pathway construction – decision version: Π_{Edec})
Instance: Aromatic/anomeric region *S* of the 2D–NOESY spectrum.
Question: Does *S* contain a NOE pathway?

We have the following theorem.

Theorem 5.2.1 (NOE pathway existence complexity)
The problem of NOE pathway existence in a NOESY spectrum is trivially easy.

Proof
To see this let us note that experimental NOE pathway construction problem Π_E belongs to a class of promise problems, i.e. it consists of an instance domain, a question and a promise of solution existence. According to this promise, each instance of the NOE pathway problem Π_E has at least one solution (since for an ideal spectrum coming from an experiment at least one NOE pathway must exist). In fact, $D_{\Pi_{Edec}}=Y_{\Pi_{Edec}}$, where $D_{\Pi_{Edec}}$ is a set of all instances of the decision problem Π_{Edec} and $Y_{\Pi_{Edec}}$ is a set of all instances of Π_{Edec} with a "yes" answer. Thus, the computational complexity of the above problem is polynomial and the Theorem 5.2.1 is true.■

Promise of solution existence, however, does not necessarily influence the computational complexity of problem's search version, which is defined below:

Definition 5.2.2 (NOE pathway construction – search version: $\Pi_{Esearch}$)
Instance: Aromatic/anomeric region *S* of 2D–NOESY spectrum.
Goal: Find a NOE pathway in *S*.
Promise: *S* contains at least one NOE pathway.

Searching the NOESY spectrum in order to determine the NOE pathway seemed to be a hard problem and thus, the following theorem is proposed:

Theorem 5.2.2 (NOE pathway construction complexity)
The problem of finding NOE pathway in a NOESY spectrum is NP–hard.

Proof
It is evident, that $D_{\Pi_{Esearch}}= D_{\Pi_{Edec}}$. In order to prove that $\Pi_{Esearch}$ is in NP–hard class of search problems, the general variant of NOE pathway will be introduced. This problem, called quasi-NOE pathway construction is defined as follows:

Definition 5.2.3 (quasi-NOE pathway construction – decision version: Π_{Qdec})
Instance: Any NOESY graph $G=(V,E)$ such that $|V|=m$, $Y_{\Pi_{Qdec}} = Y_{\Pi_{Edec}}$ and $D_{\Pi_{Edec}} \subset D_{\Pi_{Qdec}}$.
Question: Does G contain a NOE pathway?

Let us notice, that problem Π_{Qdec} is more complicated that Π_{Edec}, since apart from the set of yes-instances, the same as for Π_{Edec} ($Y_{\Pi_{Qdec}} = Y_{\Pi_{Edec}}$), it contains the instances with an answer "no" (i.e. $D_{\Pi_{Qdec}} = Y_{\Pi_{Qdec}} + N_{\Pi_{Qdec}}$). Thus, quasi-NOE pathway Π_{Qdec} is not a promise problem. It is also true that Π_{Qdec} belongs to NP–complete class. Its computational intractability results from the NP–completeness of the theoretical problem of NOE path Π_{dec} (Definition 5.1.2) which has been proved in Section 5.1. NOESY graph can or cannot contain NOE path, thus, the domain of Π_{dec} consists of both, yes-instances and no-instances, similarly to problem Π_{Qdec}. Therefore, NP–completeness of NOE path problem Π_{dec} implicates the NP–completeness of quasi-NOE pathway problem Π_{Qdec}.

From the computational complexity of Π_{Qdec} it follows, that $\Pi_{Esearch}$ is NP–hard search problem. Let us prove this implication analogously to (Johnson, 1985). Assume, there exists an algorithm A that, given a NOESY spectrum (a NOESY graph representing a spectrum) containing NOE pathway (problem $\Pi_{Esearch}$), guarantees to find one in polynomial time. Having such an algorithm we could use it to tell in polynomial time whether an arbitrary NOESY graph has NOE pathway (problem Π_{Qdec}). Let p denote the polynomial that bounds A's running time on graphs with NOE pathways. Apply algorithm A to an arbitrary NOESY graph G. If G contains NOE pathway, A will find one in time $p(|G|)$. If G does not have such a pathway, then after $p(|G|)$ steps A cannot have found one, and we will know that none exists. Thus, problem Π_{Qdec} could be solved in polynomial time by algorithm A. Since, it has been proved that Π_{Qdec} is NP–complete, such an algorithm A cannot exist. Thus, there is no algorithm solving problem $\Pi_{Esearch}$ in polynomial time (unless P=NP) and, consequently, constructing NOE pathway in the graph representing NOESY spectrum, and the same – constructing NOE pathway in the spectrum ($\Pi_{Esearch}$) is the NP–hard problem.■

At this point, the discussion on computational complexity of NOE pathway problem in experimental model is finished. It has been shown that the decision version of the problem is polynomially solvable but constructing a solution remains hard, thus, making a search variant computationally intractable. Hence, no polynomial time algorithm is likely to exist for the problem of NOE pathway construction in the spectrum. Consequently,

two approximate algorithms will be proposed in the next chapter solving an optimization variant Π_{Eopt} of the problem in question. The latter is defined as follows:

Definition 5.2.4 (NOE pathway construction – optimization version: Π_{Eopt})

Instance: NOESY graph G representing an aromatic/anomeric region S of 2D–NOESY spectrum.

Goal: In G find a NOE pathway, that is an ordering $<v_1,v_2, ..., v_n>$ of the vertices of G, such that $\{v_i,v_{i+1}\} \in E$ for all $i \in \langle 1,n)$ and:

1. $v_1 \in V_1$,
2. every vertex $v_i \in V$ and every edge $e_j \in E$ of G occurs in the path at most once,
3. vertices with different weights appear alternately in the path,
4. every two neighboring edges of the path are perpendicular,
5. no two edges of the path are collinear,
6. length l of the path is maximum.

Since the experimental (spectral) data prepared for the first tests has been generated for short RNA chains and constructed NOESY graphs have been *sparse graphs* (in a sparse graph the number of edges is much smaller than the number of vertices squared), an exact algorithm, enumerating all feasible solutions has been designed as well and will be used for test purposes. Chapter 6 will be devoted to a presentation of all the mentioned algorithms, while in Chapter 7 the results of computational experiments will be shown.

VI. ALGORITHMS FOR NOE PATHWAY CONSTRUCTION

NOE pathway construction in the 2D–NOESY spectrum of RNA molecules has appeared to be NP–hard search problem (see Section 5.2). Even the promise of solution existence (as it is a priori known, that a spectrum contains NOE pathway) has not changed computational complexity of the problem search version. NP–hardness of the problem enforces a way of its algorithmic solution. In order to obtain an exact algorithm one should apply a branch and bound method running in time bounded by an exponential function. When an approximate solution of the problem is satisfactory, heuristics based on tabu or genetic search, performing in polynomial time, can be used. All of these approaches will be presented in the chapter, while the results of their application to the real NMR spectra will be given in Chapter 7.

Due to the suggested theoretical model of the problem (see Chapter 4), all the presented algorithms will use a graph representation of the spectrum being the input data. A description of input files with spectral and supplemental data, used for graph construction, will be described in Section 6.1. Both proposed heuristics, tabu search and genetic algorithm, use the same optimization criteria in order to find optimal solution. These criteria, as well as a description of solution feasibility will be given in Section 6.2. Section 6.3 of the chapter will be devoted to an exact algorithm enumerating all feasible solutions of the problem. In Section 6.4 the tabu search method will be presented. Finally, the second heuristic, based on the genetic approach, will be introduced in Section 6.5.

VI.1 INPUT DATA AND ITS PREPROCESSING

In this section the description of input data files, used by the algorithms, will be presented. The data are processed by procedures generating a search space for the NOE pathway reconstruction. Search space characteristics and construction process will be also described in the section.

Let us recall, that the input data can be divided into two sets of, respectively, the spectral and the supplemental data (Section 3.2). The first set, which is defined for each instance of the problem, contains information about every cross–peak in the analyzed spectrum region. The supplemental data are not always demanded, but when defined, they must be considered. Giving supplemental data one can define: analyzed molecule sequence, NOE pathway expected length, inter and intra volume intervals, spectrum

resolution, region with overlapping signals, distance between cross-peaks of a doublet, additional signals rejection, pathway potential starting points, known signal positions in the pathway, H5–H6 interactions. A selection of these data is considered when the *search space* for the problem of NOE pathway construction is defined.

Spectral data are included in a text file **.list* generated by Accelrys FELIX software from the 2D–NOESY spectrum after peak-picking procedure. The file contains the following information about each cross-peak: its number (*No*), two coordinates of the cross-peak center (*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 spectrometer frequency, which is helpful in converting units (ppm to Hz). Figure 6.1.1 illustrates an example of input **.list* file.

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

Figure 6.1.1. An exemplary input file rcgcgcg.list with spectral data.

Supplemental data about the spectrum and NOE pathway that contain the domain expert knowledge and are consequently used to extract correct pathways, are 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 pathway construction;
- section <SEQUENCE> includes the sequence of RNA (both strands in case of non-selfcomplementary duplexes);
- in section <PATH_LENGTH>, a number of cross-peaks in the expected NOE pathway 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 of the pathway;
- 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 pathway.

Information given in the sections <VOLUMES> through <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 to reduce the number of potential pathways in the solution set. An exemplary supplemental data file is listed in Figure 6.1.2.

```
<VOLUMES>
INTER: 0.035 0.2
INTRA: 0.035 0.2
</VOLUMES>

<RESOLUTION>
</RESOLUTION>

<OVERLAPPING>
</OVERLAPPING>

<SEQUENCE>
CGCGCG
</SEQUENCE>

<PATH_LENGTH>
11
</PATH_LENGTH>

<H5-H6_SIGNALS>
1 13 15
</H5-H6_SIGNALS>
```

Figure 6.1.2. An exemplary input file rcgcgcg.inf with supplemental data.

Before any of the presented algorithms starts computation all the available information is looked through to generate an appropriate graph representation (i.e. set of vertices V and set of edges E) of the problem. First set of vertices V is constructed. All the cross-peaks from the given spectrum are represented as vertices in the set. Each of these vertices is described by its number, two center coordinates, two widths and a volume, which are taken from a description of the corresponding cross-peak. Next, if some cross-peaks have been indicated for rejection (specified in section <REJECT_SIGNALS>), the corresponding vertices are removed from V. Finally, if the spectrum contains doublets and the distance between cross-peaks forming a doublet have been determined (in section <DOUBLETS>), all the doublets are identified in the spectrum. For each doublet in the spectrum, there exist two vertices in the vertex set V. Let us call them *doublet vertices*. These vertices are removed from the set and one vertex replaces them. The new vertex is assigned coordinates, widths and volume computed out of the appropriate parameters of the two doublet vertices: new coordinates as well as new volume are calculated as an average of the old ones, new widths are defined so that their range covers the whole of both doublet vertices. A number of the new vertex is usually not a number exactly, but a composition of two numbers of doublet vertices written as a pair. With these operations, creation of the vertex set V is completed. Next, an edge set E should be generated on the basis of V. Again, selected supplemental data as well as the definition of the NOE pathway (Definition 3.2.1) serve as an instruction for edges generation. The procedure takes every

pair of vertices v_i, v_j from the vertex set V, i, j=1..n, and creates an edge (v_i,v_j) in E if the following conditions are satisfied:

- v_i and v_j have exactly one coordinate in common (edge will be horizontal or vertical),
- if spectrum resolution is defined, exactly one coordinate of v_i is within the error range of one coordinate of v_j,
- if volume intervals are defined, v_i and v_j have volumes from different intervals (one is intra and one is inter).

Consequently, the edge set for the given instance of the problem, as well as the vertex set are defined. An availability of the supplemental information on H5–H6 signals in the spectrum (section <H5–H6_SIGNALS>) raises a necessity of storing the information about these signals in a separate structure. Thus, let us define one more set of objects denoted by $V_{\text{H5–H6}}$. All the H5–H6 cross-peaks from the given spectrum are represented as objects in $V_{\text{H5–H6}}$. Each of these objects is described by its number, two center coordinates, two widths and a volume, which are taken from a description of the corresponding H5–H6 cross-peak. With this, creation of the search space is completed and algorithms can start path construction process.

