Institute of Computing Science
Poznan University of Technology

Bioinformatics methods of motif analysis in RNA structures

M.Sc. Joanna Miśkiewicz

Supervised by

Prof. Dr Eng. Marta Szachniuk

Poznań, 2021

Podziękowanie

Chciałbym podziękować wszystkim,
bez których pomocy i wsparcia
niniejsza praca nie mogłaby powstać.

Serdecznie dziękuję mojej promotorce,
prof. dr hab. inż. Marcie Siechinskiej, za całą
okazaną pomoc, wsparcie, oraz cennych uwag
i przede wszystkim - za opiekę przez całość
moich studiów, nie tylko doktoranckich.

Dziękuję mojej rodzinie i bliskim za ogromne
wsparcie duchowe, a w szczególności mojemu mężowi
i przyjaciołom, którzy również obrali ścieżkę naukową.
Wasze rady i troska pomagały mi w sposób nieoceniony.

Abstract

Bioinformatics is one of the most needed and fastest developing scientific branch of the century. Bioinformaticians generate and process life science-related data, perform calculations, and solve problems in a time-efficient manner. One of the problems they challenge is searching and analyzing motifs in biological data. In structural biology and bioinformatics, motif discovery and analysis help reveal the relationship between molecule structures and their functions within living organisms. This, in turn, impacts the development of molecular medicine – medical diagnosis, development of targeted therapies, drug design – and biotechnology.

This doctoral dissertation focuses on motif analysis in RNA molecules of various organisms. It addresses several motif-related problems concerning different levels of structure organization – sequence, secondary, and tertiary structure. The first problem targeted motifs in precursor microRNAs (pre-miRNAs) of *Arabidopsis thaliana*, one of the plant model organisms. We discovered a repetitive pattern (small internal loops) in the close vicinity of miRNA:miRNA* duplex – a potential recognizable site for the cleavage machinery. It led to further study of pre-miRNAs in all green plants (*Viridiplantae*), for which we analyzed all three structural levels. Results of this research confirmed previous observations for *Arabidopsis thaliana*. The next problem required an analysis of the primary transcript of human miRNA (pri-miR-125a). The transcript was tested in two variants with

the single nucleotide polymorphism (SNP). Bioinformatic analysis indicated variant-related protein binding to the pri-miR-125a sequence as well as variant-related 2D conformations. The third issue addressed within the scope of the doctoral thesis concerned quadruplexes, one of the most complex and least recognized structural motifs occurring in nucleic acids. These motifs were explored on all structural levels. Deep insight into their features led to proposing a new classification based on the secondary structure topology. The secondary structure of tetrads and quadruplexes can now be represented in an extended top-down dot-bracket notation and drawn in a top-down arc diagram – both representations we developed within the scope of the presented work. We analyzed all available bioinformatics resources to evaluate their usefulness for studying RNA quadruplexes. We also investigated human miRNA potential to form quadruplexes by applying a regular expression matching algorithm. Finally, we developed a new database named ONQUADRO to collect and analyze data on experimentally determined quadruplex structures.

All the presented *in silico* analysis was performed on publicly available data using third-party and own computational methods.

List of publications

A1. **Miskiewicz J**, Tomczyk K, Mickiewicz A, Sarzynska J, Szachniuk M (2017) Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *BioMed Research International* 2017: 6783010 (doi:10.1155/2017/6783010).

A2. **Miskiewicz J**, Szachniuk M (2018) Discovering Structural Motifs in miRNA Precursors from the Viridiplantae Kingdom. *Molecules* 23(6): 1367 (doi:10.3390/molecules23061367).

A3. Lehmann T, **Miskiewicz J**, Szostak N, Szachniuk M, Grodecka-Gazdecka S, Jagodzinski PP (2020) In Vitro and in Silico Analysis of miR-125a with rs12976445 Polymorphism in Breast Cancer Patients. *Applied Sciences* 10(20): 7275 (doi:10.3390/app10207275).

A4. *Popenda M, ***Miskiewicz J**, Sarzynska J, Zok T, Szachniuk M (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics* 36(4): 1129–1134 (doi:10.1093/bioinformatics/btz738).

* joint first authorship

A5. **Miskiewicz J**, Sarzynska J, Szachniuk M (2021) How bioinformatics resources work with G4 RNAs. *Briefings in Bioinformatics* 22(3): bbaa201 (doi:10.1093/bib/bbaa201).

A6. Zok T, Kraszewska N, **Miskiewicz J**, Pielacinska P, Zurkowski M, Szachniuk M (2021) ONQUADRO: a database of experimentally determined quadruplex structures. *submitted for publication.*

Table 1: Bibliometric parameters.

Article ID	PY[1]	IF (PY[1])	5-IF (2021)	MEiN[2] (PY[1])	MEiN[2] (2021)	Quartile (WoS[3])	Rank (WoS[3])
A1	2017	3.411	3.620	25	70	Q3	80/140
A2	2018	4.411	4.587	30	100	Q2	63/178
A3	2020	2.679	2.736	70	70	Q2	38/91
A4	2020	6.937	8.470	200	200	Q1	3/58
A5	2021	11.622	10.288	140	140	Q1	1/66
A6	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Total		29.060	29.701	465	580	–	–

Rank is given for the journal in the field of computational biology and bioinformatics if possible, otherwise in the multidisciplinary area.

Table 2: Number of citations and H-index.

Article ID	Web of Science (all citations)	Web of Science (without self-citations)	Scopus	Google Scholar
A1	12	10	13	16
A2	8	7	8	9
A3	0	0	0	0
A4	6	6	6	8
A5	1	1	1	4
A6	n/a	n/a	n/a	n/a
Total	27	24	28	37
H-index	3	3	3	4

[1]Publication Year
[2]The Ministry of National Education (Poland)
[3]Web of Science

Contribution to publications

I declare the following contributions to the publications underlying my doctoral dissertation:

A1. I collected and preprocessed the datasets for analysis, developed the *PatternSearch* algorithm together with K. Tomczyk, performed motif search, analyzed the sequences and secondary structures, and compiled the results. I was responsible for the preparation of all the graphics and plots, participated in manuscript writing, and made all additional works related to the revision process.

A2. I performed all the research presented in this publication under the supervision of prof. M. Szachniuk. I also compiled the results and prepared the initial version of the manuscript, including figures.

A3. I was responsible for the investigation and *in silico* methodology selection, sequence analysis, testing the secondary structure prediction algorithms to select the appropriate ones, secondary structure prediction and selection of best models, structure visualizations. I also participated in the results' interpretation and compilation, and preparation of the manuscript.

A4. I contributed to the development of a new topology-based classification for tetrads (ONZ) and quadruplexes (ONZM), analyzed the relationship of ONZM classes with Webba da Silva nomenclature, conducted a multi-faceted and multi-level statistical analysis of tetrad- and G4 structural features, analyzed the distribution of ONZ classes, compiled experimental results, wrote parts of the publication, and visualized statistical data.

A5. I performed all the research presented in this publication under the supervision of both co-authors. I compiled the results together with dr J. Sarzynska. I also prepared the initial version of the manuscript.

A6. I designed the database system named ONQUADRO in cooperation with dr T. Zok and prof. M. Szachniuk. I supervised the implementation of the interface's prototype (made by P. Pielacinska), extensively tested a web application implemented by N. Kraszewska, and prepared the user tutorial. I wrote the first version of the manuscript and prepared most of the figures.

.....

(Signature)

Contents

Acknowledgements **i**

Abstract **ii**

List of publications **iii**

Chapter 1 Introduction **1**

1.1 RNA structure and role 3

1.2 RNA structure representations and formats 6

1.3 Motif discovery and analysis 12

Chapter 2 Main results **22**

2.1 Motifs in plant pre-miRNA 22

2.2 Motifs in human pri-miRNA and miRNA 25

2.3 Quadruplex motifs 27

Bibliography **30**

Publication reprints **48**

Co-author declarations **153**

Extended abstract in Polish **165**

Appendices **167**

AppendixA Participation in research projects **168**

AppendixB Conference presentations **169**

AppendixC Awards and distinctions **172**

CHAPTER 1

Introduction

Nucleic acid structures are complex designs that encode crucial biological functions. Knowledge about them is primarily gained from empirical studies performed in laboratory experiments by life science researchers. Although such an approach allows for the retrieval of reliable information, at the same time, it is time-consuming and resource-dependent. The need for alternative, more efficient solutions has led to the development of computer-based methods dedicated to life science problems. These have become part of bioinformatics - one of the youngest branches of interdisciplinary sciences. Bioinformatics aims to create models and algorithms that complement or replace experimental studies and solve problems originating in biosciences at correspondingly lower costs. Today, support from bioinformatics is present wherever life science research goes on.

The idea of applying computers in biological research dates back to the 1960s [Gauthier et al. (2018); Hagen (2000)]. At that time, Margaret Oakley Dayhoff, a researcher interested in molecular evolution, started using programming to establish the sequences of protein molecules [Gauthier et al. (2018); Hagen (2000)]. With the increasing number of determined sequences, the number of molecular databases was growing. In the 1960s, biophysicist Cyrus Levinthal and his group built a 3D *in silico* model of

cytochrome c [Hagen (2000)]. In the same decade, new algorithms for sequence alignment (Saul Needleman and Christian Wunsch's algorithm) appeared, and computers became indispensable in phylogenetic analysis, significantly reducing the research time [Hagen (2000); Ouzounis & Valencia (2003)]. Since that time, bioinformatics has reached more scientific fields, including biophysics and biochemistry [Ouzounis & Valencia (2003)].

Bioinformatics owes its success to computer science, statistics, and a solid knowledge of biological processes, molecular, and structural biology [Claverie (2000); Hagen (2000); Fenstermacher (2005); Szachniuk (2019)]. A strong understanding of processes, which take place in living organisms gives us the ability to approach them from a multi-level and multidimensional perspective. Incoming biological data, especially nucleic acid sequences, contribute to collaborations between specialists of different disciplines and introduce scientists to a computational biology field [Searls (2010)]. Throughout the years, such collaborations resulted in many methods and computational tools dedicated to biological problems. Scientists can now choose between various sequencing approaches, structural prediction methods, and modeling programs. The vastness of choice led to the opening of computational competitions. The CASP (critical assessment of structure prediction) [Kryshtafovych et al. (2019)] and RNA-Puzzles [Miao & Westhof (2017); Miao et al. (2020)] projects constitute an example. CASP addresses scientists involved in protein structure prediction, whereas RNA-Puzzles is for RNA structure prediction groups. Structure prediction methods are in constant improvement – competitions like CASP and RNA-Puzzles help laboratories adjust their predicting algorithms through established challenges. Predicted models, which with each subsequent competition are more and more resembling the original structure, testify that structural issues can be solved by applying structural bioinformatics. Within a given

sequence, bioinformatics provides *in silico* analysis, motif search, and discovery, revealing homological sequences and helping to construct phylogenetic trees. Throughout the years, bioinformatics delivered programs and provided multiple approaches to predict 2D and 3D structures from nucleic acid sequences. In this dissertation, the structural bioinformatics abilities and challenges in the context of RNA motifs are presented and discussed.

1.1. RNA structure and role

RNA (ribonucleic acid) is a strand composed of modules. They are called nucleotides and connect by phosphodiester bonds. Each nucleotide is built of sugar (ribose), a nucleic base (adenine, guanine, cytosine, uracil), and a phosphate group. In an RNA chain, the nucleic base is the only variable within nucleotides. Thus, the chain is often described in a simplified form by enumerating the bases. This composition – a sequence of bases – is the primary structural level of the nucleic acid.

The secondary structure of RNA represents pairings between its bases. RNA usually exists as a single-stranded molecule. However, bases within a strand may bind together, which results in forming single- and double-stranded fragments in the structure [Hames & Hooper (2010); Hoehndorf et al. (2011)]. Within the double-stranded regions, one strand (one part of an RNA chain) is directed from the 5'- to 3'-end, whereas the second one is oriented in the opposite direction, giving the anti-parallel double helix segment [Hoehndorf et al. (2011)]. The binding between nucleotides of this helix is governed by the rules established by Watson and Crick [Watson & Crick (1974)]. In RNA, Watson-Crick pairings – also called

canonical – include A-U base pair sealed with two hydrogen bonds and G-C pair with three hydrogen bonds [Halder & Bhattacharyya (2013); Šponer et al. (2005)]. The canonical pairings in RNA also include the G-U wobble pair (with two hydrogen bonds) that does not follow the Watson-Crick rule. In addition, many non-canonical base pairs form in RNA structures [Leontis & Westhof (2001); Hoehndorf et al. (2011)]. They were classified by Leontis and Westhof, who established twelve different families of edge-to-edge interactions based on the types of interacting edges and the glycosidic bond orientation (cis or trans) [Leontis & Westhof (2001); Laing & Schlick (2011)]. Each nucleotide can interact along one of three edges: Watson-Crick (W), Hoogsteen (H), or Sugar edge (S) [Leontis & Westhof (2001); Laing & Schlick (2011)]. A detailed description of the secondary structure includes both a list of base pairs and their classification.

In the appropriate, stable environment, i.a. specific temperature and ion homeostasis, RNA can fold into the biologically active tertiary form [Eric & Pascal (2006); Hoehndorf et al. (2011); Zemora & Waldsich (2010)]. Molded functional conformation of RNA takes a form of a helical structure stabilized by ions and hydrogen interactions between the structural elements [Halder & Bhattacharyya (2013); Draper (2004); Rother et al. (2011)]. Experimentally solved 3D RNA structures are deposited in the RCSB PDB repository [Burley et al. (2020)]. However, many RNAs have unknown structures that were not solved experimentally. They become a great challenge for bioinformaticians who try to predict the 3D shape using computational methods. RNA molecule modeling has to face many obstacles – it is worth notice that similar sequences may not fold into similar 3D structures [Wiedemann & Miłostan (2017)]. On the other hand, even if sequences of the same molecule family diverged in time, their 3D motifs may be conserved [Hoehndorf et al. (2011)]. Thus, despite identifying

secondary structure elements of a given sequence and having multiple homologous sequences, modeling the 3D structure of nucleic acids in a proper way is still challenging [Leontis & Westhof (2012)].

RNA structure reflects the molecule function – both in the context of the overall structure and from the perspective of selected motifs that formed in the molecule [Conn & Draper (1998); Li et al. (2020)]. The RNA structure element may constitute a binding site for proteins [Mortimer et al. (2014); Brosius & Raabe (2016)]. Protein recognizes specific motif in the RNA molecule by the motif's domain [Corley et al. (2020)]. Formed RNA-protein complexes (RNPs) can be further transported into destined cellular locations and fulfill their biological purpose [Brosius & Raabe (2016)]. Depending on the type (i.a. mRNA, tRNA, miRNA, sRNA), RNA plays important roles in cellular processes [Breaker & Joyce (2014); Mortimer et al. (2014); Miao & Westhof (2017); Doudna (2000)]. They include regulation of gene expression (on each level of this process), gene silencing, and catalytic functions [Breaker & Joyce (2014); Mortimer et al. (2014); Doudna (2000); Kun et al. (2015); Leontis & Westhof (2012)]. The list of RNA functions is not limited to the above ones. RNAs are involved in various diseases' pathways and machineries, such as cancer and neurodegenerative diseases [Cammas & Millevoi (2016); Cooper et al. (2009); Wu & Kuo (2020)]. Understanding RNA structure and the related functions may help in the development of RNA therapies against these disorders [Li et al. (2020); Cooper et al. (2009); Kim (2020)]. There is also a hypothesis that considers RNAs as the first information carriers that functioned similarly to current enzymes, proteins, and DNA molecules [Robertson & Joyce (2010); Higgs & Lehman (2014); Kun et al. (2015)]. The concept of RNA World, in which the origins of life on Earth have emerged from RNA molecules, has a growing number of supporters [Robertson & Joyce

(2010); Pearce et al. (2017); Synak et al. (2020); Szostak et al. (2017); Wasik et al. (2019)]. Among them, bioinformaticians work on simulation models of how RNA molecules may have acted on Earth billions of years ago [Higgs & Lehman (2014); Synak et al. (2020); Szostak et al. (2017); Wu et al. (2017)]. The molecular dynamics system may help to reveal if the RNA world hypothesis is a valid assumption of how we originated.

1.2. RNA structure representations and formats

On every level of organization, RNA structure can be represented in various manners, depending on how much and what structural information we highlight and how we process the structural data. There are many machine representations – textual and graphical – developed for storing and visualizing molecular structures. Thus, algorithmic scientists can choose the proper one to design an efficient algorithm for solving a structural problem. Not all representations are equally suitable for automatic search and analysis of structural motifs. The following paragraphs present the basic models to represent molecular structures and file formats to store them.

Sequence

The sequence of nucleic acid is usually represented as a word constructed based on a four-letter alphabet – {A,U,G,C} for RNA or {A,T,G,C} for DNA. A corresponds to Adenine, C represents Cytosine, G is for Guanine, U for Uracil, and T for Thymine. Alternatively, 3-letter identifiers can be used to encode nucleotides: ADE, URA, GUA, CYT, THY. In the nucleic acid chain, we distinguish two ends: 5'-end and 3'-end [Hames & Hooper (2010)]. The sequence is always written from the 5'- to the 3'-end, which

can or cannot be indicated. For example, the following sequences encode the same RNA molecule:

- 5'-GGAACCGGUGCGCAUAACCACCUCAGUGCGAGCAA-3'
- GGAACCGGUGCGCAUAACCACCUCAGUGCGAGCAA
- GUA GUA ADE ADE CYT CYT GUA GUA URA GUA CYT GUA CYT ADE URA ADE ADE CYT CYT ADE CYT CYT URA CYT ADA GUA URA GUA CYT GUA ADE GUA CYT ADE ADE

The above suggestions do not exhaust the subject of sequence encoding. Experimental determination of a sequence of the studied molecule does not always give an unambiguous read. Sometimes we expect a pyrimidine at some position in the sequence, but we do not know which one. This information, however, is also worth encoding. In such cases, one can apply the extended alphabet to encode molecule sequence proposed by IUPAC (International Union of Pure and Applied Chemistry) [Johnson (2010); Hendrix et al. (2005)]. This extended alphabet allows putting a symbol representing two, three, or four bases in one sequence position. The full IUPAC alphabet for RNA sequences is presented in Table 1.1.

Sequences are usually written down in the FASTA or FASTQ file [Mills (2014); Pearson (2016)]. FASTA is a simple text-based format that contains a sequence and additional comment line at the beginning of each file [Mills (2014); Pearson (2016)]. The comment line is used to store a molecule description. FASTQ format stores information about sequence and quality scores. It is mainly used for handling sequence reads [Mills (2014)].

Secondary structure

The secondary structure of nucleic acid includes information on interactions between nucleotides. One of the simplest and most common representation

Table 1.1: IUPAC codes for RNA nucleotides.

Nucleotide(s)		Symbol
A	Adenine	A
G	Guanine	G
C	Cytosine	C
U	Uracil	U
A or G	puRine	R
C or U	pYrimidine	Y
A or C	aMino	M
G or U	Ketone	K
C or G	Strong interaction	S
A or U	Weak interaction	W
A, C or U	not G	H
C, G or U	not A	B
A, C or G	not U	V
A, G or U	not C	D
A, C, G or U	aNy nucleotide	N

of the secondary structure is a dot-bracket notation [Hofacker et al. (1994); Ponty & Leclerc (2014); Mattei et al. (2014)]. In this format, each unpaired nucleotide is represented as a single dot, whereas two paired bases are encoded as opening and closing parenthesis [Ponty & Leclerc (2014)]. Opening bracket refers to the nucleotide closer to the 5'-end, closing bracket reflects the nucleotide closer to the 3'-end [Ponty & Leclerc (2014)]. This simple notation can be extended by adding more symbols to encode pseudoknots [Ramlan & Zauner (2008)] or adding the second line [Popenda et al. (2019)] to represent quadruplex motifs. G-quadruplex secondary structures can also be presented as an additional $'+'$ sign in dot-bracket notation (such representation is used in RNAfold – one of the Vienna services [Lorenz et al. (2011)]). A 2D structure can also be presented as a special list of characters [Liao et al. (2006); Zhang et al. (2016)]. The secondary structure of nucleic acids uses IUPAC encoding [Johnson (2010) for a non-paired nucleotides, and uses $'$ symbol for paired nucleotide [Liao et al. (2006); Zhang et al. (2016)]. Thus, for example, the G' is encoding G-C pairing, U' encodes U-A pairing.

A 2D structure is often represented as a list of base pairs. Every element in the list is a pair of numbers representing two paired nucleotides (numbers refer to their order in the strand). The list can include additional information about nucleotides or their neighbors. Depending on the amount of additional data, we can have different formats to write down the secondary structure data. BPSeq format is characterized by an information section, which contains comments on encoded data, and a structure section, divided into three columns [Ponty & Leclerc (2014)]. The first column represents the position of the nucleotide in the sequence (begins with 1), the second is a base encoded according to IUPAC standards [Johnson (2010)], the third one is a position of a paired nucleotide [Ponty & Leclerc (2014)]. If nucleotide is unpaired, third column contains 0 value [Ponty & Leclerc (2014)]. Extension to the BPSeq is a Connectivity Table (CT) format [Ponty & Leclerc (2014)]. The header of the CT file has sequence length and additional information of the sequence, the main body of the CT file is a table that has a base position, its IUPAC-coded symbol, position of the previous base, position of next base, paired nucleotide position (if any, if not - 0), and the original number of nucleotide [Ponty & Leclerc (2014)]. The next example of a 2D format is BEAR, a proposition by [Mattei et al. (2014)]. This representation encodes secondary structure elements (loops, stems, bulges) with a specified character from the predefined structural alphabet [Mattei et al. (2014)]. Another way to present nucleic acids 2D structures is using a squiggle plot or a dot plot [Churkin & Barash (2013)]. The first representation places bases along the line that runs in 5' to 3' direction and sets base-pairing interactions as a straight line [Churkin & Barash (2013)]. Following representation, the dot plot, presents base pairings by a dot in the two-dimensional matrix where a sequence is in both x- and y-axis [Churkin & Barash (2013)]. A graphical representation of a 2D structure can take a form of an arc diagram [Wattenberg (2002)].

Arc diagram can match not only the base pairs but it can also be used for showing repetitive subsequences [Wattenberg (2002)]. Additional bottom arcs might be added to reflect the quadruplex motif [Popenda et al. (2019)]. Quadruplexes can also be visualized using VARNA [Darty et al. (2009)]. VARNA (Visualization Applet for RNA) displays sequence in a circular, linear, or planar graph format that encodes non-canonical base pairs using Leontis-Westhof nomenclature [Darty et al. (2009); Leontis & Westhof (2001)]. 2D structures of nucleic acids can also be represented in XML format, i.a. RNAML [Waugh et al. (2002); Ponty & Leclerc (2014)], as 2D curves ([Yao et al. (2005)]), as 3D graphics ([Zhang et al. (2016); Fu et al. (2018)]), and projected as a graph [Schlick (2018); Laing & Schlick (2011)]. Graph models apply to both 2D and 3D structure representations. Tree graphs can represent secondary structure motifs, excluding pseudoknots [Schlick (2018); Laing & Schlick (2011); Gan (2003)]. Stems can be visualized as edges, while loops, bulges, and junctions are mapped to graph vertices [Schlick (2018); Laing & Schlick (2011); Gan (2003)]. Dual and secondary structure graphs represent RNA motifs including pseudoknots [Schlick (2018); Laing & Schlick (2011); Gan (2003)]. Dual graphs can have reversed assignments for vertices and edges in comparison to tree graphs – thus, the vertices present stems and edges present the remaining set of the secondary structure motifs [Schlick (2018); Laing & Schlick (2011); Gan (2003)]. In the secondary structure graphs, vertices are reserved for nucleotides and edges for base pairs [Laing & Schlick (2011)]. In the 3D graph representation, vertices represent nucleotides of a nucleic acid and edges correspond to interactions between them [St-Onge et al. (2007)]. The type of interaction can be denoted as a label of an edge [St-Onge et al. (2007)]. It is worth to notice that the graph theory used in nucleic acids representations may also help discovering nucleic acid motifs and structure comparisons [Laing & Schlick (2011); Gan (2003)].

Tertiary structure

Textual representations of the 3D structure include algebraic, geometric [Ryu et al. (2020); Gong & Fan (2019)], trigonometric [Zok et al. (2013)], and probabilistic [Frellsen et al. (2009)] models. The first one is most commonly used. It describes the 3D fold by listing all the atoms from the structure along with their coordinates in three-dimensional space. In the geometric representation, distances between nucleotides (or atoms – depending on the model's resolution) are given [Gong & Fan (2019)]. Trigonometric representation is defined by a set of dihedral angles [Zok et al. (2013)], whereas probabilistic is based on atom distribution [Frellsen et al. (2009)].

The most popular file formats to store the three-dimensional structure of a molecule are PDB (Protein Data Bank) and mmCIF (macromolecular Crystallographic Information File) [Westbrook & Fitzgerald (2005)]. Both are dedicated to algebraic representation. PDB format describes the 3D characteristics of a molecule, including atomic coordinates and hydrogen bonds, divided by record types (i.a. HEADER or ATOM) [Westbrook & Fitzgerald (2005)]. mmCIF is a dictionary-structured format that contains information of i.a. crystallographic experimental details and atom sites [Westbrook & Fitzgerald (2005)].

There are also many graphical models designed for the tertiary structures of biomolecules. One of the most popular is a ball-and-stick model. It presents atoms as spheres (balls) and bonds between them as sticks [Roy et al. (2015)]. The surface representation is a merged area of atoms' spaces that are available for interactions [Goodsell (2005)]. The ribbon model shows a molecule backbone [Kuttel et al. (2006)]. It is one of the most common representation used for all types of molecules [Goodsell (2005)]. Another popular visualization mode is called a cartoon. It renders the structure

into a simplified model composed of arrows and ribbons [Kozlíková et al. (2016)].

Table 1.2: Popular representations of a molecule structure.

	Textual	Graphical
Sequence	single-letter coding three-letter coding IUPAC	
2D structure	dot-bracket notation extended dot-bracket notation extended IUPAC encoding BP (Base Pairs) BPSeq (Base Pairs & Sequence) CT (Connectivity Table) BEAR XML	squiggle plot dot plot classical diagram arc diagram tree graph circle graph mountain plot 2D curves
3D structure	algebraic geometric trigonometric probabilistic	ball-and-stick surface ribbon cartoon

1.3. Motif discovery and analysis

A definition of a motif – a repetitive pattern, a reoccurring theme – varies depending on the research area. Even among structural motif-search resources, restrictions of motif characteristics can be program-specific [Jossinet et al. (2007)]. Motifs appear in art, literature, science, they can be found in various aspects of our lives, even when we do not notice them at the first sight. Motif recognition can help find defects in patterns, i.a. during the texture analysis [Ngan et al. (2008)] or support establishing a potential function of a molecule [Nicodème et al. (2002)]. Motif analysis can lead to a better understanding of the world and ourselves but finding or discovering

a motif also raises challenges and comprises non-trivial problems [Lacroix et al. (2006); Yu et al. (2020); Xiao et al. (2019)]. This section refers to motifs in different scientific fields with particular emphasis on structural motifs in terms of nucleic acids structures.

1.3.1. Motifs in life sciences

The level of difficulty in finding motifs depends on the used methods, dataset, and the skills of laboratory researchers. In particular cases, the time is another crucial component of the investigation. It is a key parameter while searching for a time series motif during a behavior monitoring in a specified time frame [Mueen (2014); Torkamani & Lohweg (2017)]. This method is used in medicine (EEG, ECG) and – among others – in seismology, telecommunication, entomology, and ornithology studies [Mueen (2014); Torkamani & Lohweg (2017)].

In biochemistry and systems biology a motif can be described as a series of interactions, activatory or inhibitory, within components of a biochemical reaction network [Tyson & Novák (2010); Alon (2007); Jazayeri & Yang (2020)]. Each processed information in a living organism is a part of a reaction chain that can be divided into small interaction motifs that play a specific function in a system [Tyson & Novák (2010)]. The interaction network may concern transcription factors proteins [Alon (2007)], gene regulatory processes [Hallinan & Jackway (2005)], metabolic pathways [Lacroix et al. (2006)], and brain networks [Battiston et al. (2017)]. In the relation to the recent threat situation caused by SARS-CoV-2 virus, we can also search for epidemic motifs in the epidemic dynamics network, using the method based on ordinary differential equations, proposed by [House et al. (2009)].

Knowledge of these network motifs may help in better understanding of

an end-to-end process that happens in a cell or a system [Tyson & Novák (2010)]. An example of one of the tools that was employed by bioinformaticians for interaction motif discovery and analysis in reaction networks are Petri nets modeling [Liu & Heiner (2013)]. The chemical reactions modeled by this method can contribute to find crucial dependencies and positive/negative factors in diseases, i.a. atherosclerosis [Formanowicz et al. (2018)]. Other useful tools for network representations and network motif search are graph models [Yu et al. (2020); Lacroix et al. (2006); Angulo et al. (2015)]. In graph-coded biological networks, motifs are described as a repetitive subgraph patterns [Yu et al. (2020)]. The decomposition of a network enables revealing the functional modules in which motifs can be searched for [Lacroix et al. (2006)]. Discovered network motifs might help scientists reveal characteristics and key processes in analyzed biological (regulatory or metabolic) networks [Yu et al. (2020); Lacroix et al. (2006)]. Algorithms for motif discovery in biological networks using graphs have already been implemented in several programs (i.a. NetMODE [Li et al. (2012)], MFinder [Kashtan et al. (2004)]) which can be downloaded and used on users local machines [Yu et al. (2020)]. In gene regulatory networks, motifs are recognized as DNA binding sites for transcription factors (proteins) [Tompa et al. (2005); Li & Tompa (2006); Zambelli et al. (2012); Sandve & Drabløs (2006)]. For *in silico* discovery of such motifs, scientists have been using i.a. machine learning [Zhang et al. (2017); He et al. (2020)] and discriminative algorithms [Redhead & Bailey (2007); Grau et al. (2013)] or combination of those two methods [Hu et al. (2019)].

An interesting example of motif search and analysis are particularly the ones that can be explored by different techniques. G-quadruplexes and i-motifs are good representatives of such group. These structural motifs can be under investigation with analytical chemistry methods, e.g. using NMR

spectroscopy [Alba et al. (2016); Lin et al. (2019)]. In the crystal structure, the G4 motif can conform a pattern of interactions, i.a. hydrogen bonds [Rehman et al. (2018)]. On the other hand, the i-motif and G-quadruplexes can be a part of structural bioinformatics or computational biology research [Zok et al. (2020); Belmonte-Reche & Morales (2019)]. Structural motifs are discussed in the following sections.

1.3.2. Structural motifs

A biological motif can be described as an interaction (i.a. hydrogen bond), a regulatory element (i.a. transcription factor binding site), sorting signal (i.a. mRNA signal for directing it into hepatic extracellular vesicles [Szostak et al. (2014)] or protein signal that helps recognizing to which cellular compartment it should be destined [Sancho-Andrés et al. (2016)]), or as a part of a network (i.a. chemical reaction in a metabolic pathway). In molecular structures, motifs can be analyzed by structural levels: sequence, secondary, and tertiary structure. RNA motifs allow correct cell functioning by enabling the controlled release of information [Dandekar (2002)]. Discovering and understanding the role of structural motifs in biological units help to understand the whole process they are a part of and apply this knowledge for medical treatments.