In the section, all the aspects of generation of the search space for algorithms constructing NOE pathways has been presented. Solutions that are a subject of our interest should satisfy some predefined feasibility conditions. These conditions will be given in the following section of the chapter. This section will also contain a presentation of optimization criteria used by heuristic algorithms. Algorithms themselves will be described in sections 6.3 through 6.5.

VI.2 FEASIBILITY AND OPTIMIZATION CRITERIA

In the chapter, automatic methods of NOE pathways construction in 2D–NOESY spectra generated for RNA molecules are discussed. The first proposed method, based on exact algorithm, enumerates all solutions being feasible paths. This feasibility property will be defined in the first part of this section. Since it has been proved (see Chapter 5), that NOE pathway construction is NP–hard problem, two heuristics solving it have been designed as well. Both use the same goal function, which serves evaluation of the generated solutions. The function will be also described in the section.

Let us start from the description of the feasible pathway notion, which will be used in the algorithms' design. As it has been mentioned, exact algorithm (see Section 6.3) will be looking for all feasible pathways. The number of pathways consistent with the NOE pathway definition (Definition 3.2.1), that can be found in the spectrum as well as their lengths depend on RNA tertiary structure and some properties of the acquired NOESY spectrum (doublets, overlapping, missing signals influence this in a large degree). Computational analysis has shown that the number of all NOE pathways in the NOESY graph reaches $2^{-(n-3)} \cdot n!$ for $n>2$, where n denotes the number of graph vertices (Adamiak *et al.*, 2004). These pathways will be called feasible solutions to the problem. Only one (or at most two) of them will correspond to the original pathway (see Section 3.2), which is correct from the biochemical point of view and is looked for in the spectrum. More formally, the feasible solution can be defined in the following way:

Definition 6.2.1 (*feasible NOE pathway*)
Let $P_F=v_1, v_2, \ldots, v_l$ be a sequence of vertices of the NOESY graph $G=(V,E)$ representing spectrum S for RNA molecule M. We will call P_F the *feasible NOE pathway* in G, if the following conditions are satisfied:

1. every vertex and every edge of G occurs in pathway P_F at most once,
2. every two neighboring edges of P_F are perpendicular,
3. no two edges of P_F are collinear,
4. if some vertex positions within the pathway are predefined, P_F contains these vertices as specified,
5. if $V_{H5-H6} \neq \varnothing$ (i.e. H5–H6 signals in S are specified) and M contains citidine and / or uridine, every object from V_{H5-H6} has exactly one coordinate in common with one vertex representing citidine / uridine intranucleotide signal in G.

Each pathway generated by any of the proposed algorithms is tested according to feasibility conditions 1–5 from Definition 6.2.1. If the conditions are satisfied, the pathway is accepted as feasible solution. Such a solution is next returned by exact algorithm (Section 6.3). In 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 random factor also slightly influences the evaluation of solutions. It has been introduced in order to increase the probability of leaving the local optimum and differentiate solutions with the same scores. A feasible solution with the best score is defined as the *optimal solution*.

The global *criterion function f* has been defined as a weighted sum combining a set of different criteria:

$$f = \frac{1}{l}\left(\sum_{i=1}^{7} w_i y_i + r\right). \tag{6.2.1}$$

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

- $l \geq 1$,

  l denotes pathway length; if the length has not been predefined by the user then $l=1$;

- $r \in \langle 0, 0.0001 \rangle$,

  r denotes a random factor;

- $y_1 \in \{0, 1\}$,

  $y_1=1$ if the predefined starting cross-peak is not present on the first/last position of solution x,

  $y_1=0$ if starting cross-peak is not predefined or the predefined starting cross-peak is present on the first/last position of x;

- $y_2 = \sum_{i=1}^{l-1} a_{i,i+1}$, where $a_{i,i+1} \in \{0, 1\}$,

  $a_{i,i+1}=1$ if the i–th and the $(i+1)$–st cross-peaks have intensities in the same interval, $a_{i,i+1}=0$ otherwise;

- $y_3 = \sum_{j=2}^{l-2} b_{j-1,j,j+1}$, where $b_{j-1,j,j+1} \in \{0, 0.8, 1\}$,

  the value of $b_{j-1,j,j+1}$ depends on deviation of edges between j–th and $(j-1)$–st as well as j–th and $(j+1)$–st cross-peaks from horizontal/vertical position: $b_{j-1,j,j+1}=1$ if both $(j-1)$–st and j–th edges are horizontal or vertical, $b_{j-1,j,j+1}=0.8$ in three cases: if both $(j-1)$–st and j–th edge are double edges (an edge is called double if its vertices are located so close to each other and they are so large that could be connected by both horizontal and vertical edge), if $(j-1)$–st edge is vertical, j–th is double edge and $(j+1)$–st is horizontal edge, or if $(j-1)$–st edge is

horizontal, j–th is double edge and (j+1)–st is vertical edge, $b_{j-1,j,j+1}$=0 in the opposite cases;

- $y_4 = \sum_{j=1}^{l-1} \sum_{i=1}^{l-1} c_{ji}$, where $c_{ji} \in \{0, 1\}$,

  c_{ji}=1 if the j–th and the i–th edges are collinear,

  c_{ji}=0 otherwise;

- $y_5 = \sum_{i=1}^{l} d_i$, where $d_i \in \{0, 1\}$,

  d_i=1 if the i–th cross-peak does not correspond to any predefined H5–H6 cross-peak,

  d_i=0 otherwise;

- $y_6 = \sum_{i=1}^{l-1} e_{i,i+1}$,

  where $e_{i,i+1}$ has a value corresponding to an acceptable horizontal/vertical deviation of an edge between the i–th and the (i+1)–st cross-peak in the solution;

- $y_7 = n - l$.

In the above formulas n denotes a total number of cross-peaks in the considered region of the spectrum, l stands for the length of the current solution x.

Weighting factors in function f have been set to the following values: w_1=100000, w_2=10000, w_3=10000, w_4=10000, w_5=1000, w_6=1, w_7=1.

Optimization performed by heuristic algorithms, tabu search and genetic algorithm, means minimization of the global criterion function value. Thus, solution with minimum value of function f is returned as the *optimal pathway*, which, however, does not mean, that it is the original NOE pathway.

In the section, feasible and optimal solutions have been defined. Finding them is the goal of procedures used by enumerative, tabu and genetic algorithms. All of these algorithms will be presented in the following sections of the chapter.

VI.3 ENUMERATIVE ALGORITHM

Since introducing automatic procedures for RNA assignments, which start the process of tertiary structure determination with NMR, remains a crucial necessity, a new method is proposed in this section. It performs a complete search on the search space of the problem and enumerates all feasible NOE pathways in the graph representing an input

NOESY spectrum (see Chapter 4). The method is based on a Hamiltonian path construction procedure. It uses domain expert knowledge to introduce additional constraints that limit the search space to the reasonable proportions.

The algorithm starts computations from building a search space for the given instance. The detailed description of search space generation has been given in Section 6.1. Goal of the algorithm performance is listing of all the pathways, which satisfy feasibility conditions (Definition 6.2.1), i.e. all feasible pathways. A proposed enumerative algorithm builds NOE pathways from a chosen vertex $v_i \in V$ (i=1..$|V|$), 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. The current pathway is verified according to the feasibility conditions. If the conditions are satisfied and no more edges can be added, the pathway is remembered as a feasible solution in the solution set F. Afterwards, the algorithm goes back removing the edges from the current pathway and tries to add the other edges in place of the removed ones. An algorithm stops after looking through the whole search space. Finally, the solution set F is verified according to pathways inclusion. If one feasible pathway includes the whole of the other, shorter feasible pathway, the latter one is removed from F. The main procedure of the enumerative algorithm given in pseudo-code is as follows:

Enumerative Algorithm

```
1. read input spectral and supplemental data;
2. construct a set of vertices V;
3. construct a set of edges E upon the set of vertices V;
4. if (H5–H6 signals are determined) then construct a set of
   objects V_H5–H6;
5. for i:=0 to ⟨number of edges⟩ do
6. begin
7.    empty the stack storing current solution;
8.    current pathway := the i–th edge;
9.    if (current pathway satisfies condition 4 from Definition
      6.2.1) then
10.      begin
11.         stack ← current pathway;  //put  vertices of the i–th
            edge onto the stack
12.         current vertex := current pathway.last vertex;
13.         call procedure recursive search(current vertex);
14.      end;
15.end;
```

```
16. verify solution set and remove pathways included in other
    solutions;
17. return ⟨set of solutions⟩.
```

In step 13, *Enumerative Algorithm* executes the procedure *recursive search* finding a NOE pathway starting from the current vertex being the last vertex of current solution. At this step of the search process, current solution is a pathway composed of one edge only (the *i*–th edge), accepted as the first edge in the NOE pathway. Procedure *recursive search*, given in pseudo-code is presented below:

procedure recursive search (current vertex)

```
1. for k:=0 to ⟨number of edges⟩ do
2. begin
3.    current edge := the k–th edge;
4.    current solution := pathway stored in the stack;
5.    if (current edge ∉ current solution) and (current edge.first
      vertex = current vertex) then
6.    begin
7.       current pathway ← current edge;    //add current edge to
         the pathway
8.       if (current pathway satisfies conditions 2–5 from
         Definition 6.2.1) then
9.       begin
10.         current vertex := current pathway.last vertex;
11.         stack ← current vertex;
12.         call procedure recursive search(current vertex);
            //recursion
13.      end;
14.      remove current vertex from the stack;
15.   end;
16.  end;
17.  set of solutions ← current solution from the stack;
18. return.
```

Enumerative algorithm starts from reading the input information, which is next used for generating basic data structures: a set of vertices V, a set of edges E and a set of H5–H6 objects V_{H5-H6}. The appropriate data are placed in two input files – one storing spectral data and the other with supplemental data. A process of pathways construction is mainly

based on the edge set, generated upon the set of vertices, while pathways correctness is verified with the use of V_{H5-H6} and a collection of rules defining NOE pathway feasibility. Taking each single edge $e_i \in E$ (i=1..$|E|$) algorithm begins a new route in the search tree, trying to build a pathway starting with e_i. The following edges are added to the pathway with regard to feasibility conditions. If no more edges can be added in the route, it may be said that a leaf of the search tree has been achieved and the generated pathway is placed in the solution set. Then, algorithm traces back removing edges from the end of the pathway and trying to replace them with other unused edges. Thus, the new branches appear in the search tree. After tracing all the possible routes to the search tree leaves and back, the algorithm finishes the search. Finally, it verifies whether solution set includes pathways being parts of the other (longer) solutions and removes them. Solution set processed in this way is returned as an output of *Enumerative Algorithm*.

Let us now examine an example of algorithm performance for 2D–NOESY spectrum of r(CGCGCG)$_2$ molecule. An aromatic/anomeric region of this spectrum has been displayed in Figure 3.2.3.

NMR software (e.g. Accelrys Felix) produces a graphic (e.g. Figure 3.2.1) as well as a text version of spectral data. An appropriate text file *rcgcgcg.list* with spectral information for the considered example has been listed in Figure 6.1.1.

Since some additional expert knowledge has been available, file *rcgcgcg.inf* containing these supplemental data has been generated for the analysed example (Figure 6.1.2).