Sequence motifs

RNA sequence motif can be described as a reoccurring word with a particular biological significance. The sequence of a motif may not always be reflected as a simple combination of four bases. The consensus motif sequence can take a form of a word from an extended alphabet by IUPAC [Johnson (2010); Hendrix et al. (2005)]. For instance, the GNRA tetraloop motif for receptors interactions [Fiore & Nesbitt (2013)], always start with

guanine and ends with adenine. On the second position can appear any of four nucleobases (N), the third position is a purine (R), which means it is occupied by either the guanine or adenine.

Instead of IUPAC coding, positions occupied with more than one possible base can also be presented in square brackets. Therefore, the GNRA loop can be coded as follows: $G[AUGC][AG]A$. If a motif constitutes a repetition of nucleobase, the number indicator is applied. This format is often utilized in G-quadruplex (G4) motifs. G-quadruplexes appear in G-rich sequences and have the unique composition of four tracts of guanines (typically two, three, or more) divided by subsequences of equal or unequal length (depending on a G-quadruplex) [Griffin & Bass (2018)]. Algorithms for finding putative G4s within given sequence or set of sequences are usually arbitrary searching for a specified motif (regular expression), i.a. $G_3N_{[1-7]}G_3N_{[1-7]}G_3N_{[1-7]}G_3$ [Maizels (2012); Takahashi et al. (2012)]. Discovering novel sequence motifs even when using IUPAC coding or regular expression is a non-trivial problem, considered to be qualified as an NP-hard problem, and thus, constituting a challenge for bioinformaticians and computational biologists [Li et al. (2010); Rajasekaran & Dinh (2011); Rampášek et al. (2016); Ashraf & Shafi (2020)].

Secondary structure motifs

RNA is usually a single-stranded molecule with the ability to form double-stranded regions [Chheda & Gupta (2014)]. Its secondary structure describes paired and unpaired fragments [Batey et al. (1999)] that compose into basic secondary structure elements and their layout. The list of secondary structure elements include:

- single strand (between duplexes or at 5' or 3' end),
- duplex, double helix, stem (double-stranded region),

- loop
  - apical loop (loop closed by one canonical base pair),
  - internal loop (loop closed by two canonical base pairs),
    - symmetric internal loop (number of unpaired nucleotides is equal in both single-stranded fragments),
      - mismatch (one unpaired nucleotide on each strand),
    - asymmetric internal loop (number of unpaired nucleotides differs between strands) - for example, bulge (unpaired nucleotide(s) only on one of the strands),
  - mulitbranched loop, n-way junction (intersection of three or more stems, loop closed by at least three canonical base pairs),
- pseudoknot (base-pairings between single-stranded region and hairpin loop) [Chastain & Tinoco (1991); Batey et al. (1999); Eric & Pascal (2006); Chheda & Gupta (2014)].

A schematic view of basic secondary structure motifs is presented in Figure 1.1. Complete RNA secondary structure is a combination of the listed elements. One of the most often RNA 2D structure motif is a hairpin composed of an apical loop and adjacent duplex. The other common motif is a "cloverleaf" that appears in transfer RNA (tRNA) [Hames & Hooper (2010); Batey et al. (1999)]. The cloverleaf in tRNA consists of four arms: three hairpins (anticodon, D-, and T-arm) and the acceptor stem [Hames & Hooper (2010); Batey et al. (1999)]. Even though some tRNAs have additional (variable) arm, the cloverleaf structure remains highly conserved [Hames & Hooper (2010); Zell et al. (2002); Westhof & Auffinger (2012)].

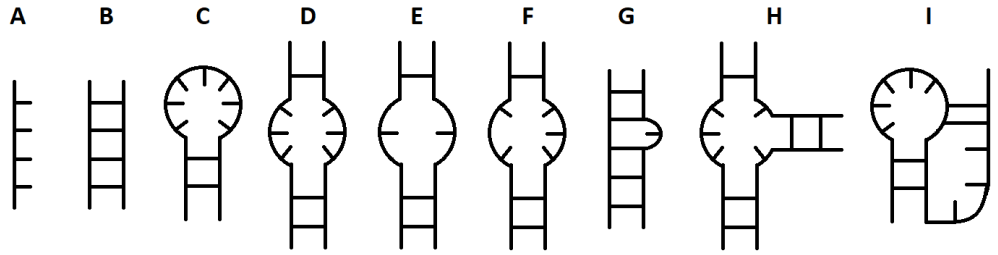


Figure 1.1: Secondary structure building blocks: (A) single strand, (B) duplex, (C) hairpin loop, (D) symmetric internal loop, (E) mismatch, (F) asymmetric internal loop, (G) bulge, (H) junction, and (I) pseudoknot.

Tertiary structure motifs

RNA 3D structure represents a complete, biologically active molecule [Batey et al. (1999); Hoehndorf et al. (2011)]. The folded RNA architecture presents positions of all of the molecular atoms, as well as the tertiary interactions that bind the structural motifs (including non-Watson-Crick base pairings) [Halder & Bhattacharyya (2013); Picardi (2015); Miao & Westhof (2017)]. RNA folding is often stabilized by hydrogen bonds and metal ion bindings [Butcher & Pyle (2011); Hendrix et al. (2005)]. These determined 3D structures are typically solved by NMR spectroscopy or X-ray crystallography methods, however, bioinformatics allows to predict these structures using *in silico* algorithms (under the condition of properly identified 2D structural building blocks) [Hendrix et al. (2005); Leontis & Westhof (2012)]. As a secondary structure can be determined by the combination of structural elements, similarly 3D structure can be described. The following set can constitute an example of such 3D structural elements [Batey et al. (1999); Hendrix et al. (2005); Brunel et al. (2002); Miao & Westhof (2017)]:

- double helix,
- loop-loop interaction,
- ribose zipper,
- coaxial stacking,
- U-turn,
- S-turn,
- A-minor interaction,
- base triples,
- tetrad,
- quadruplex,
- tetraloop.

The above list is just a subset of 3D elements that can be found in nucleic acids. It can be simplified into three types of interactions: between two helices (double-stranded helical regions), between two unpaired regions (non-helical regions), and between one unpaired fragment and a double-stranded helix [Eric & Pascal (2006); Batey et al. (1999)]. An example motif of tertiary interactions between helical and unpaired regions is a GNRA tetraloop motif [Batey et al. (1999); Halder & Bhattacharyya (2013)]. This 4-nucleotide loop has a sheared base pairing between G and A with hydrogen bonds between G_{N3} and A_{N6} atoms and G_{N2} and A_{N7} atoms [Batey et al. (1999)]. The resulting modifications of a second (N) and third (R) nucleotides may lead to divergent hydrogen bonds but thermodynamic stability remains similar between all GNRA variants [Batey et al. (1999)]. In Figure 1.2 the example GNRA tetraloop is presented.

1.3.3. Motif searching methods

Determining the motif existence is a demanding venture and still may not lead to the eligible results. Nevertheless, the possible outcome of discovering the recognizable pattern in a process or a structure is worth explo-

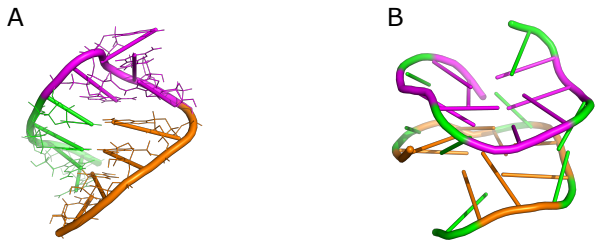


Figure 1.2: (A) GAGA (i.e. GNRA) tetraloop (in magenta) in 1ZIG RNA structure, 5'-end is colored in green, 3'-end is orange; (B) 2RQJ RNA structure with a quadruplex motif; guanines of the first chain (top tetrad) are colored in magenta, guanines of the second chain (bottom tetrad) are orange. Both structures were visualized in PyMOL.

ration and years of research. During the time of exploring motifs, several approaches, methods, and models were developed and established. They become useful in searching motifs in different areas.

For nucleic acid sequence motif search, these methods can be divided into two types: word-based (enumeration) and probabilistic methods [Das & Dai (2007); Hashim et al. (2019)]. Methods that rely on word enumeration are typically used for searching short motifs and require several user input parameters, i.a., the number and/or type of mismatches that could appear in a motif [Das & Dai (2007); Hashim et al. (2019)]. These algorithms search the dataset for predefined or non-defined expressions that could be identified as a potential motif [Das & Dai (2007); Hashim et al. (2019)]. The [Brāzma et al. (1998)] and [Karaboga & Aslan (2016)] research groups proposed word-based methods for the motif search [Das & Dai (2007); Hashim et al. (2019)]. The probabilistic approach requires fewer input parameters and relies on the position weight matrix of motifs [Das & Dai (2007); Hashim et al. (2019)]. Example methods which are using probabilistic methods for motif discovery are proposed in [Bailey et al. (2015)] and [Lawrence & Reilly (1990); Das & Dai (2007); Hashim et al. (2019)].

As the number of motif finding methods grew, scientists started to combine and modify the existing approaches, to receive more accurate solutions for

the motif discovery problem. One of the example methods that could be used in motif exploration process are evolutionary algorithms. In [Shao et al. (2009)], the authors combine metaheuristic (tabu search algorithm) with the bacterial foraging optimization algorithm. In comparison to other known methods, researchers tested their solution against known sequence motifs, obtaining satisfying results. Other algorithms for motif discovery are based on the established (l, d)-motif model. The (l, d)-motif model is one of the most popular models used for motif search (called LDMS or PMS) [Xiao et al. (2019); Mohanty et al. (2018); Davila et al. (2007)]. The foundation of the LDMS model is to find a substring of length l that appears repeatedly in the input sequences with the maximum of d errors [Davila et al. (2007)]. The LDMS problem belongs to the NP-hard problems [Xiao et al. (2019)]. The [Xiao et al. (2019)] group proposed an LDDMS modification in which, additionally to the condition established in (l, d)-motif model, the motif should be present in at least one of the input sequences and hamming distance of $d2$. The new LDDMS algorithm was successfully tested against synthetic DNA sequences and real sequence datasets [Xiao et al. (2019)].

The number of algorithms and methods for motif searching is still expanding, as the problem is not limited to sequence motif analysis. Moreover, it is also related to structural representations. Each motif search algorithm is based on a specific data format, thus we need to bear in mind that outputs (patterns), for even the same data set, can be different.

CHAPTER 2

Main results

The research described in this dissertation was conducted in collaboration with scientists from the Institute of Bioorganic Chemistry PAS and Poznan University of Medical Sciences. It concerned the structure of nucleic acids at different levels of the organization and for different living organisms (Figure 2.1). Most of the research was performed *in silico*, using external or in-house computational algorithms. In one case [A3], *in vitro* experiments were also carried. In all computational experiments, we used data available in public resources – data repositories and scientific papers. The obtained results were published in open access articles in five JCR-indexed journals; the sixth paper was submitted for publication. The following subsections briefly summarize the research conducted and the results obtained. Full texts of the articles are included in the next section.

2.1. Motifs in plant pre-miRNA

Work focusing on motifs in plant pre-miRNAs has been reported in [A1] and [A2]. In [A1], we looked at the genesis of miRNA in plant organisms, which has not yet been recognized completely. It involves the Dicer-like 1 (DCL1)

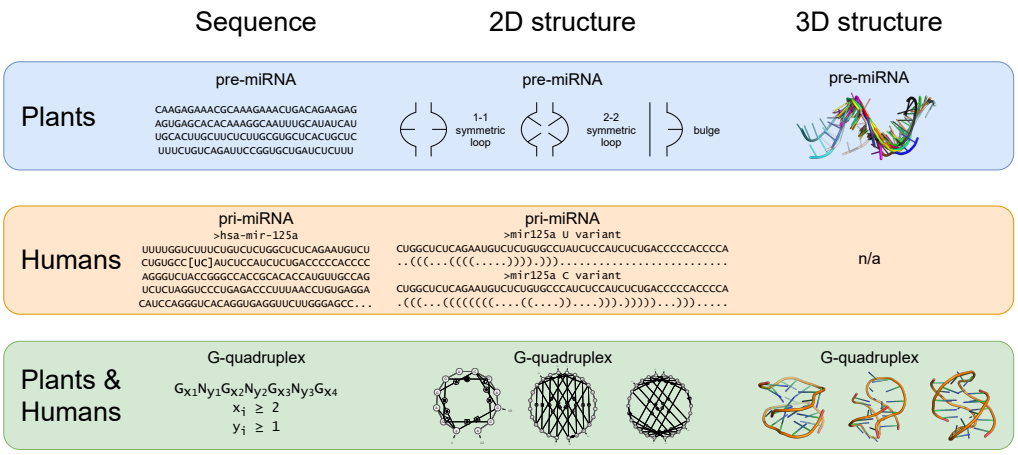


Figure 2.1: Topics addressed in the doctoral study.

enzyme cutting out the miRNA:miRNA* duplex from pre-miRNAs. We assumed that the motif guiding the enzyme to the cleavage site should be located near the duplex area. Thus, we searched for sequence and secondary structure patterns in the vicinity of miRNA:miRNA*. The size of the vicinity to explore (on both sides of miRNA:miRNA*) was arbitrarily set up to 22 nucleotides. The test dataset contained sequences of 50 micro RNA precursors from *Arabidopsis thaliana*, for which the experiments already confirmed the base-to-loop or loop-to-base cleavage mechanism. The results of the first computational experiment revealed four potential sequence motifs – UCUC, AACA, GUGG, and ACGG – and indicated pyrimidine dominance in certain positions of the analyzed regions.

For further analysis, we ran RNAstructure [Reuter & Mathews (2010)] and mfold [Zuker (2003)] to predict the secondary structures for all studied pre-miRNA sequences. Out of multiple variants generated by both algorithms, 50 most compact ones (one structure per every input sequence) were selected for motif searching. The latter process was conducted using a self-developed program looking for predefined patterns – bulges and internal loops. We expected to find diverse patterns in the lower and upper stem of the same pre-miRNA and similar patterns between the first-cut regions

– i.e., lower stem in structures with base-to-loop mechanism and upper in structures with loop-to-base cutting direction. The results showed more symmetric internal loops in the first-cut region compared to the second-cut place. We also found small symmetric internal loops (1-2 unpaired nucleotides per strand) in the closest vicinity of miRNA:miRNA* ($\leq$5nt from the duplex). We concluded that symmetric internal loops could constitute the motif recognized by DCL1 to perform the first cut releasing miRNA:miRNA* duplex.

In [A2], continuing the subject of motifs recognizable by DCL1, we extended the study to pre-miRNAs of all green plants – the *Viridiplantae* kingdom. The sequences of miRNA precursors were downloaded from miRBase [Kozomara et al. (2018)], and processed with WebLogo [Crooks (2004)] and self-developed scripts for purine-pyrimidine patterns. This time, the cutting mechanism was unknown for all miRNA:miRNA* duplexes. Based on the results obtained in [A1], we reduced the motif search area to 8nt fragments neighboring miRNA on both 5'- and 3'-side. Thus, we created two subsets of vicinity regions, named VS8-5' and VS8-3'. In both sets, we found uracil to be the most frequent nucleotide on most positions of analyzed sequences. In VS8-3', higher pyrimidine occupancy appeared in the first position of the analyzed region. In VS8-5', a similar observation was made for the second and fifth positions. We then further reduced the searched sequence fragments by taking four nucleotides adjacent to the miRNA for the next round of analysis. In the 4nt-long vicinity, we looked for 16 predefined purine-pyrimidine patterns. The statistical analysis resulted in the five most frequent motifs starting with pyrimidine: YYRY, YRRR, YRRY, YYYY, and YYRR (according to IUPAC, R stands for purine, Y - pyrimidine).

We applied ContextFold [Zakov et al. (2011)] to predict the secondary

structures for all downloaded sequences of *Viridiplantae* pre-miRNAs. The obtained data were preprocessed with RNApdbee [Antczak et al. (2014); Zok et al. (2018)] to adjust the format. A self-developed Python program called MotifSeeker searched for the secondary structure motifs – bulges and internal loops – in the closest vicinity of miRNA:miRNA* duplex. The output data confirmed the results obtained for *Arabidopsis thaliana* in [A1]. Most structures (76%) contained small symmetric internal loops, 1-1 and 2-2, close to the miRNA:miRNA* duplex.

Motif analysis in [A2] extended to the third structural level. We randomly selected 40 pre-miRNA sequences – ten per each phylum, Chlorophyta, Coniferophyta, Embryophyta, and Magnoliophyta – and predicted their three-dimensional structures using RNAComposer [Antczak et al. (2017); Purzycka et al. (2015)]. In the generated 3D models, we studied the 8nt fragments (4nt beyond miRNA and 4nt within miRNA itself). As both sides of miRNAs were considered, the analyzed set consisted of 20 substructures per phylum. The 3D fragments were superimposed using PyMOL software. Their RMSD and eRMSD values were computed with PyMOL and baRNAba [Bottaro et al. (2014)], respectively. Low RMSD and eRMSD values – in most cases, RMSD$<$1.5Å and eRMSD$<$1Å – confirmed the high similarity of miRNA vicinity within all the phyla. We concluded that the region closest to the miRNA:miRNA* duplex is highly conserved in the *Viridiplantae* kingdom.

2.2. Motifs in human pri-miRNA and miRNA

[A3] summarizes the study of pri-miR-125a with rs12976445 polymorphism in breast cancer patients. At the beginning of work, we performed *in*

vitro investigation on rs12976445 SNP frequency in pri-miR-125a among both cancerous and non-cancerous samples. The comparison of SNP genotypes revealed that CT genotype appeared less frequently in cancerous samples than in the control samples. The second part of the research focused on *in silico* analysis. 51nt-long fragment of pri-miR-125a sequence containing SNP (single nucleotide polymorphisms) was an input to RNAstructure [Reuter & Mathews (2010)] and RNAfold [Lorenz et al. (2011)] – two secondary structure prediction tools – with C- and U-variant of the amplicon. The generated 2D models were subjected to comparative analysis among variants. The C-variant formed a base-pairing, whereas the U-variant showed the SNP as an unpaired nucleotide included in a large single-stranded region. The diverse folding of the secondary structure of C- and U-variants made us anticipate the SNP-dependent protein binding and – thus – a different expression of miR-125a. Potential RNA binding proteins, specific to or shared in both variants, were investigated using the RBPmap web server [Paz et al. (2014)] and the ATtRACT database [Giudice et al. (2016)]. We were most interested in proteins that would bind exclusively to one of the variants. The analysis revealed several variant-specific binding proteins. Among them, one protein (NOVA1) appeared to bind only to the C-variant. The results suggest diverse functionality of the pri-miR-125a in molecules with different SNPs.

The second project focusing on human miRNAs examined their sequences for susceptibility to quadruplex formation. Data for analysis were acquired from the miRBase database [Kozomara et al. (2018)]. In the set of 1,917 human pre-miRNA sequences, we found 2,879 miRNAs located on the 5'- and 3'-sides. From this collection, we selected miRNAs containing at least 8 Guanines – a prerequisite for the existence of a G-quadruplex (G4). Nearly 30% of sequences met this condition. We examined them for two- or three-

tetrad G4s with uninterrupted G-tracts. The search algorithm used regular expression matching to identify motifs according to one of the following patterns: $G_2N_{1-7}G_2N_{1-7}G_2N_{1-7}G_2$ and $G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$, where $N \in \{\text{A,U,G,C}\}$. 194 miRNAs fulfilled the criteria and showed potential to form the two-Guanine tracts; 5 miRNAs could fold into the three-tetrad motif. The results were not published. A continuation of the study is planned with the other patterns to search.

2.3. Quadruplex motifs

Articles [A4], [A5], and [A6] refer to the third project carried as part of the doctoral dissertation. Its subject is the quadruplex - a particular structural motif composed of stacked nucleotide quartets (tetrads), which can form in DNA and RNA molecules. In [A4], we introduced a new classification (ONZ) of tetrads and quadruplexes. It reflected a view of these motifs from the perspective of the secondary structure, unlike the only existing classification to date based on glycosidic bond angles and loop topologies [Webba da Silva (2007); Dvorkin et al. (2018)]. We proposed to model the secondary structure of tetrad T as a cyclic graph $G = (V, E)$, where $|V| = |E| = 4$. Each $v \in V$ represented one nucleotide from the tetrad, and every $e \in E$ corresponded to a hydrogen-bonding interaction between respective nucleotides. If the vertices of G are located at equal distances on a circle clockwise, in the order imposed by the sequence, the graph took the shape of a square (O-shaped), a bow tie (N-shaped), or an hourglass (Z-shaped). This observation allowed distinguishing 3 groups of tetrads and defining their ONZ taxonomy:

- T $\in$ O if T = (N1,N2), (N2,N3), (N3,N4), (N4,N1),

- T ∈ N if T = (N1,N2), (N2,N4), (N4,N3), (N3,N1),
- T ∈ Z if T = (N1,N3), (N3,N2), (N2,N4), (N4,N1),

where N1, N2, N3, and N4 denote nucleotides that form tetrad T.

Classification of tetrads carries over to the quadruplex that comprises them. If all component tetrads (note that a quadruplex contains ≥2 tetrads) are of type O, then the quadruplex belongs to the O class. The same rule applies to N and Z classes. Thus, a quadruplex composed of homogeneous quartets belongs to O, N, or Z class. Heterogeneous G4, i.e., quadruplex built from tetrads of different types, is assigned to class M (mixed) – additional class defined for quadruplexes. The specificity of ONZ nomenclature makes it applicable to unimolecular quadruplexes only.

The ONZ classification is accompanied by dedicated textual and graphical representations proposed in [A4]. We introduced a two-line dot-bracket notation to encode tetrads and quadruplexes unambiguously. In accordance with the adjusted dot-bracket, we developed a top-down arc diagram to clearly visualize both motifs' secondary structures. Finally, we implemented an optimization algorithm to automatically create both representations of tetrads and quadruplexes based on a basic notation of the secondary structure (e.g., a list of base pairs). The algorithm is available within the functionality of the ElTetrado software [Zok et al. (2020)].

After introducing the ONZ classification, we performed statistical analyses to learn the distribution of each class in structures determined experimentally. For this purpose, in April 2019, we downloaded all PDB-deposited three-dimensional structures of nucleic acids. Using ElTetrado [Zok et al. (2020)], we selected 188 instances containing tetrads and unimolecular quadruplexes, and we assigned them to ONZ classes. Class 0

proved to be the most numerous, with 75% of tetrads and 56% of quadruplexes assigned to it. The least numerous was the Z class, to which only 2% of tetrads and 1% of quadruplexes belonged. The results also showed that structures sharing the same sequence could have diverse secondary structure topologies. Thus, they belong to different ONZ classes.

The interest in quadruplexes was a motivation to analyze all existing bioinformatics resources for their applications in the study of RNA G4s. The [A5] paper summarizes a multi-faceted analysis of resources – databases and computer programs for putative quadruplex-forming sequence (PQS) analysis; prediction, modeling, annotation, and visualization of quadruplex structures. We found 16 repositories to store the quadruplex-related data, 14 tools to predict quadruplex location within the nucleic acid sequence, one program to anticipate quadruplex within the secondary structure, and 4 tools to analyze and visualize the secondary and tertiary structure. Tools predicting the G4 in sequence and secondary structure were tested on 532 non-redundant sequences downloaded from the G4RNA database [Garant et al. (2015)] and 10,218 instances from miRBase [Kozomara et al. (2018)]. The first dataset contained experimentally confirmed positive and negative cases (i.e., sequences confirmed to form or not form quadruplexes), while the second included sequences with quadruplex folding propensity. Computational experiments followed by statistical analysis of their results revealed the superior performance of G4Catchall (motif-based algorithm) [Doluca (2019)]: $\geq$90% correct predictions for positive cases, $\geq$60% correct predictions for negative cases. Right behind was RNAfold [Lorenz et al. (2011)], the tool for secondary structure prediction enriched with the quadruplex annotation option. It correctly predicted the presence or the lack of quadruplex in over 70% of sequences. Four existing structure-based tools addressing G4s were tested on two RNA structures confirmed to fold

into a quadruplex motif. Two of them, ElTetrado [Zok et al. (2020)] and DSSR [Lu et al. (2015)], succeeded in retrieving multiple structural information of tetrads and quadruplexes. 3D-NuS [Patro et al. (2017)] required a lot of user involvement at the preprocessing stage (i.a., selecting the 3D model to which the structure folds) and generated unsatisfying results.

The analysis carried in [A4] pointed out the lack of a dedicated database collecting information about experimentally-determined quadruplex structures, along with their parameters, classification, and visualization of structural models. Therefore, we developed a new, self-updating repository and made it available under the name ONQUADRO (https://onquadro.cs.put.poznan.pl/) – publication [A6]. The database gathers data on structures – on the sequence, secondary, and tertiary structure level – of quadruplexes, tetrads, G4 helices, and PDB-deposited nucleic acids containing G4s. As of September 2021, it stored 1,661 tetrads, 518 quadruplexes, 30 G4-helices, and 467 structures folding into a quadruplex motif. ONQUADRO (i) allows searching the data; (ii) visualizes secondary and tertiary structure models (classic diagram, arc diagram, layer diagram, ball-and-stick model, surface model); (iii) provides detailed reports on structure properties (i.a., rise, twist, planarity, chi angles, ions, Webba da Silva and ONZ classifications, loop characteristics, strand directionality); (iv) enables quantitative data analysis through statistics available in tabular and graphical form. The newsletter facility automatically informs subscribers about new G4s delivered to the database.

Bibliography

Alba, J. J., Sadurní, A., & Gargallo, R. (2016). Nucleic acid i-motif structures in analytical chemistry. *Critical Reviews in Analytical Chemistry*, 46(5), 443–454.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450–461.

Angulo, M. T., Liu, Y.-Y., & Slotine, J.-J. (2015). Network motifs emerge from interconnections that favour stability. *Nature Physics*, 11(10), 848–852.

Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R. W., & Szachniuk, M. (2017). New functionality of RNAComposer: application to shape the axis of miR160 precursor structure. *Acta Biochimica Polonica*, 63(4), 737—744.

Antczak, M., Zok, T., Popenda, M., Lukasiak, P., Adamiak, R., Blazewicz, J., & Szachniuk, M. (2014). RNApdbee - a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Research*, 42(W1), W368–W372.

Ashraf, F. B. & Shafi, M. S. R. (2020). MFEA: An evolutionary approach for motif finding in DNA sequences. *Informatics in Medicine Unlocked*, 21, 100466.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, 43(W1), W39–W49.

Batey, R. T., Rambo, R. P., & Doudna, J. A. (1999). Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition*, 38(16), 2326–2343.

Battiston, F., Nicosia, V., Chavez, M., & Latora, V. (2017). Multilayer motif analysis of brain networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4), 047404.

Belmonte-Reche, E. & Morales, J. C. (2019). G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. *NAR Genomics and Bioinformatics*, 2(1), lqz005.

Bottaro, S., Palma, F. D., & Bussi, G. (2014). The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Research*, 42(21), 13306–13314.

Brāzma, A., Jonassen, I., Vilo, J., & Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Research*, 8(11), 1202–1215.

Breaker, R. R. & Joyce, G. F. (2014). The Expanding View of RNA and DNA Function. *Chemistry & Biology*, 21(9), 1059–1065.

Brosius, J. & Raabe, C. A. (2016). What is an RNA? A top layer for RNA classification. *RNA Biology*, 13(2), 140–144.

Brunel, C., Marquet, R., Romby, P., & Ehresmann, C. (2002). RNA loop–loop interactions as dynamic functional motifs. *Biochimie*, 84(9), 925–944.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, J. D., Young, J. Y., Zardecki, C., & Zhuravleva, M. (2020). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1), D437–D451.

Butcher, S. E. & Pyle, A. M. (2011). The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Accounts of Chemical Research*, 44(12), 1302–1311.

Cammas, A. & Millevoi, S. (2016). RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Research*, (pp. gkw1280).

Chastain, M. & Tinoco, I. (1991). Structural elements in RNA. In *Progress in Nucleic Acid Research and Molecular Biology* (pp. 131–177). Elsevier.

Chheda, N. & Gupta, M. K. (2014). RNA as a permutation. *arXiv: Biomolecules*.

Churkin, A. & Barash, D. (2013). RNA dot plots: an image representation for RNA secondary structure analysis and manipulations. *Wiley Interdisciplinary Reviews: RNA*, 4(2), 205–216.

Claverie, J.-M. (2000). From bioinformatics to computational biology. *Genome Research*, 10(9), 1277–1279.

Conn, G. L. & Draper, D. E. (1998). RNA structure. *Current Opinion in Structural Biology*, 8(3), 278–285.

Cooper, T. A., Wan, L., & Dreyfuss, G. (2009). RNA and Disease. *Cell*, 136(4), 777–793.

Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell*, 78(1), 9–29.

Crooks, G. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6), 1188–1190.

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

Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15), 1974–1975.

Das, M. K. & Dai, H.-K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(S7), S21.

Davila, J., Balla, S., & Rajasekaran, S. (2007). Fast and practical algorithms for planted (l, d) motif search. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4), 544–552.

Doluca, O. (2019). G4Catchall: A G-quadruplex prediction approach considering atypical features. *J Theor Biol*, 463(1), 92–98.

Doudna, J. A. (2000). Structural genomics of RNA. *Nature Structural Biology*, 7, 954–956.

Draper, D. (2004). A guide to ions and RNA structure. *RNA*, 10(3), 335–343.

Dvorkin, S. A., Karsisiotis, A. I., & da Silva, M. W. (2018). Encoding canonical DNA quadruplex structure. *Science Advances*, 4(8), eaat3007.

Eric, W. & Pascal, A. (2006). *RNA Tertiary Structure*, chapter Nucleic Acids Structure and Mapping. John Wiley & Sons, Ltd.

Fenstermacher, D. (2005). Introduction to bioinformatics. *Journal of the American Society for Information Science and Technology*, 56(5), 440–446.

Fiore, J. L. & Nesbitt, D. J. (2013). An RNA folding motif: GNRA tetraloop–receptor interactions. *Quarterly Reviews of Biophysics*, 46(3), 223–264.

Formanowicz, D., Gutowska, K., & Formanowicz, P. (2018). Theoretical studies on the engagement of Interleukin 18 in the immunoinflammatory processes underlying Atherosclerosis. *International Journal of Molecular Sciences*, 19(11), 3476.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K. V., Ferkinghoff-Borg, J., & Hamelryck, T. (2009). A probabilistic model of RNA conformational space. *PLoS Computational Biology*, 5(6), e1000406.

Fu, X., Liao, B., Zhu, W., & Cai, L. (2018). New 3D graphical representation for RNA structure analysis and its application in the pre-miRNA identification of plants. *RSC Advances*, 8(54), 30833–30841.

Gan, H. H. (2003). Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*, 31(11), 2926–2943.

Garant, J., Luce, M., Scott, M., & Perreault, J. (2015). G4RNA: an RNA G-quadruplex database. *Database*, 2015(1), bav059.

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.

Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT—a database of RNA-binding proteins and associated motifs. *Database*, 2016, baw035.

Gong, W. & Fan, X.-Q. (2019). A geometric characterization of DNA sequence. *Physica A: Statistical Mechanics and its Applications*, 527, 121429.

Goodsell, D. S. (2005). Visual methods from atoms to cells. *Structure*, 13(3), 347–354.

Grau, J., Posch, S., Grosse, I., & Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21), e197–e197.