Both files, with spectral and supplemental data, are used by *Enumerative Algorithm* in steps 1 through 4 and in step 9, as well as by *procedure recursive search* in step 8. The detailed description of these files structure has been given in Section 6.1. Let us, now, say more about supplemental data provided for the considered instance of the problem, which have been placed in file *rcgcgcg.inf*. Volume intervals for inter– and intranucleotide signals have been defined as: 0.035 – 0.2. This will make *Enumerative Algorithm* reject four cross-peaks (with numbers: 1, 13, 15, 17) during construction of a vertex set. These cross-peaks do not represent the interactions of our interest (H6–H1', H8–H1') and so they should not be taken into the pathway. Additionally, these four cross-peaks may be also specified in section <REJECT_SIGNALS>. The difference between volumes of inter– and intranucleotide signals was hard to specify, thus, the intervals for both sets are equal. Since NOE pathway length is known for the instance (11 cross-peaks), it has been specified in section <PATH_LENGTH> of the file. Finally, H5–H6 signals have been enumerated in section <H5–H6_SIGNALS> and the RNA sequence has been given in section

<SEQUENCE>. The latter information instructs algorithm to accept the pathways consistent with the primary sequence, so that the peaks corresponding to citidine (these are: 1st, 5th and 9th signals in the pathway) 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.

Algorithm starts from constructing a set of vertices. Since four cross-peaks (1, 13, 15, 17) have been enumerated in section <REJECT_SIGNALS> of the supplemental file, vertex set is constructed out of the remaining thirteen vertices V={2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16}. Next, all possible edges are created: E={(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 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 starts searching for feasible pathways and finds 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

Since NOE pathway length has been defined, verifying procedure rejects all pathways that consisted of less than 11 cross-peaks (the longer pathways could not be found because the algorithm stops searching in the current direction if the pathway achieved the defined length) and two NOE pathways are 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 pathways are symmetrical, so only one of them is correct from the biochemical point of view. The information about H5–H6 cross-peaks given in section <H5–H6_SIGNALS> of *rcgcgcg.inf* file can help to choose the right NOE pathway. Thus, *Enumerative Algorithm* verifies pathway consistency with 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 pathway is consistent, so it is returned as the only solution of the considered instance:

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

Figure 6.3.1 illustrates the above pathway 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 *rcgcgcg.inf* file and one can see that they have the same D2 coordinates as citidine signals, respectively: 9,7,11 within the NOE pathway.

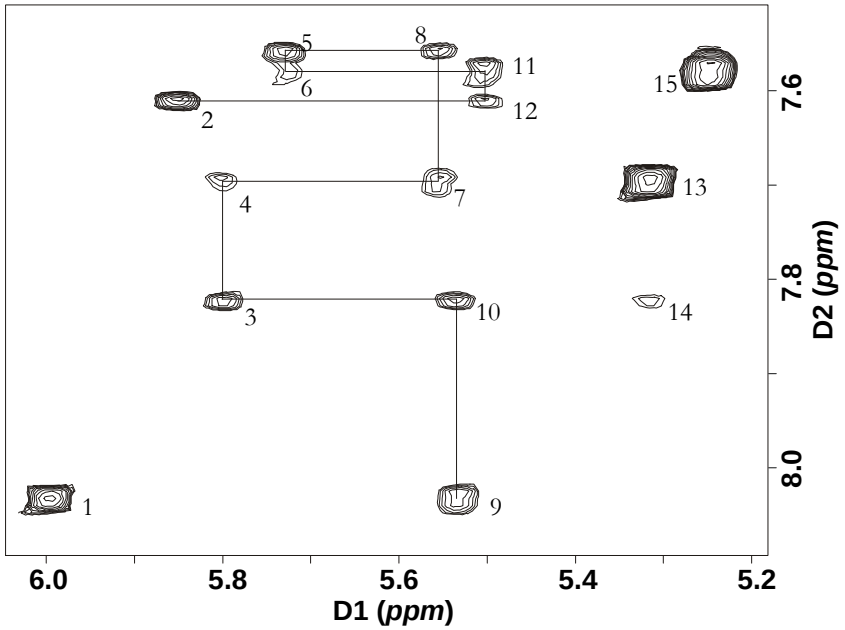


Figure 6.3.1. NOE pathway found in the spectrum of r(CGCGCG)$_2$.

The solutions in the simple form, i.e. arrangement of the vertices (like in the above example) are written to file *pathways.out*. Additionally, the program creates detailed assignment files with solutions. Figure 6.3.2 shows such a file for the analysed example.

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*

Figure 6.3.2. Assignment file for r(CGCGCG)$_2$.

In the section, an enumerative algorithm constructing all the feasible NOE pathways in the graph representation of 2D–NOESY spectrum for RNA molecules has been presented. Since the algorithm performs a complete search through the search space of the NP–hard problem of NOE pathways, the algorithm complexity is exponential. However, in case of an analysis of short RNA chains, it appears very helpful for their tertiary structure determination. Appropriate computational tests measuring execution time of enumerative algorithm have been performed and their results will be presented in Chapter 7. Since application of this algorithm for longer nucleic acid chains becomes less effective, two polynomial time heuristics dedicated to NOE pathways problem will be introduced in the following sections of this chapter. In Section 6.4, a tabu search approach will be presented, whereas Section 6.5 will be devoted to the presentation of a genetic algorithm.

VI.4 TABU SEARCH ALGORITHM

This section introduces a tabu search algorithm designed to solve the problem of NOE pathways for RNA molecules. It has been proved, that this problem in its search version is strongly NP–hard and hence, no polynomial-time exact algorithm is likely to exist for it. However, since NOESY spectra are represented as sparse graphs, an exact algorithm

enumerating all feasible solutions has been proposed (Section 6.3). This algorithm has appeared to perform quite fast for small RNA molecules (see Chapter 7 for the test results). Applied for the long nucleic chains, enumerative algorithm may construct too many feasible solutions, preventing from performance of the next steps in the determination process and its computation time grows exponentially. Because of these drawbacks of the exact method, a need has arisen to try another approach which 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 has been proposed. As it is crucial to generate the pathways as close as possible to the original one, an application of metaheuristics has been considered.

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

The proposed tabu search algorithm based on the above general tabu approach is extended by adding an *elite* structure that stores the most promising solutions. They are used as base solutions in the succeeding iterations of the search if the neighborhood of

a new solution appears to be worse. The elite structure stores whole solutions together with their versions of tabu list.

Before an algorithm starts computation all the available information is looked through to generate an appropriate graph representation of the problem. Next, the *search space* X is constructed on the basis of spectral and supplemental input data (Section 6.1). Experimental (spectral) data come from 2D–NOESY experiment and describe cross-peaks in the spectrum, i.e. cross-peak center coordinates, widths in two dimensions and NOE signal intensity. The supplemental data concern the analyzed structure and the obtained spectral information, and include analyzed molecule sequence, NOE pathway length, intensity intervals, spectrum resolution, overlapping signals, doublets, additional signal rejection, pathway potential starting points, known signal positions within the pathway, H5–H6 interactions. Every solution $x \in X$ is a vector of at most l cross-peaks, where $l=2N-1$ and N is a number of residues in the analyzed RNA molecule. Solution x in X, being a NOE pathway, has the following properties: each cross-peak is unique within x, every two neighboring edges of x are perpendicular, no two edges of x lie on the same horizontal nor vertical line, solution satisfies some predefined conditions (Definition 6.2.1). A move of tabu method is *feasible* if it constructs a solution obeying these constraints, i.e. feasible solution.

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

Initial solution is generated by a random procedure limited by the general feasibility rules (Definition 6.2.1) or by a greedy algorithm constructing the solution by adding cross-peaks one by one and starting from various cross-peaks of a NOESY spectrum aromatic/anomeric region. Finally, if the pathway we look for should have a maximum length (due to some predefined conditions) and the greedy procedure returns a solution shorter than $2M-1$, then the vector is complemented with random cross-peaks.

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

Every solution is evaluated according to the global criterion function *f* (see Section 6.2, formula 6.2.1). Solution with the best (minimum) value of function *f* is selected in the following step of the searching procedure. The algorithm tends 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. Detailed description of criterion function has been presented in Section 6.2.

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

IF $f(x_T^{'}) \leq f(x_{nT}^{'}) < f(best)$ OR $f(best) < f(x_T^{'}) \leq f(x_{nT}^{'})$

THEN new base solution = $x_{nT}^{'}$

ELSE IF $f(x_T^{'}) < f(best) \leq f(x_{nT}^{'})$

THEN new base solution = $x_T^{'}$.

In the above statement *f* denotes the global criterion function and *f*(*best*) stands for the value of the best solution found so far. The remaining cases for aspiration criterion are typical and the same as for a standard version of the tabu search algorithm.

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

The tabu search algorithm, based on the combinatorial model of the problem, has been proposed and applied to the collection of spectral data gathered from the NMR experiments for different RNA molecules (see Chapter 7). Its efficiency has been compared to the other heuristics based on evolutionary approach that will be presented in the next section.

VI.5 EVOLUTIONARY ALGORITHM

This section is devoted to an evolutionary algorithm which has been designed for the problem of NOE pathways construction. Since the problem in question in its search version is strongly NP–hard, no exact algorithm can solve it in time bounded by

a polynomial. However, since NOESY spectra are represented as sparse graphs, an enumerative algorithm based on exact search has been proposed (Section 6.3) and appeared to perform quite fast for small RNA molecules (tests results are given in Chapter 7). Exponential time of computations and an enormous number of feasible solutions enumerated make exact method ineffective in case of long nucleic chains. Thus, an application of metaheuristics has been considered. One heuristic, based on tabu search method has been presented in the previous section. For a purpose of comparison and – perhaps – improving some results of computation, the second metaheuristic, designed with the use of evolutionary procedure, has been also proposed. The latter will be presented in the section.

The proposed evolutionary algorithm for NOE pathways construction is based on genetic method introduced by John Holland in 1975. Typical algorithm of this kind employs biologically derived techniques such as inheritance, mutation, natural selection and recombination to evolve solutions to the combinatorial problem. Basic components of the method are: 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. Usually, evolution starts from a population of completely random individuals (solutions), represented by chromosomes, and happens in generations. 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 the hope, that the new population will be better than the old one. In such a manner, an approximation algorithm evolves toward 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.

Computational efficiency of evolutionary algorithm depends on the values given to algorithm parameters (population size, initial population, genetic operators, fitness and stopping criteria, etc.). The proposed implementation of the method 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.

Search space of the problem is constructed on the basis of spectral and supplemental input data (Section 6.1), which describe an aromatic/anomeric region of 2D–NOESY spectrum resulting from the NMR experiment performed for RNA molecule. Main subject of interest are cross-peaks, which represent NOE interactions between protons H6, H8, H1', H5. Every cross-peak is characterized by the two coordinates of its center, widths in two dimensions and the volume (intensity) of NOE signal (spectral data). Additionally, algorithm considers supplemental data which concern the analyzed structure and the obtained spectral information, that is analyzed molecule sequence, NOE pathway length, intensity intervals, spectrum resolution, overlapping signals, doublets, additional signal rejection, pathway potential starting points, known signal positions within the pathway, H5–H6 interactions.

Population size p is kept constant through the generations. It is the parameter, which can be changed if necessary. Arbitrarily its value has been set between 250 and 1000 individuals.

The *initial population P*=0 partially consists of, respectively, the individuals generated randomly (but satisfying predefined conditions of feasible solution from Definition 6.2.1) and the solutions generated by the greedy algorithm starting the search from various starting points.

One of characteristics dependent on the problem is *individual encoding* rule. In the presented evolutionary algorithm an individual is represented by a vector of size l, where l is a maximum length of the NOE pathway for a given molecule. The value of l can be derived from the molecule primary structure and equals $2N–1$, where N denotes the number of nucleotides (residues) in the RNA chain. The vector is composed of the sequence of vertex numbers written in the order of their occurrence in the pathway.

During evolution process, each individual is evaluated and assigned a *fitness* value. Fitness of the individual is determined by the associated value of the goal function, being one of the most crucial components of the algorithm.

Goal function f (formula 6.2.1) comprises the knowledge about desirable features of the problem solutions, thus assuring metaheuristic efficiency. It assembles the criteria, which serve the evaluation of individuals (pathways). Value of the goal function is to be minimized during the search process. In particular, the algorithm tends 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. Detailed description of the goal function has been presented in Section 6.2.

Selection is the first step taken in the process of generating the new population. The aim of the selection step is to eliminate bad solutions and transfer 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. The roulette wheel selection strategy (RWS) has been adopted for selection. The strategy 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, the algorithm draws $p/2$ (where p is population size) values out of the interval <0,1> and removes from the current population all individuals having cumulative distribution values corresponding to the fated ones. Thus, worse goal function values result in a greater probability of being removed from the population with no chance to get into the mating pool.

In the *reproduction* step new solutions (offspring) are generated with the use of crossover and mutation operators applied to the mating pool. All of these operators will be presented in the following paragraphs.

Crossover phase in the presented algorithm is based on the two operators *OX* and *merge* applied to the individuals selected according to the roulette mechanism. Solutions having better values of the goal function are chosen for a reproduction with a 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. The following operators have been introduced:

- **OX operator**. In the problem of the NOE pathway reconstruction, like in the other problems of that kind (Homaifar and Guan, 1991), 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 (Davis, 1985). 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.

- **Merge operator**. If the length of the NOE pathway is not known a priori, two operators are proposed to be used: 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, the operator copies the subsequent vertices of the second parent to the offspring sequence. Copying from the second parent stops, when the procedure finds a vertex which has been already put into the generated solution.

The next technique used by the algorithm is *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 unknown. Otherwise, only three of them mutate. In theory, the probability of utilization equals 0.1 for each mutation operator and the probability of mutation equals 0.271-0.40951 for each solution. Practically, likelihood values are smaller. The following mutation operators have been defined:

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

The method used for *creating the new generation* is an important aspect of evolution. In the proposed algorithm, the next generation is formed out of the best parent solutions and all individuals from the offspring population.

Algorithm stops when the *stopping criterion* has been satisfied. In the presented evolutionary algorithm stopping criterion has been defined as the number of succeeding generations without improvement of the best individuals. Its value has been set experimentally to 250 iterations.

The proposed algorithm follows the steps of the typical evolutionary procedure and performs steps described below.

Evolutionary Algorithm

1. generate initial population *P*=0: create *p* individuals, where each individual is a permutation of *n* signals given in the input;
2. calculate fitness *F* of each individual in population *P*;
3. find the best individual in population *P*;
4. repeat
5. basing on the fitness values of the individuals select parents for the new population *P+1* according to the roulette system;
6. for each pair of the selected parents apply crossover operators *OX* and *Merge*;
7. mutate individuals: considering offspring and parents populations pick an individual and apply the chosen mutation operators;
8. evaluate fitness for each individual in the current offspring and parent population;
9. create the new generation out of the best parents and all the offspring solutions;
10. until the stopping criteria are not satisfied.

Three algorithms for NOE pathways construction have been presented in the chapter: an exact algorithm enumerating all feasible solutions to the problem as well as a tabu search and evolutionary metaheuristics, both looking for optimal solutions. All of these methods have been applied for a set of spectral data gathered from NMR experiments for different RNA molecules. The results of computational tests performed with the use of the designed algorithms will be presented in Chapter 7. Also the input data including spectra, supplemental information and the details of NMR experiments from which the spectra have been obtained, will be described in the following chapter.

VII. COMPUTATIONAL EXPERIMENTS

An assignment of H6/H8–H1' NOE connectivity has been one of the hardest steps in a process of RNA tertiary structure determination with NMR techniques. In the thesis, the problem has been modeled (Chapter 4) with the use of tools provided by graph as well as computational complexity theories. On the basis of the proposed model, three algorithms for an automatic assignment have been designed (Chapter 6): the exact enumerative algorithm, the tabu search and the evolutionary algorithm. All the methods have been implemented in ANSI C programming language and applied to a set of 2D–NOESY spectra. The detailed description of the spectral as well as the supplemental data used in tests will be given in Section 7.1. Results of computational experiments performed with all the algorithms will be presented in Section 7.2 of the chapter.

VII.1 OVERVIEW OF EXPERIMENTAL DATA SET

The problem of NOE pathways assignment, being one of the first steps in the process of RNA tertiary structure determination with NMR techniques, has been considered in the previous chapters of the thesis. On the basis of the proposed theoretical model of the problem (Chapter 4), three methods for the automatic assignment have been designed (Chapter 6). Their efficiencies have been tested on the real data acquired from NMR experiments and will be discussed in the next section. This section is devoted to a presentation of testing data set which has been used for experimental NOE assignment performed with exact, tabu search and evolutionary algorithms.

The manual assignment of NOE resonances is very tedious and time-consuming due to a large number of cross-peaks and possible large number of existing alternative pathways. As a result, RNA chains analyzed in labs are generally rather small. Since author's intension was to verify computational results according to the real solutions, small molecules, with already known assignments, have been selected for test purposes.

Experimental data set consisted of NMR spectra acquired for the following molecules: $r(CGCGCG)_2$, $2'–OMe(CGCGCG)_2$, $r(CGCG^FCG)_2$, $d(GACTAGTC)_2$. Spectra of $r(CGCGCG)_2$, $2'–OMe(CGCGCG)_2$ and $r(CGCG^FCG)_2$ in D_2O at 30^oC were recorded on Varian Unity+ 500 MHz spectrometer. Standard pulse sequence (Jeener et al., 1979) $\pi/2–t_1–\pi/2–\tau_m–\pi/2–t_2$ was applied with mixing time τ_m=150 ms. Spectra were acquired with 1K complex data points in t_2 and 1K real points in the t_1 dimension, with

spectral width set to 3.7 kHz. After digital filtration by Gaussian functions, filling zero in t_1 dimension and base correction in t_2, data were collected in 1K×1K matrixes with the final digital resolution of 3.5 Hz/point in both dimensions.

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

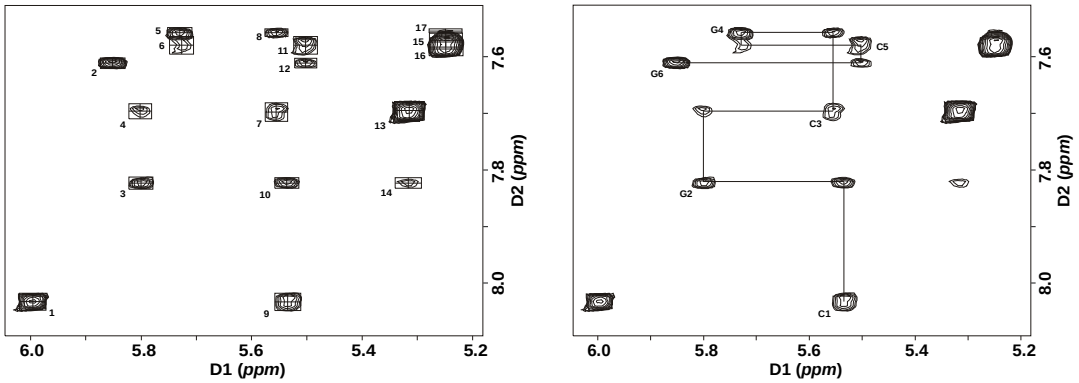
Note that for $r(CGCGCG)_2$ and $2'–OMe(CGCGCG)_2$ and $d(GACTAGTC)_2$ molecules, [5–6]×[7–8] region of the spectrum has been analyzed. In case of $r(CGCG^FCG)_2$ two regions, [5–6]×[7–8] and [7–8]×[5–6], respectively, have been considered, each storing the other information about correlation signals.

Inaccessibility of a large number of NMR spectral data for RNA molecules led us to a back-track simulation of some spectra from structures solved with NMR spectroscopy methods. 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 ^1H chemical shifts (McDowell and Turner, 1996; Wu et al., 1997; McDowell et al., 1997; SantaLucia Jr. and Turner, 1993) and three dimensional structures from Protein Data Bank (http://www.rcsb.org/pdb). Volumes of NOE cross-peaks for τ_m=0.3 ms were calculated from the Full Relaxation Matrix, where a correlation time was set to 2 ns. The Lorentzian line shape functions were used for simulated NOE cross-peaks. The widths of these functions depended on the sums of coupling constants calculated from the duplex structures based on Karplus equation using Lankhorst and Haasnoot parameters (Lankhorst et al. 1984, Haasnoot et al. 1980).

In case of $r(GGCAGGCC)_2$ and $r(GAGGUCUC)_2$, [5–6]×[7–8] region of the spectrum has been analyzed. Two regions, [5–6]×[7–8] and [7–8]×[5–6], respectively, have been considered in the analysis of $r(GGAGUUCC)_2$ and in case of $r(GGCGAGCC)_2$, [7–8]×[5–6] region has been processed.

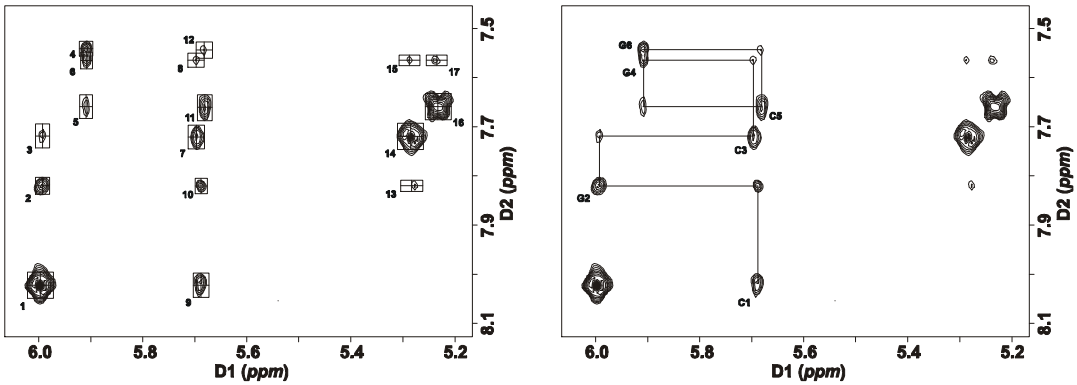
Since all the instances had been already solved manually, it was possible to verify whether or not each algorithm found an original or a feasible solution. It was also possible to examine the way supplementary expert knowledge influences qualifying solutions and building the final solution set.

All molecules formed self-complementary chains, thus, one biochemically correct pathway (an original path) existed for each of them. These original NOE paths drafted in the aromatic/anomeric region of the 2D–NOESY spectrum as well as the considered fragments of the spectrum for the selected cases are demonstrated in Figures 7.1.1 through 7.1.7.



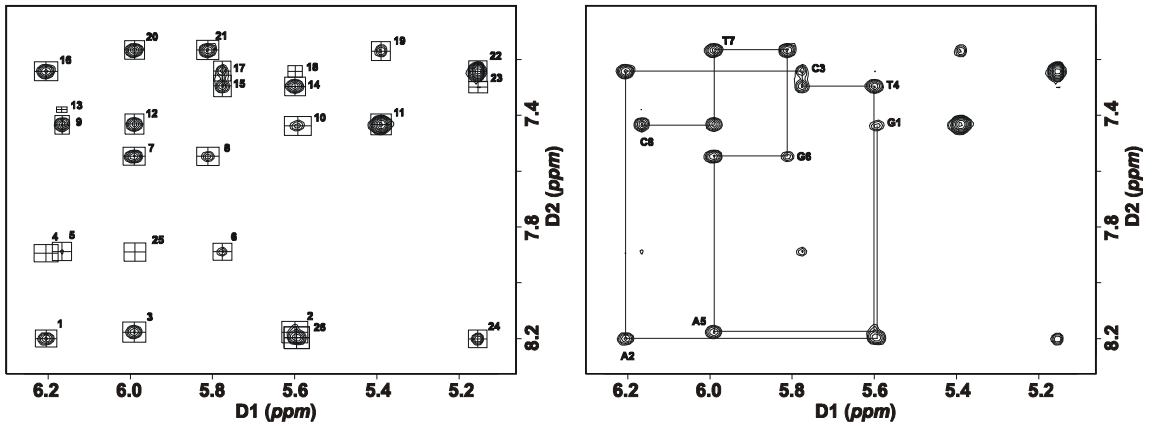
(a) (b)

Figure 7.1.1. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for r(CGCGCG)$_2$ and the NOE path (b).