Griffin, B. D. & Bass, H. W. (2018). Review: plant G-quadruplex (G4) motifs in DNA and RNA; abundant, intriguing sequences of unknown function. *Plant Science*, 269, 143–147.

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics*, 1(3), 231–236.

Halder, S. & Bhattacharyya, D. (2013). RNA structure and dynamics: A base pairing perspective. *Progress in Biophysics and Molecular Biology*, 113(2), 264–283.

Hallinan, J. & Jackway, P. (2005). Network motifs, feedback loops and the dynamics of genetic regulatory networks. In *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*: IEEE.

Hames, D. & Hooper, N. (2010). *Krótkie wykłady. Biochemia.* Warszawa: Wydawnictwo Naukowe PWN.

Hashim, F. A., Mabrouk, M. S., & Al-Atabany, W. (2019). Review of different sequence motif finding algorithms. *Avicenna Journal of Medical Biotechnology*, 11(2), 130–148.

He, Y., Shen, Z., Zhang, Q., Wang, S., & Huang, D.-S. (2020). A survey on deep learning in DNA/RNA motif mining. *Briefings in Bioinformatics*, (pp. bbaa229).

Hendrix, D. K., Brenner, S. E., & Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. *Quarterly Reviews of Biophysics*, 38(3), 221–243.

Higgs, P. G. & Lehman, N. (2014). The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16(1), 7–17.

Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, R., Mungall, C. J., Richardson, J. S., Stombaugh, J., Westhof, E., & et al. (2011). The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*, 6(1), 53–89.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly*, 125(2), 167–188.

House, T., Davies, G., Danon, L., & Keeling, M. J. (2009). A motif-based approach to network epidemics. *Bulletin of Mathematical Biology*, 71(7), 1693–1706.

Hu, J., Wang, J., Lin, J., Liu, T., Zhong, Y., Liu, J., Zheng, Y., Gao, Y., He, J., & Shang, X. (2019). MD-SVM: a novel SVM-based algorithm for the motif discovery of transcription factor binding sites. *BMC Bioinformatics*, 20(S7), 200.

Jazayeri, A. & Yang, C. C. (2020). Motif discovery algorithms in static and temporal networks: A survey. *Journal of Complex Networks*, 8(4), cnaa031.

Johnson, A. D. (2010). An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics*, 26(10), 1386–1389.

Jossinet, F., Ludwig, T. E., & Westhof, E. (2007). RNA structure: bioinformatic analysis. *Current Opinion in Microbiology*, 10(3), 279–285.

Karaboga, D. & Aslan, S. (2016). A discrete artificial bee colony algorithm for detecting transcription factor binding sites in DNA sequences. *Genetics and Molecular Research*, 15(2), gmr8645.

Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11), 1746–1758.

Kim, Y.-K. (2020). RNA Therapy: Current Status and Future Potential. *Chonnam Medical Journal*, 56(2), 87.

Kozlíková, B., Krone, M., Falk, M., Lindow, N., Baaden, M., Baum, D., Viola, I., Parulek, J., & Hege, H.-C. (2016). Visualization of biomolecular structures: State of the art revisited. *Computer Graphics Forum*, 36(8), 178–204.

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2018). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1), D155–D162.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, 87(12), 1011–1020.

Kun, Á., Szilágyi, A., Könnyű, B., Boza, G., Zachar, I., & Szathmáry, E. (2015). The dynamics of the RNA world: insights and challenges. *Annals of the New York Academy of Sciences*, 1341(1), 75–95.

Kuttel, M., Gain, J., Burger, A., & Eborn, I. (2006). Techniques for visualization of carbohydrate molecules. *Journal of Molecular Graphics and Modelling*, 25(3), 380–388.

Lacroix, V., Fernandes, C., & france Sagot, M. (2006). Motif search in graphs: application to metabolic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4), 360–368.

Laing, C. & Schlick, T. (2011). Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology*, 21(3), 306–318.

Lawrence, C. E. & Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Genetics*, 7(1), 41–51.

Leontis, N. & Westhof, E. (2012). Modeling RNA Molecules. In *Nucleic Acids and Molecular Biology* (pp. 5–17). Springer Berlin Heidelberg.

Leontis, N. B. & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4), 499–512.

Li, B., Cao, Y., Westhof, E., & Miao, Z. (2020). Advances in RNA 3D

Structure Modeling Using Experimental Data. *Frontiers in Genetics*, 11, 1147.

Li, G., Chan, T.-M., Leung, K.-S., & Lee, K.-H. (2010). A cluster refinement algorithm for motif discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4), 654–668.

Li, N. & Tompa, M. (2006). Analysis of computational approaches for motif discovery. *Algorithms for Molecular Biology*, 1(1), 8.

Li, X., Stones, R. J., Wang, H., Deng, H., Liu, X., & Wang, G. (2012). NetMODE: network motif detection without nauty. *PLoS ONE*, 7(12), e50093.

Liao, B., Luo, J., Li, R., & Zhu, W. (2006). RNA secondary structure 2D graphical representation without degeneracy. *International Journal of Quantum Chemistry*, 106(8), 1749–1755.

Lin, C., Dickerhoff, J., & Yang, D. (2019). NMR studies of G-quadruplex structures and G-quadruplex-interactive compounds. In *Methods in Molecular Biology* (pp. 157–176). Springer New York.

Liu, F. & Heiner, M. (2013). Petri nets for modeling and analyzing biochemical reaction networks. In *Approaches in Integrative Bioinformatics* (pp. 245–272). Springer Berlin Heidelberg.

Lorenz, R., Bernhart, S. H., zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1).

Lu, X.-J., Bussemaker, H. J., & Olson, W. K. (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, (pp. e142).

Maizels, N. (2012). G4 motifs in human genes. *Annals of the New York Academy of Sciences*, 1267(1), 53–60.

Mattei, E., Ausiello, G., Ferrè, F., & Helmer-Citterich, M. (2014). A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Research*, 42(10), 6146–6157.

Miao, Z., Adamiak, R. W., Antczak, M., Boniecki, M. J., Bujnicki, J., Chen, S.-J., Cheng, C. Y., Cheng, Y., Chou, F.-C., Das, R., Dokholyan, N. V., Ding, F., Geniesse, C., Jiang, Y., Joshi, A., Krokhotin, A., Magnus, M., Mailhot, O., Major, F., Mann, T. H., Piątkowski, P., Pluta, R., Popenda, M., Sarzynska, J., Sun, L., Szachniuk, M., Tian, S., Wang, J., Wang, J., Watkins, A. M., Wiedemann, J., Xiao, Y., Xu, X., Yesselman, J. D., Zhang, D., Zhang, Y., Zhang, Z., Zhao, C., Zhao, P., Zhou, Y., Zok, T., Żyła, A., Ren, A., Batey, R. T., Golden, B. L., Huang, L., Lilley, D. M., Liu, Y., Patel, D. J., & Westhof, E. (2020). RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8), 982–995.

Miao, Z. & Westhof, E. (2017). RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 46(1), 483–503.

Mills, L. (2014). Common file formats. *Current Protocols in Bioinformatics*, 45(1), A.1B.1–A.1B.18.

Mohanty, S., Mohanty, S., & Roy, S. (2018). Exact planted (l, d) motif search algorithms: a review. In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*: IEEE.

Mortimer, S. A., Kidwell, M. A., & Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7), 469–479.

Mueen, A. (2014). Time series motif discovery: dimensions and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(2), 152–159.

Ngan, H. Y., Pang, G. K., & Yung, N. H. (2008). Motif-based defect detection for patterned fabric. *Pattern Recognition*, 41(6), 1878–1894.

Nicodème, P., Salvy, B., & Flajolet, P. (2002). Motif statistics. *Theoretical Computer Science*, 287(2), 593–617.

Ouzounis, C. A. & Valencia, A. (2003). Early bioinformatics: the birth of a discipline–a personal view. *Bioinformatics*, 19(17), 2176–2190.

Patro, L. P. P., Kumar, A., Kolimi, N., & Rathinavelan, T. (2017). 3D-NuS: A Web Server for Automated Modeling and Visualization of Non-Canonical 3-Dimensional Nucleic Acid Structures. *Journal of Molecular Biology*, 429(16), 2438–2448.

Paz, I., Kosti, I., Ares, M., Cline, M., & Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*, 42(W1), W361–W367.

Pearce, B. K. D., Pudritz, R. E., Semenov, D. A., & Henning, T. K. (2017). Origin of the RNA world: The fate of nucleobases in warm little ponds. *Proceedings of the National Academy of Sciences*, 114(43), 11327–11332.

Pearson, W. R. (2016). Finding protein and nucleotide similarities with FASTA. *Current Protocols in Bioinformatics*, 53(1), 3.9.1–3.9.25.

Picardi, E., Ed. (2015). *RNA Bioinformatics*. Springer New York.

Ponty, Y. & Leclerc, F. (2014). Drawing and Editing the Secondary Structure(s) of RNA. In *Methods in Molecular Biology* (pp. 63–100). Springer New York.

Popenda, M., Miskiewicz, J., Sarzynska, J., Zok, T., & Szachniuk, M. (2019). Topology-based classification of tetrads and quadruplex structures. *Bioinformatics*, 36(4), 1129–1134.

Purzycka, K., Popenda, M., Szachniuk, M., Antczak, M., Lukasiak, P., Blazewicz, J., & Adamiak, R. (2015). Automated 3D RNA Structure Prediction Using the RNAComposer Method for Riboswitches. In *Methods in Enzymology* (pp. 3–34). Elsevier.

Rajasekaran, S. & Dinh, H. (2011). A speedup technique for (l, d)-motif finding algorithms. *BMC Research Notes*, 4(1), 54.

Ramlan, E. I. & Zauner, K.-P. (2008). An extended dot-bracket notation for functional nucleic acids. In *International Workshop on Computing With Biomolecules* (pp. 75–86).

Rampášek, L., Jimenez, R. M., Lupták, A., Vinař, T., & Brejová, B. (2016). RNA motif search with data-driven element ordering. *BMC Bioinformatics*, 17(1), 216.

Redhead, E. & Bailey, T. L. (2007). Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, 8(1), 385.

Rehman, A., Delori, A., Hughes, D. S., & Jones, W. (2018). Structural studies of crystalline forms of triamterene with carboxylic acid, GRAS and API molecules. *IUCrJ*, 5(3), 309–324.

Reuter, J. & Mathews, D. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1), 129.

Robertson, M. P. & Joyce, G. F. (2010). The Origins of the RNA World. *Cold Spring Harbor Perspectives in Biology*, 4(5), a003608–a003608.

Rother, K., Rother, M., Boniecki, M., Puton, T., & Bujnicki, J. M. (2011). RNA and protein 3D structure modeling: similarities and differences. *Journal of Molecular Modeling*, 17(9), 2325–2336.

Roy, K., Kar, S., & Das, R. N. (2015). Computational chemistry. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (pp. 151–189). Elsevier.

Ryu, M. W., Oh, S. M., Kim, M. J., Cho, H. H., Son, C. B., & Kim, T. H. (2020). Algorithm for generating 3D geometric representation based on indoor point cloud data. *Applied Sciences*, 10(22), 8073.

Sancho-Andrés, G., Soriano-Ortega, E., Gao, C., Bernabé-Orts, J. M., Narasimhan, M., Müller, A. O., Tejos, R., Jiang, L., Friml, J., Aniento, F., & Marcote, M. J. (2016). Sorting motifs involved in the trafficking and localization of the PIN1 Auxin Efflux carrier. *Plant Physiology*, 171(3), 1965–1982.

Sandve, G. & Drabløs, F. (2006). A survey of motif discovery methods in an integrated framework. *Biology Direct*, 1(1), 11.

Schlick, T. (2018). Adventures with RNA graphs. *Methods*, 143, 16–33.

Searls, D. B. (2010). The Roots of Bioinformatics. *PLoS Computational Biology*, 6(6), e1000809.

Shao, L., Chen, Y., & Abraham, A. (2009). Motif discovery using evolutionary algorithms. In *2009 International Conference of Soft Computing and Pattern Recognition*: IEEE.

Šponer, J. E., Špačková, N., Leszczynski, J., & Šponer, J. (2005). Principles of RNA Base Pairing: Structures and Energies of the Trans Watson-Crick/Sugar Edge Base Pairs. *The Journal of Physical Chemistry B*, 109(22), 11399–11410.

St-Onge, K., Thibault, P., Hamel, S., & Major, F. (2007). Modeling RNA tertiary structure motifs by graph-grammars. *Nucleic Acids Research*, 35(5), 1726–1736.

Synak, J., Rybarczyk, A., & Blazewicz, J. (2020). Multi-agent approach to sequence structure simulation in the RNA World hypothesis. *PLOS ONE*, 15(8), e0238253.

Szachniuk, M. (2019). RNApolis: Computational Platform for RNA Structure Analysis. *Foundations of Computing and Decision Sciences*, 44(2), 241–257.

Szostak, N., Royo, F., Rybarczyk, A., Szachniuk, M., Blazewicz, J., del Sol, A., & Falcon-Perez, J. M. (2014). Sorting signal targeting mRNA into hepatic extracellular vesicles. *RNA Biology*, 11(7), 836–844.

Szostak, N., Synak, J., Borowski, M., Wasik, S., & Blazewicz, J. (2017). Simulating the origins of life: The dual role of RNA replicases as an obstacle to evolution. *PLOS ONE*, 12(7), e0180827.

Takahashi, H., Nakagawa, A., Kojima, S., Takahashi, A., Cha, B.-Y., Woo, J.-T., Nagai, K., Machida, Y., & Machida, C. (2012). Discovery of novel rules for G-quadruplex-forming sequences in plants by using bioinformatics methods. *Journal of Bioscience and Bioengineering*, 114(5), 570–575.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1), 137–144.

Torkamani, S. & Lohweg, V. (2017). Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), e1199.

Tyson, J. J. & Novák, B. (2010). Functional motifs in biochemical reaction networks. *Annual Review of Physical Chemistry*, 61(1), 219–240.

Wasik, S., Szostak, N., Kudla, M., Wachowiak, M., Krawiec, K., & Blazewicz, J. (2019). Detecting life signatures with RNA sequence similarity measures. *Journal of Theoretical Biology*, 463, 110–120.

Watson, J. D. & Crick, F. H. C. (1974). Molecular structure of nucleic acids: a structure for Deoxyribose Nucleic Acid. *Nature*, 248(5451), 765–765.

Wattenberg, M. (2002). Arc Diagrams: Visualizing Structure in Strings. In *Proceedings of the IEEE Symposium on Information Visualization*.

Waugh, A., Gendron, P., Altman, R., Brown, J. W., Case, D., Gautheret, D., Harvey, S. C., Leontis, N., Westbrook, J., Westhof, E., Zuker, M., & Major, F. (2002). RNAML: A standard syntax for exchanging RNA information. *RNA*, 8(6), 707–717.

Webba da Silva, M. (2007). Geometric Formalism for DNA Quadruplex Folding. *Chemistry - A European Journal*, 13(35), 9738–9745.

Westbrook, J. D. & Fitzgerald, P. M. D. (2005). The PDB format, mmCIF formats, and other data formats. In *Structural Bioinformatics* (pp. 159–179). John Wiley & Sons, Inc.

Westhof, E. & Auffinger, P. (2012). *Transfer RNA Structure*. John Wiley & Sons.

Wiedemann, J. & Miłostan, M. (2017). StructAnalyzer - a tool for sequence vs. structure similarity analysis. *Acta Biochimica Polonica*, 63(4), 753—-757.

Wu, S., Yu, C., Zhang, W., Yin, S., Chen, Y., Feng, Y., & Ma, W. (2017). Tag mechanism as a strategy for the RNA replicase to resist parasites in the RNA world. *PLOS ONE*, 12(3), e0172702.

Wu, Y.-Y. & Kuo, H.-C. (2020). Functional roles and networks of non-coding RNAs in the pathogenesis of neurodegenerative diseases. *Journal of Biomedical Science*, 27(1), 49.

Xiao, P., Schiller, M., & Rajasekaran, S. (2019). Novel algorithms for LDD motif search. *BMC Genomics*, 20(S5), 424.

Yao, Y.-H., Nan, X.-Y., & Wang, T.-M. (2005). A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them. *Journal of Computational Chemistry*, 26(13), 1339–1346.

Yu, S., Feng, Y., Zhang, D., Bedru, H. D., Xu, B., & Xia, F. (2020). Motif discovery in networks: A survey. *Computer Science Review*, 37, 100267.

Zakov, S., Goldberg, Y., Elhadad, M., & Ziv-Ukelson, M. (2011). Rich Parameterization Improves RNA Structure Prediction. *Journal of Computational Biology*, 18(11), 1525–1542.

Zambelli, F., Pesole, G., & Pavesi, G. (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2), 225–237.

Zell, R., Sidigi, K., Bucci, E., Stelzner, A., & Görlach, M. (2002). Determi-

nants of the recognition of enteroviral cloverleaf RNA by coxsackievirus B3 proteinase 3C. *RNA*, 8(2), 188–201.

Zemora, G. & Waldsich, C. (2010). RNA folding in living cells. *RNA Biology*, 7(6), 634–641.

Zhang, H., Zhu, L., & Huang, D.-S. (2017). WSMD: weakly-supervised motif discovery in transcription factor ChIP-seq data. *Scientific Reports*, 7(1), 3217.

Zhang, Y., Huang, H., Dong, X., Fang, Y., Wang, K., Zhu, L., Wang, K., Huang, T., & Yang, J. (2016). A Dynamic 3D Graphical Representation for RNA Structure Analysis and Its Application in Non-Coding RNA Classification. *PLOS ONE*, 11(5), e0152238.

Zok, T., Antczak, M., Zurkowski, M., Popenda, M., Blazewicz, J., Adamiak, R., & Szachniuk, M. (2018). RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Research*, 46(W1), W30–W35.

Zok, T., Popenda, M., & Szachniuk, M. (2013). MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22(3), 457–473.

Zok, T., Popenda, M., & Szachniuk, M. (2020). ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *BMC Bioinformatics*, 21(1), 40.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406—-3415.

Publication reprints

Hindawi
BioMed Research International
Volume 2017, Article ID 6783010, 8 pages
https://doi.org/10.1155/2017/6783010

Research Article

Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors

J. Miskiewicz,[1] K. Tomczyk,[1] A. Mickiewicz,[2] J. Sarzynska,[2] and M. Szachniuk[1,2]

[1]*Institute of Computing Science and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Poznan, Poland*
[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

Correspondence should be addressed to M. Szachniuk; marta.szachniuk@cs.put.poznan.pl

Received 10 August 2016; Revised 18 December 2016; Accepted 12 January 2017; Published 9 February 2017

Academic Editor: Yudong Cai

According to the RNA world theory, RNAs which stored genetic information and catalyzed chemical reactions had their contribution in the formation of current living organisms. In recent years, researchers studied this molecule diversity, i.a. focusing on small non-coding regulatory RNAs. Among them, of particular interest is evolutionarily ancient, 19–24 nt molecule of microRNA (miRNA). It has been already recognized as a regulator of gene expression in eukaryotes. In plants, miRNA plays a key role in the response to stress conditions and it participates in the process of growth and development. MicroRNAs originate from primary transcripts (pri-miRNA) encoded in the nuclear genome. They are processed from single-stranded stem-loop RNA precursors containing hairpin structures. While the mechanism of mature miRNA production in animals is better understood, its biogenesis in plants remains less clear. Herein, we present the results of bioinformatics analysis aimed at discovering how plant microRNAs are recognized within their precursors (pre-miRNAs). The study has been focused on sequential and structural motif identification in the neighbourhood of microRNA.

1. Introduction

For many years, computational methods have been applied to support wet-lab experiment in biologically oriented study. *In silico* methods can help in discriminating ineffective experimental approaches or indicate the most promising ones. On the other hand, *in vivo* and *in vitro* experiments enable validation of hypotheses proposed in computational phase and may confirm or contradict them.

Solving biological problems with the use of computational methods appeared successful in various domains [1–11]. In our work, we consider an issue of microRNA recognition and we apply bioinformatics methods aiming to explain how this molecule is formed within living organisms of the plant kingdom.

MicroRNAs constitute a group of small, non-coding single-stranded RNAs of ~22 nt in length, involved in post-transcriptional regulation of gene expression [6, 12–15]. These molecules are widespread in genomes of animals and plants [6, 16]. Their biogenesis is a complex process which differs between organisms. The biogenesis of miRNA starts in the nucleus, with transcription process of primary miRNA (pri-miRNA) performed by RNA polymerase II [17, 18]. The transcript forms long hairpin loop structure which consists of single- and double-stranded regions. In double-stranded regions, single mismatches (i.e., noncomplementary nucleotides), internal loops, and bulges can be found in numerous locations [19, 20]. In animals, pri-miRNA is further processed by Drosha enzyme (ribonuclease III endonuclease). In collaboration with the RNA-binding protein, Drosha cleaves primary transcript of miRNA to the ~70 nt long precursor (pre-miRNA) [17, 21, 22]. For animals, this is the last step of the nuclear stage. Created stem-loop structure is transferred to the cytoplasm where another enzyme, called Dicer, cleaves miRNA:miRNA* duplex out of pre-miRNA [22–24]. Ribonuclease Dicer acts as a molecular ruler: it counts the distance between 3′ or 5′ end and the cleavage site; next, it performs the cut which releases the miRNA:miRNA* duplex [25–27]. In plants, the maturation process of miRNA is slightly different. After creation of

plant pri-miRNA, a complex consisting of Dicer Like 1 enzyme (DCL1), the double-stranded RNA-binding protein HYL1 (hyponastic leaves 1), and the zinc-finger protein SE (serrate), with an assistance of the nuclear cap-binding complex, cleaves pri-miRNA to pre-miRNA [22, 28]. Before being transported to the cytoplasm, DCL1 performs at least two cleavages to release miRNA:miRNA* duplex from the precursor structure. There are two miRNA cleavage mechanisms in plants. In base-to-loop cleavage mechanism, the first cut is done in the lower stem region and the second cut in the upper (loop) region. In loop-to-base mechanism, DCL1 cuts in the opposite direction. After being cleaved, the duplex is transferred outside the nucleus. This double-stranded miRNA:miRNA* duplex is the result of a primary transcript maturation process. miRNA constitutes one strand and miRNA* is located on the complementary one. These two strands are next separated by Argonaute (AGO) protein, the main part of RNA-induced silencing complex (RISC) [17, 23, 28]. In most cases, miRNA* is degraded after its separation from miRNA [17, 28]. Thus, activated molecular version of miRNA molecule is a single-stranded RNA. After creation of single-stranded mature miRNA and embedding it to the complex (given a mi-RISC complex in result), miRNA guides this complex to mRNA with the complementary sequence. mi-RISC enables degradation of the target mRNA or inhibition of the translation process [29–32].

Compared to the biogenesis of miRNA in animals, which is better understood, maturation of plant miRNA still has some unresolved issues. One of them is recognition of the miRNA:miRNA* duplex in miRNA precursor by a microprocessor complex, containing DCL1, HYL1, and SE proteins. This microprocessor complex performs at least two cuts in order to release the miRNA:miRNA* duplex (miRNA on one strand and complementary sequence of miRNA* on the other strand) from pre-miRNA molecule. Yet, it has not been discovered how this duplex is recognized within the precursor structure. It is supposed that some structural motifs, appearing in the vicinity of microRNA [20, 33–35], guide the DCL1 enzyme where to perform the cutting. The importance of miRNA neighbouring regions where the DCL1 enzyme starts cleaving has been already experimentally confirmed on the secondary structure level [20, 33–38]. Thus, we assume that irregulations in primary and secondary structures may be the signal for DCL1 where to starts cleaving.

Recent research concerning miRNAs has suggested the role of proteins in microprocessor complex (HYL1 and SE) in miRNA recognition. The proper selection of cutting sites is poorly understood in both pri-miRNA and pre-miRNA, but most probably it depends on HYL1 [39–41]. The importance of mismatches occurring in double-stranded regions of miRNA:miRNA* duplex was also revealed. Mismatches can influence the length of the mature microRNA, producing either longer [42–44] or shorter molecules [19]. It has been proven that miRNA genes can contain introns which are strictly correlated with biogenesis and proper functioning of their host miRNAs [12]. However, in spite of all this information about plant miRNAs, we still do not know how microprocessor complex enzyme recognizes the miRNA:miRNA* duplex within its precursor.

The presented analysis, aimed at helping in answering this issue, was performed according to the following steps: (i) creating a set of plant miRNA sequences with experimentally confirmed cleavage mechanisms, (ii) downloading precursor sequences of selected miRNAs and supplementing them to the desired length, (iii) analysing sequences using WebLOGO [45] and MEME Suite [46], (iv) predicting secondary structures of miRNA precursors by RNAstructure [4] and mfold [9], (v) choosing the best secondary structures for further analysis, and (vi) analysing predicted secondary structures in the search for structural patterns. In the paper, we present consecutive analytical steps, starting from Section 2. In Section 3, we present the results of our work. Finally, the discussion of results and future plans are presented.

2. Materials and Methods

In our research, we have decided to analyse microRNA precursors in plants on two structural levels, the sequence and the secondary structure, using selected bioinformatic tools. The sequences of pre-miRNAs were derived from publicly available data sources. First, 50 plant miRNAs with experimentally confirmed cleavage mechanism (base-to-loop or loop-to-base) were selected based on [37]. This preliminary dataset $S1$ consisted of

(i) 38 miRNAs with base-to-loop mechanism: mir164b, mir165a, mir165b, mir167a, mir167b, mir167d, mir168a, mir168b, mir169a, mir170, mir171a, mir172a, mir172b, mir172d, mir172e, mir390a, mir390b, mir391, mir393a, mir393b, mir395a, mir395b, mir395c, mir396b, mir397a, mir398b, mir398c, mir399b, mir399c, mir408, mir827, ymir158a, ymir403, ymir771, ymir824, ymir864, ymir161, ymir400, ymir825, and mir164c,

(ii) 12 miRNAs with loop-to-base mechanism: ymir400, ymir825, mir156a, mir156b, mir156c, mir156d, mir156h, mir160a, mir160b, mir160c, mir171b, and mir171c.

Next, precursors including sequences from the preliminary set $S1$ were searched in miRBase [47]. Sequences of these pre-miRNAs were downloaded for further analysis and collected in set $S2$. Taking into account previous results [20, 33–35], we have assumed that a region recognized by the microprocessor complex is located in the closest vicinity of the miRNA:miRNA* duplex. Thus, we planned to analyse ca 22 nt-long fragments neighbouring miRNA from both the base and the loop side, and we needed all pre-miRNA sequences in $S2$ to have at least 22 nucleotides in every strand of the region between miRNA:miRNA* and $3'/5'$ ends. Some sequences that did not satisfy this condition were supplemented based on [37] and *Arabidopsis thaliana* genome stored completely in the TAIR database [48]. After collecting in $S2$ sequences of the required length, miRNA vicinity was annotated in each instance. In lower stem, we distinguished *regA* region in $5'$ strand and *regD* in the opposite strand. Upper stem included *regB* region in $5'$ strand and *regC* in $3'$ strand (Figure 1).

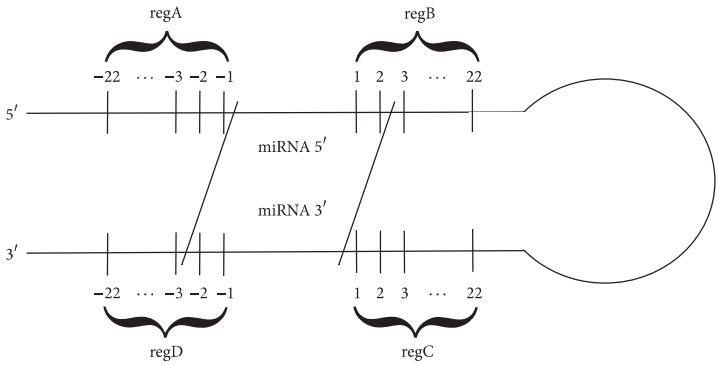


Figure 1: Schematic view of pre-miRNA with annotated miRNA:miRNA* vicinity regions. *regA* and *regD* are located in the lower stem and *regB* and *regC* in the upper stem.

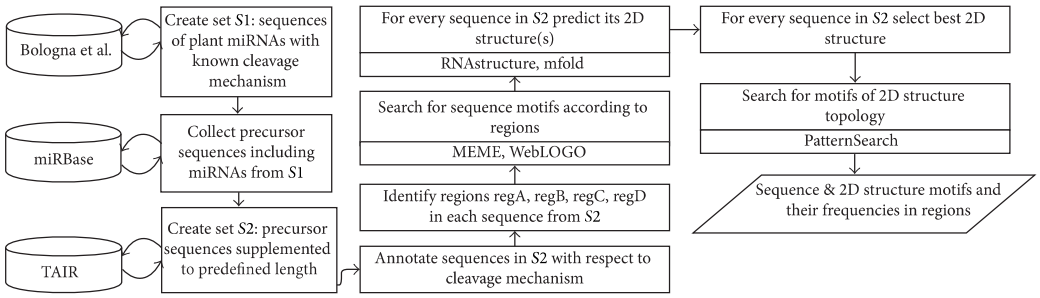


Figure 2: Consecutive steps of bioinformatics analysis.

The first analytical step was done using WebLOGO [45] and MEME Suite [46]. WebLOGO tool finds the relative frequency of particular nucleotide type at each position of the sequence within multiple sequence alignment. MEME Suite is a motif-based sequence analysis tool. MEME was run with default parameter values, except for minimum width set to 3 and maximum width set to 4. Thus, MEME was tuned to look for 3-4-nucleotide-long sequence motifs [49, 50]. Next, the secondary structures of pre-miRNAs from *S2* were predicted using RNAstructure [4] and mfold [9]. For both programs default parameter values were applied, except for the temperature in RNAstructure that was set to 295.15 K (22°C). Finally, the secondary structures were processed using own script that searched for predefined structural patterns, like bulges and symmetric and asymmetric loops. Structural pattern (motif) of our interest was defined as double-stranded structure fragment closed by canonical base-pairs on both ends, having up to 10 nucleotides in each strand and including at least one unpaired nucleotide (in any strand).

Figure 2 presents the main steps of our bioinformatic analysis.

3. Results

In the preparation step preceding an analysis, we have collected basic information about primary structures of plant miRNAs. Next, set *S2* was created and subjected to processing by bioinformatics tools. The set contained precursor sequences derived from *Arabidopsis thaliana* for which cleavage sites and cleavage mechanisms were identified and confirmed experimentally [37]. As a reference, we prepared a set of 50 random sequences, each of 22-nucleotide length, which correspond to miRNA:miRNA* vicinity regions.

The analysis started from running MEME Suite aimed at searching for sequence patterns. Sequences were encoded using 4-letter alphabet A,C,G,U representing four nucleotides building up RNA molecule: Adenine, Cytosine, Guanine, and Uracil, respectively. It has been decided to look for 3-4-nucleotide-long sequence motifs located between the 1st and the 22nd nucleotide beyond the miRNA:miRNA* duplex. Motifs were searched in the first-cut regions, that is, *regA* and *regD* in sequences with base-to-loop mechanism and *regC* and *regB* in loop-to-base sequences. We

Table 1: MEME results for *regAC*, *regBD*, and random sequences.