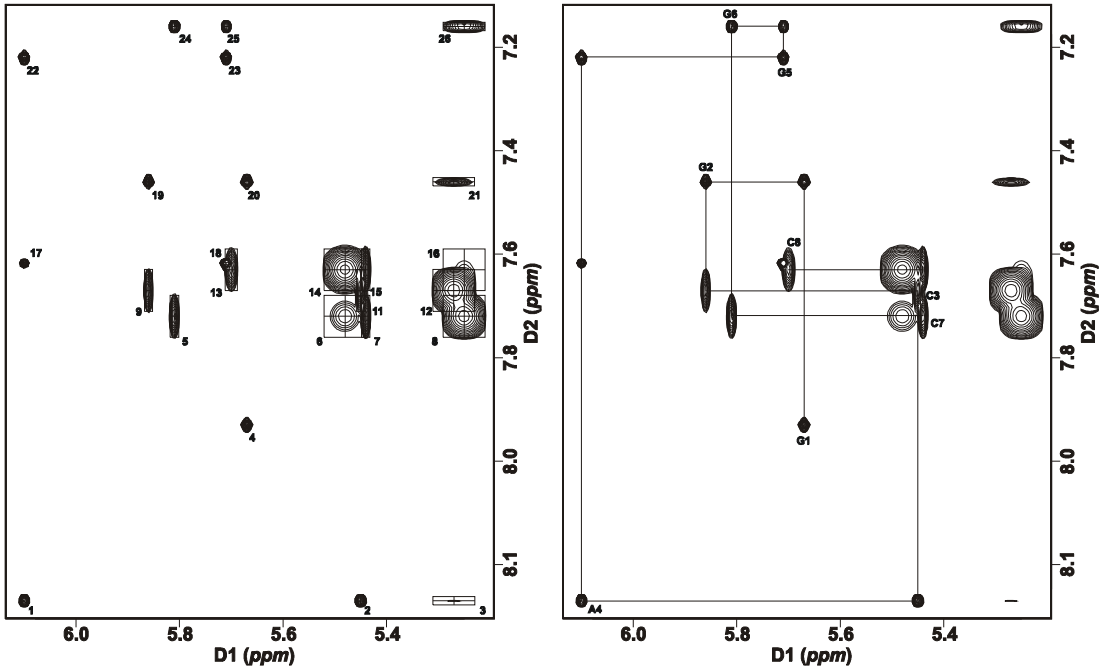
(a) (b)

Figure 7.1.2. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for 2'–OMe(CGCGCG)$_2$ and the NOE path (b).



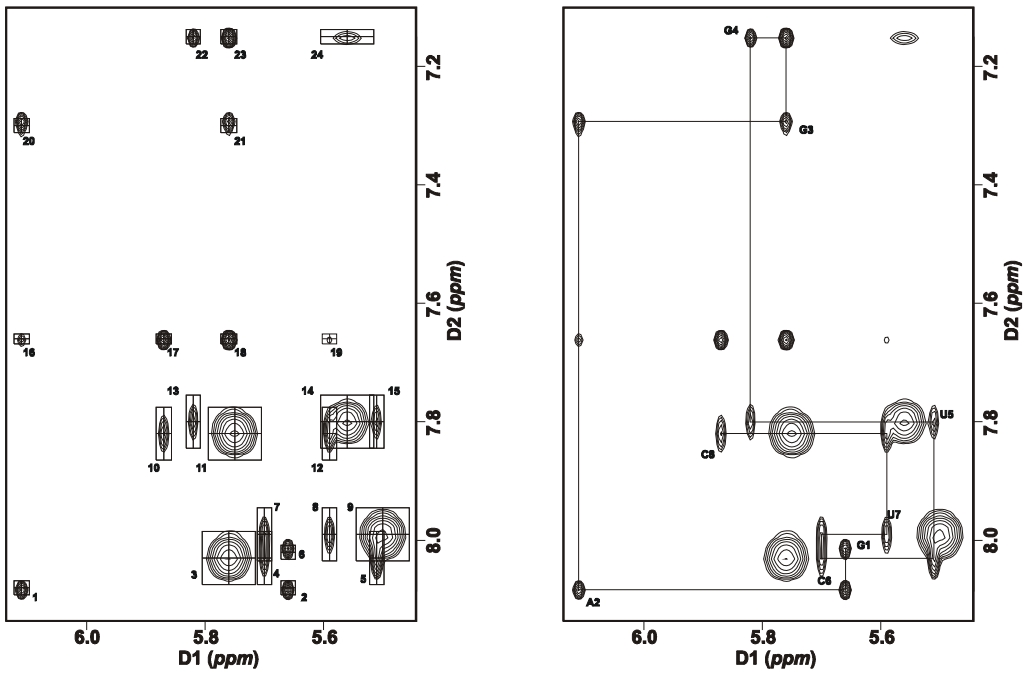
(a) (b)

Figure 7.1.3. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for d(GACTAGTC)$_2$ and the NOE path (b).



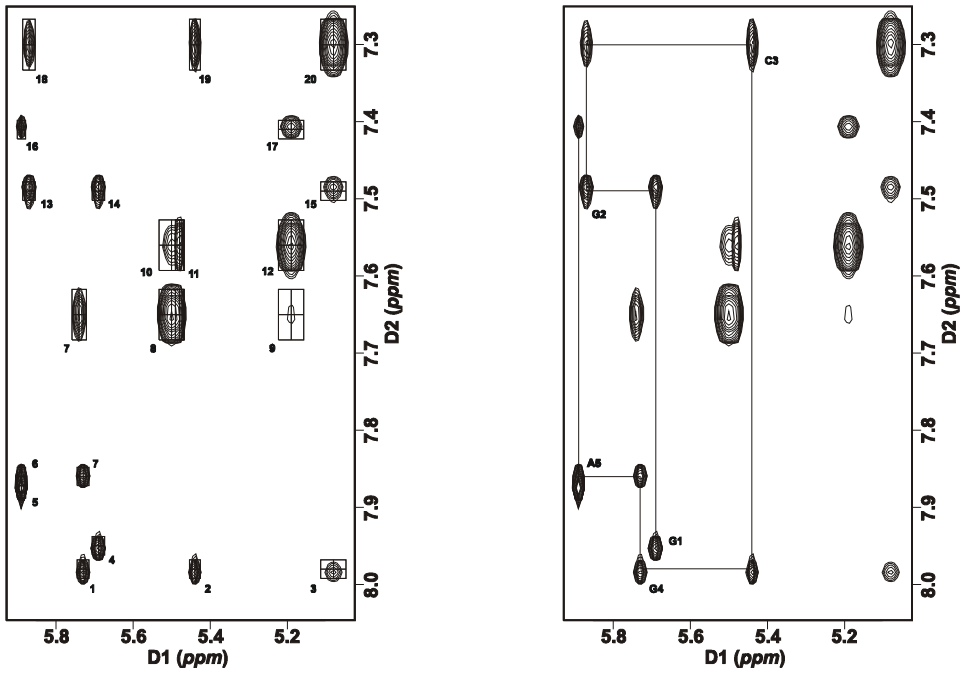
(a) (b)

Figure 7.1.4. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for r(GGCAGGCC)$_2$ and the NOE path (b).



(a) (b)

Figure 7.1.5. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for r(GAGGUCUC)$_2$ and the NOE path (b).



(a) (b)

Figure 7.1.6. [7–8]×[5–6] region of 2D–NOESY spectrum (a) for r(GGCGAGCC)$_2$ and the NOE path (b).

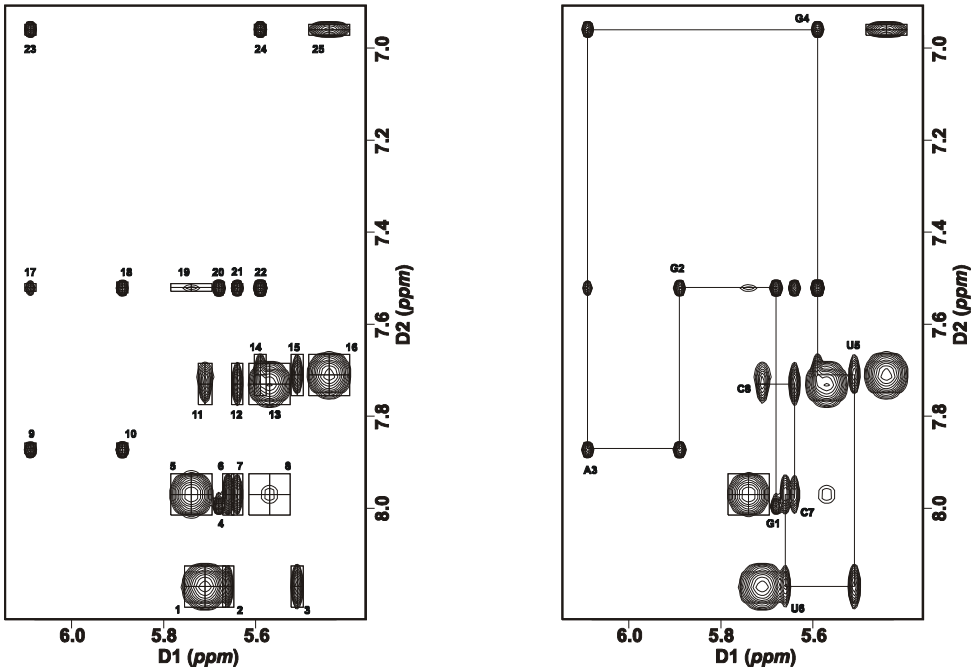


Figure 7.1.7. [5–6]×[7–8] region of 2D–NOESY spectrum (a) for r(GGAGUUCC)$_2$ and the NOE path (b).

The numerical data for computational experiments were obtained from experimental and simulated spectra (Figures 7.1.1 through 7.1.7) after peak-picking procedure of Felix Accelrys. These data are referred to as spectral data (see Chapter 3 and Chapter 5). Additionally, a set of supplemental data can be defined for each analyzed instance. Table 7.1.1 shows a complete set of supplemental information that can be provided for each instance, basic information about instance and instance unique identifier. Identifiers will be used in Section 7.2 in the description of computational results.

Table 7.1.1.Test instances description.

Id	Basic information	Supplemental information
I	molecule: r(CGCGCG)$_2$ experimental spectrum spectral region [5–6]×[7–8] instance size: 17 cross-peaks	5 signals rejected; path length; 3 H5–H6 signals enumerated;
II	molecule: 2'–OMe(CGCGCG)$_2$ experimental spectrum spectral region [5–6]×[7–8] instance size: 17 cross-peaks	volume intervals (inter/intra); interval with overlapping signals; path length; 3 H5–H6 signals enumerated;

Id	Basic information	Supplemental information
III	molecule: r(CGCG^{F}CG)$_2$ experimental spectrum spectral region [7–8]×[5–6] instance size: 15 cross-peaks	2 signals rejected; resolution; path length; 2 H5–H6 signals enumerated;
IV	molecule: r(CGCG^{F}CG)$_2$ experimental spectrum spectral region [5–6]×[7–8] instance size: 22 cross-peaks	volume intervals (inter/intra); distance between doublets; resolution; path length;
V	molecule: d(GACTAGTC)$_2$ experimental spectrum spectral region [5–6]×[7–8] instance size: 26 cross-peaks	7 signals rejected; path length; 4 H5–H6 signals enumerated;
VI	molecule: r(GGCAGGCC)$_2$ simulated spectrum (PDB ID: 1MWG) spectral region [5–6]×[7–8] instance size: 26 cross-peaks	8 signals rejected; path length; 5 H5–H6 signals enumerated;
VII	molecule: r(GAGGUCUC)$_2$ simulated spectrum (PDB ID: 1GUC) spectral region [5–6]×[7–8] instance size: 24 cross-peaks	6 signals rejected; path length; 4 H5–H6 signals enumerated;
VIII	molecule: r(GGCGAGCC)$_2$ simulated spectrum (PDB ID: 1YFV) spectral region [7–8]×[5–6] instance size: 20 cross-peaks	5 signals rejected; path length (broken chain);
IX	molecule: r(GGAGUUCC)$_2$ simulated spectrum (PDB ID: 1QES) spectral region [5–6]×[7–8] instance size: 25 cross-peaks	6 signals rejected; path length; 4 H5–H6 signals enumerated;
X	molecule: r(GGAGUUCC)$_2$ simulated spectrum (PDB ID: 1QES) spectral region [7–8]×[5–6] instance size: 25 cross-peaks	9 signals rejected; path length; 4 H5–H6 signals enumerated;

In computational experiments, supplemental data influence on the results has been tested. Thus, not all available data has been used in the tests. However, let us notice that in some cases supplemental knowledge is necessary for an appropriate interpretation of the input spectral data. For example, without additional information algorithms cannot find original

solution in the spectrum containing doublets or overlapping signals or in the case where spectrum resolution should be considered. All the results of computational experiments will be discussed in Section 7.2.