Region	Sequence motif	Occurrence
regAC	UCUC	11 (22%)
regBD	AACA	9 (18%)
	GUGG	6 (12%)
	ACGG	5 (10%)
Random	GUGU, GUUC, GUUU...	2 (4%)

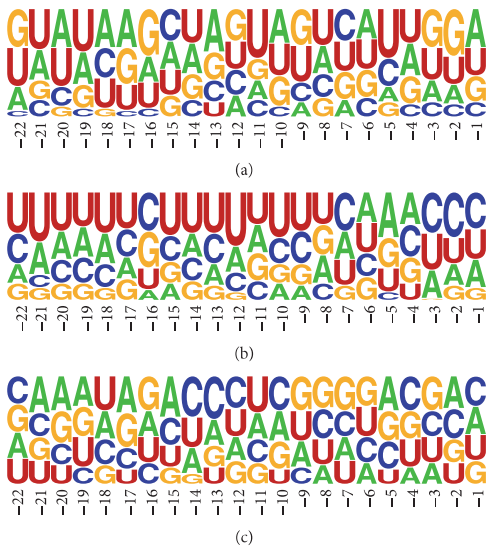


Figure 3: WebLOGO plots of nucleotide frequency in miRNA:miRNA* vicinity. First-cut region: (a) *regAC*, (b) *regBD*, and (c) random sequences.

combined these regions based on the cleavage sites and 5′-3′ strand orientation. The numbering of nucleotides in miRNA vicinity goes from −1 to −22, where nucleotide −1 is the first nucleotide beyond miRNA. The number of each motif at particular position of the sequence is shown in Figure S2 (in Supplementary Material available online at https://doi.org/10.1155/2017/6783010).

MEME results obtained for the sequences including first-cut regions are displayed in Table 1. It appears that no motif occurs in more than 25% of these sequences, but a comparison to random sequences shows significant differences between the sets.

The next analytical step concerned again regA and regC (further treated as single set and denoted as regAC), regB and regD (further denoted as regBD), and random sequences. We used WebLOGO tool to receive the information about nucleotide frequencies at particular position in miRNA vicinity. Nucleotides represented by respective letters positioned on top are the most frequent on associated positions (Figure 3 for first-cut regions, Figure S1 (Supplementary Material) for the second-cut regions). The general information concerning particular nucleotide occurrence in entire region(s) is provided in Table 2. From diagram *a* in Figure 3, showing the results for *regAC*, we can observe that U and C dominate at position −5. In the analogical region, for second cutting, the same position is occupied by A and U. Diagram *b* in Figure 3 with the results for *regBD* shows many pyrimidines (U or C). Particularly, C and U dominate positions 1–3 of the 3′ overhang. Purines (A and G) are overrepresented only at position 5. Despite these observations, WebLOGO results for *regAC* and *regBD* seemed similar to the results obtained for random sequences. Thus, two Student's *t*-tests of paired-samples were applied for

(i) each nucleotide occurrence (percentage values) at positions −22 to −1 in *regAC* and in random sequences,

(ii) each nucleotide occurrence (percentage values) at positions −22 to −1 in *regBD* and in random sequences.

The resulting *p* values obtained for both tests were equal to 1.00. This revealed no significant difference between values for *regAC*/*regBD* and values for the random set. Percentages of each nucleotide occurrence in the first-cut regions provided by WebLOGO are displayed in Tables S1–S3 (Supplementary Material).

In the next step, the secondary structures were predicted from sequences of 50 pre-miRNAs from set *S2*. Every sequence was processed by RNAstructure and mfold which generated several output structures. The most compact structure was selected for every input sequence and passed for further analysis. In the majority of cases, the most compact was the structure displaying the minimum free energy.

Consequently, bulges and internal loops were searched and subjected to an analysis by self-developed script *PatternSearch*. Bulge is a structural motif formed in a double-stranded fragment where at least one nucleotide of one strand is unpaired. Internal loop has unpaired nucleotides in both strands. If the number of unpaired nucleotides is equal for both strands, the motif is known as symmetric internal loop. Otherwise, asymmetric internal loop is formed [51, 52]. In the manuscript, we use the following notation to encode secondary structure motif. Each motif is described by a pair of numbers U-W, which specify how many unpaired nucleotides are found in each strand of the double-stranded fragment. If U is equal to 0 (no unpaired nucleotide in one of the strands) and W is between 1 and 10, the corresponding motif is a W-nucleotide bulge. If both U and W are greater than 0, the corresponding motif is an internal loop. For example, 2-3 loop describes a motif composed of two strands, where there are 2 unpaired nucleotides in one of the strands (either 5′-3′ or 3′-5′) and 3 unpaired nucleotides in the other strand.

In this paper, we focused on regions where DCL1 performs the first cutting within the precursor structure, that is, lower stem (*regAD*) in structures with base-to-loop mechanism and upper stem (*regBC*) in case of loop-to-base mechanism. With regard to these vicinity regions, we have searched 50 secondary structures of pre-miRNAs for

Table 2: Percentage of each nucleotide occurrence in *regAC*, *regBD*, and random sequences.

	First-cut region		Second-cut region		
	regAC	regBD	regAC	regBD	Random
A	26%	24%	30%	29%	26%
C	17%	26%	14%	21%	26%
G	26%	18%	21%	17%	25%
U	31%	33%	35%	34%	23%

Table 3: Number of relevant secondary structure motifs found in miRNA vicinity.

	Secondary structure motif								
	1-1	2-2	3-3	1-2	1-3	2-3	0-1	0-2	0-3
Base-to-loop									
Lower stem	42	12	8	13	3	8	16	6	3
Upper stem	26	9	5	7	10	2	9	4	7
Loop-to-base									
Lower stem	10	1	1	5	2	4	3	0	0
Upper stem	14	3	2	2	1	1	2	1	4

an occurrence of bulges and internal loops with different sizes (up to 10 nt on one of the strands). Table 3 presents most numerous motifs found by the script. The numbers of particular secondary structure motifs (bulges and internal loops), with respect to the first-cut and the second-cut region, are presented in Figure S3 (Supplementary Material). The other motifs' total occurrence has not exceeded 10 in the whole dataset of secondary structures; thus, they were considered irrelevant.

An analysis of *PatternSearch* output suggested that an arrangement of bulges (0-1, 0-2, and 0-3 in Table 3) within pre-miRNA was random. Our study also shows that significantly more symmetric internal loops occur in the first-cut regions than in the region of the second cutting. Symmetric internal 1-1 loops (one unpaired nucleotide in each strand of the loop) appeared 3 times more often than bulges. In contrast to bulge arrangement, small internal loops demonstrated the tendency to locate in specific regions of the structure. In 90% of the analysed structures, we found symmetric internal 1-1 and 2-2 loops in the closest vicinity of miRNA:miRNA* duplex, that is, 1-5 nucleotides beyond the duplex (c.f. Table 4).

We investigated the number of unpaired regions in *regAC* and *regBD* from first-cut mechanism (Figure 4). Thus, in the case of *regAC* analysis, we used *regA* of base-to-loop structures and *regC* of loop-to-base structures. In the case of *regBD*, we took *regB* of loop-to-base structures and *regD* of base-to-loop structures. In *regAC,* the most paired are positions −8, −9, and −15 (80%). On the three farthest positions mismatches are most frequently occurring (60%). On the other positions, the frequency of mismatches is approximately 30%. In contrast to *regAC*, where either regions with high paring or high mismatch level appear, in *regBD* the frequency of mismatches is very similar at each position.

An occurrence of small symmetric loops complies with the small number of mismatches distorting the stem in the region of the first cut performed by DCL1. This indicates potential structural pattern recognized by this enzyme. Occurrence of unpaired residues located further than 5 nt beyond miRNA:miRNA* duplex can also indicate potential position for the first cut.

4. Conclusions

In the paper, we focused on discovering motifs in primary and secondary structures of selected plant pre-microRNAs in order to answer the question how microprocessor composed of DCL1, HYL1, and SE recognizes the borders of microRNA:microRNA* duplex. The set of 50 sequences with experimentally confirmed cleavage mechanism was tested by selected bioinformatic tools. Sequence analysis was done using MEME Suite and WebLOGO tool. The results from MEME suggest that potential sequence motifs are UCUC in *regAC* and AACA, GUGG, and ACGG in *regBD*. This indicates that the sequence motifs could consist of either pyrimidines only (in *regAC*) or three purines and only one pyrimidine (in *regBD*). The results from WebLOGO tool were considered nonsignificant. An analysis of the secondary structure shows that the region in the vicinity of the first cut forms well defined stem comparing to the region of the second cut. However, it has been found that small symmetric internal loops 1-1 and 2-2 appear in up to 5 nt distance from the duplex. This constitutes a derogation from the results obtained for the experimentally solved RNA structures where 0-1 bulges are more common than internal loops 1-1 [53, 54]. These defined sequence and secondary structure patterns can play a key role in recognizing the location of miRNA:miRNA* duplex by DCL1 enzyme. To verify this theory, biochemical experiments involving artificially designed pre-miRNA [11] should be performed, which is planned to be done in the nearest future. Moreover, our future plans include prediction of the 3D structures of pre-miRNAs and their analysis with respect to characteristic structural features. For this purpose, computational tools like RNAComposer [3], MCQ4Structres [8], and PyMOL [55] will be applied. The generated 3D models will be evaluated based on their adjustment to the model of DCL1 structure. Finally, the analysis concerning three structural levels is going to be extended for all sequences of plant pre-miRNAs deposited in publicly available databases. However, it should be mentioned that for the majority of these sequences the cleavage mechanism has not been recognized yet.

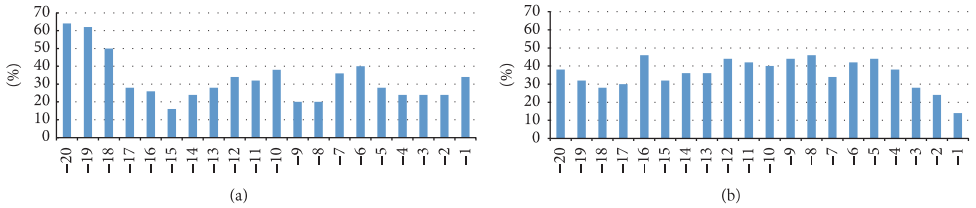


Figure 4: Percentage of unpaired nucleotides at specified positions in the first-cut regions: (a) *regAC* and (b) *regBD*.

Table 4: Small internal loops found in close miRNA:miRNA* vicinity. The table shows the number and percentage of structures that have 1-1 and 2-2 loops at specified positions in miRNA:miRNA* vicinity.

Secondary structure motif	1-1 loop					2-2 loop		
Distance from miRNA [nt]	1	2	3	4	5	1	2	3
Number and percentage of structures with motif	19 38%	5 10%	6 12%	4 8%	2 4%	4 8%	2 4%	3 6%

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was carried out in the European Centre for Bioinformatics and Genomics (Poznan, Poland) and supported by the Leading National Research Centre Program (KNOW) granted by the Polish Ministry of Science and Higher Education.

References

[1] M. Antczak, M. Popenda, T. Zok et al., "New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure," *Acta Biochimica Polonica*, vol. 63, no. 4, pp. 737–744, 2016.

[2] J. Blazewicz, W. Frohmberg, P. Gawron et al., "DNA sequence assembly involving an acyclic graph model," *Foundations of Computing and Decision Sciences*, vol. 38, no. 1, pp. 25–34, 2013.

[3] M. Popenda, M. Szachniuk, M. Antczak et al., "Automated 3D structure composition for large RNAs," *Nucleic Acids Research*, vol. 40, no. 14, article e112, 2012.

[4] J. S. Reuter and D. H. Mathews, "RNAstructure: software for RNA secondary structure prediction and analysis," *BMC Bioinformatics*, vol. 11, article no. 129, 2010.

[5] M. Szachniuk, M. Malaczynski, E. Pesch, E. K. Burke, and J. Blazewicz, "MLP accompanied beam search for the resonance assignment problem," *Journal of Heuristics*, vol. 19, no. 3, pp. 443–464, 2013.

[6] K. L. Tkaczuk, A. Obarska, and J. M. Bujnicki, "Molecular phylogenetics and comparative modeling of HEN1, a methyltransferase involved in plant microRNA biogenesis," *BMC Evolutionary Biology*, vol. 6, article 6, 2006.

[7] P. Wojciechowski, W. Frohmberg, M. Kierzynka, P. Zurkowski, and J. Blazewicz, "G-MAPSEQ—a new method for mapping reads to a reference genome," *Foundations of Computing and Decision Sciences*, vol. 41, no. 2, pp. 123–142, 2016.

[8] T. Zok, M. Popenda, and M. Szachniuk, "MCQ4Structures to compute similarity of molecule structures," *Central European Journal of Operations Research*, vol. 22, no. 3, pp. 457–473, 2014.

[9] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.

[10] J. Wiedemann and M. Miłostan, "StructAnalyzer—a tool for sequence versus structure similarity analysis," *Acta Biochimica Polonica*, vol. 63, no. 4, pp. 753–757, 2016.

[11] A. Mickiewicz, A. Rybarczyk, J. Sarzynska, M. Figlerowicz, and J. Blazewicz, "AmiRNA Designer—new method of artificial miRNA design," *Acta Biochimica Polonica*, vol. 63, no. 1, pp. 71–77, 2016.

[12] D. Bielewicz, M. Kalak, M. Kalyna et al., "Introns of plant pri-miRNAs enhance miRNA biogenesis," *EMBO Reports*, vol. 14, no. 7, pp. 622–628, 2013.

[13] M. S. Ebert and P. A. Sharp, "Roles for microRNAs in conferring robustness to biological processes," *Cell*, vol. 149, no. 3, pp. 515–524, 2012.

[14] R. Sunkar, Y.-F. Li, and G. Jagadeeswaran, "Functions of microRNAs in plant stress responses," *Trends in Plant Science*, vol. 17, no. 4, pp. 196–203, 2012.

[15] K. Wu, C. Zhu, Y. Yao, X. Wang, J. Song, and J. Zhai, "MicroRNA-155-enhanced autophagy in human gastric epithelial cell in response to Helicobacter pylori," *The Saudi Journal of Gastroenterology*, vol. 22, no. 1, pp. 30–36, 2016.

[16] T. Conrad and U. A. Ørom, "Insight into miRNA biogenesis with RNA sequencing," *Oncotarget*, vol. 6, no. 29, pp. 26546–26547, 2015.

[17] T. Du and P. D. Zamore, "microPrimer: the biogenesis and function of microRNA," *Development*, vol. 132, no. 21, pp. 4645–4652, 2005.

[18] M. Ha and V. N. Kim, "Regulation of microRNA biogenesis," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 8, pp. 509–524, 2014.

[19] W.-C. Lee, S.-H. Lu, M.-H. Lu, C.-J. Yang, S.-H. Wu, and H.-M. Chen, "Asymmetric bulges and mismatches determine 20-nt

microRNA formation in plants," *RNA Biology*, vol. 12, no. 9, pp. 1054–1066, 2015.

[20] B. C. Meyers, S. A. Simon, and J. Zhai, "MicroRNA processing: battle of the bulge," *Current Biology*, vol. 20, no. 2, pp. R68–R70, 2010.

[21] Y. Wang, R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch, "DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal," *Nature Genetics*, vol. 39, no. 3, pp. 380–385, 2007.

[22] J.-K. Zhu, "Reconstituting plant miRNA biogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 9851–9852, 2008.

[23] U. Rosani, A. Pallavicini, and P. Venier, "The miRNA biogenesis in marine bivalves," *PeerJ*, vol. 4, article no. e1763, 2016.

[24] Z. Szweykowska-Kulinska, A. Jarmolowski, and F. Vazquez, "The crosstalk between plant microRNA biogenesis factors and the spliceosome," *Plant Signaling & Behavior*, vol. 8, no. 11, Article ID e26955, 2013.

[25] R. W. Carthew and E. J. Sontheimer, "Origins and mechanisms of miRNAs and siRNAs," *Cell*, vol. 136, no. 4, pp. 642–655, 2009.

[26] T. Doran and C. Helliwell, *RNA Interference: Methods for Plants and Animals*, CABI, Oxfordshire, UK, 2009.

[27] J. Starega-Roslan, P. Galka-Marciniak, and W. J. Krzyzosiak, "Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer," *Nucleic Acids Research*, vol. 43, no. 22, pp. 10939–10951, 2015.

[28] H. J. Debat and D. A. Ducasse, "Plant microRNAs: recent advances and future challenges," *Plant Molecular Biology Reporter*, vol. 32, no. 6, pp. 1257–1269, 2014.

[29] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[30] K. J. Beezhold, V. Castranova, and F. Chen, "Microprocessor of microRNAs: regulation and potential for therapeutic intervention," *Molecular Cancer*, vol. 9, article no. 134, 2010.

[31] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl, "Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs," *Molecular Cell*, vol. 15, no. 2, pp. 185–197, 2004.

[32] Z. Xie, K. D. Kasschau, and J. C. Carrington, "Negative feedback regulation of *Dicer-Like1* in *Arabidopsis* by microRNA-guided mRNA degradation," *Current Biology*, vol. 13, no. 9, pp. 784–789, 2003.

[33] J. T. Cuperus, T. A. Montgomery, N. Fahlgren et al., "Identification of MIR390a precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 1, pp. 466–471, 2010.

[34] J. L. Mateos, N. G. Bologna, U. Chorostecki, and J. F. Palatnik, "Identification of microRNA processing determinants by random mutagenesis of *Arabidopsis MIR172a* precursor," *Current Biology*, vol. 20, no. 1, pp. 49–54, 2010.

[35] H. Zhu, Y. Zhou, C. Castillo-González et al., "Bidirectional processing of pri-miRNAs with branched terminal loops by Arabidopsis Dicer-like1," *Nature Structural & Molecular Biology*, vol. 20, no. 9, pp. 1106–1115, 2013.

[36] L. Song, M. J. Axtell, and N. V. Fedoroff, "RNA secondary structural determinants of miRNA precursor processing in *Arabidopsis*," *Current Biology*, vol. 20, no. 1, pp. 37–41, 2010.

[37] N. G. Bologna, A. L. Schapire, J. Zhai et al., "Multiple RNA recognition patterns during microRNA biogenesis in plants," *Genome Research*, vol. 23, no. 10, pp. 1675–1689, 2013.

[38] W. Kim, H.-E. Kim, A. R. Jun et al., "Structural determinants of miR156a precursor processing in temperature-responsive flowering in Arabidopsis," *Journal of Experimental Botany*, vol. 67, no. 15, pp. 4659–4670, 2016.

[39] X. Yang, W. Ren, Q. Zhao, P. Zhang, F. Wu, and Y. He, "Homodimerization of HYL1 ensures the correct selection of cleavage sites in primary miRNA," *Nucleic Acids Research*, vol. 42, no. 19, pp. 12224–12236, 2014.

[40] S. W. Yang, H. Y. Chen, J. Yang, S. Machida, N. H. Chua, and Y. A. Yuan, "Structure of Arabidopsis HYPONASTIC LEAVES1 and its molecular implications for miRNA processing," *Structure*, vol. 18, no. 5, pp. 594–605, 2010.

[41] S. Baranauske, M. Mickute, A. Plotnikova et al., "Functional mapping of the plant small RNA methyltransferase: HEN1 physically interacts with HYL1 and DICER-LIKE 1 proteins," *Nucleic Acids Research*, vol. 43, no. 5, pp. 2802–2812, 2015.

[42] H.-M. Chen, L.-T. Chen, K. Patel, Y.-H. Li, D. C. Baulcombe, and S.-H. Wu, "22-nucleotide RNAs trigger secondary siRNA biogenesis in plants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 34, pp. 15269–15274, 2010.

[43] J. T. Cuperus, A. Carbonell, N. Fahlgren et al., "Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in *Arabidopsis*," *Nature Structural & Molecular Biology*, vol. 17, no. 8, pp. 997–1003, 2010.

[44] J. Starega-Roslan, J. Krol, E. Koscianska et al., "Structural basis of microRNA length variety," *Nucleic Acids Research*, vol. 39, no. 1, pp. 257–268, 2011.

[45] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLOGO: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[46] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME suite," *Nucleic Acids Research*, vol. 43, no. 1, pp. W39–W49, 2015.

[47] A. Kozomara and S. Griffiths-Jones, "MiRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D68–D73, 2014.

[48] P. Lamesch, T. Z. Berardini, D. Li et al., "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1202–D1210, 2012.

[49] W. Fang and D. P. Bartel, "The menu of features that define primary MicroRNAs and enable de novo design of MicroRNA genes," *Molecular Cell*, vol. 60, no. 1, pp. 131–145, 2015.

[50] V. C. Auyeung, I. Ulitsky, S. E. McGeary, and D. P. Bartel, "Beyond secondary structure: Primary-sequence determinants license Pri-miRNA hairpins for processing," *Cell*, vol. 152, no. 4, pp. 844–858, 2013.

[51] B. Tian, P. C. Bevilacqua, A. Diegelman-Parente, and M. B. Mathews, "The double-stranded-RNA-binding motif: interference and much more," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 12, pp. 1013–1023, 2004.

[52] M. Popenda, M. Szachniuk, M. Blazewicz et al., "RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures," *BMC Bioinformatics*, vol. 11, article no. 231, 2010.

[53] P. L. Vanegas, G. A. Hudson, A. R. Davis, S. C. Kelly, C. C. Kirkpatrick, and B. M. Znosko, "RNA CoSSMos: characterization of secondary structure motifs—a searchable database of secondary structure motifs in RNA three-dimensional structures," *Nucleic Acids Research*, vol. 40, no. 1, pp. D439–D444, 2012.

[54] M. Popenda, M. Błażewicz, M. Szachniuk, and R. W. Adamiak, "RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures," *Nucleic Acids Research*, vol. 36, no. 1, pp. D386–D391, 2008.

[55] "The PyMOL Molecular Graphics System," Version 1.8 Schrodinger, LLC.

Supplementary material

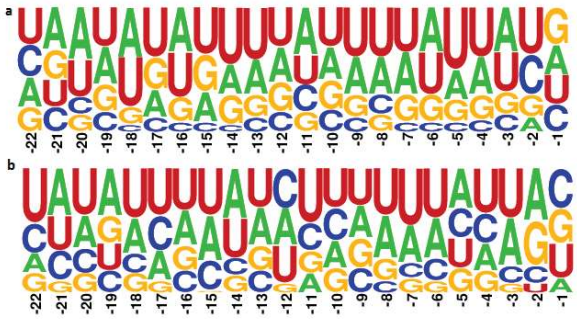


Figure S1. WebLOGO plots of nucleotide frequency in miRNA:miRNA* vicinity on the side of the second cut. (a)regAC, (b)regBD.

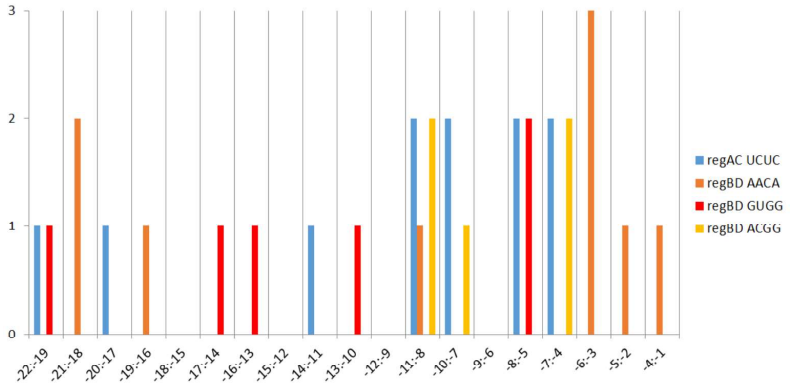


Figure S2. A number of particular sequence motifs in the first-cut regions *regAC* and *regBD*.

Table S1. Percentage of each nucleotide occurrence in the first-cut region *regAC* provided by WebLOGO.

Position	A	C	G	U
-22	22	6	36	36
-21	28	16	18	38
-20	36	18	10	36
-19	30	8	22	40
-18	36	28	8	28
-17	36	6	34	24
-16	24	8	44	24
-15	26	30	20	24
-14	28	16	24	32
-13	40	20	26	14
-12	20	24	30	26
-11	18	16	20	46
-10	32	10	28	30
-9	14	18	34	34
-8	26	22	18	34
-7	16	30	26	28
-6	28	18	26	28
-5	14	26	14	46
-4	30	12	26	32
-3	20	14	36	30
-2	18	12	44	26
-1	36	14	24	26

Table S2. Percentage of each nucleotide occurrence in the first-cut region *regBD* provided by WebLOGO.

Position	A	C	G	U
-22	18	28	16	38
-21	30	12	12	46
-20	26	22	16	36
-19	28	24	16	32
-18	28	22	12	38
-17	20	26	20	34
-16	10	36	34	20
-15	16	24	20	40
-14	22	22	18	38
-13	20	28	14	38
-12	24	18	8	50
-11	26	18	26	30
-10	16	24	18	42
-9	14	30	20	36
-8	26	16	28	30
-7	26	36	14	24
-6	28	22	22	28
-5	42	8	28	22
-4	30	30	24	16
-3	26	38	2	34
-2	20	42	14	24
-1	26	34	12	28

Table S3. Percentage of each nucleotide occurrence in random sequences provided by WebLOGO.

Position	A	C	G	U
-22	24	26	26	24
-21	34	22	22	22
-20	32	20	30	18
-19	28	16	28	28
-18	26	24	18	32
-17	34	20	32	14
-16	24	20	32	24
-15	38	26	16	20
-14	24	38	10	28
-13	24	40	20	16
-12	22	30	22	26
-11	24	22	20	34
-10	28	32	26	14
-9	24	16	32	28
-8	18	28	30	24
-7	22	24	32	22
-6	20	24	28	28
-5	32	26	28	14
-4	16	32	26	26
-3	20	24	32	24
-2	30	28	22	20
-1	26	28	22	24

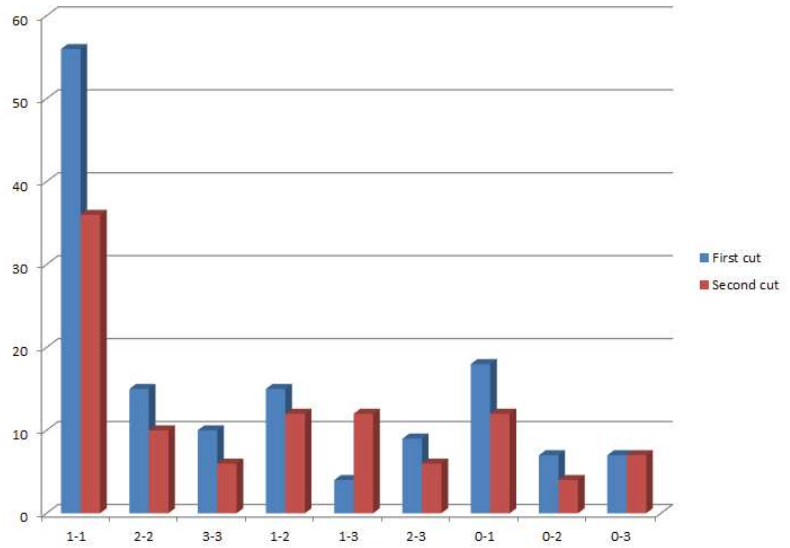


Figure S3. A number of particular secondary structure motifs (bulges and internal loops) in the first-cut (blue) and the second-cut (red) region.

molecules

MDPI

Article

Discovering Structural Motifs in miRNA Precursors from the *Viridiplantae* Kingdom

Joanna Miskiewicz [1] and Marta Szachniuk [1,2,*]

[1] Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland; joanna.miskiewicz@cs.put.poznan.pl
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland
* Correspondence: marta.szachniuk@cs.put.poznan.pl; Tel.: +48-616-653-030

Received: 29 April 2018; Accepted: 4 June 2018; Published: 6 June 2018

Abstract: A small non-coding molecule of microRNA (19–24 nt) controls almost every biological process, including cellular and physiological, of various organisms' lives. The amount of microRNA (miRNA) produced within an organism is highly correlated to the organism's key processes, and determines whether the system works properly or not. A crucial factor in plant biogenesis of miRNA is the Dicer Like 1 (DCL1) enzyme. Its responsibility is to perform the cleavages in the miRNA maturation process. Despite everything we already know about the last phase of plant miRNA creation, recognition of miRNA by DCL1 in pre-miRNA structures of plants remains an enigma. Herein, we present a bioinformatic procedure we have followed to discover structure patterns that could guide DCL1 to perform a cleavage in front of or behind an miRNA:miRNA* duplex. The patterns in the closest vicinity of microRNA are searched, within pre-miRNA sequences, as well as secondary and tertiary structures. The dataset consists of structures of plant pre-miRNA from the *Viridiplantae* kingdom. The results confirm our previous observations based on *Arabidopsis thaliana* precursor analysis. Hereby, our hypothesis was tested on pre-miRNAs, collected from the miRBase database to show secondary structure patterns of small symmetric internal loops 1-1 and 2-2 at a 1–10 nt distance from the miRNA:miRNA* duplex.

Keywords: miRNA biogenesis; structural patterns; DCL1

1. Introduction

MicroRNAs (miRNAs) represent a group of small noncoding RNAs (sRNA) that consist of about 21–24 nucleotides [1–8]. They are present in animals, plants, and single-cell eukaryotes. The key role of miRNA is to regulate gene expression via degrading or blocking the targeted mRNA transcript [9,10]. With the ability to silence various genes, microRNA can modulate the homeostasis of the organism by interfering with specific mRNAs, as well as by preventing further expression of genes engaged in development, metabolism, or differentiation [3,11–14]. Mis-regulation of miRNAs, which are involved in different biological processes, is believed to be a major contributor to various diseases [15]. The recognition of targeted transcripts comes through nearly complete (in plants) or partially complete (in animals) base pair complementarity [6,16]. The multistep miRNA biogenesis differs between plants and animals, mainly in the cell location where each stage of the process is held and in the contributing proteins. The transcribed miRNA gene (pri-miRNA) in animals is cleaved into a precursor (pre-miRNA) structure by a microprocessor. The microprocessor primarily consists of two enzymes: RNAse III Drosha and DiGeorge Syndrome Critical Region 8 (DGCR8) (in several organisms DGCR8 is replaced by Pasha) [17–19]. At this phase, pre-microRNA is transported from the nucleus to the cytoplasm by Exportin 5 protein (XPO5). Next, Dicer (the other RNase III type enzyme), performs cleavages in pre-miRNA to release the duplex of microRNA (miRNA:miRNA*) [19].

In plants, all endonucleolytic cleavages of pri-miRNA and pre-miRNA are performed in the nucleus by Dicer-Like 1 (DCL1), being a homologue of Dicer. The process of plant miRNA maturation also requires engagement of HYPONASTIC LEAVES 1 (HYL1), a protein that contains a dsRNA-binding domain, and SERRATE (SE), a protein containing a zinc-finger domain. After creation, pre-miRNA is exported to the cytoplasm by the HASTY enzyme, a homologue of XPO5 [5,12,20,21]. In both, animal and plant cells, the miRNA:miRNA* duplex consists of a guide and a passenger strand. During incorporation of the duplex into the RNA-induced silencing complex (RISC), the passenger strand is discarded, while the guide strand leads the complex toward the target mRNA [22–24]. The passenger strand (miRNA*) is either degraded or used as a guide for other transcripts. Besides miRNA, which determines the targeted mRNA via base pair complementarity, RISC includes an ARGONAUTE (AGO) protein, the effector molecule with slicing activity [7,25]. The RISC enables degradation of the target mRNA or inhibition of the translation process by several mechanisms, including ARGONAUTE endonuclease activity, which enables slicing of targeted mRNA [3,5,10,25]. Biogenesis of animal miRNAs can be classified as a well-known process. The cleavages performed on animal pre-miRNA by the molecular ruler Dicer are measured from the pre-miRNA terminus, either the 3′ or the 5′ end, to the RNase III domain-dependent cleavage site [26,27]. In plants, it is still a mystery how the DCL1 enzyme recognizes miRNAs within pre-miRNA structures to perform cuts and release the miRNA:miRNA* duplex. Therefore, we have decided to analyze a set of available pre-miRNA structures and look for structural patterns occurring in miRNA vicinity. It is assumed that some motifs should exist and guide DCL1. Herein, we present a broad approach to pattern searching within pre-miRNAs. We have applied it to structures from four phyla of the *Viridiplantae* kingdom. We drew from our previous research concerning structural motifs in precursor microRNAs of *Arabidopsis thaliana*.

2. Results

2.1. A Scheme of Data Processing

Our research project has followed several steps (Figure 1). At first, the data for an analysis was collected and pre-processed. After dataset preparation, a semi-automated processing of pre-miRNAs followed. It was conducted at three structure levels. We started by investigating the sequences, and going through secondary structure studies, we ended up with a three-dimensional (3D) structure analysis. A detailed description of these steps is provided in the following paragraphs.

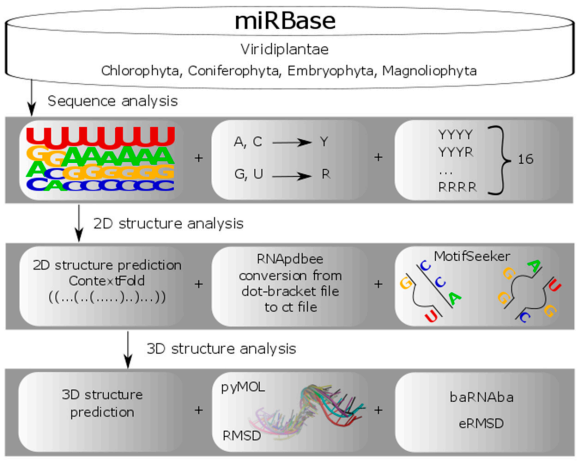


Figure 1. Precursor microRNA (pre-miRNA) analysis workflow.

2.2. Dataset Preparation

In order to find structural motifs in plant pre-miRNA, which could help understand DCL1 performance, we prepared a dataset based on sequences stored in the miRBase database [4]. We considered records under the *Viridiplantae* kingdom assigned to the following phyla: *Magnoliophyta* (6547 sequences), *Coniferophyta* (108 sequences), *Chlorophyta* (50 sequences), and *Embryophyta* (287 sequences). Altogether, our initial collection contained 6992 sequences. The Table S1 from Supplementary Materials contains number of sequences extracted from miRBase website [4] distributed by phylum, clade, family and species. Next, we extracted the relevant information of collected *Viridiplantae* from the miRBase [4] website, and shaped it to adjust to further processing. This was done using self-prepared scripts written in Python language. The prepared data files contained an accession number for each pre-miRNA (in accordance with the miRBase nomenclature) assigned to the sequence, and an miRNA position within its appropriate precursor. From the miRBase [4] database, we also collected evidence about every miRNA found within the set of 6992 sequences, which could be experimental (by similarity) or not experimental. In our research, we planned to focus the analysis on the miRNA vicinity. Thus, we needed to have the sequences and structures of miRNA precursors containing miRNAs with sufficiently large neighbouring regions. It had been decided that eight nucleotides per strand constituted a sufficient size for the vicinity sequence to be analyzed. In the initial collection of 6992 sequences, we identified 5345 pre-miRNA sequences in which miRNAs were surrounded by at least 8 nt on their 5′ and 3′ ends: 4956 from *Magnoliophyta*, 80 from *Coniferophyta*, 38 from *Chlorophyta*, and 271 from *Embryophyta*. These sequences were selected to form the basic *S8* set used in the majority of forthcoming experiments. Within this set, at least one miRNA per each sequence was confirmed experimentally (in the subset of 4388 sequences) or by similarity (within the subset of 343 sequences). In the remaining 614 sequences of the *S8* set (<11.5%), miRNAs were confirmed non-experimentally (i.e., the miRNA sequence was revealed by sequencing, and not used in any experiment yet).

Further, we found it also necessary to limit the miRNA vicinity size to 4 nt. To meet this requirement, from the initial 6992 sequences, we picked 5975 pre-miRNAs with at least 4 nucleotides on both sides of miRNA: 5555 from *Magnoliophyta*, 99 from *Coniferophyta*, 41 from *Chlorophyta*, and 280 from *Embryophyta*. These were collected in the *S4* set, which included 5345 sequences from the *S8* set (vicinity size $\geq$8 nt) and 630 sequences with vicinity size between 4 and 7 nt. These sequence collections allowed us to properly define the search space for our computational experiments. Within the *S4* set, at least one miRNA per sequence was confirmed experimentally (in the subset of 4890 sequences) or by similarity (within the subset of 389 sequences). In the remaining 696 sequences of the *S4* set (<12%), miRNAs were not confirmed experimentally (i.e., miRNA sequence was revealed by sequencing, and not used in any experiment yet).

2.3. Primary Structure-Based Analysis

In the first computational experiment, we have used the *S8* set of the pre-miRNA sequences. In every sequence from *S8*, either one or two miRNAs were found. We identified an 8 nt-long vicinity sequence on the 5′ and 3′ end of each of these miRNAs. These sequence fragments were extracted to form *VS8-5′* and *VS8-3′* subsets of a large *VS8* collection, including 12802 vicinity sequences with the length equal to 8 nt exactly. Subset *VS8-5′* contains 6401 vicinity sequences occurring in the miRNA vicinity on the 5′ end, and subset *VS8-3′* has 6401 sequences from the 3′ end vicinity. Both subsets, *VS8-5′* and *VS8-3′*, were processed using WebLogo tool versions 2.8.2 (https://weblogo.berkeley.edu/logo.cgi) [28] and 3.0 (http://weblogo.threeplusone.com/create.cgi) [28]. WebLogo allowed us to obtain a diagram showing the most- and the least-frequent nucleotides occurring on each of the eight positions of miRNA vicinity sequence. The first position in each sequence is the first nucleotide behind the microRNA, counting towards the 3′ end (in the *VS8-5′* subset) or towards the 5′ end (in the *VS8-3′* subset). The most frequent nucleotides are shown at the top of the stack, while the least frequent ones are at the bottom (Figure 2). Detailed information about nucleotides occupying the following

positions within vicinity sequences is provided in Table 1 (for the *VS8-5′* subset) and Table 2 (for the *VS8-3′* subset).

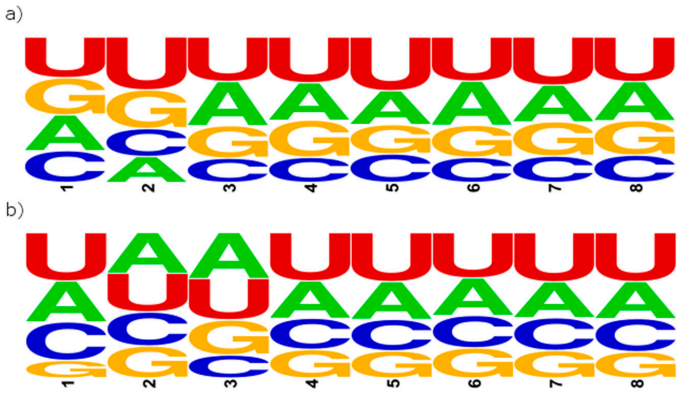


Figure 2. WebLogo 2.8.2 [28] diagram for sequences from the (**a**) *VS8-5′* and (**b**) *VS8-3′* subsets.

Table 1. WebLogo 3.0 [28] results for vicinity sequences in the *VS8-5′* subset.

Position	A [%]	C [%]	G [%]	U [%]	R [%]	Y [%]
1	24.48	21.87	25.62	28.03	50.10	49.90
2	17.78	19.28	26.56	36.38	44.34	55.66
3	30.32	16.67	22.37	30.64	52.69	47.31
4	25.68	17.31	25.46	31.54	51.15	48.85
5	23.18	19.56	20.26	36.99	43.45	56.55
6	30.21	17.23	22.11	30.45	52.32	47.68
7	25.17	18.12	23.54	33.17	48.71	51.29
8	26.71	18.75	23.76	30.78	50.48	49.52

Table 2. WebLogo 3.0 [28] results for vicinity sequences in the *VS8-3′* subset.

Position	A [%]	C [%]	G [%]	U [%]	R [%]	Y [%]
1	27.98	26.51	12.19	33.32	40.17	59.83
2	27.90	23.37	22.09	26.64	49.99	50.01
3	31.48	15.92	24.14	28.46	55.62	44.38
4	25.56	21.45	19.72	33.28	45.27	54.73
5	25.84	21.67	18.73	33.76	44.57	55.43
6	25.73	22.26	20.81	31.20	46.54	53.46
7	24.57	22.54	19.00	33.89	43.57	56.43
8	25.31	22.81	18.15	33.73	43.46	56.54

It can be observed that Uracil is the most frequent nucleotide on almost every position of each vicinity sequence. In sequences from the *VS8-5′* subset, the second position is heavily occupied by Uracil (36.38% of sequences in *VS8-5′* have Uracil on the second position), and rather poorly by Adenine (17.78%). This can indicate an unpairing in the structure, which occurs exactly on this position. In the *VS8-3′* subset, bigger differences are observed between Cytosine and Guanine occupation. The biggest difference reaches 14.22%, and concerns the first position of the vicinity sequence. In the *VS8-5′* subset, nucleotides on the first position are almost evenly distributed, while the second position seems to create an unpaired region. The *VS8-3′* subset seems to be contrary to this. It shows almost equally distributed values on the second position and highly varied distribution in the first position. Thus, it is possible that in the region of the first two positions beyond the miRNA sequence, one could find a small

mismatch, revealed as a bulge or a loop in the structure. In the second experiment, aimed to search for sequential motifs in miRNA vicinity, we decided to represent each nucleotide in nucleotide ambiguity code (IUPAC) [29], based on the number of carbon-nitrogen rings, as a purine (R) or pyrimidine (Y). At first, this experiment was run on the previously created *VS8-3′* and *VS8-5′* subsets. In every vicinity sequence from these subsets, we changed the representation of adenines (A) and guanines (G) into purines (R) and uracils (U) and cytosines (C) into pyrimidines (Y). Next, we searched for exactly 8 nt-long patterns that were also encoded using the two-letter alphabet {R, Y}. All permutations for eight positions with two possible variants, purine or pyrimidine, gave us 256 possible patterns. We did not observe any significant results in this experiment. Therefore, we decided to restrict the search space and run the experiment for shorter vicinity sequences. We have taken the *S4* set of 5975 pre-miRNAs, containing miRNAs with neighbouring regions having at least 4 nucleotides on both the 5′ and 3′ end next to the miRNA region. From this collection, we extracted 14300 vicinity sequences 4 nt long, and divided them into two subsets, *VS4-5′* and *VS4-3′*, in the same manner as *VS8*. Each of these subsets contained 7150 short sequences. Every vicinity sequence from *VS4-5′* and *VS4-3′* was next represented with the two-letter alphabet {R, Y}, and the search for 4 nt-long patterns was performed, providing the results as presented in Table 3.

Table 3. Pattern occurrence in the *VS4-5′* and *VS4-3′* subset.

Pattern	*VS4-5′* [%]	*VS4-3′* [%]	Total [%]
RRYR	4.36	3.82	4.09
YRYR	4.41	4.57	4.49
RYYR	6.22	3.90	5.06
RRRY	5.43	5.92	5.67
RYRY	6.08	5.33	5.71
RRYY	6.13	5.30	5.71
YRYY	4.98	6.78	5.88
RYYY	6.90	4.98	5.94
YYYR	6.77	5.45	6.11
RYRR	7.50	5.29	6.39
RRRR	7.43	5.64	6.53
YYRY	6.77	6.67	6.72
YRRR	6.38	7.40	6.89
YRRY	4.83	10.10	7.46
YYYY	7.29	9.17	8.23
YYRR	8.55	9.68	9.11

The first symbol of a pattern corresponds to the nucleotide on the first position beyond miRNA sequence. From these statistics, we can observe that five of the most frequent motifs start with pyrimidine: YYRY, YRRR, YRRY, YYYY, and YYRR. This suggests that many sequences which encounter miRNA involve uracil or cytosine right before the first nucleotide of miRNA sequence.

2.4. Secondary Structure-Based Analysis

The second part of our analysis concerned the secondary structures. Since our input data collection contained sequences only, we decided to predict their secondary structures using ContextFold version 1.0 [30] installed on a local computer. The software was chosen based on the CompaRNA benchmark [31]. All 5975 sequences from the *S4* set were processed by ContextFold [30] to predict their secondary structures. Predicted structures were encoded in dot-bracket notation. For the facilitation of further analysis, we used RNApdbee program (http://rnapdbee.cs.put.poznan.pl/) [32–34] to transform two-dimensional (2D) structures from dot-bracket to CT (Connect) format. Next, we applied a script called MotifSeeker implemented in Python language. The MotifSeeker processes CT files, and searches for bulges and internal loops in the vicinity of the miRNA:miRNA* duplex (up to four nucleotides beyond the miRNA on both sides). The generated output file contains brief information

about what motif has been found, on which strand, and how far it was from the microRNA. Guided by our previous study of the pre-miRNA sequences of *Arabidopsis thaliana* [5] and current WebLogo [28] results, we expected an accumulation of mismatches between the first and fourth position beyond miRNA. Although it is known that similar sequences do not always maintain the similarities at higher structural levels [35], we supposed that in our case, the analyzed structures would share some of their pattern in the short fragment beyond the miRNA:miRNA* duplex at the secondary or tertiary structural level. MotifSeeker allowed us to identify the most frequently occurring secondary structure pattern, along with its distance from the miRNA:miRNA* duplex, and a number of structures in which the motif was found. According to our assumptions, the first eight most frequent patterns had small mismatches: symmetric internal loops 1-1 (single unpaired nucleotide on every strand of the vicinity region) and 2-2 (two unpaired nucleotides on every strand of the vicinity region). We have found that in 21.56% of the 5975 secondary structures, the first nucleotides beyond the miRNA:miRNA* duplex were unpaired and formed symmetric 1-1 internal loops. The same 1-1 pattern was shared by 13.82% of the secondary structures, starting from the second position, and 16.55% of the structures starting from the third position beyond the miRNA:miRNA* duplex. This means that over 50% (exactly 51.93%) of the analyzed secondary structures contain the 1-1 motif at the maximum distance of three positions beyond miRNA. In Table 4, we present the exact number of motifs found within the structures in which we discovered the pattern. All motifs identified by MotifSeeker are represented in Figure 3, where each position is defined by the pattern type (1-1 or 2-2) and the distance between the motif and the miRNA, from 1 nt (D:1) up to 4 nt beyond miRNA (D:4). The MotifSeeker code and input files can be found here: http://bio.cs.put.poznan.pl/fileserver/.

Table 4. Motif occurrence in the *S4* set. The number of motifs was calculated based on the number of specific patterns in defined locations, referring to structures which contain at least one motif.

Motif/Distance	Number of Motifs	Number of Structures with at Least One Motif
1-1/D:1	1397	1288
1-1/D:3	1043	989
1-1/D:2	861	826
1-1/D:4	807	769
2-2/D:3	221	219
2-2/D:1	190	187
2-2/D:2	149	147
2-2/D:4	118	117

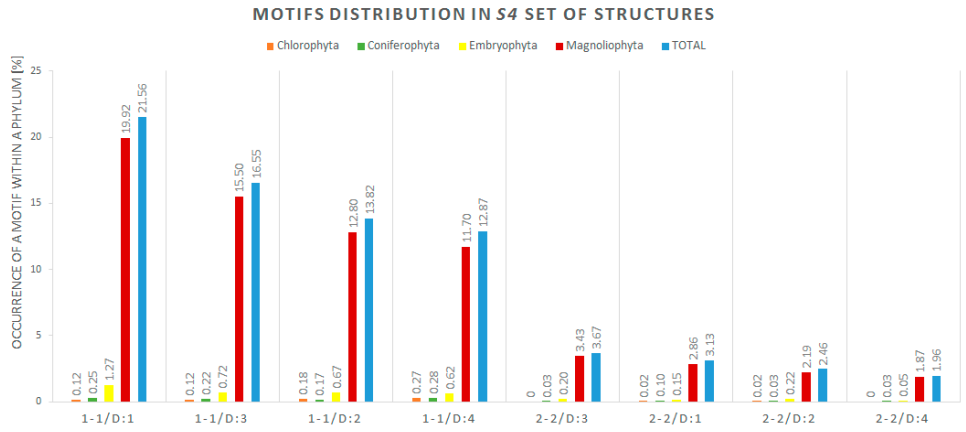


Figure 3. Distribution of eight most-occurring two-dimensional (2D) motifs in 5975 structures by phyla. The results are arranged from the least frequent motif to the most common one.

2.5. Tertiary Structure-Based Analysis

In the third stage of analysis, the tertiary structures of miRNA vicinity were analyzed using bioinformatics tools. Over many years, lots of methods for RNA 3D structure analysis have been developed [36,37]. In our experiments, we decided to focus on three of them: RNAComposer [38,39], PyMOL [40], and baRNAba [41]. First, we predicted 40 tertiary structures by using RNAComposer [38,39]. The input set for the prediction process included 10 sequences for each phylum picked randomly from S4 dataset. The obtained models were next processed by using the PyMOL program [40]. From each predicted tertiary structure, the closest vicinity regions of miRNA were cut out for alignment. Due to the shift between the 5′ and 3′ miRNA, we decided to use regions that were overlapping the miRNA:miRNA* duplex for 4 nt beyond the duplex and 4 nt within the duplex. This resulted in obtaining 8 nt-long structures from both sides of the miRNA:miRNA*. For each phylum, we have generated 20 short 3D fragments. Among them, one random structure was chosen as a reference—the remaining ones were aligned to it. Thus, we created four different alignments (Figure 4), with root mean square deviation (RMSD) values measured by PyMOL [40] and eRMSD values computed by the baRNAba software [41]. RMSD allowed us to measure the similarity between the superimposed atomic coordinates [42] whereas eRMSD facilitated to measure the distance between structures based only on the relative positions and orientations of nucleobases [41].

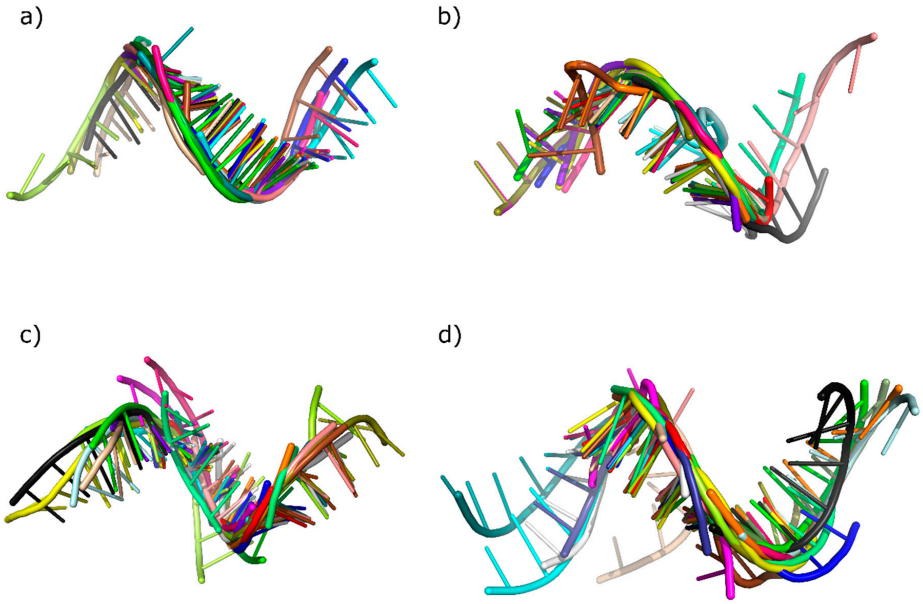


Figure 4. Aligned three-dimensional (3D) substructures within each phylum: (**a**) *Chlorophyta*, (**b**) *Coniferophyta*, (**c**) *Embryophyta*, and (**d**) *Magnoliophyta*.

The RMSD values presented in Table 5 do not exceed 2.5 Å, while the average values are not higher than 1.5 Å. Relatively low values are also found in Table 6, representing eRMSD. The highest value in Table 6 is 1.101 Å, and all four calculated averages are below 0.90 Å. In both situations, the results indicate high 3D structure similarity between the four phyla. Thus, the closest region to the miRNA:miRNA* duplex seems to be highly conserved between the phyla in *Viridiplantae* kingdom.

Table 5. RMSD values of 3D fragments from each phylum.

Fragment Id	RMSD [Å]			
	Chlorophyta	*Coniferophyta*	*Embryophyta*	*Magnoliophyta*
1	2.112	0.463	1.882	2.245
2	0.278	0.430	0.290	2.270
3	0.256	1.194	2.058	1.135
4	0.117	0.381	1.626	0.679
5	0.467	0.258	2.351	0.352
6	2.209	1.228	1.810	0.567
7	0.257	0.469	1.966	0.123
8	0.560	1.226	1.587	0.449
9	0.142	1.018	1.773	2.171
10	0.864	0.412	1.247	1.672
11	0.502	0.461	0.910	0.845
12	0.547	0.444	1.573	0.607
13	0.034	1.377	0.974	1.171
14	0.389	0.846	1.546	0.963
15	1.155	1.036	0.944	0.836
16	0.139	0.481	0.837	1.094
17	0.686	1.210	1.839	0.597
18	0.637	0.390	1.730	1.344
19	2.159	0.266	0.330	2.304
Average	0.711	0.715	1.435	1.128

Table 6. eRMSD values of 3D fragments from each phylum.

Fragment Id	eRMSD [Å]			
	Chlorophyta	*Coniferophyta*	*Embryophyta*	*Magnoliophyta*
1	0.459	0.765	0.802	0.554
2	0.788	0.771	0.434	0.503
3	0.587	0.436	0.725	0.730
4	0.291	1.047	0.776	1.101
5	0.477	1.047	0.868	0.325
6	0.432	0.746	0.858	0.444
7	0.561	1.025	0.868	0.832
8	0.442	0.799	0.817	0.365
9	0.459	0.675	0.767	0.455
10	0.438	0.800	0.842	0.643
11	0.386	0.749	1.080	0.390
12	0.251	0.753	0.841	0.398
13	0.605	0.745	0.906	0.457
14	0.410	0.680	0.791	0.447
15	0.410	0.891	0.883	0.394
16	0.463	0.729	0.901	0.467
17	0.564	1.023	0.788	0.331
18	0.528	1.058	0.764	0.604
19	0.453	0.712	0.810	0.604
Average	0.474	0.813	0.817	0.529

3. Discussion

MicroRNA research has become increasingly popular since these molecules were discovered [43,44]. Nowadays, it is not only in-vivo or in-vitro methods that are used to examine the nature of miRNAs. In-silico approaches allow us to predetermine the direction of experiments, and help to narrow the search space to answer the questions raised. Here, we focused on plant microRNAs and performed a series of computational experiments using bioinformatic methods and programs. At each level

of the RNA structure, we searched for specific motifs that could guide the DCL1 enzyme to the cutting position of the miRNA:miRNA* duplex. Every analytical step we carried out led to us finding small mismatches placed in the closest vicinity of the 5′ and 3′ ends of the miRNA. Although the results of sequence analysis did not unequivocally indicate the specific unpairing in this area, the secondary structure study proved this hypothesis. In the phase of 2D structure analysis, we discovered a high number of symmetric 1-1 and 2-2 internal loops occurring no further than four nucleotides behind the miRNA:miRNA* duplex. This supports the results of our previous research on *Arabidopsis thaliana*, where we also found a significant number of such motifs in the direct vicinity of miRNA [10]. Additionally, we examined tertiary structures by aligning predicted 3D models of the miRNA neighbourhood and calculating two distance measures (RMSD and eRMSD) between them, divided by phyla. The results confirmed the appearance of a conserved region close to the duplex. In conclusion, the taken bioinformatic pathway helped us to discover potential motifs recognized by the DCL1 enzyme. By examining each structural level, we managed to extract the necessary information and draw proper conclusions. Obtained via in-silico methods, the results clearly point out the significance of closest vicinity of miRNA and mismatches occurring in this region.

4. Materials and Methods

The research focused on three structural levels of RNA architecture: sequence, secondary, and tertiary structure. Sequences were obtained from miRBase (http://www.mirbase.org/), a repository of pre-microRNAs of various organisms [4]. Based on experimental data, this database includes not only sequences, but also positions of miRNA on the 5′ and 3′ strand. Annotation and sequence data for each entry are displayed on the website, along with the proposed secondary structure model of the pre-miRNA.

4.1. WebLogo

Sequence analysis was performed using WebLogo [28], aimed to discover the most frequent nucleotide on each position of miRNA vicinity area. WebLogo version 2.8.2 [28] (https://weblogo.berkeley.edu/logo.cgi) produced diagrams showing the frequency of nucleotides at each analyzed position. The first position is marked as the closest one to miRNA. WebLogo version 3.0 [28] (http://weblogo.threeplusone.com/create.cgi) was used to generate numerical values of nucleotide frequencies. WebLogo 2.8.2 [28] was used with the following settings for image format and size: *Image format* as eps (vector), and Logo Size per line equals to 18 × 5 cm. For advanced logo options, the settings were as follows: *Sequence Type* was automatic detection; *First Position Number* was 1; *Small Sample Correction* was true; *Frequency Plot* was true; *Logo Range* was none; *Multiline Logo (Symbols per Line)* was false. The advanced image options were set as follows: *Bitmap Resolution* at 96 pixels/inch (dpi); *Antialias Bitmaps* was set to true; *Title* was none; *Y-Axis Height* was none; *Show Y-Axis* was true; *Show X-Axis* was true; *Y-Axis Label* was none; *X-Axis Label* was none; *Show Error Bars* was false; *Boxed/Boxed Shrink Factor* was false; *Show Fine Print* was true; *Label Sequence Ends* was false; *Outline Symbols* was false; and *Y-Axis Tic Spacing* was 1 bit. Colors settings were selected as default. In the WebLogo 3.0 tool [28], we used following parameters: *Title* was none; *Output Format* was data (plain text); *Sequence type* was auto; *Logo size* was medium; *Stacks per Line* was 40; *Ignore lower case* was false; *Units* were probability; *First position number* was 1; *Logo range* was none; *Figure label* was none; *Scale stack widths* was true; *Composition* was auto; *Error bars* were false; *Show Sequence Ends labels* was false; *Version Fine Print* was true; *X-axis* was true; *Y-axis* was true; *Y-axis scale* was auto; *Y-axis tic spacing* was 1.0; and *Color Scheme* was auto.

4.2. Purine–Pyrimidine Patterns

The next phase of the study required changes in miRNA vicinity sequences. Adenine and guanine were represented as R (which denotes purines), while cytosine and uracil were represented as Y (which denotes any pyrimidine). These substitutions were applied by self-created script in Python

language. Again, sequence patterns were searched in the modified sequences with using self-developed Python script.

4.3. ContextFold

In the second analytical step, the secondary structures were predicted via the ContextFold program [30]. This program, installed on a local computer, produces files which contain 2D structures defined in dot-bracket notation. In this format, each unpaired nucleotide (mismatch or gap) is represented as a single dot, and a paired nucleotide as an opening or closing bracket. The command used, *java-cp bin contextFold.app.Predict in: input_file.txt out:output_file.txt*, enabled prediction of the secondary structures for all RNA sequences in the input file, using the (default) supplied StHighCoHigh trained model, and saving the result to the output file [45].

4.4. RNApdbee

To facilitate further research, we used the RNApdbee webserver [32,34] (http://rnapdbee.cs.put.poznan.pl/) to convert dot-bracket representation into CT format. The latter data format describes the position of nucleotide in the sequence, nucleobase encoding, the position of the previous and next nucleotides in the sequence, and the index of the paired nucleotide. If the nucleotide is unpaired, the index equals 0. On the RNApdbee website, we chose the third mode of analysis (i.e., third tab page, selecting "(. . . .) → image"). After uploading the structures in dot-bracket notation, we selected the options to (1) identify the structural elements by treating pseudoknots as paired residues, and (2) visualize the secondary structure using the VARNA-based procedure. When the computation was finished, we downloaded the results in CT file format.

4.5. MotifSeeker

The secondary structures were examined by self-developed script named MotifSeeker. MotifSeeker reads CT files and additional information from the pre-miRNA id and its microRNA positions at the 5′ and 3′ ends. Next, the script searches for bulges and internal loops, providing information about the type of mismatch and its distance from miRNA.

4.6. RNAComposer

The last phase of our research involved the prediction of tertiary structures of RNA. We selected 10 secondary structures from each phylum, and used them to predict their 3D structures using RNAComposer (http://rnacomposer.cs.put.poznan.pl/), running it in batch mode [38,39]. RNAComposer allows us to automatically predict tertiary RNA structures, up to 500 nt per structure, based on their secondary structure in dot-bracket format. It is possible for the user to choose one of the six secondary structure prediction methods incorporated into the system. For our analysis, we set the *Select secondary structure prediction method* option to "true", and from the drop-down list we chose the ContextFold method [30]. The same can be done in the interactive mode of RNAComposer, where the user can either select the secondary structure prediction method by selecting it from drop-down list or by typing the method name in the next line after the sequence (no dot-bracket notation is required in this case), e.g.,:

```
#zma_MIR168a
>example
GAAGCCGCGCCGCCUCGGGCUCGCUUGGUGCAGAUCGGGACCCGCCGCCCGGCCGACGG
GACGGAUCCCGCCUUGCACCAAGUGAAUCGGAGCCGGCGGAGCGA
ContextFold
```

Since we have used the batch mode, we could generate more than one 3D structure per secondary structure input. However, we decided to generate a single 3D structure model, and the *Maximum number of generated 3D models* was set to 1.

4.7. PyMOL

The obtained 3D structures were processed in PyMOL [40]. PyMOL software enables molecular visualization, measurement, processing, and model comparison. We used it to align structures within each phylum, and to measure the RMSD values between them. RMSD (root mean square deviation) is one of the standard measures that calculates an average distance between the atoms.

4.8. BaRNAba

Finally, the baRNAba tool was applied to calculate eRMSD values, which refer to the distance considering only the relative positions and orientations of nucleobases [46]. The command applied for baRNAba tool was *./baRNAba –name output_file.txt ERMSD –pdb reference.pdb -f 1_structure.pdb 2_structure.pdb ... 19_structure.pdb*.

Supplementary Materials: The following are available online. Table S1. Number of sequences extracted from miRBase website [4] distributed by phylum, clade, family and species.

Author Contributions: Marta Szachniuk conceived and supervised the study. Joanna Miskiewicz prepared the dataset, designed and carried the experiments. Both authors analyzed the results and participated in manuscript writing. Joanna Miskiewicz prepared the figures.

Funding: The authors acknowledge partial support from the National Science Center, Poland (grant 2016/23/B/ST6/03931), and Institute of Bioorganic Chemistry, Polish Academy of Sciences, within an intramural financing program.

Acknowledgments: This research was carried in the European Centre for Bioinformatics and Genomics, Poznan University of Technology, and the Institute of Bioorganic Chemistry PAS, Poland (granted HR Excellence in Research).