VII.2 EXPERIMENTAL RESULTS

This section is devoted to a presentation of the results obtained from computational experiments dealing with automatic reconstruction of the NOE path in the 2D–NOESY spectra of RNA molecules. The set of spectral as well as expert supplemental data used in the experiments, has been described in Section 7.1.

Three methods, tabu search (TS), evolutionary algorithm (EA) and exact enumerative algorithm have been considered in the tests. All of them have been coded in ANSI C and tested on Indigo 2 Silicon Graphics workstation (1133 MHz, 64 MB) in IRIX 6.5 environment.

Two tests T1, T2 have been performed for every instance and each algorithm. In the first test (T1) algorithms used all available supplemental data. In the second case (T2) a minimum amount of expert supplemental information has been considered. First, numbers of feasible solutions found by enumerative algorithm for tests T1 and T2 have been computed (Table 7.2.1). Next, the time of computations by all the algorithms has been considered (Table 7.2.2). Finally, solutions generated by tabu search and evolutionary algorithm have been analyzed according to their quality (Table 7.2.3, Figures 7.2.1 through 7.2.7). The value of a solution precision has been calculated as a percentage of the original path covered by the best solution generated by heuristic algorithm (TS or EA). Genetic algorithm has been tested for four population sizes: 250, 500, 750 and 1000.

Table 7.2.1 shows numbers of feasible paths generated by enumerative algorithm for test T1 and T2.

Table 7.2.1. Numbers of feasible solutions found by enumerative algorithm.

Instance		I	II	III	IV	V	VI	VII	VIII	IX	X
feasible solutions number	T1	1	2	3	2	4	2	1	4	1	1
	T2	140	776	72	63	240	3192	160	64	843	1134

The tests have shown that exact enumerative algorithm is very useful in case of supplemental data availability. For all the instances a number of feasible solutions generated in test T1 has been small. Of course, in all the cases original solution has been

found. However, a number of feasible solutions generated in test T2 is huge, thus, making enumerative algorithm hardly efficient for these cases. Results of the latter test deserve a special attention. In T2, the algorithms have operated on the minimum amount of supplemental data, what means that the information required for a proper interpretation of the input spectral data only, has been supplied. Thus, the information about spectral resolution, doublets or overlapping has been given, while any additional information, like path length, volume intervals, H5–H6 signals, known signal positions within the path, signals rejections has not been provided. Additional information (not provided in T2) 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 the spectrum obtained in the NMR experiment is. This results in many overlapping signals. An extreme number of cross-peaks located within the same spectral region often causes problems in supplying additional information to the algorithms. Consequently, the experimenters try rather - the less risky - algorithms without the expert information. Unfortunately, computational analysis with the use of enumerative algorithm in such cases appears rather inefficient and disqualifies this method here. Of course, one may be sure that the enumerative algorithm finds the original solution, but looking through the generated set of thousands feasible paths in order to situate this original one is a hopeless job and harder than a manual reconstruction of the NOE path. It seems possible to analyze manually the set of up to 20 solutions in order to find the original one among them. But a greater number of paths discourages an ordinary researcher to look through them. Thus, it seems a good idea to apply a heuristic algorithm, generating one optimal solution, for solving such instances of the problem. Even if the heuristic method finds approximate solution covering only half of the original NOE path it facilitates the problem to a very large degree. Having the partial assignment an experimenter is able to complete the NOE pathway in a reasonable time without an extreme effort.

Let us then consider enumerative algorithm and two heuristic methods, tabu search and evolutionary algorithm, as applied to the problem of NOE paths construction. Table 7.2.2 presents time of computations (given in seconds) by all these algorithms for tests T1 and T2. The evolutionary algorithm has been tested with different values of population size parameter, $p \in \{250, 500, 750, 1000\}$.

Table 7.2.2. Time of computations [s].

Instance	Test	Time of computations [s]					
		Enum. algorithm	TS	EA			
				p=250	p=500	p=750	p=1000
I	T1	1	0.5	1	3	8	11
	T2	5	1	3	10	21	48
II	T1	1	0.7	1	3	5	13
	T2	60	1	2	12	38	63
III	T1	2	0.6	2	4	16	17
	T2	4	1.1	4	6	17	45
IV	T1	1	0.5	1	4	5	19
	T2	4	1	2	4	7	20
V	T1	1	0.5	7	22	74	129
	T2	5	0.9	4	6	27	90
VI	T1	1	0.4	4	6	10	16
	T2	2453	1	2	8	27	62
VII	T1	1	0.7	4	18	44	80
	T2	30	1.3	4	17	73	132
VIII	T1	1	0.4	1	2	4	12
	T2	5	0.8	2	5	21	24
IX	T1	1	0.6	5	18	44	80
	T2	170	1	2	6	22	63
X	T1	1	0.6	5	19	43	80
	T2	573	1	3	5	34	31

Analyzing Table 7.2.2, we can observe that all the algorithms, work quite fast. In many

cases the results are obtained after 1–5 seconds (enumerative algorithm for T1, TS, EA for p=250). 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. Generally it can be said that tabu search is the fastest method. Enumerative algorithm for test T1 as well as evolutionary algorithm for small population size are also fast. Obviously, evolutionary algorithm performs slower for bigger populations. Computational time of enumerative algorithm for test T2 appears quite big for more problematic instances, i.e. when a number of feasible paths is also big (cf. Table 7.2.1). Since computational times are mostly short, it can be concluded that, so far, time of computations plays a peripheral role in solving the problem of NOE pathway assignments with automatic methods.

Finally let us analyze the quality of optimal solutions found by heuristic algorithms. The quality will be given as a precision, calculated as a percentage of an original path covered by the generated optimal solution. Let us show an example. Assume we have original NOE path that can be written as the following sequence of cross-peaks: 1–2–3–4–5–6–7–8–9–10. Let us calculate precision of the near–optimal solution, found by the heuristic algorithm, given as a sequence 6–7–8–9–10–3–2–4–1–5. We see that the original path consists of 9 edges: 1–2, 2–3, 3–4, 4–5, 5–6, 6–7, 7–8, 8–9, 9–10. 5 of these edges are found in the optimal solution (2–3, 6–7, 7–8, 8–9, 9–10). Thus, the precision of the optimal solution equals c.a. 55 %.

Table 7.2.3 presents results of the precision calculation for optimal solutions generated by tabu search and evolutionary algorithm (for different population sizes), respectively, in both tests, T1 and T2.

Table 7.2.3. Precision of optimal solution found by heuristic algorithms.

Instance	Test	Optimal solution precision				
		TS	EA			
			p=250	p=500	p=750	p=1000
I	T1	100 %	82 %	100 %	100 %	100 %
	T2	91 %	36 %	91 %	91 %	91 %
II	T1	100 %	100 %	100 %	100 %	100 %
	T2	91 %	45 %	100 %	100 %	64 %
III	T1	91 %	100 %	82 %	100 %	82 %
	T2	100 %	82 %	82 %	64 %	64 %
IV	T1	91 %	82 %	82 %	82 %	82 %
	T2	73 %	82 %	82 %	55 %	73 %
V	T1	73 %	47 %	80 %	53 %	40 %
	T2	33 %	33 %	33 %	47 %	33 %
VI	T1	100 %	87 %	100 %	87 %	100 %
	T2	40 %	27 %	27 %	53 %	47 %
VII	T1	93 %	93 %	93 %	93 %	93 %
	T2	47 %	40 %	33 %	33 %	33 %
VIII	T1	100 %	100 %	100 %	100 %	100 %
	T2	100 %	60 %	100 %	80 %	80 %
IX	T1	93 %	93 %	100 %	93 %	100 %
	T2	46 %	33 %	80 %	87 %	40 %
X	T1	93 %	93 %	93 %	93 %	100 %
	T2	46 %	53 %	33 %	27 %	33 %

The precisions of optimal solutions have been also displayed as charts. The first chart

(Figure 7.2.1) shows precisions of solutions generated by tabu search algorithm. Figure 7.2.2 presents precisions of solutions found by evolutionary algorithm run for different sizes of population (p=250, 500, 750, 1000).

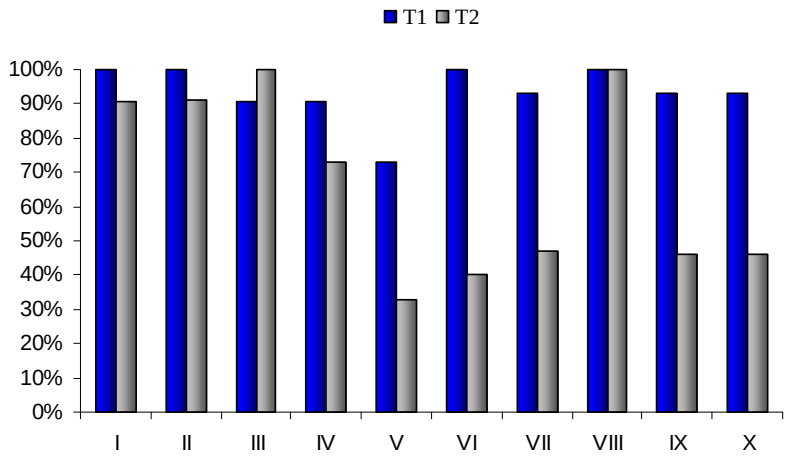


Figure 7.2.1. Precisions of optimal solutions found by tabu search.

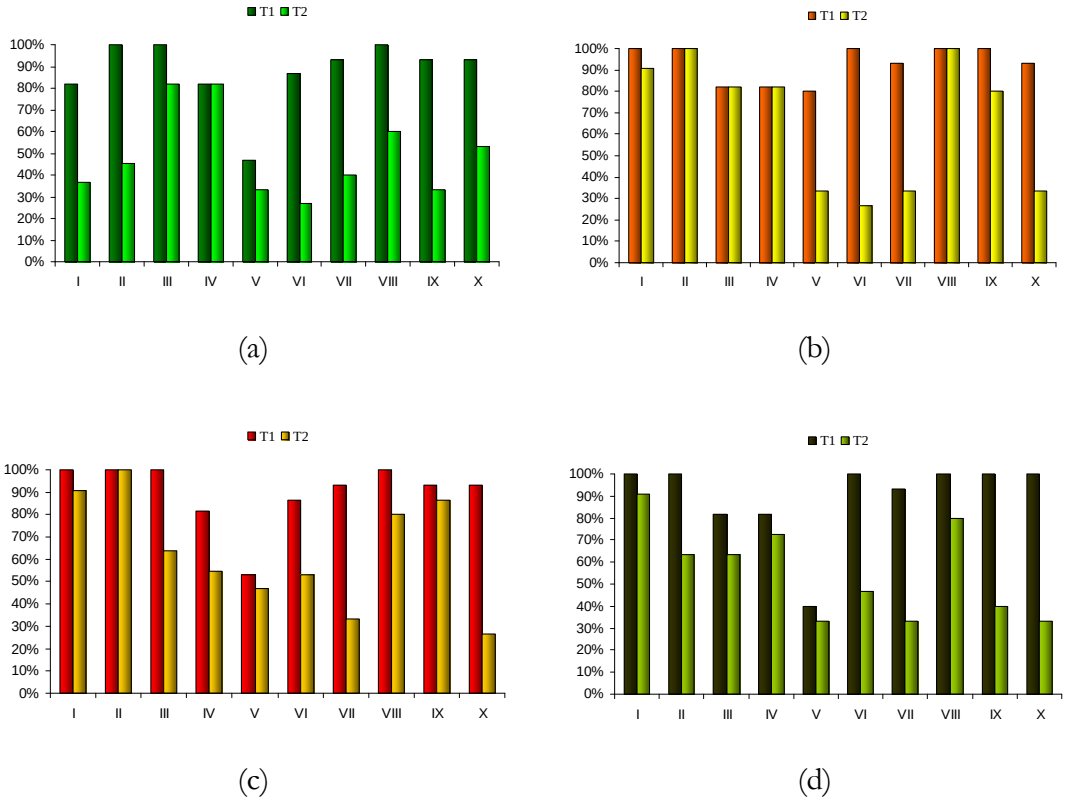


Figure 7.2.2. Precisions of optimal solutions found by evolutionary algorithm for p=250 (a), p=500 (b), p=750 (c), p=1000 (d).