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Iorio, M.V.; Croce, C.M. microRNA involvement in human cancer. *Carcinogenesis* **2012**, *33*, 1126–1133. [CrossRef] [PubMed]
2. Tutar, Y. miRNA and cancer; computational and experimental approaches. *Curr. Pharm. Biotechnol.* **2014**, *15*, 429. [CrossRef] [PubMed]
3. Rogers, K.; Chen, X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell* **2013**, *25*, 2383–2399. [CrossRef] [PubMed]
4. Kozomara, A.; Griffiths-Jones, S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **2014**, *42*, D68–D73. [CrossRef] [PubMed]
5. Stepien, A.; Knop, K.; Dolata, J.; Taube, M.; Bajczyk, M.; Barciszewska-Pacak, M.; Pacak, A.; Jarmolowski, A.; Szweykowska-Kulinska, Z. Posttranscriptional coordination of splicing and miRNA biogenesis in plants. *Wiley Interdiscip. Rev. RNA* **2017**, *8*, 1–23. [CrossRef] [PubMed]
6. Carthew, R.W.; Sontheimer, E.J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **2009**, *136*, 642–655. [CrossRef] [PubMed]
7. Bartel, D.P. MicroRNA Target Recognition and Regulatory Functions. *Cell* **2009**, *136*, 215–233. [CrossRef] [PubMed]
8. Axtell, M.J. Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* **2013**, *64*, 137–159. [CrossRef] [PubMed]
9. Mickiewicz, A.; Rybarczyk, A.; Sarzynska, J.; Figlerowicz, M.; Blazewicz, J. AmiRNA Designer—New method of artificial miRNA design. *Acta Biochim. Pol.* **2016**, *63*, 71–77. [CrossRef] [PubMed]
10. Miskiewicz, J.; Tomczyk, K.; Mickiewicz, A.; Sarzynska, J.; Szachniuk, M. Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *BioMed Res. Int.* **2017**, *2017*, 6783010. [CrossRef] [PubMed]

11. Achkar, N.P.; Cambiagno, D.A.; Manavella, P.A. miRNA Biogenesis: A Dynamic Pathway. *Trends Plant Sci.* **2016**, *21*, 1034–1044. [CrossRef] [PubMed]
12. Cho, S.K.; Ryu, M.Y.; Shah, P.; Poulsen, C.P.; Yang, S.W. Post-Translational Regulation of miRNA Pathway Components, AGO1 and HYL1, in Plants. *Mol. Cells* **2016**, *39*, 581–586. [CrossRef] [PubMed]
13. Bartel, D.P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **2004**, *116*, 281–297. [CrossRef]
14. Chávez Montes, R.A.; de Fátima Rosas-Cárdenas, F.; De Paoli, E.; Accerbi, M.; Rymarquis, L.A.; Mahalingam, G.; Marsch-Martínez, N.; Meyers, B.C.; Green, P.J.; de Folter, S. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat. Commun.* **2014**, *5*, 3722. [CrossRef] [PubMed]
15. Tarver, J.E.; Donoghue, P.C.; Peterson, K.J. Do miRNAs have a deep evolutionary history? *BioEssays* **2012**, *34*, 857–866. [CrossRef] [PubMed]
16. Drusin, S.I.; Suarez, I.P.; Gauto, D.F.; Rasia, R.M.; Moreno, D.M. dsRNA-protein interactions studied by molecular dynamics techniques. Unravelling dsRNA recognition by DCL1. *Arch. Biochem. Biophys.* **2016**, *15*, 118–125. [CrossRef] [PubMed]
17. Dolata, J.; Bajczyk, M.; Bielewicz, D.; Niedojadlo, K.; Niedojadlo, J.; Pietrykowska, H.; Walczak, W.; Szweykowska-Kulinska, Z.; Jarmolowski, A. Salt stress Reveals a New Role for ARGONAUTE1 in miRNA Biogenesis at the Transcriptional and Posttranscriptional Levels. *Plant Physiol.* **2016**, *172*, 297–312. [CrossRef] [PubMed]
18. Conrad, T.; Orom, U.A. Insight into miRNA biogenesis with RNA sequencing. *Oncotarget* **2015**, *6*, 26546–26547. [CrossRef] [PubMed]
19. Zhu, H.; Zhou, Y.; Castillo-González, C.; Lu, A.; Ge, C.; Zhao, Y.T.; Duan, L.; Li, Z.; Axtell, M.J.; Wang, X.J.; et al. Bidirectional processing of pri-miRNAs with branched terminal loops by Arabidopsis Dicer-like1. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1106–1115. [CrossRef] [PubMed]
20. Starega-Roslan, J.; Krol, J.; Koscianska, E.; Kozlowski, P.; Szlachcic, W.J.; Sobczak, K.; Krzyzosiak, W.J. Structural basis of microRNA length variety. *Nucleic Acids Res.* **2011**, *39*, 257–268. [CrossRef] [PubMed]
21. Voinnet, O. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell* **2009**, *136*, 669–687. [CrossRef] [PubMed]
22. Mickiewicz, A.; Sarzynska, J.; Milostan, M.; Kurzynska-Kokorniak, A.; Rybarczyk, A.; Lukasiak, P.; Kulinski, T.; Figlerowicz, M.; Blazewicz, J. Modeling of the catalytic core of Arabidopsis thaliana Dicer-like 4 protein and its complex with double-stranded RNA. *Comput. Biol. Chem.* **2017**, *66*, 44–56. [CrossRef] [PubMed]
23. Beezhold, K.J.; Castranova, V.; Chen, F. Microprocessor of microRNAs: Regulation and potential for therapeutic intervention. *Mol. Cancer* **2010**, *9*, 1–9. [CrossRef] [PubMed]
24. Søkilde, R.; Newie, I.; Persson, H.; Borg, Å.; Rovira, C. Passenger strand loading in overexpression experiments using microRNA mimics. *RNA Biol.* **2015**, *12*, 787–791.
25. Zha, X.; Xia, Q.; Yuan, Y.A. Structural insights into small RNA sorting and mRNA target binding by Arabidopsis Argonaute Mid domains. *FEBS Lett.* **2012**, *586*, 3200–3207. [CrossRef] [PubMed]
26. Starega-Roslan, J.; Galka-Marciniak, P.; Krzyzosiak, W.J. Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer. *Nucleic Acids Res.* **2015**, *43*, 10939–10951. [CrossRef] [PubMed]
27. Flores-Jasso, C.F.; Arenas-Huertero, C.; Reyes, J.L.; Contreras-Cubas, C.; Covarrubias, A.; Vaca, L. First step in pre-miRNAs processing by human Dicer. *Acta Pharmacol. Sin.* **2009**, *30*, 1177–1185. [CrossRef] [PubMed]
28. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190. [CrossRef] [PubMed]
29. Nucleotide Ambiguity Code (IUPAC). Available online: http://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html (accessed on 1 February 2017).
30. Zakov, S.; Goldberg, Y.; Elhadad, M.; Ziv-Ukelson, M. Rich parameterization improves RNA structure prediction. *J. Comput. Biol.* **2011**, *18*, 1525–1542. [CrossRef] [PubMed]
31. Puton, T.; Kozlowski, L.P.; Rother, K.M.; Bujnicki, J.M. CompaRNA: A server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.* **2013**, *41*, 4307–4323. [CrossRef] [PubMed]
32. Antczak, M.; Zok, T.; Popenda, M.; Lukasiak, P.; Adamiak, R.W.; Blazewicz, J.; Szachniuk, M. RNApdbee—A webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Res.* **2014**, *42*, W368–W372. [CrossRef] [PubMed]

33. Rybarczyk, A.; Szostak, N.; Antczak, M.; Zok, T.; Popenda, M.; Adamiak, R.W.; Blazewicz, J.; Szachniuk, M. New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinform.* **2015**, *16*, 276. [CrossRef] [PubMed]
34. Zok, T.; Antczak, M.; Zurkowski, M.; Popenda, M.; Blazewicz, J.; Adamiak, R.W.; Szachniuk, M. RNApdbee 2.0: Multifunctional tool for RNA structure annotation. *Nucleic Acids Res.* **2018**, *46*. [CrossRef] [PubMed]
35. Wiedemann, J.; Milostan, M. StructAnalyzer—A tool for sequence versus structure similarity analysis. *Acta Biochim. Pol.* **2016**, *63*, 753–757. [CrossRef] [PubMed]
36. Wiedemann, J.; Zok, T.; Milostan, M.; Szachniuk, M. LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinform.* **2017**, *18*, 456. [CrossRef] [PubMed]
37. Blazewicz, J.; Szachniuk, M.; Wojtowicz, A. RNA tertiary structure determination: NOE pathways construction by tabu search. *Bioinformatics* **2005**, *21*, 2356–2361. [CrossRef] [PubMed]
38. Antczak, M.; Popenda, M.; Zok, T.; Sarzynska, J.; Ratajczak, T.; Tomczyk, K.; Adamiak, R.W.; Szachniuk, M. New functionality of RNAComposer: An application to shape the axis of miR160 precursor structure. *Acta Biochim. Pol.* **2016**, *63*, 737–744. [CrossRef] [PubMed]
39. Purzycka, K.J.; Popenda, M.; Szachniuk, M.; Antczak, M.; Lukasiak, P.; Blazewicz, J.; Adamiak, R.W. Automated 3D RNA structure prediction using the RNAComposer method for riboswitches. *Methods Enzymol.* **2015**, *553*, 3–34. [PubMed]
40. *The PyMOL Molecular Graphics System*, version 1.8; Schrodinger, LLC: New York, NY, USA, 2015.
41. Bottaro, S.; Palma, F.D.; Bussi, G. The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.* **2014**, *42*, 13306–13314. [CrossRef] [PubMed]
42. Kufareva, I.; Abagya, R. Methods of protein structure comparison. *Methods Mol. Biol.* **2012**, *857*, 231–257. [PubMed]
43. Almeidaa, M.I.; Reisb, R.M.; Calin, G.A. MicroRNA history: Discovery, recent applications, and next frontiers. *Mutat. Res.* **2011**, *717*, 1–8. [CrossRef] [PubMed]
44. Varani, G. Twenty years of RNA: The discovery of microRNAs. *RNA* **2015**, *21*, 751–752. [CrossRef] [PubMed]
45. Context Fold 1.00. Available online: https://www.cs.bgu.ac.il/~negevcb/contextfold/readme.pdf (accessed on 1 February 2017).
46. eRMSD. Available online: https://plumed.github.io/doc-master/user-doc/html/_e_r_m_s_d.html (accessed on 1 February 2017).

Sample Availability: Samples of the compounds are not available from the authors.

SUPPLEMENTARY MATERIAL

Phylum	Clade	Family	Species	Sequences
Chlorophyta			*Chlamydomonas reinhardtii (cre)*	50
Coniferophyta			*Cunninghamia lanceolata (cln)*	4
			Picea abies (pab)	40
			Pinus densata (pde)	29
			Pinus taeda (pta)	35
Embryophyta			*Physcomitrella patens (ppt)*	229
			Selaginella moellendorffii (smo)	58
Magnoliophyta			*Amborella trichopoda (atr)*	124
	eudicotyledons	*Araliaceae*	*Panax ginseng (pgi)*	29
		Asteraceae	*Cynara cardunculus (cca)*	48
			Helianthus annuus (han)	6
			Helianthus argophyllus (har)	3
			Helianthus ciliaris (hci)	3
			Helianthus exilis (hex)	2
			Helianthus paradoxus (hpa)	3
			Helianthus petiolaris (hpe)	3
			Helianthus tuberosus (htu)	16
		Brassicaceae	*Arabidopsis lyrata (aly)*	205
			Arabidopsis thaliana (ath)	325
			Brassica napus (bna)	90
			Brassica oleracea (bol)	10
			Brassica rapa (bra)	96
		Caricaceae	*Carica papaya (cpa)*	79
		Cucurbitaceae	*Cucumis melo (cme)*	120
		Euphorbiaceae	*Hevea brasiliensis (hbr)*	31
			Manihot esculenta (mes)	153
			Ricinus communis (rco)	63
		Fabaceae	*Acacia auriculiformis (aau)*	7
			Acacia mangium (amg)	3
			Arachis hypogaea (ahy)	23
			Glycine max (gma)	573
			Glycine soja (gso)	13
			Lotus japonicus (lja)	62
			Medicago truncatula (mtr)	672
			Phaseolus vulgaris (pvu)	8
			Vigna unguiculata (vn)	18
		Lamiales	*Avicennia marina (ama)*	2
			Digitalis purpurea (dpr)	13
			Rehmannia glutinosa (rgl)	32
			Salvia sclarea (ssl)	18
		Linaceae	*Linum usitatissimum (lus)*	124
		Malvaceae	*Gossypium arboreum (gar)*	1

			Gossypium herbaceum (ghb)	1
			Gossypium hirsutum (ghr)	78
			Gossypium raimondii (gra)	296
			Theobroma cacao (tcc)	82
		Ranunculaceae	*Aquilegia caerulea (agc)*	45
		Rhizophoraceae	*Bruguiera cylindrica (bcy)*	4
			Bruguiera gymnorhiza (bgy)	4
		Rosaceae	*Malus domestica (mdm)*	206
			Prunus persica (ppe)	180
		Rutaceae	*Citrus clementina (ccl)*	5
			Citrus reticulata (crt)	4
			Citrus sinensis (csi)	60
			Citrus trifoliata (ctr)	6
		Salicaceae	*Populus euphratica (peu)*	4
			Populus trichocarpa (ptc)	352
		Solanaceae	*Nicotiana tabacum (nta)*	162
			Solanum lycopersicum (sly)	77
			Solanum tuberosum (stu)	224
		Vitaceae	*Vitis vinifera (vvi)*	163
	monocotyledons		*Aegilops tauschii (ata)*	88
			Brachypodium distachyon (bdi)	317
			Elaeis guineensis (egu)	6
			Festuca arundinacea (far)	15
			Hordeum vulgare (hvu)	69
			Oryza sativa (osa)	592
			Saccharum officinarum (sof)	16
			Saccharum sp. (ssp)	19
			Sorghum bicolor (sbi)	205
			Triticum aestivum (tae)	116
			Triticum turgidum (ttu)	1
			Zea mays (zma)	172

Table 1S. Number of sequences extracted from miRBase website [3] distributed by phylum, clade, family and species.

applied sciences

MDPI

Article

In Vitro and in Silico Analysis of miR-125a with rs12976445 Polymorphism in Breast Cancer Patients

Tomasz P. Lehmann [1,*], Joanna Miskiewicz [2], Natalia Szostak [3], Marta Szachniuk [2,3,*], Sylwia Grodecka-Gazdecka [4] and Paweł P. Jagodziński [1]

[1] Department of Biochemistry and Molecular Biology, Poznan University of Medical Sciences, Swiecickiego 6, 60-781 Poznan, Poland; pjagodzi@ump.edu.pl
[2] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland; jmiskiewicz@cs.put.poznan.pl
[3] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland; nszostak@ibch.poznan.pl
[4] Department of Surgery, Chair and Clinic of Oncology Poznan University of Medical Sciences, Szamarzewskiego 82/84, 61-001 Poznan, Poland; sylwia.grodecka-gazdecka@oncology.am.poznan.pl
* Correspondence: tlehmann@ump.edu.pl (T.P.L.); mszachniuk@cs.put.poznan.pl (M.S.); Tel.: +48-618-546-513 (T.P.L.); +48-616-653-030 (M.S.); Fax: +48-854-6510 (T.P.L.); +48-618-771-525 (M.S.)

Received: 23 September 2020; Accepted: 15 October 2020; Published: 17 October 2020

Abstract: *Background*: Breast cancer affects over 2 million women yearly. Its early detection allows for successful treatment, which motivates to research factors that enable an accurate diagnosis. miR-125a is one of them, correlating with different types of cancer. For example, the miR-125a level decreases in breast cancer tissues; polymorphisms in the miR-125a encoding gene are related to prostate cancer and the risk of radiotherapy-induced pneumonitis. *Methods*: In this work, we investigated two variants of rs12976445 polymorphism in the context of breast cancer. We analyzed the data of 175 blood samples from breast cancer patients and compared them with the control data from 129 control samples. *Results*: We observed the tendency that in breast cancer cases TT genotype appeared slightly more frequent over CC and CT genotypes (statistically nonsignificant). The TT genotype appeared also to be more frequent among human epidermal growth factor receptor 2 (HER2) positive patients, compared to HER2 negative. In silico modelling showed that the presence of uridine (U) diminished the probability of pri-miR-125a binding to NOVA1 and HNRNPK proteins. We demonstrated that U and C -variants could promote different RNA folding patterns and provoke alternative protein binding. *Conclusions*: U-variant may imply a lower miR-125a expression in breast cancer.

Keywords: microRNA; RNA protein binding; RNA folding; breast cancer; polymorphism

1. Introduction

Constant investigation of the genetic background of breast cancer is a crucial endeavour since only 15–36% of hereditary breast cancers have known genetic background [1]. Many current studies refer to estimate the association of polymorphisms in genes encoding miRNA with cancer and to find how single nucleotide polymorphisms (SNP) modulate miRNA expression [2]. SNPs in genes encoding miRNAs have been revealed in several different types of cancer, including breast cancer [3–5]. It has been proven that SNPs in miRNA genes affect their expression in breast cancer and modulate the expression of miRNA target genes (e.g., miR-27a, miR-196a2, miR-559) [6–8]. SNPs contribute to abnormal expression of miRNAs in breast cancers either as upregulated oncomirs or downregulated suppressors [9]. Therefore, understanding of SNPs impact on miRNA levels might be helpful in the selection of breast cancer diagnostic markers.

One of such potentially useful miRNA marker is miR-125a-the gene is located on chromosome 19, in the proximity of miR-99b and let-7e. This miRNA cluster is situated within the first intron of the transcription unit of sperm acrosomal membrane-associated protein 6 (SPACA6) variant 1 [10]. The genomic DNA sequence located between -3875 bp and -3006 bp from pre-miR-99b drives the expression of SPACA6 gene and its intronic miR-99b/let-7e/miR-125a cluster [10]. miR-125a is involved in the development of various types of cancer, including lung, gastric, and breast cancers [6,11–13]. miR-125a functions as an oncomir or a tumour suppressor miRNA depending on the cellular context [6,14]. In breast cancer tissues, the miR-125a level is decreased when comparing to normal adjacent tissue. Several SNPs have been already described in pri-miR-125a and mature miR-125a: rs41275794 (A/G) and rs12976445 (C/T) SNPs in the pri-miR-125a; rs12975333 (G/T) SNP on the 5′ strand and -rs143525573 (A/G) SNP on the 3′ strand of mature miR-125a [15–17]. rs12976445 C and T haplotypes are correlated with cancerous and noncancerous diseases [18,19]. One of the SNPs in miR-125a, rs12976445, is considerably associated with lymph node metastasis, tumour stage (Classification of Malignant Tumors, TNM), estrogen receptor status, and progesterone receptor status [19]. The TT genotype of rs12976445 significantly increases the risk of mortality in breast cancer patients compared with those carrying the CC genotypes of the SNP [19]. Moreover, patients carrying the TT or CT genotype ofrs12976445 have a higher risk of radiotherapy-induced pneumonitis [20,21]. The CC-variant of the mentioned SNP has been shown to correlate with higher expression compared with C/T and T/T [20]. Homozygous CC-variant of rs12976445 was shown to increase the risk of prostate cancer [22]. In a study of autoimmune thyroid diseases, it was found that the C allele is significantly increased in Hashimoto's thyroiditis patients [18]. In the same study, a protecting role of the T allele was suggested to decrease the risk for Hashimoto's thyroiditis for about 31.6% [18].

Although it has been suggested that SNPs have an impact on miRNA expression and are associated with various diseases, the precise mechanism of the rs12976445 effect on miR-125a processing remains unknown. In vitro and in silico analysis of the SNP and the recognition of proteins interacting with rs12976445 SNP can reveal a detailed picture of the underlying mechanism. In this research, we analyzed the frequency of rs12976445 alleles and genotypes in breast cancer in Polish patients and estimated the genotype frequency with the receptor status. Following the methodology from our previous studies [23,24], we used in silico analysis and computational tools to generate and compare the secondary structure of miRNA of two alternative variants, C and U. Using in silico approach, we found potential proteins interacting with sequence variants of rs12976445 pre-miRNA.

2. Materials and Methods

2.1. Patients and Samples

27 Polish patients (women) were undergoing surgery for breast cancer in the Department of Surgery, Chair of Oncology of Poznan University of Medical Sciences (PUMS). The study protocol was approved by the bioethics board of PUMS (number 839/09 8 October 2009, 690/10 2 September 2010). Blood samples for experiments were collected from participants with their consent.

2.2. Blood Sampling and Measurements

A total 175 blood samples were obtained by antecubital vein puncture. Further, 129 control blood samples were collected from patients with the negative diagnoses of breast cancer, and not diagnosed with other types of cancer. All 304 samples were collected from 14–88 years old patients of the Department of Surgery, Chair of Oncology at PUMS.

2.3. DNA Extraction, PCR, Restriction Analysis

DNA from blood was extracted using GenElute™ Mammalian Genomic DNA Miniprep kit (nr G1N70 Sigma-Aldrich, St. Luis, MO, USA). The quantity of obtained nucleic acid was assessed using a BioPhotometer™ (Eppendorf, Hamburg, Germany).

2.4. Polymerase Chain Reaction, Sequencing, and Restriction Analysis

DNA specimens were amplified using the standard PCR protocols. DreamTaq DNA Polymerase 0.02 U/µl (EP0701, Thermo Fisher Scientific, Waltham, MA, USA), dNTP Mix 1 mM (U151A, Promega, Madison, WI, USA), primers 0.5 µM were applied. The temperature profile was as follows: (1) the preheating phase, 95 °C for 10 min, (2) the amplification cycle of 35 repeats, 95 °C for 30 s and 55 °C for 30 s, and finishing the cycle with 72 °C for 30 s, (3) the final elongation process, 72 °C for 7 min. PCR primers: 5′-TTTTGGTCTTTCTGTCTCTGG-3′ and 5′-TGGAGGAAGGGTATGAGGAGT-3′, which were spanning the sequence of pre-miR-125a (accession number ENSG00000208008) were designed using the Oligo software (Molecular Biology Insights, Inc., DBA Oligo, Inc., Colorado Springs, CO, USA). This amplicon was used for the restriction analysis and MIR125A sequencing. PCR products of 27 samples collected from patients were purified using the gel-out purification kit (A&A Biotechnology, Gdynia, Poland) and sequenced at the DNA Sequencing and Oligonucleotide Synthesis Laboratory of the Institute of Biochemistry and Biophysics of the Polish Academy of Sciences (Warsaw, Poland). The sequencing results were analyzed using BioEdit Sequence Alignment Editor [25]. The rs12976445 SNP was genotyped using the BaeGI restriction enzyme (R0708S, New England Biolabs, Ipswich, MA, USA). The genotyping required 17.8 µL of the PCR product and 2 U of a restriction enzyme that were kept in 37 °C for 12 h. BaeGI cuts GKGCM|C K = G or T, M = A or C (T-variant TGTGCCTA, C-variant TGTGCCCA). Upon preliminary sequencing samples of 27 patients, we supposed that the uncut sequence would contain T-variant, although G- and A-variant could not be excluded [26]. The enzyme HpyF3I (DdeI) (ER 1881, Thermo Scientific, Waltham, MA, USA), which cuts C|TNAG, was used to analyze rs143525573 and rs12975333 SNP in the miR-125a amplicon. The applied procedure was analogous to the genotyping with the BaeGI restriction enzyme.

2.5. Preparing Sequence for in Silico Analysis

From the amplicon comprising pre-miR-125a sequence, we selected a fragment which consisted of the SNP nucleotide, the 25 nt upstream region of SNP, and the 25 nt downstream region of SNP. This subsequence was named as 51-miR-125a and it was analyzed in two variants, C and U, of the genotype.

2.6. Predicting Protein Interactions

RBPmap web server [27] and ATtRACT database [28] were used to find possible protein interactions within 51-miR-125a, for both U- and C-variants. RBPmap was designed in 2014 as a tool for finding potential binding sites of RNA-interacting proteins, especially in human, mouse, and *Drosophila melanogaster* genomes. Searching for potential binding sites of RNA-interacting proteins was performed using RBPmap server with three different stringency levels. ATtRACT is a database with an implemented search engine that applies a fast algorithm dedicated to finding motifs corresponding to RNA-interacting proteins. The database contains 370 RBPs and 1583 RBP consensus binding motifs that can be searched throughout different organisms (including humans).

2.7. 2D Structure Modelling

The secondary structure of 51 nt long RNA fragment of pri-miR-125a sequence (51-miR-125a) was predicted using bioinformatictools -RNAstructure [29] and RNAfold programs [30]. RNAfold is a part of the Vienna RNA package that collects tools specialised for analysing single-stranded nucleic acid sequences. RNAfold generates 2D models optimizing either minimum free energy (MFE mode) or minimal base-pair distance (centroid mode) parameters. RNAstructure is the second program that we used for predicting 2D structures. It can handle both RNA and DNA sequences and allows users to predict biomolecular secondary structures. Based on a given sequence, RNAstructure generates the lowest (minimum) free energy 2D structure model (MFE mode) and a 2D structure composed of highly probable base pairs (MaxExpect mode). Predicted models in this research for both RNAfold and RNAstructure were obtained using default settings.

2.8. Statistical Analysis

The expected genotype and allele frequencies for the observed variations were calculated for all 175 positive cases (Ca) and 129 control samples (Co). These frequencies were tested for Hardy-Weinberg equilibrium. Statistical analysis of genotypes frequency was conducted using the online tool SNPstats [31]. The association of rs12976445 in miR-125a in breast cancer cells was calculated using odds ratio (OR) at a 95% confidence interval (CI). The association of the genotypes with the receptor status was calculated using χ square test and the software GraphPadInStat version 7.00 (GraphPad Software, San Diego, California, USA).

3. Results

3.1. In Vitro Analysis

rs12976445 SNP in pri-miR-125a occurring in three genotypes TT, CT, and CC has been associated with cancer and other diseases [32–34]. We studied the pri-miR-125a amplicon (Figure 1) by sequencing and restriction analysis to determine the genotype frequency and its association with breast cancer.

Version U
uuuuggucuuucugucu***cuggcucucagaaugucucugugccuaucuccaucucugacccccc***
accccagggucuaccgggccaccgcacaccauguUGCCAGUCUCUAGG**UCCCUGAGACCCUU**
UAACCUGUGAGGACAUCCAGGGUC**ACAGGUGAGGUUCUUGGGAG**CCUGGCGUCUGGCCcaac
cacacaccuggggaauugcuggccugacuucugaccccugacuccucauacccuuccucca

Version C
uuuuggucuuucugucu***cuggcucucagaaugucucugugccCaucuccaucucugacccccc***
accccagggucuaccgggccaccgcacaccauguUGCCAGUCUCUAGG**UCCCUGAGACCCUU**
UAACCUGUGAGGACAUCCAGGGUC**ACAGGUGAGGUUCUUGGGAG**CCUGGCGUCUGGCCcaac
cacacaccuggggaauugcuggccugacuucugaccccugacuccucauacccuuccucca

Figure 1. Whole amplicon (247 nt) comprising pre-miR-125a sequence, the fragment of the sequence from Ensembl Gene ID: ENSG00000208008. Lower-case corresponds to pri-miR-125a fragments removed in pre-miR-125a processing; grey upper-case is pre-miR-125a; in bold upper-case we marked mature miR-125a 5p and 3p, respectively; bold underlined lower-case nucleotides are the analyzed U-, C-variants; bold italic lower-case is a 51 nt fragment (51-miR-125a) that was analyzed using computational methods. BaeGI restriction site is located directly before rs12976445 SNP, dividing sequence into 42 nt and 205 nt subsequences.

We amplified DNA obtained from 27 breast cancer tumours to assess the frequency of several known SNPs in pri-miR-125a sequence. The sequencing of 27 pri-miR-125a amplicons showed that only nrs12976445 polymorphic site was highly variable in our samples. Subsequently, we estimated the association of rs12976445 with breast cancer using DNA from 304 blood samples, 175 positive (cases, Ca) and 129 negative (controls, Co) samples. We digested the pri-miR-125a amplicon using BeaGI restriction enzyme producing 42 and 205 fragments if C or non-digesting if T variant was present. We performed statistical analysis with SNPStats on-line software. The allele and genotype frequencies analysis are presented in Table 1. We found that rs12976445 in pri-miR-125a followed Hardy-Weinberg equilibrium in breast cancer cases (Ca) set (Table 1).

Table 1. The allele, genotype frequencies, the exact test for Hardy–Weinberg equilibrium analysis, and the single nucleotide polymorphism (SNP) association with breast cancer status of rs12976445 miR-125a gene polymorphism. Ca—cases, Co—controls, OR—odds ratio, CI—confidence interval.

SNP Allele Frequencies (n = 304)						
	All subjects		**Ca**		**Co**	
Allele	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**
T	414	0.68	241	0.69	173	0.67
C	194	0.32	109	0.31	85	0.33
SNP Genotype Frequencies (n = 304)						
	All subjects		**Ca**		**Co**	
Genotype	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**
C/C	24	0.08	14	0.08	10	0.08
T/C	146	0.48	81	0.46	65	0.5
T/T	134	0.44	80	0.46	54	0.42
SNP Exact Test for Hardy–Weinberg Equilibrium (n = 304)						
	TT	**TC**	**CC**	**T**	**C**	***p*-value**
All subjects	134	146	24	414	194	0.086
Ca	80	81	14	241	109	0.38
Co	54	65	10	173	85	0.16
SNP Association with Response STATUS (n = 304, Crude Analysis)						
Model	**Genotype**	**Ca**	**Co**	**OR (95% CI)**		***p*-value**
Codominant	T/T	80 (45.7%)	54 (41.9%)	1.00		0.77
	C/T	81 (46.3%)	65 (50.4%)	1.19 (0.74–1.91)		
	C/C	14 (8%)	10 (7.8%)	1.06 (0.44–2.56)		
Dominant	T/T	80 (45.7%)	54 (41.9%)	1.00		0.5
	C/T+ C/C	95 (54.3%)	75 (58.1%)	1.17 (0.74–1.85)		
Recessive	T/T+ C/T	161 (92%)	119 (92.2%)	1.00		0.94
	C/C	14 (8%)	10 (7.8%)	0.97 (0.41–2.25)		
Over-dominant	T/T+ C/C	94 (53.7%)	64 (49.6%)	1.00		0.48
	C/T	81 (46.3%)	65 (50.4%)	1.18 (0.75–1.86)		
Log-additive	n/a	n/a	n/a	1.10 (0.76–1.58)		0.62

The statistical analysis, comprising odds ratio (OR) at 95% confidence interval (CI), are presented in Table 1. In the co-dominant model (TT vs. TC vs. CC), the heterozygous CT genotype of rs12976445 SNP was slightly more frequent in control (Co) set (50.4%) compared to cases (Ca) (46.3%) with OR Co/Ca = 1.19, 95%, CI = 0.74–1.91 and p = 0.77 (Table 1). In the dominant model (TT vs. TC + CC), we observed slightly higher level of CT + CC in controls over cases with OR = 1.17, 95% (CI = 0.74–1.85) and p = 0.5. These calculations indicated the tendency that in breast cancer cases, TT genotype appeared slightly more frequent than CC and CT genotypes.

Additionally, we analyzed two other known SNPs, rs143525573 and rs12975333, located in the mature miR-125a coding sequence. The DdeI enzyme was applied to assess these two SNPs. We did not observe the variability of rs143525573 and rs12975333 in the studied group of patients with breast cancer.

Since SNP in pri-miR-125a could modulate the level of mature miRNA, we assessed the relationship of rs12976445 variability with the status of human epidermal growth factor receptor 2 (HER2), ER1α and PR receptors. mRNA encoding receptor HER2 (human epidermal growth factor receptor 2, receptor tyrosine-protein kinase erbB-2, encoded by ERBB2 gene) is a target of miR-125a [33]. HER2 together with ER1α and PR receptors are used in the diagnosis of breast cancer. Therefore, we decided to perform tests of the relationship of rs12976445 variation with the status of these receptors. HER2 receptor status 0, 1, and 2 without amplification was classified as the negative group (N), while status

2 with amplification and status 3 were ranked as the positive group (P). Patients were also divided into two groups according to ER1α or PR status, 0–10% as the negative group and 10–75% as the positive group. We analyzed data using χ square test. The lowest *p*-value, 0.0606, was obtained for TT, CT, CC genotypes analysis for HER2 receptor. Our analysis revealed a tendency that T allele predominated in HER2 positive samples (Table 2). These results are concomitant with genotype frequency, which TT-variant slightly predominates in cases over controls. To determine how C- and T(U)-variants could regulate miR-125a expression, we decided to analyze the 2D structure of pre-miRNA usingin silicomethods.

Table 2. The association analysis between miR-125a rs12976445 polymorphism and HER2, ER and PR receptors status in breast cancer patients. The association was calculated using χ square test. The status of receptors was divided into two groups-positive (P) and negative (N).

Receptor	Case	Variant	Group Status		*p*-Value
			P	N	
HER2	Genotype	TT	14	30	0.0606
		CT	7	48	
		CC	1	6	
		TT+CT	21	78	0.2609
		CT+CC	8	54	
	Alleles frequency	T	35	108	0.0814
		C	9	60	
ER	Genotype	TT	36	8	0.2001
		CT	37	17	
		CC	6	1	
		TT+CT	73	25	0.2891
		CT+CC	33	18	
	Alleles frequency	T	109	33	0.5702
		C	49	19	
PR	Genotype	TT	30	14	0.4922
		CT	35	20	
		CC	6	1	
		TT+CT	65	34	1.0000
		CT+CC	41	21	
	Alleles frequency	T	95	48	0.9297
		C	47	22	

3.2. In Silico Analysis of RNA Binding Proteins (RBPs)

To find potential binding proteins to the pri-miR-125a containing C- and U-variants, we analyzed the 51-nucleotide fragment of pri-miR-125a using RBPmap web server and ATtRACT database [27,28]. Both, the C- and U-variants were analyzed using RBPmap with three different stringency levels: low, medium, and high. Low stringency level corresponds to significant *p*-value < 0.01 and suboptimal *p*-value < 0.02; medium stringency level thresholds were at significant *p*-value < 0.005 and suboptimal *p*-value < 0.01; high stringency level was set to <0.001 and <0.01 for significant *p*-value and suboptimal *p*-value, respectively. We searched for any available human motif stored in RBPmap. Low stringency algorithm mapped SRSF5, PTBP1, and PCBP1 proteins in both C-/U-variants. BRUNOL4 and BRUNOL5 proteins were only mapped in U-variant, whereas SRSF3, HNRNPK, and NOVA1 proteins were only

found in C-variant. In medium and high stringency levels, RBPmap found only PTBP1 protein motif in U-variant. For the C-variant, medium stringency algorithm found following potentially binding proteins: SRSF3, HNRNPK, NOVA1, and PTBP1. High stringency level for C-variant found the same RBPs as the medium algorithm, excluding HNRNPK protein. All RBPmap-predicted proteins that potentially bind to the region containing C-/U-variant in 51-miR-125a sequence are shown in Table 3.

Table 3. Motif sequences of possible protein-binding sites in C- variant and U-variant found by RBPmap tool. SNPs are marked as bold.

Variant	Protein	Mode			Motif Containing Variant
		Low	Medium	High	
C	HNRNPK	✓	✓		C**C**AUCUC
	NOVA1	✓	✓	✓	C**C**AU
	PCBP1	✓			**C**AUCUCC
	PTBP1	✓	✓	✓	C**C**AUCU
	SRSF3	✓	✓	✓	CC**C**AUCU
	SRSF5	✓			**C**AUCUCC
U	BRUNOL4	✓			UGUGCC**U**
	BRUNOL5	✓			UGUGCC**U**
	PCBP1	✓			CC**U**AUCU
	PTBP1	✓	✓	✓	C**U**AUCU
	SRSF5	✓			CC**U**AUCU, **U**AUCUCC

We focused on proteins that were variant-specific. Less frequent C-variant interacts potentially with HNRNPK and NOVA1—the proteins involved in the RNA processing [35,36]. The C-variant was also mapped with SRSF3 protein, which was not found by RBPmap in U-variant on any stringency level of the algorithm execution and appeared on every stringency level of C-variant. On the contrary, BRUNOL4 and BRUNOL5, proteins that regulate alternative splicing of pre-mRNA and are possibly connected with mRNA editing and translation, were only mapped in U-variant of the analyzed sequence.

In the next step, we used ATtRACT database to search for potential RBP motifs in 51-miR-125a sequence. In the U-variant, ATtRACT predicted PTBP1 motif corresponding to the results from RBPmap, where PTBP1 was mapped in all three stringency levels of the algorithm. In the C-variant, ATtRACT predicted YBX1 and NOVA1 binding motifs. YBX1 protein is involved in cellular processes, including pre-mRNA splicing, transcriptional and translational regulation. It is also potentially involved in miRNA processing [37]. What brings particular interest, is the mapping of the NOVA1 protein in the C-variant. It complies to the results from RBPmap presented earlier. Similarly, as PTBP1, RBPmap predicted NOVA1 binding to 51-miR-125a for all stringency levels (low, medium, high). The results suggest that C-variant in 51-miR-125a is connected with different RBP (PTBP1) than the U-variant (NOVA1). Binding motifs for C-/U-variants from ATtRACT database are presented in Table 4.

Table 4. Motif sequences of possible protein-binding sites in C-variant and U-variant found in ATtRACT database. SNPs are marked as bold.

Variant	Protein	Motif Containing Variant
C	NOVA1	C**C**AU
	YBX1	**C**AUC
U	PTBP1	C**U**AU, CC**U**AU

3.3. In Silico Modelling of pri-miR-125a Folding

Computational modelling of the 2D structure of 51-miR-125a was performed using RNAfold (Figure 2) and RNAstructure (Figure 3) [29,30].

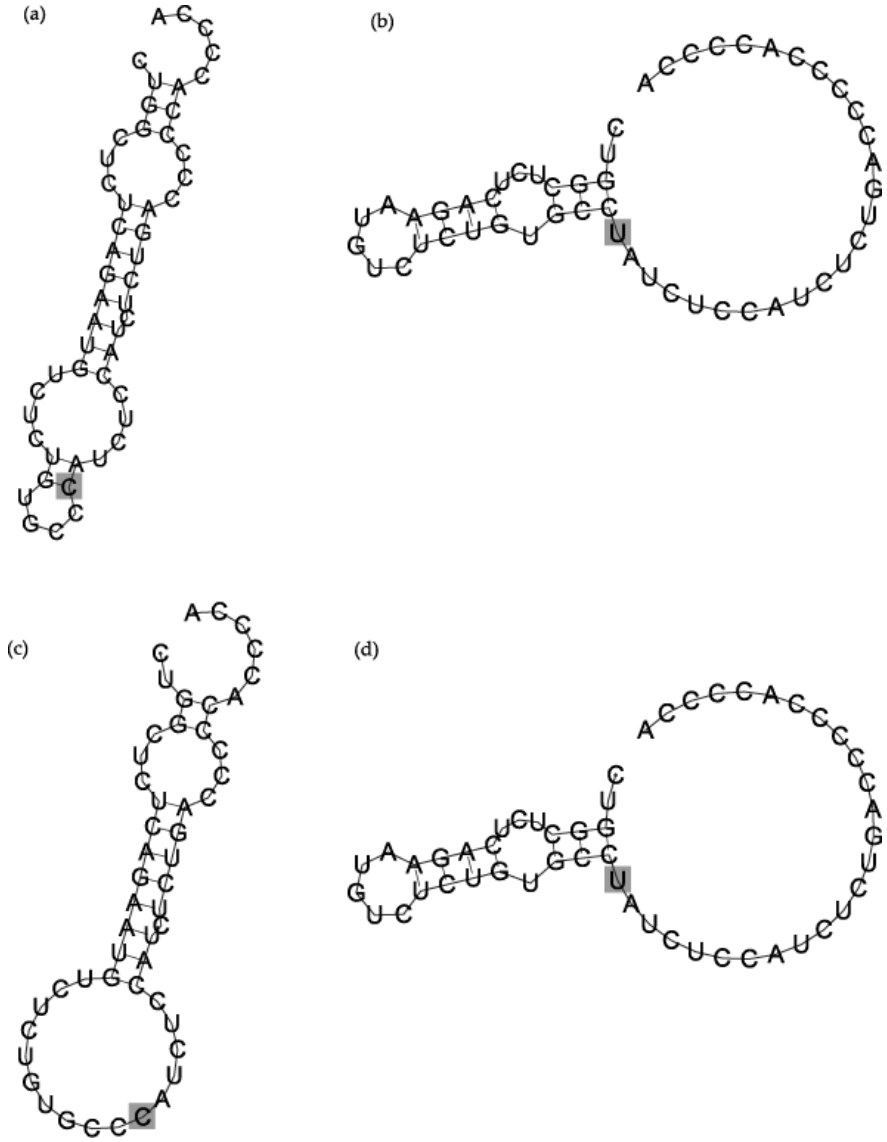


Figure 2. The 2D structures of 51-miR-125a with C-variant and U-variant predicted by RNAfold. In grey squares we marked SNPs. (**a**) C-variant MFE model (minimizing free energy), (**b**) U-variant MFE model (minimizing free energy) (**c**) C-variant variant centroid model (minimizing base-pair distance) and (**d**) U-variant centroid model (minimizing base-pair distance).

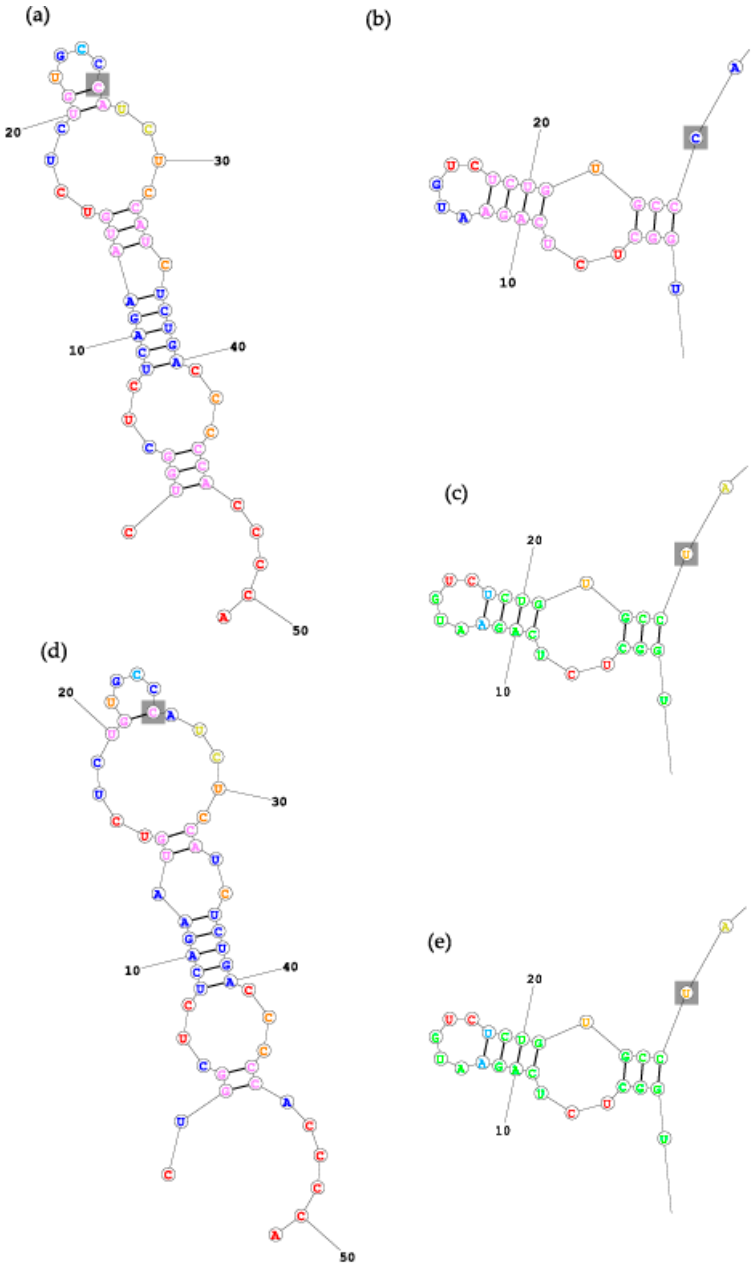


Figure 3. The 2D structures of 51-miR-125a with C-variant (**a**,**b**,**d**) and U-variant (**c**,**e**) predicted by RNAstructure. In grey squares we marked SNPs. (**a**) C-variant MFE model (minimizing free energy), (**b**) C-variant MFE model, and (**c**) U-variant MFE model, (**d**) C-variant MaxExpect model (most probable base-pairing) (**e**) U-variant MaxExpect model. (**b**,**c**,**e**) models were cut for better visibility of SNP fragment.

Predicted models revealed variant-dependent folding of the 51 nucleotides long pri-miR- 125a region comprising rs12976445. The modelling of C-variant 51-miR-125a with RNAfold resulted in two different structures when we applied two modes of analysis: minimizing free energy (MFE) and minimizing base-pair distance (centroid). For the U-variant using MFE and centroid modes, we obtained the same model of RNA structure. In the RNAfold MFE model of the C-variant (Figure 2a), SNP created a base-pairing near the hairpin loop. In the RNAfold centroid model of the C-variant structure, SNP was unpaired within the hairpin loop. In both RNAfold modes MFE and centroid from U-variant, SNP was the first unpaired nucleotide of a long unpaired subsequence located on 3' end of 51-miR-125a.

Predicting the 2D structure of 51-miR-125a using RNAstructure resulted in two MFE structures for C-variant (Figure 3a,b) and a single model for U-variant. In both, MFE and centroid (Figure 3a,d, respectively), structures predicted for 51-miR-125a C-variant by RNAstructure (Figure 3), we observed the same location of SNP as in the model generated by the RNAfold in the MFE mode (Figure 2a). Moreover, all predicted U-variant foldings (Figure 2b,d and Figure 3c,e) have the same structure, independent of the 2D structure prediction software used or mode. In both structures generated with RNAstructure MFE (Figure 3a) and centroid (Figure 3d) modes, SNP was a part of base-pairing near the hairpin loop. In the remaining three models, RNAstructure MFE C-variant (in Figure 3b) and U-variant RNAstructure MFE and centroid (3c,e), SNP is a part of a long unpaired region starting with SNP (26th nucleotide) and moving along till the last (51st) nucleotide. The differences observed between U- and C- variant models in both RNAfold and RNAstructure may suggest SNP-dependent folding of the structure.

4. Discussion

The analysis of our results concerning rs12976445 SNP in miR-125a revealed that the TT genotype was slightly more frequent in breast cancer patients and HER2 positive patients. Our study revealed only a tendency, and we obtained p-values above 0.05. One reason for this is the participation of miR-125a with numerous other factors in the development of breast cancer. The correlation of such multigenic diseases with single SNP requires a greater number of individuals and controls. Our findings suggested that rs12976445 has the potential to be a predictive biomarker for cancer risk, but a meta-analysis of a greater number of cases is required. Several studies have described the association of rs12976445 genotypes TT, CT, and CC in miR-125a with cancer and other diseases. In the Chinese study of Jiao et al. TT genotype has been significantly related to increased risk of mortality in breast cancer patients compared with those carrying the CC genotypes [19]. miR-125a rs12976445 was significantly associated with survival in codominant, recessive, and dominant models. However, only an association under the codominant model remained significant after adjustment for lymph node metastasis, TNM stage, estrogen receptor, and progesterone receptor [19]. It has been shown in previous studies that the miR-125a level is decreased in breast cancer [32–34], however, the level of miR-125a-5p is significantly higher in younger patients than in the older ones [38]. It has been shown that SNPs located in miR-125a are associated with breast cancer tumorigenesis [39] and rs12976445 SNP in miR-125a may serve as a prognostic biomarker for breast cancer [19]. Although rs12976445 is associated with breast and prostate cancer, the impact of this SNP has been rarely studied in a functional assay. The prominent study of the effect of rs12976445 on miR-125a expression was evaluated in the context of recurrent pregnancy loss [15,40]. In embryonic kidney cells, HEK293T miR-125a expression level of C haplotype was nearly four-fold higher than T haplotype [15,40]. The question remained how the rs12976445 genotype impacts the miR-125a level.

In the previous studies, rs12976445 SNP in miR-125a has been associated with the risk of pneumonitis [20]. The expression level of miR-125a mRNA has been significantly downregulated in the CT and TT groups, and CC genotype samples demonstrated upregulated miR-125a expression [20,21]. Rs12976445 polymorphism, also associated with the risk of diabetic nephropathy, showed that the expression levels of miR-125a were approximately three times lower in patients carrying TT and

CT than in the CC [41]. Other studies showed that concomitantly with miR-125a the expression of miR-205 in breast cancer is also downregulated due to SNP variations in the miR-205 sequence [42]. Studies in various cell lines identified the differential expression of miR-205 in breast cancer cell lines in correlation with the missing number of AGC repeats [42].

Herein, we decided to analyze the association of rs12976445 genotypes with breast cancer. Moreover, we performed computational research to find potential protein interacting with the SNP region and to predict and compare the 2D structure of C-/U-variant sequence. In both cases, we used two independent tools, RBPmap and ATtRACT for RNA binding protein (RBP) search, and RNAstructure and RNAfold programs to build 2D models. We observed different RBP mapped to C- and U-variants of the sequence. Both programs that we used indicated the interaction between U-variant and PTBP1 protein. In C-variant, NOVA1 protein appears in the results obtained from both ATtRACT and RBPmap. It should be underlined that NOVA1 was found only in the C-variant sequence, whereas connection with PTBP1 protein was established in U- and C-variant by RBPmap program. Another interesting fact was the appearance of HNRNPK protein in the results for C-variant obtained from RBPmap for low and medium stringency level of the algorithm search criteria. This protein was not recognized as a possible binding protein to the U-variant. This indicated that RNA processing proteins have a different affinity to C-variant and U-variant potentially modulating the probability of pre-miR-125a RNA maturation.

We also predicted the 2D structures of 51 nt sequence with C-/U-variant. We applied the same RNAfold system as Hu et al. [15]. These authors have modelled the 2D structures of rs12976445 allele T, revealing that the rare allele T can neither change the predicted secondary structure nor the predicted ΔG [15]. Hu used 1016 nt fragment of RNA, whereas we used RNAfold with 51 nt, revealing two different 2D structures and different hairpin structures in U-variant and C-variant. The shorter sequence used in our study strengthened the obtained results of modelling as in silico methods deal better with shorter sequences. In a later paper, Hu et al. admitted the difference in 2D structure between variants of rs12976445 rare allele T using the same software as previously and the same 1016 nt RNA fragment [40]. For RNAfold-predicted structures we received different models per each variant. C-variants predicted by this tool were either a part of a hairpin loop (as an unpaired nucleotide) or as a base pair near the hairpin loop. On the contrary, U-variants in RNAfold were part of a long unpaired region on the 3′ end of the sequence. RNAstructure program predicted 3 models for the C-variant sequence and 2 for the U-variant. In the Figure 3b,c,e structures were identical (except for the SNP in A1.2). Let us notice, that in Figures 2a and 3a are the same structures, and Figure 3c,d models are identical, despite using different predicting tools.

Using two computational programs, we revealed the potential differences in RNA-binding proteins between the analyzed C- and U-variant. We found that NOVA1 and HRNPK RNA-binding protein may interact with the C-variant and PTBP1 with the U-variant. Polypyrimidine tract binding protein 1 (PTBP1) binds to mRNA and regulates alternative splicing patterns [43]. In the previous reports, it has been shown that PTBP1 enhances miR-101-guided AGO2 (Argonaute) interaction with MCL1, thereby regulating miR-101-induced apoptosis and cell survival [44]. NOVA1 stimulates miRNA function by different mechanisms that converge on Argonaute proteins, a core component of the miRNA-induced silencing complex (miRISC). NOVA1 physically interacts with Ago proteins, and control neuronal miRISC function at the level of Ago proteins, with possible implications for the regulation of synapse development and plasticity [35,45]. Heterogeneous nuclear ribonucleoprotein K (hnRNPK), a ubiquitously occurring RNA-binding protein (RBP), can interact with many nucleic acids and various proteins and is involved in several cellular functions including transcription, translation, splicing, chromatin remodelling, etc. [36].

The limitation of the current study is the number of cases and controls. In the future study, a larger group of breast cancer patients should be analyzed to confirm the tendency of TT genotype association with breast cancer and to test the significance of the observed associations.

5. Conclusions

Understanding of the miRNA processing in cancer cells is an important step towards global fight against cancer. Among others, it could explain the reason why miRNA-125a level is decreased in breast cancer. Moreover, it can also help to focus the investigation on solving the upstream cascade of factors participating in abnormal regulation of miRNA in cancer. We were attempting to reveal potential features of rs12976445 SNPs of miRNA-125aby using all available resources, in vitro analysis on experimental data and using bioinformatics resources. The combination of these two approaches, in vitro and in silico, is more efficient and allows for a broader research perspective. Our analysis showed that the TT genotype of miRNA-125a was slightly more frequent in breast cancer patients and HER2 positive patients. We also demonstrated that the U-variant of rs12976445 diminished the probability of pri-miR-125a binding to NOVA1 and HNRNPK proteins. Our in silico analysis revealed that C- and U-variants could promote different RNA folding patterns that may further affect protein binding. Altogether, these may imply a lower miR-125a expression in breast cancer. These results may not only be useful for diagnostic purposes but can also contribute to the research into novel therapies for breast cancer. In a wider perspective, our experimental protocol and findings may be extended into different types of cancer or other diseases, where miRNA expression level is affected. Following this path, in future work, we would like to consider the other molecular types of cancer to extend the understanding of the impact of rs12976445 on miR-125a expression and verified the obtained in silico results by in vivo methods

Author Contributions: T.P.L.: conceptualization, methodology, data curation, supervision, formal analysis. J.M., N.S., M.S.: methodology, investigation, visualization, writing—review and editing, supervision. S.G.-G.: writing—original draft preparation. P.P.J.: writing—original draft preparation, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Centre, Poland (NN 403598538, 2016/23/B/ST6/03931, 2019/35/B/ST6/03074). NS was supported by the FNP START 2019 program.

Acknowledgments: The authors thank Bogumila Ratajczak for her assistance during the preparation of this manuscript.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Keeney, M.G.; Couch, F.J.; Visscher, D.W.; Lindor, N.M. Non-BRCA familial breast cancer: Review of reported pathology and molecular findings. *Pathology* **2017**, *49*, 363–370. [CrossRef] [PubMed]
2. Malhotra, P.; Read, G.H.; Weidhaas, J.B. Breast Cancer and miR-SNPs: The Importance of miR Germ-Line Genetics. *Noncoding RNA* **2019**, *5*, 27. [CrossRef] [PubMed]
3. Mosallaei, M.; Simonian, M.; Esmaeilzadeh, E.; Bagheri, H.; Miraghajani, M.; Salehi, A.R.; Mehrzad, V.; Salehi, R. Single nucleotide polymorphism rs10889677 in miRNAs Let-7e and Let-7f binding site of IL23R gene is a strong colorectal cancer determinant: Report and meta-analysis. *Cancer Genet.* **2019**, *239*, 46–53. [CrossRef]
4. Yan, Z.; Zhou, Z.; Li, C.; Yang, X.; Yang, L.; Dai, S.; Zhao, J.; Ni, H.; Shi, L.; Yao, Y. Polymorphisms in miRNA genes play roles in the initiation and development of cervical cancer. *J. Cancer* **2019**, *10*, 4747–4753. [CrossRef]
5. Yin, Z.; Cui, Z.; Ren, Y.; Xia, L.; Wang, Q.; Zhang, Y.; He, Q.; Zhou, B. Association between polymorphisms in pre-miRNA genes and risk of lung cancer in a Chinese non-smoking female population. *Lung Cancer* **2016**, *94*, 15–21. [CrossRef] [PubMed]
6. Bahreini, F.; Ramezani, S.; Shahangian, S.S.; Salehi, Z.; Mashayekhi, F. miR-559 polymorphism rs58450758 is linked to breast cancer. *Br. J. Biomed. Sci.* **2019**, *77*, 29–34. [CrossRef]
7. Linhares, J.J.; Azevedo, M., Jr.; Siufi, A.A.; de Carvalho, C.V.; WolgienMdel, C.; Noronha, E.C.; Bonetti, T.C.; da Silva, I.D. Evaluation of single nucleotide polymorphisms in microRNAs (hsa-miR-196a2 rs11614913 C/T) from Brazilian women with breast cancer. *BMC Med. Genet.* **2012**, *13*, 119. [CrossRef] [PubMed]

8. Zhang, N.; Huo, Q.; Wang, X.; Chen, X.; Long, L.; Jiang, L.; Ma, T.; Yang, Q. A genetic variant in pre-miR-27a is associated with a reduced breast cancer risk in younger Chinese population. *Gene* **2013**, *529*, 125–130. [CrossRef]
9. Loh, H.Y.; Norman, B.P.; Lai, K.S.; Rahman, N.; Alitheen, N.B.M.; Osman, M.A. The Regulatory Role of MicroRNAs in Breast Cancer. *Int. J. Mol. Sci.* **2019**, *20*, 4940. [CrossRef]
10. Potenza, N.; Panella, M.; Castiello, F.; Mosca, N.; Amendola, E.; Russo, A. Molecular mechanisms governing microRNA-125a expression in human hepatocellular carcinoma cells. *Sci. Rep.* **2017**, *7*, 10712. [CrossRef]
11. Li, G.; Ao, S.; Hou, J.; Lyu, G. Low expression of miR-125a-5p is associated with poor prognosis in patients with gastric cancer. *Oncol. Lett.* **2019**, *18*, 1483–1490. [CrossRef] [PubMed]
12. Sun, C.; Zeng, X.; Guo, H.; Wang, T.; Wei, L.; Zhang, Y.; Zhao, J.; Ma, X.; Zhang, N. MicroRNA-125a-5p modulates radioresistance in LTEP-a-2 non-small cell lung cancer cells by targeting SIRT7. *Cancer Biomark* **2020**, *27*, 39–49. [CrossRef] [PubMed]
13. Tang, L.; Zhou, L.; Wu, S.; Shi, X.; Jiang, G.; Niu, S.; Ding, D. miR-125a-5p inhibits colorectal cancer cell epithelial-mesenchymal transition, invasion and migration by targeting TAZ. *Onco Targets Ther.* **2019**, *12*, 3481–3489. [CrossRef] [PubMed]
14. Ma, J.; Zhan, Y.; Xu, Z.; Li, Y.; Luo, A.; Ding, F.; Cao, X.; Chen, H.; Liu, Z. ZEB1 induced miR-99b/let-7e/miR-125a cluster promotes invasion and metastasis in esophageal squamous cell carcinoma. *Cancer Lett.* **2017**, *398*, 37–45. [CrossRef] [PubMed]
15. Hu, Y.; Liu, C.M.; Qi, L.; He, T.Z.; Shi-Guo, L.; Hao, C.J.; Cui, Y.; Zhang, N.; Xia, H.F.; Ma, X. Two common SNPs in pri-miR-125a alter the mature miRNA expression and associate with recurrent pregnancy loss in a Han-Chinese population. *RNA Biol.* **2011**, *8*, 861–872. [CrossRef]
16. Morales, S.; De Mayo, T.; Gulppi, F.A.; Gonzalez-Hormazabal, P.; Carrasco, V.; Reyes, J.M.; Gomez, F.; Waugh, E.; Jara, L. Genetic Variants in pre-miR-146a, pre-miR-499, pre-miR-125a, pre-miR-605, and pri-miR-182 Are Associated with Breast Cancer Susceptibility in a South American Population. *Genes* **2018**, *9*, 427. [CrossRef]
17. Peterlongo, P.; Caleca, L.; Cattaneo, E.; Ravagnani, F.; Bianchi, T.; Galastri, L.; Bernard, L.; Ficarazzi, F.; Dall'olio, V.; Marme, F.; et al. The rs12975333 variant in the miR-125a and breast cancer risk in Germany, Italy, Australia and Spain. *J. Med. Genet.* **2011**, *48*, 703–704. [CrossRef]
18. Cai, T.; Li, J.; An, X.; Yan, N.; Li, D.; Jiang, Y.; Wang, W.; Shi, L.; Qin, Q.; Song, R.; et al. Polymorphisms in MIR499A and MIR125A gene are associated with autoimmune thyroid diseases. *Mol. Cell. Endocrinol.* **2017**, *440*, 106–115. [CrossRef]
19. Jiao, L.; Zhang, J.; Dong, Y.; Duan, B.; Yu, H.; Sheng, H.; Huang, J.; Gao, H. Association between miR-125a rs12976445 and survival in breast cancer patients. *Am. J. Transl. Res.* **2014**, *6*, 869–875.
20. Huang, X.; Zhang, T.; Li, G.; Guo, X.; Liu, X. Regulation of miR-125a expression by rs12976445 single-nucleotide polymorphism is associated with radiotherapy-induced pneumonitis in lung carcinoma patients. *J. Cell. Biochem.* **2019**, *120*, 4485–4493. [CrossRef]
21. Quan, H.Y.; Yuan, T.; Hao, J.F. A microRNA125a variant, which affects its mature processing, increases the risk of radiationinduced pneumonitis in patients with nonsmallcell lung cancer. *Mol. Med. Rep.* **2018**, *18*, 4079–4086.
22. Damodaran, M.; Paul, S.F.D.; Venkatesan, V. Genetic Polymorphisms in miR-146a, miR-196a2 and miR-125a Genes and its Association in Prostate Cancer. *Pathol. Oncol. Res.* **2018**, *26*, 193–200. [CrossRef] [PubMed]
23. Miskiewicz, J.; Tomczyk, K.; Mickiewicz, A.; Sarzynska, J.; Szachniuk, M. Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *Biomed. Res. Int.* **2017**, *2017*, 6783010. [CrossRef]
24. Miskiewicz, J.; Szachniuk, M. Discovering Structural Motifs in miRNA Precursors from the Viridiplantae Kingdom. *Molecules* **2018**, *23*, 1367. [CrossRef] [PubMed]
25. Hall, T.A. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **1999**, *41*, 95–98.
26. Xu, K.; Cai, H.; Shen, Y.; Ni, Q.; Chen, Y.; Hu, S.; Li, J.; Wang, H.; Yu, L.; Huang, H.; et al. [Management of corona virus disease-19 (COVID-19): The Zhejiang experience]. *Zhejiang Da Xue Xue Bao Yi Xue Ban* **2020**, *49*, 147–157. [PubMed]
27. Paz, I.; Kosti, I.; Ares, M., Jr.; Cline, M.; Mandel-Gutfreund, Y. RBPmap: A web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **2014**, *42*, W361–W367. [CrossRef]