To compare the efficiency of heuristic algorithms, quality (precision) of optimal solution found by each algorithm has been displayed in additional charts. Thus, chart in Figure

7.2.3 compares precisions for tabu search and evolutionary algorithm, respectively, for test T1, while chart in Figure 7.2.4 shows the same comparison for test T2.

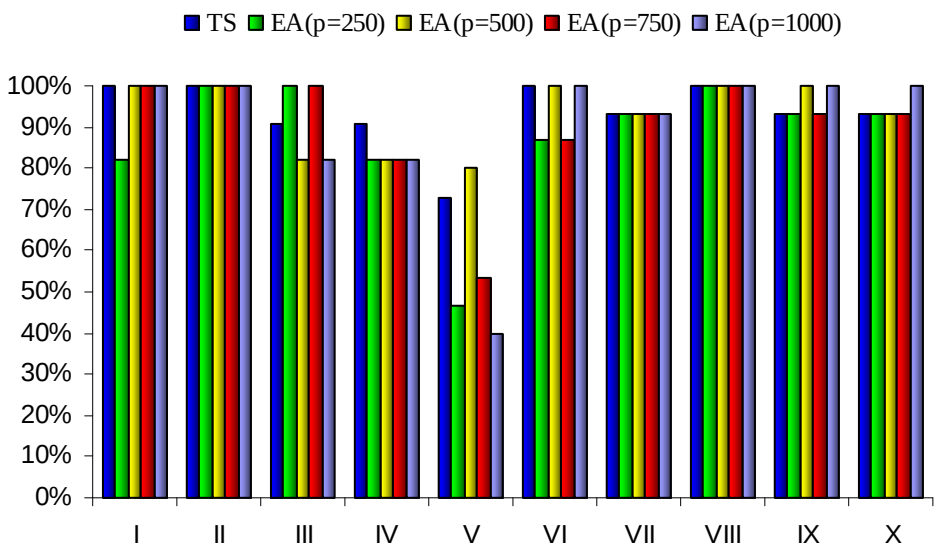


Figure 7.2.3. Precisions of optimal solutions found by TS and EA for test T1.

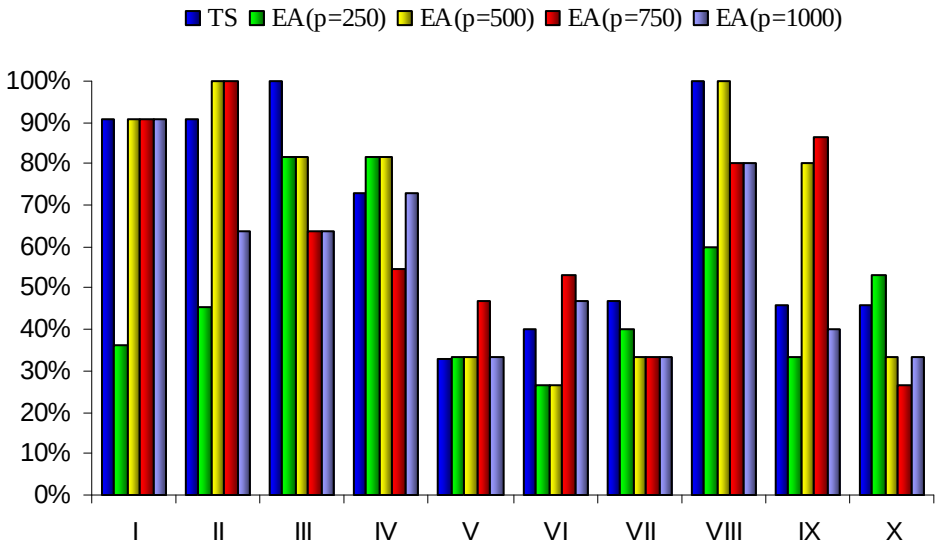


Figure 7.2.4. Precisions of optimal solutions found by TS and EA for test T2.

Computational experiments show advantages and disadvantages of all the proposed algorithms. Exact enumerative algorithm has appeared successful for the instances with supplemental data provided. However, it is hard to estimate this algorithm efficiency for the longer nucleic chains, for which no algorithms have been tested yet. We can only expect problems in defining supplemental data and, thus, also less efficient performance of

the exact algorithm. We see that harder instances demand an application of heuristics rather than the exact approach. The proposed tabu search and evolutionary algorithms in most cases give very good results. Their times of computations are short and precision of these methods appears high enough to consider them as alternative algorithms in solving the problem of NOE pathway reconstruction. This is especially true for the instances, for which an enumerative algorithm is hardly effective because of a huge number of feasible solutions generated. The resulting qualities of optimal solutions found by both algorithms are similar for test T1. Test T2 shows many differences in the performance of both algorithms and it is rather hard to declare which algorithm is generally better here.

The chapter has discussed all the aspects of computational experiments performed on the real data with the use of three algorithms for automatic NOE assignment. The data used in tests as well as the results of computation have been presented. Final discussion on the results and the proposed methods of solving the problem of NOE path reconstruction will be given in the following chapter.

VIII. CONCLUSIONS

In the thesis, the problem of automatic resonance assignment in 2D–NOESY NMR spectra of RNA duplexes 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. This manual assignment of NOE resonances is very tedious and time-consuming due to a large number of cross-peaks present in the NOESY spectra of biomolecules and a possible large number of existing alternative pathways. Thus, manual assignment limits structural studies to small molecules of nucleic acids.

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. The model has been presented in Chapter 4 of the thesis.

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 (Chapter 5) 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* (Chapter 6), *implemented and applied to a set of real data*. Analyzing results of computational experiments performed with this algorithm (Chapter 7), one can notice that *it constructed surprisingly small number of alternative pathways in tests with supplemental data provided* (see test T1, Table 7.2.1), thus, proving its high accuracy in these cases. However, supplemental knowledge deficiency causes a great increase of solution set generated by enumerative algorithm, disqualifying this method in such a situation (see test T2, Table 7.2.1). Since analysis of bigger molecules implies processing of more crowded spectra and problems with defining supplemental information, designing the other method that could deal with the problem of NOE assignment basing just on spectral data (like in test T1), 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 also 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. Even enumerative algorithm does not drop behind, despite its computational complexity which equals $O(m^m)$, where m is the number of graph edges. 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 (Chapter 4), 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.

INDEX OF BASIC NOTIONS

algorithm, 31
  branch-and-bound, 36
  exponential, 31
  genetic, 37
  polynomial, 31
  pseudo-polynomial, 31
  tabu search, 36
amino acid, 11
base, 12
biopolymer, 11
chemical shift, 25
complexity class
  NP, 33
  NP-complete, 33
  NP-hard, 34
  PROMISE-P, 35
  strongly NP-complete, 33
conformation, 18
  anti, 17
  syn, 17
correlation signal, 39
cross-peak, 43
data
  spectral, 47
  supplemental, 48
digraph, 29
  arc, 29
DNA, 13
edge
  adjacent, 29
  incident, 29
Eulerian
  circuit, 30
  graph, 30
  trail, 30
feasible pathway, 72
gene expression, 15
graph, 29
  acyclic, 30
  circuit, 30
  connected, 30
  cycle, 30
  directed, 29
  edge, 29
  loop, 29
  nontrivial, 29
  order, 29
  path, 30
  size, 29
  sparse, 66
  trail, 30
  trivial, 29
  vertex, 29
  walk, 30
Hamiltonian
  cycle, 30
  graph, 30
  path, 30
input size, 31
instance, 31
  size, 31

model
experimental, 53
real, 53
theoretical, 53
monomer, 11
MRI, 22
NMR
spectroscopy, 21
NOE
path, 54
pathway, 45, 54
signal, 43
NOESY
graph, 52
spectrum, 43
nucleic acid, 11
nucleotide, 12
oligonucleotide, 13
optimal
pathway, 74
solution, 73
original pathway, 49
path
Hamiltonian, 30
NOE, 54
pathway
feasible, 72
optimal, 74
polymer, 11
polynucleotide, 13
ppm, 25
precession, 23
problem
combinatorial, 32
decision, 32
intractable, 31
optimization, 32
promise, 35
search, 32
uniquely promised, 35
protein biosynthesis, 15
proteins, 13
resonance signal, 39
retrovirus, 15
RNA, 14
mRNA, 14
non-coding RNA, 15
RNA genes, 15
RNAi, 15
rRNA, 14
small RNA, 15
tRNA, 15
RNA World, 16
search space, 68
search tree, 36
sequencing, 17
signal
correlation, 39
internucleotide, 45
intranucleotide, 45
NOE, 43
resonance, 39
strand, 11

structure
native, 19
primary, 17
quaternary, 20
secondary, 18
tertiary, 18
time complexity function, 31
transcription, 15
transformation
parsimonious, 35
polynomial, 33, 60
polynomial Turing, 34
pseudo-polynomial, 34
T-transformation, 34
translation, 15
vertex
adjacent, 29
degree, 30
incident, 29
isolated, 30
walk
closed, 30
length, 30
open, 30

BIBLIOGRAPHY

Aarts, E.H.L., Lenstra, J.K. (1997) *Local Search in Combinatorial Optimization*. John Willey&Sons, Chichester, UK.

Adamiak, R.W., Błażewicz, J., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M., Szachniuk, M. (2004) An algorithm for an automatic NOE pathways analysis of 2D NMR spectra of RNA duplexes. *Journal of Computational Biology* 11/1, 163-180.

Akmaev, V.R., Kelley, S.T., Stormo, G.D. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics* 6, 501-512.

Atreya, H.S., Sahu, S.C., Chary, K.V., Govil, G. (2000) A tracked approach for automated NMR assignments in protein (TATAPRO). *Journal of Biomolecular NMR* 17*,* 125-36.

Balley-Kellog, C., Chainraj, S., Pandurangan, G. (2004) A random graph approach to NMR sequential assignment. *Currents in Computational Molecular Biology*, 58-67.

Bartels, C., Guntert, P., Billeter, M., Wuthrich, K. (1997) GARANT – a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry* 18/1, 139-149.

Bax, A., Kontaxis, G., Tjandra, N. (2001) Dipolar couplings in macromolecular structure determination. *Methods in Enzymology* 339, 127-174.

Błażewicz, J. (1988) *Złożoność obliczeniowa problemów kombinatorycznych*. Wydawnictwa Naukowo-Techniczne, Warszawa.

Błażewicz, J., Cellary, W., Słowiński, R., Węglarz, J. (1983) *Badania operacyjne dla informatyków*. Wydawnictwa Naukowo-Techniczne, Warszawa.

Błażewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., Świercz, A. (2004c) Tabu search algorithm for DNA sequencing by hybridization with isothermic libraries. *Computational Biology and Chemistry* 28, 11-19.

Błażewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., Węglarz, J. (1999) DNA sequencing with positive and negative errors. *Journal of Computational Biology* 6, 113-123.

Błażewicz, J., Kaczmarek, J., Kasprzak, M., Markiewicz, W.T., Węglarz, J. (1997) Sequential and parallel algorithms for DNA sequencing. *Computer Applications in the Biosciences* 13(2), 151-158.

Błażewicz, J., Szachniuk, M., Wójtowicz, A. (2004a) Evolutionary approach to NOE paths assignment in RNA structure elucidation. *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 206-213.

Błażewicz, J., Szachniuk, M., Wójtowicz, A. (2004b) Genetic algorithm for a reconstruction of NOE paths in NMR spectra of RNA chains. *Bulletin of the Polish Academy of Sciences. Technical Sciences* 52 (3), 221-230.

Błażewicz, J., Szachniuk, M., Wójtowicz, A. (2005) RNA tertiary structure determination: NOE pathways construction by tabu search. *Bioinformatics* 21 (10), 2356-2361.

Bourne, P.E., Weissig, H. (eds) (2003) *Structural bioinformatics*. John Wiley&Sons, Hoboken, New Jersey.

Caplen, N.J., Parish, S., Imani, F., Fire, A., Morgan, R.A. (2001) Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proceedings of the National Academy of Science* 98, 9742-9747.

Case, D.A. (1998) NMR Refinement. *Technical report of the Scripps Research Institute*, La Jolla, USA.

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

Cech, T.R. (1990) Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. *Bioscience Reports* 10, 236-261.

Cech, T.R. (1993) Catalytic RNA: structure and mechanism. *Biochemical Society Transactions* 21, 229-234.

Chartrand, G., Lesniak, L. (1986) *Graphs & Digraphs*. Wadsworth & Brooks/Cole, Belmont.

Chen, Z., Blanc, E., Chapman, M.S. (1999) Real-space molecular-dynamics structure refinement. *Acta Crystallographica* D55, 464-468.

Clore, G.M., Gronenborn, A.M. (1998) New methods of structure refinement for macromolecular structure determination by NMR. *Proceedings of the National Academy of Science* 95, 5891-5898.

Croft, D., Kemmink, J., Neidig, K.J., Oschkinat, H. (1997) Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. *Journal of Biomolecular NMR* 10, 207-219.

Dandekar, T. (2002) *RNA Motifs and Regulatory Elements*. Springer-Verlag, Berlin, Heidelberg.

Davis, L. (1985) Applying adaptive algorithms for epistatic domains. *Proceedings of the International Joint Conference on Artificial Intelligence*, 162-164.

Drozdowski, M. (1997) Selected problems of scheduling tasks in multiprocessor computer systems. *Habilitation thesis. Poznan University of Technology Press* 321, Poznan.

Ejchart, A., Kozerski, L. (1981) *Spektrometria magnetycznego rezonansu jądrowego ^{13}C.* Wydawnictwo PWN, Warszawa.

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

Garrett, R.H., Grisham, C.M. (1995) *Biochemistry.* Saunders College Publishing, Orlando, Florida.

Gilbert, W. (1986) The RNA World. *Nature* 319, 618-619.

Glover, F., Laguna, F. 1997. *Tabu Search.* Kluwer Academic Publishers, Norwell, MA, USA.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA.

Gross, J. L., Yellen, J. (2004) *Handbook of Graph Theory.* CRC Press, London.

Gulko, B., Haussler, D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. *Pacific Symposium on Biocomputing*, 350-367.

Guntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics* 31, 145-237.

Gunther, H. (1996) *NMR Spectroscopy.* John Wiley&Sons, New York.

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

Hausser, K.H., Kalbitzer, H.R. (1993) *NMR w biologii i medycynie.* Wydawnictwo Naukowe UAM, Poznań.

Hilbers, C.W., Wijmenga, S.S. (1996) Nucleic Acids: Spectra, Structures, & Dynamics. *Encyclopedia of Nuclear Magnetic Resonance.* John Wiley&Sons, Sussex, UK, 3346-3359.

Holland, J. (1975) *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbour.

Homaifar, A., Guan, S. (1991) *A new approach to the traveling salesman problem by genetic algorithm.* Technical Report. North Carolina A&T State University.

Ieong, S., Kao, M.-Y., Lam, T.-W., Sung, W.-K., Yiu, S.-M. (2003) Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *Journal of Computational Biology* 10 (6), 981-995.

Jacobson, A.B., Zuker, M. (1993) Structural analysis by energy dot plot of a large mRNA. *Journal of Molecular Biology* 233, 261-269.

Janiak, A. (1999) *Wybrane problemy i algorytmy szeregowania zadań I rozdziału zasobów*. Akademicka Oficyna Wydawnicza PLJ, Warszawa.

Jardetzky, O., Schmitt, T.H. (1996a) Biological Macromolecules: NMR Parameters. *Encyclopedia of Nuclear Magnetic Resonance* 2 (A-COM), John Wiley&Sons, Sussex, UK, 921-932.

Jardetzky, O. (1996b) Biological Macromolecules. *Encyclopedia of Nuclear Magnetic Resonance* 2 (A-COM), John Wiley&Sons, Sussex, UK, 901-920.

Jardetzky, O. (1996c) NMR in Molecular Biology – a History of Basic Ideas. *Encyclopedia of Nuclear Magnetic Resonance* 1. *Historical Perspectives*, John Wiley&Sons, Sussex, UK, 402-408.

Jeener, J., Meier, B.H., Bachmann, P., Ernst, R.R. (1979) Investigation of exchange processes by 2-D NMR spectroscopy. *Journal of Chemical Physics* 71, 4546-4593.

Johnson, D.S. (1985) The NP–completeness column: an ongoing guide, *Journal of Algorithms* 6, 291-305.

Juan, V., Wilson, C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *Journal of Molecular Biology* 289, 935-947.

Karp, R.M. (1972) Reducibility among combinatorial problems. *Complexity of Computer Computations*, New York: Plenum, 85-103.

Kanehisa, M. (2000) *Post-Genome Informatics*. Oxford University Press, New York.

Kraulis, P.J. (1989) ANSIG: A program for the assignment of protein ^{1}H 2D NMR spectra by interactive graphics. *Journal of Magnetic Resonance* 24*,* 627-633.

Koradi, R., Billeter, M., Engeli, M., Guntert, P., Wuthrich, K. (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance* 135, 288-297.

Langmead, C.J., Yan, A., Lielien, R., Wang, L., Donald, B.R. (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *Journal of Computational Biology* 11/2-3, 277-298.

Lankhorst, P.P, Haasnoot, C.A.G., Erkelens, C., Altona, C. (1984) Carbon-13 NMR in conformational analysis of nucleic acid fragment. *Journal of Biomolecular Structure and Dynamics* 1, 1387-1405.

Lesk, A. M. (2002) *Introduction to Bioinformatics*. Oxford University Press, Oxford.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G., Kessler, H. (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR* 11*,* 31-43.

Linge, J.P., Habeck, M., Rieping, W., Nilges, M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19, 315-316.

Luck, R., Graf, S., Steger, G. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structures. *Nucleic Acids Research* 27, 4208-4217.

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

McDowell, J.A., He, L., Chen, X., Turner, D.H. (1997) Investigation of the structural basis for thermodynamic stabilities of tandem GU wobble pairs: NMR structure of $(rGGAGUUCC)_2$ and $(rGGAUGUCC)_2$. *Biochemistry* 36, 8030-8038.

McDowell, J.A., 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., Pardi, A. (2000) NMR solution structure determination of RNAs. *Current Opinion in Structural Biology* 10*,* 298-302.

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

Moseley, H.N.B., Montelione, G.T. (1999) Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology* 9, 635-642.

Mumenthaler, C., Guntert, P., Braun, W., Wuthrich, K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *Journal of Biomolecular NMR* 10, 351-362.

Neidle, S. (1999) *Oxford Handbook of Nucleic Acid Structure*. Oxford University Press, London, New York.

Neidle, S. (2002) *Nucleic Acid Structure and Recognition*. Oxford University Press, London, New York.

Neuhaus, D., Williamson, M. (1989) *The Nuclear Overhauser Effect in Structural and Conformational Analysis*. VCH Publishers Inc., New York, Weinheim, Cambridge.

Nilges, M. (1999) *Applications of Molecular Modelling in NMR Structure Determination*. Technical Report. European Molecular Biology Laboratory, Heidelberg, Germany.

Osman, I.H., Kelly, J.P. (1995) *Theory and Applications*. Kluwer Academic Publishers, Boston.

Papadimitriou, C.H., Steiglitz, K. (1982) *Combinatorial Optimization. Algorithms and Complexity*. Prentice Hall, New Jersey.

Perkel, J.M., (2004) Technology. How it works. Nuclear Magnetic Resonance Spectrometer. *The Scientist* September 13, 32-33.

Popenda, M. (1998) *Zastosowanie metod magnetycznego rezonansu jądrowego oraz modelowania molekularnego w analizie strukturalnej RNA*. PhD thesis. Institute of Bioorganic Chemistry of Polish Academy of Sciences, Poznan.

Popenda, M., Biała, E., Milecki, J., 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 Acids Research* 25, 4589-4598.

Reeves, C.R. (1993) *Modern Heuristic Techniques for Combinatorial Problems*. McGraw-Hill, London.

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

Ruan, J., Stormo, G., Zhang, W. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 20, 58-66.

SantaLucia Jr., J., Turner, D.H. (1993) Structure of (rGGCGAGCC)2 in solution from NMR and restrained molecular dynamics. *Biochemistry* 32, 12612-12623.

Sattler, M., Schleucher, J., Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in NMR Spectroscopy* 34*,* 93-158.

Setubal, J., Meidanis, J. (1997) *Introduction to Computational Biology*. PWS Publishing Company, Boston.

Stryer, L. (2000) *Biochemia*. Wydawnictwo Naukowe PWN, Warszawa.

Szachniuk, M., Adamiak, R., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M., Błażewicz, J. (2003) A combinatorial analysis of 2D NMR spectra of RNA duplexes. *Currents in Computational Molecular Biology*, 345-346.

Szachniuk, M., Popenda, M., Adamiak R., Błażewicz, J. (2004) The method of an assignment of the magnetization transfer pathway between H6/H8-H1' protons in the 2-dimensional NOESY spectra in nuclear magnetic resonance spectroscopy of the nucleic acids. *Polish Patent Pending Application P364736*.

Tabaska, J., Cary, R., Gabow, H., Stormo. G. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics* 14, 691-699.

Varani, G., Aboul-ela, F., Allain, F.H.T. (1996) NMR investigation of RNA structure. *Progress in NMR Spectroscopy* 29, 51-127.

Varani, G., Tinoco Jr., I. (1991) RNA structure and NMR spectroscopy. *Quarterly Reviews of Biophysics* 24*,* 479-532.

Watson, J.D., Crick, F.H.C. (1953) A structure for deoxyribose nucleic acid. *Nature* 171, 737-738.

Westhof, E., Auffinger, P. (2000) RNA Tertiary Structure. *Encyclopedia of Analytical Chemistry*, John Wiley&Sons, Chichester, UK, 5222-5232.

Wijmenga, S.S., van Buuren, B.N.M. (1998) The use of NMR methods for conformational studies of nucleic acids. *Progress in NMR Spectroscopy* 33*,* 287-387.

Williams, D.H., Fleming, I. (1996) *Spectroscopic Methods in Organic Chemistry*. McGraw-Hill, New York.

Wójtowicz, A. (1994) *Słownik terminów rezonansu magnetycznego*. Ośrodek Wydawnictw Naukowych, Poznań.

Wu, M., SantaLucia Jr., J., 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 Willey&Sons, New York.

Zidek, L., Stefl, R., Sklenar, V. (2001) NMR methodology for the study of nucleic acids. *Current Opinion in Structural Biology* 11, 275-281.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C-Y., Powers, R., Montelione, G.T. (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology* 269*,* 592-610.