28. Giudice, G.; Sanchez-Cabo, F.; Torroja, C.; Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**, *2016*. [CrossRef]
29. Reuter, J.S.; Mathews, D.H. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinform.* **2010**, *11*, 129. [CrossRef]
30. Gruber, A.R.; Lorenz, R.; Bernhart, S.H.; Neubock, R.; Hofacker, I.L. The Vienna RNA websuite. *Nucleic Acids Res.* **2008**, *36*, W70–W74. [CrossRef]
31. Sole, X.; Guino, E.; Valls, J.; Iniesta, R.; Moreno, V. SNPStats: A web tool for the analysis of association studies. *Bioinformatics* **2006**, *22*, 1928–1929. [CrossRef]
32. Guo, X.; Wu, Y.; Hartley, R.S. MicroRNA-125a represses cell growth by targeting HuR in breast cancer. *RNA Biol.* **2009**, *6*, 575–583. [CrossRef]
33. Scott, G.K.; Goga, A.; Bhaumik, D.; Berger, C.E.; Sullivan, C.S.; Benz, C.C. Coordinate suppression of ERBB2 and ERBB3 by enforced expression of micro-RNA miR-125a or miR-125b. *J. Biol. Chem.* **2007**, *282*, 1479–1486. [CrossRef] [PubMed]
34. Serpico, D.; Molino, L.; Di Cosimo, S. microRNAs in breast cancer development and treatment. *Cancer Treat. Rev.* **2014**, *40*, 595–604. [CrossRef] [PubMed]
35. Xin, Y.; Li, Z.; Zheng, H.; Ho, J.; Chan, M.T.V.; Wu, W.K.K. Neuro-oncological ventral antigen 1 (NOVA1): Implications in neurological diseases and cancers. *Cell Prolif.* **2017**, *50*, e12348. [CrossRef]
36. Xu, Y.; Wu, W.; Han, Q.; Wang, Y.; Li, C.; Zhang, P.; Xu, H. Post-translational modification control of RNA-binding protein hnRNPK function. *Open Biol.* **2019**, *9*, 180239. [CrossRef] [PubMed]
37. Wu, S.L.; Fu, X.; Huang, J.; Jia, T.T.; Zong, F.Y.; Mu, S.R.; Zhu, H.; Yan, Y.; Qiu, S.; Wu, Q.; et al. Genome-wide analysis of YB-1-RNA interactions reveals a novel role of YB-1 in miRNA processing in glioblastomamultiforme. *Nucleic Acids Res.* **2015**, *43*, 8516–8528. [CrossRef]
38. He, H.; Xu, F.; Huang, W.; Luo, S.Y.; Lin, Y.T.; Zhang, G.H.; Du, Q.; Duan, R.H. miR-125a-5p expression is associated with the age of breast cancer patients. *Genet. Mol. Res.* **2015**, *14*, 17927–17933. [CrossRef]
39. Li, W.; Duan, R.; Kooy, F.; Sherman, S.L.; Zhou, W.; Jin, P. Germline mutation of microRNA-125a is associated with breast cancer. *J. Med. Genet.* **2009**, *46*, 358–360. [CrossRef]
40. Hu, Y.; Huo, Z.H.; Liu, C.M.; Liu, S.G.; Zhang, N.; Yin, K.L.; Qi, L.; Ma, X.; Xia, H.F. Functional study of one nucleotide mutation in pri-miR-125a coding region which related to recurrent pregnancy loss. *PLoS ONE* **2014**, *9*, e114781. [CrossRef]
41. Li, C.; Lei, T. Rs12976445 Polymorphism is Associated with Risk of Diabetic Nephropathy Through Modulating Expression of MicroRNA-125 and Interleukin-6R. *Med. Sci. Monit.* **2015**, *21*, 3490–3497. [CrossRef]
42. Zhang, J.; Wei, B.; Hu, H.; Liu, F.; Tu, Y.; He, F. The association between differentially expressed micro RNAs in breast cancer cell lines and the micro RNA-205 gene polymorphism in breast cancer tissue. *Oncol. Lett.* **2018**, *15*, 2139–2146. [CrossRef]
43. Keppetipola, N.; Sharma, S.; Li, Q.; Black, D.L. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit. Rev. Biochem. Mol. Biol.* **2012**, *47*, 360–378. [CrossRef] [PubMed]
44. Cui, J.; Placzek, W.J. PTBP1 enhances miR-101-guided AGO2 targeting to MCL1 and promotes miR-101-induced apoptosis. *Cell Death Dis.* **2018**, *9*, 552. [CrossRef] [PubMed]
45. Storchel, P.H.; Thummler, J.; Siegel, G.; Aksoy-Aksel, A.; Zampa, F.; Sumer, S.; Schratt, G. A large-scale functional screen identifies Nova1 and Ncoa3 as regulators of neuronal miRNA function. *EMBO J.* **2015**, *34*, 2237–2254. [CrossRef] [PubMed]

Bioinformatics, 36(4), 2020, 1129–1134
doi: 10.1093/bioinformatics/btz738
Advance Access Publication Date: 7 October 2019
Original Paper

Structural bioinformatics

Topology-based classification of tetrads and quadruplex structures

Mariusz Popenda[1,†], Joanna Miskiewicz[2,†], Joanna Sarzynska[1], Tomasz Zok[2,3] and Marta Szachniuk[1,2,*]

[1]Department of Structural Bioinformatics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan 61-704, Poland, [2]Institute of Computing Science and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Poznan 60-965, Poland and [3]Poznan Supercomputing and Networking Center, Poznan 61-139, Poland

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally.
Associate Editor: Alfonso Valencia

Received on May 15, 2019; revised on August 12, 2019; editorial decision on September 19, 2019; accepted on September 25, 2019

Abstract

Motivation: Quadruplexes attract the attention of researchers from many fields of bio-science. Due to a specific structure, these tertiary motifs are involved in various biological processes. They are also promising therapeutic targets in many strategies of drug development, including anticancer and neurological disease treatment. The uniqueness and diversity of their forms cause that quadruplexes show great potential in novel biological applications. The existing approaches for quadruplex analysis are based on sequence or 3D structure features and address canonical motifs only.

Results: In our study, we analyzed tetrads and quadruplexes contained in nucleic acid molecules deposited in Protein Data Bank. Focusing on their secondary structure topology, we adjusted its graphical diagram and proposed new dot-bracket and arc representations. We defined the novel classification of these motifs. It can handle both canonical and non-canonical cases. Based on this new taxonomy, we implemented a method that automatically recognizes the types of tetrads and quadruplexes occurring as unimolecular structures. Finally, we conducted a statistical analysis of these motifs found in experimentally determined nucleic acid structures in relation to the new classification.

Availability and implementation: https://github.com/tzok/eltetrado/

Contact: mszachniuk@cs.put.poznan.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Nucleic acids have the ability to fold into a variety of configurations. One of them is quadruplex, a characteristic structural motif found in DNAs, RNAs and nucleic acid analogs, such as peptide nucleic acids (PNA) or conformationally locked nucleic acids (LNA) (Burge *et al.*, 2006; Malgowska *et al.*, 2016). Quadruplexes occur in genomes of different species, including humans (Chambers *et al.*, 2015; Marsico *et al.*, 2019; Sahakyan *et al.*, 2017; Yadav *et al.*, 2017). They are important regulators in cellular functions of nucleic acids, including telomere elongation and gene expression mechanisms. The number of discovered influences of these structures is still growing (Cammas and Millevoi, 2017; Fay *et al.*, 2017; Gudanis *et al.*, 2016; Marušič and Plavec, 2015; Tan *et al.*, 2016; Trajkovski *et al.*, 2012). The connections between regulatory processes and unique structures make quadruplexes particularly important in novel therapies for cancer and neurodegenerative disorders (Cammas and Millevoi, 2017; Fay *et al.*, 2017; Huppert, 2008).

Quadruplex is composed of at least two building blocks called N-tetrads (where N denotes any nucleotide residue) stacked upon one another at a distance of about 3.3 Å (Bhattacharya *et al.*, 2019; Fay *et al.*, 2017; Kotar *et al.*, 2019). Single tetrad is formed by four nucleotide residues, usually of the same type, found in a planar arrangement (Cammas and Millevoi, 2017; Malgowska *et al.*, 2016). Thus, one quadruplex, also referred to as N4 in this paper (N—any nucleotide), contains at least eight nucleotides organized in tetrads (Lorenz *et al.*, 2013; Pandey *et al.*, 2015). Quadruplexes are known to occur in Guanine-rich regions of nucleic acid structures. Thus, in most cases (over 90%), the residues that build the tetrad are Guanine-based. They form G-tetrads which—in turn—create the G-quadruplex abbreviated to G4. Each nucleobase in the canonical G-tetrad forms two non-canonical pairs, serving as a donor at Watson-Crick (W) edge, and an acceptor at Hoogsteen (H) edge (Fay *et al.*, 2017; Malgowska *et al.*, 2016; Sahakyan *et al.*, 2017). Therefore, four nucleobases in the G-tetrad form eight hydrogen bonds in total (Bhattacharya *et al.*, 2019; Fay

et al., 2017; Malgowska *et al.*, 2016). However, in pseudo G-tetrads and tetrads composed of other nucleotide residues, one can meet base pairs other than WH. Quadruplexes display structural diversity that depends on the primary sequence, ion type and environment.

A study of quadruplex structure often starts from the sequence level and an analysis of G-tracts. G-tract is an uninterrupted string of characters (at least two) representing consecutive nucleobases in the nucleic acid strand. Nucleobases represented in one G-tract usually belong to different tetrads. Four G-tracts define G4-stem in the structure (Dvorkin *et al.*, 2018; da Silva, 2007). Quadruplex motifs merge such structural elements as G4-stem, tetrads and loops that connect two outer tetrads.

Quadruplex may exist either as intra- or intermolecular structure. In the first case, it is formed from one strand and may be classified as a unimolecular motif. In the second case, G4 consists of two or four strands, and thus, belongs to a class of bi- or tetramolecular structures, respectively (Fay *et al.*, 2017; Huppert, 2008; Kwok and Merrick, 2017; Malgowska *et al.*, 2016). If all strands of the quadruplex are oriented in the same direction, G4 is parallel. If half of the strands have the opposite direction than the others, we call the quadruplex antiparallel. Whereas, if one strand is directed contrary to the remaining three, we have a hybrid-type motif (Burge *et al.*, 2006; Malgowska *et al.*, 2016; Rhodes and Lipps, 2015).

An increasing interest in quadruplexes has resulted in the development of computational methods to support their study. In recent years, several bioinformatics tools focusing on the sequence, secondary and tertiary structure of these motifs have been published. Computational programs like G4Hunter (Bedrat *et al.*, 2016), G4RNA screener (Garant *et al.*, 2015, 2017), QGRS Mapper (Kikin *et al.*, 2006), G4-iM Grinder (Reche and Morales, 2019) parse DNA or RNA sequence to find motifs with a potential to form G-quadruplexes. GRSDB2 (Kikin *et al.*, 2007) is a database of putative quadruplex-forming Guanine-rich sequences mapped in pre-mRNAs and mRNAs. QuadBase2 (Dhapola and Chowdhury, 2016) allows mining of G4 motifs in a genome. ViennaRNA package (Lorenz *et al.*, 2012) includes algorithms for RNA secondary structure prediction extended by the option to annotate quadruplexes in the output model. They encode the secondary structure of an RNA in dot-bracket notation, where every nucleotide in the G-tract is represented by '+' sign. The tertiary topology of canonical G-quadruplexes has been explored by Webba da Silva group (Dvorkin *et al.*, 2018; Karsisiotis *et al.*, 2013; da Silva, 2007). Their studies have resulted in proposing a classification of G-tetrads based on glycosidic bond angles between nucleotide components of the tetrads. They have also defined categories of canonical G4s following the topology of loops (diagonal, propeller, lateral) between consecutive G-tracts (Karsisiotis *et al.*, 2013; da Silva, 2007). Finally, some tools have appeared to facilitate analysis of G4-rich nucleic acid interactions with proteins (Mishra *et al.*, 2016) or searching for 3D structure motifs with the potential to form quadruplexes (Reche and Morales, 2019).

Structural and topological diversity of quadruplex structures goes far beyond the framework of canonical G4s. Even among G-quadruplexes themselves, identified nowadays, we find canonical and non-canonical cases. New quadruplex structures are constantly being solved [e.g. Z-DNA quadruplex (Bakalar *et al.*, 2019)] and the number of known non-canonical quadruplexes increases. The in-depth analysis of quadruplex features is still a challenge (Dvorkin *et al.*, 2018). It should not be restricted by molecule type, its sequence, structure canonicality, and should not focus only on one structure level, e.g. sequence or 3D structure, which is characteristic of existing computational methods.

Here, we introduce a new classification of tetrads and quadruplexes occurring in nucleic acids. It is based on the secondary structure topology of these motifs and can handle both canonical and non-canonical structures. We present two-line dot-bracket notation to represent tetrads and quadruplexes, the adjusted graphical views generated by the latest version of our RNApdbee webserver (Zok *et al.*, 2018) and arc diagrams that clearly show the differences between quadruplex topologies. Our concept is accompanied by the automated method ElTetrado that identifies tetrads and quadruplexes in the 3D structures of nucleic acids and classifies them according to newly defined categories (Zok *et al.*, 2019). It is available for download at https://github.com/tzok/eltetrado/. We show the results of the statistical analysis run on the dataset of all PDB-deposited nucleic acid structures with the use of our method.

2 Materials and methods

The research presented here was carried out on the basis of data downloaded from the Protein Data Bank (Berman, 2000) on April 18, 2019. Out of all the 3D structures present in biological assembly files acquired from RCSB PDB website, we selected those that contained quadruplexes. For this purpose, we used our own script that processed structural data and searched for these motifs.

The dataset for further analysis was created from 308 PDB structures in which quadruplexes formed. It contained 258 DNAs, 45 RNAs and 5 other molecules. The latter group included structures in which over 50% of nucleotides within quadruplexes were modified. PDB identifiers of all analyzed molecules have been listed in the Supplementary Material (Supplementary Table S1). For the subsequent analysis, in the case of NMR structures, the first model was taken and in the case of X-ray structures, all biological units were selected to the dataset. In our collection, we distinguished uni-, bi- and tetramolecular quadruplexes. All of them were considered in the statistical analysis of tetrads and quadruplexes. Whereas, the secondary structure topology-based taxonomy of these motifs covered only unimolecular cases. The latter set contained 188 PDB structures, including 160 DNAs, 26 RNAs and 2 other molecules. More detailed information on the datasets' contents has been given in the Supplementary Material (Supplementary Fig. S1).

Structures from both sets were analyzed using self-implemented programs along with DSSR software from the 3DNA suite (Lu *et al.*, 2015). From DSSR, we acquired the information about base pairs and stacking. We applied PyMOL [Schrödinger, LLC (2015)] to visualize and inspect the tertiary structures from the dataset. Arc diagrams of the secondary structure were generated using R4RNA package (Lai *et al.* (2012)) and refined in the vector graphic software Inkscape.

The preliminary classification of tetrads and quadruplexes was performed based on the results of RNApdbee 2.0 (Antczak *et al.*, 2018, 2014; Zok *et al.*, 2018). This webserver is a part of RNApolis toolset (Szachniuk, 2019). It retrieves secondary structure topology from the 3D structure data saved in PDB and mmCIF files. Additionally, we utilized a self-developed computer program ElTetrado to assign the categories to tetrads and quadruplexes identified in the analyzed molecules (Zok *et al.*, 2019).

3 Representation and classification of tetrads and quadruplexes

3.1 Two-line dot-bracket for tetrads and quadruplexes

The dot-bracket notation has been designed to encode the secondary structure topology of an RNA molecule using a sequence of dots and brackets [and letters in the extended dot-bracket notation (Antczak *et al.*, 2018)]. A dot represents an unpaired nucleotide residue while a pair of brackets (opening and closing one) encodes for a base pair. Typically, a continuous string of characters in the dot-bracket notation, written in a single line, represents the secondary structure of a single RNA strand. This nomenclature has some limitations. It has been designed to encode canonical base pairs. Therefore, it does not allow to represent multiplets [i.e. three or more nucleobases associated in a coplanar geometry through a network of hydrogen bonds (Colasanti *et al.*, 2013)]. Moreover, dot-bracket encoding depends on the strands' order. Thus, it is not unequivocal for multi-stranded structures (Popenda *et al.*, 2008). Until now, the latter reason has precluded encoding tetrads and quadruplexes in the dot-bracket nomenclature.

Here, we show how to encode tetrads and quadruplexes in dot-bracket extended to a two-line form. In the case of tetrad representation, each line holds two base pairs that do not share nucleobases.

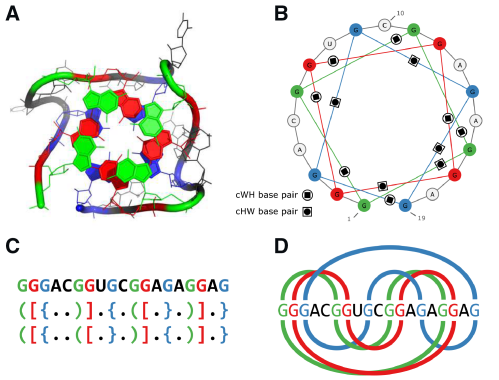


Fig. 1. Quadruplex formed in RNA Mango (PDB id: 5V3F chain B) (Trachman *et al.*, 2017) in (**A**) 3D and (**B**) secondary structure view, (**C**) dot-bracket encoding, (**D**) arc diagram. For clarity, each tetrad has a different color

Thus, if nucleobase N_i forms hydrogen bonds with N_j and N_k in the tetrad, one of these pairs (e.g. N_i-N_j) is encoded in the first line, and the other (N_i-N_k)—in the second line of dot-bracket. In canonical G-tetrad, every nucleotide is involved in one interaction along its Watson-Crick edge, and one along the Hoogsteen edge. The first nucleotide N_1 of the tetrad (the closest to 5′-end) corresponds to the leftmost opening bracket in the first line of dot-bracket representation. A base pair which N_1 forms along its Watson-Crick edge is encoded in the first line. In the other tetrads, if N_1 does not interact along its Watson-Crick edge, the first line of dot-bracket includes a base pair formed along the Hoogsteen edge of N_1. The assignment of the remaining base pairs to dot-bracket lines is determined automatically.

The quadruplex's encoding adds up dot-bracket representations of the component tetrads (Fig. 1). Thus, its first line holds brackets for nucleotides of the first G-tract (or more generally N-tract) interacting along their Watson-Crick edges.

The two-line notation is correlated with arc diagrams (Lai *et al.*, 2012) optimally adapted to visualize the secondary structure of tetrads and quadruplexes. Like the dot-bracket representation, the arc diagram consists of two parts, the upper and the lower one. The upper arcs represent the first line of the corresponding dot-bracket, while the lower part is related to the second line. So designed arc diagrams clearly show the differences between the topologies of quadruplexes and allow for their easy differentiation according to the secondary structure features.

3.2 Classification of tetrads

The secondary structure of a tetrad can be represented by a cyclic graph $G^*=(V,E)$, where $|V|=|E|=4$. Every vertex in G^*represents one nucleotide residue from the tetrad. Every edge in G^* is related to a hydrogen-bonding interaction between respective nucleotides. If we placed the vertices of G^* at equal distances on a circle clockwise, in the order along the sequence, we would see that graph can take the shape of a square (O-shaped), a bow tie (N-shaped), or an hourglass (Z-shaped). This observation has led us to define three categories of tetrads and establish the ONZ taxonomy.

ONZ classification is determined by pairings between the tetrad-forming nucleotide residues, N_1, N_2, N_3, N_4 (Fig. 2). Category O (O-shaped) contains tetrads, the nucleotides of which interact according to strand direction (from 5′- to 3′-end). It means that in the O-type tetrad, N_1 (the first nucleotide from 5′-end) interacts with N_2, N_2 with N_3, N_3 with N_4 and—finally—N_4 with N_1. The N category (N-shaped) represents tetrads stabilized by base pairs (N_1,N_2), (N_2,N_4), (N_4,N_3), (N_3,N_1). Finally, the tetrad belongs to class Z (Z-shaped), if the following interactions takes place between

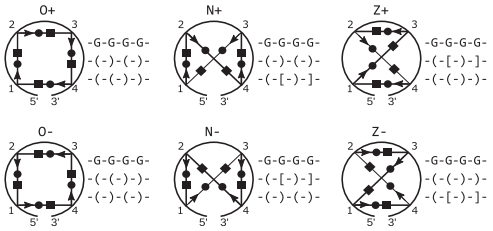


Fig. 2. ONZ classes of tetrads in diagrams of unimolecular structures and their dot-bracket representations. Arrows ease discerning between + and −. Here, black filled circle denotes Watson-Crick edge, square—Hoogsteen edge

its nucleotides: (N_1,N_3), (N_3,N_2), (N_2,N_4), (N_4,N_1). Let us note, that the classification is based on the order of tetrad-involved nucleotides in the strand. Thus, in this form, it can only be applied unambiguously to unimolecular structures.

Additionally, we can annotate the tetrad as positive (+) or negative (−), according to the arrangement of edges along which base pairs in the tetrad are formed. ElTetrado does this by analyzing two pairs, (N_1,N_i) and (N_1,N_j), in which the first nucleotide (N_1) in the tetrad is involved. Assume that we number nucleotide residues in the tetrad from 5′ to 3′ end such that $i < j$. Thus, we can set the nucleotides from two considered pairs in the following order: $N_1 < N_i < N_j$. Let us now assume that the following edge hierarchy has been established, along which N_1 binds to N_i and N_j: $W < H < S$, where W is for Watson-Crick edge, H—Hoogsteen edge, S—sugar edge. We can order nucleotides N_1, N_i, N_j according to this hierarchy. For example it means that if N_1 interacts with N_j along its Watson-Crick edge, and with N_i along Hoogsteen edge, then the order is: $N_1 < N_j < N_i$. ElTetrado checks whether the order of nucleotides applied in the first case is the same as the order in the second case. If so, the tetrad is assigned the positive type (+), otherwise, it has the negative type (−). Therefore, every class in ONZ can be divided into two subcategories: O+, O−, N+, N−, Z+, Z-. Each of these subclasses has a unique dot-bracket representation (Fig. 2).

3.3 Classification of quadruplexes

Making use of the new tetrad classification, we have proposed ONZ-based taxonomy for quadruplexes. We have assumed that the classes of component tetrads automatically determine the category of the whole quadruplex. From this it follows that a quadruplex consisting of O-type tetrads belongs to class O, a motif built from N-shaped tetrads is in class N, while if it includes only tetrads from Z category, it is assigned to class Z. Hence, we have O, N and Z classes that group regular motifs, i.e. quadruplexes composed of tetrads of the same type. However, among various quadruplex forms there are also irregular structures that contain tetrads of different types. To hold such cases, we have defined category of mixed structures denoted by M. Finally, let us recall that—preliminarily—ONZ taxonomy of tetrads has addressed unimolecular structures. Whereas, there are also bi- and tetramolecular quadruplexes. We propose to assign them to class R (remaining structures).

We can divide quadruplexes according to the order of nucleotides in N-tracts. From this perspective, we distinguish parallel (p), antiparallel (a) and hybrid (h) motifs. We can combine both approaches to define the following classes of regular quadruplexes: Op, Oa, Oh, Np, Na, Nh, Zp, Za, Zh. As for irregular quadruplexes, M category contains all the mixed motifs, without dividing them by the N-tracts' arrangement into parallel, antiparallel and hybrid cases. Eventually, bi- and tetramolecular quadruplexes are assigned to Rp, Ra and Rh.

Let us add that the structural diversity of quadruplexes is noticeable even within a single class in ONZ-based taxonomy. The best way to observe this is to analyze the arc diagrams representing the secondary structure topologies. Quadruplexes belonging to a given

Table 1. Number of tetrads by sequence and molecule type

	GG GG	UU UU	TT TT	AA AA	CC CC	Other	Total
DNA	1003	–	13	7	5	30	1058
RNA	228	29	–	9	–	9	275
OTHER	39	2	3	–	–	–	44
Total	1270	31	16	16	5	39	1377

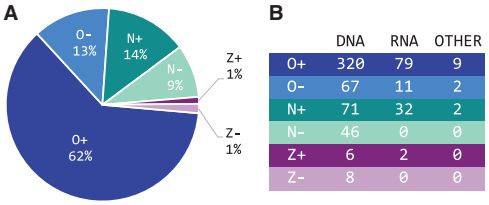


Fig. 3. ONZ class coverage by tetrads from unimolecular quadruplexes

category can have more than one topology. The number of diverse topologies depends on the number and types of tetrads in the motif. For example, in Supplementary Figure S2 of Supplementary Material, we have shown all possible topologies of regular G4s which are composed of two tetrads belonging to positive (+) categories (i.e. O+, N+, Z+).

4 Results

4.1 Analysis of the tetrad set

In the source dataset of 308 PDB entries, we have identified 1377 tetrads with different composition (Table 1). G-tetrads account for 92% of this collection. The remaining part consists of U-, C-, A-, or T-tetrads, and mixed tetrads the majority of which include Guanine as one of the contributors. Mixed tetrads have been enumerated in the Supplementary Material (Supplementary Table S2). 655 tetrads from the source dataset were unimolecular and could be categorized due to the ONZ taxonomy. We have run ElTetrado to classify them and find the coverage of every category (Fig. 3). Let us note that ONZ classification has encompassed all unimolecular tetrads, independently on their sequence. 75% of single-stranded tetrads appear to be O-type, with a significant prevalence of O+. The class of N-shaped contains 23% of tetrads from the analyzed collection (with N+ being the majority), and class Z—the remaining 2%. In case of the latter category, which is few in number, more tetrads have been found in subset Z-.

4.2 Analysis of the quadruplex set

The preliminary analysis of the initial dataset has given information about 423 quadruplexes which were identified in 308 PDB structures. Let us recall our assumption that the quadruplex is a motif consisting of at least two stacked tetrads (not necessarily G-tetrads). We have examined the structural complexity of these quadruplexes.

First, we have checked how many tetrads make up the motifs. It has appeared that 90% of quadruplexes in the set are composed of 2, 3 or 4 tetrads (with 3-tetrad motifs coming to the lead). We have also found single quadruplexes containing more component tetrads, including an exceptionally large motif consisting of up to 13 of them (Table 2).

Next, we have investigated how many strands contributed to building quadruplexes (Table 3). Our research has shown that more than half of the collection is made up of unimolecular motifs. These 242 quadruplexes could be next studied with respect to ONZ-based taxonomy. The remaining 181 motifs belong to category R, as their

Table 2. Number of quadruplexes composed of 2–13 tetrads

Number of tetrads:	2	3	4	5	6	7	9	13
DNA	75	138	102	4	5	3	1	1
RNA	29	28	4	22	2	1	–	–
OTHER	4	–	4	–	–	–	–	–
Total	108	166	110	26	7	4	1	1

Table 3. Number of uni-, bi- and tetramolecular quadruplexes

	Unimolecular	Bimolecular	Tetramolecular	Total
DNA	188	90	51	329
RNA	50	7	29	86
OTHER	4	–	4	8
Total	242	97	84	423

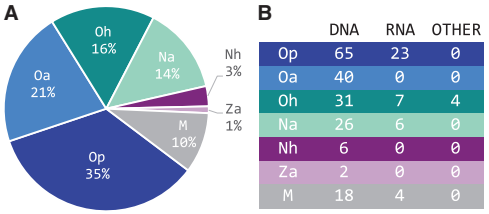


Fig. 4. ONZM class coverage by unimolecular quadruplexes

tetrads were not assigned to ONZ classes. The set of unimolecular quadruplexes has been processed to find the distribution of quadruplex topologies in ONZM categories (Fig. 4). Since, the classification of quadruplexes depends strictly on the types of component tetrads, we have expected that O-based classes would be the most represented. Indeed, 70% of unimolecular quadruplexes have been assigned to Op, Oa and Oh groups. These are regular motifs composed only of tetrads from O+ and O- categories. Regular Z-type quadruplexes constitute the least numerous group. Let us notice that parallel structures have been found only among O-type quadruplexes, as far as the regular cases are considered. As for the irregular motifs, we have identified 32 examples and we have assigned them to category M. Finally, we have found bi- and tetramolecular quadruplexes 99 of which are of type Rp, 73 ones of type Ra and 9 cases in Rh group.

4.3 Example quadruplex structures in ONZ-based taxonomy

In this section, we show the results of ONZ-based classification, as well as dot-bracket and arc representations correlated with exemplary quadruplex structures.

To the first example, we have chosen two quadruplexes that have the same nucleotide sequence but belong to different categories in the ONZ taxonomy. Both are G4s included in the human telomeric DNA. The first one (PDB id: 1KF1) is the X-ray structure (Parkinson *et al.*, 2002) composed of three tetrads of type O+. According to the G-tracts' arrangement, this motif is parallel. Thus, it has been classified as Op. The second quadruplex (PDB id: 143D) has been solved using NMR in solution (Wang and Patel, 1993). Its outer tetrads belong to N+ category, whereas the middle one is of type N-. Since this G4 is antiparallel, we have assigned it to Na group. Figure 5 presents the tertiary structures of both motifs as well as the secondary structures encoded in two-line dot-bracket notation and represented in arc diagrams. Both dot-bracket strings and arc diagrams reveal the differences between the topologies of compared

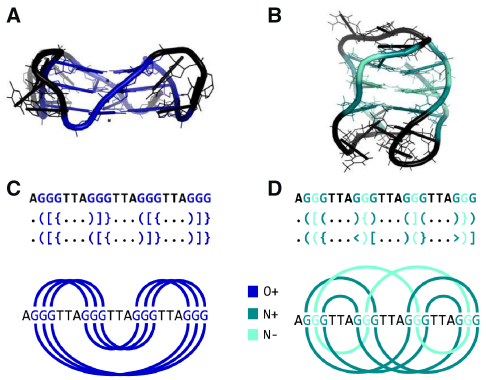


Fig. 5. 3D structure of human telomeric DNA (**A**) 1KF1 (Parkinson *et al.*, 2002) and (**B**) 143D (Wang and Patel, 1993) with colored tetrads and (**C, D**) their secondary structure topologies in dot-bracket and arc diagram representations. The colors in the secondary structure representations indicate the tetrad types

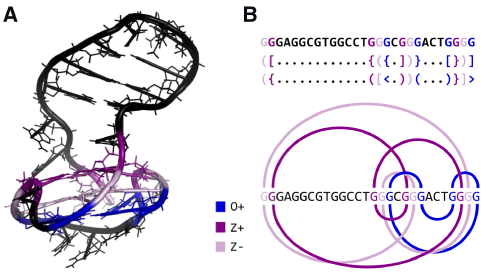


Fig. 6. (**A**) 3D structure of major G-quadruplex form in HIV-1 LTR (PDB id: 6H1K) (Butovskaya *et al.*, 2018) with colored tetrads, and (**B**) its secondary structure topology in dot-bracket and arc diagram representation. The colors in the secondary structure representations indicate the tetrad types

quadruplexes. In particular, the arc diagram makes it easy to trace which residues interact within every tetrad. Arcs have been color-coded. The color of each tetrad denotes the category: blue is for O+, dark cyan for N+ and aquamarine for N-.

The second example presents G-quadruplex from HIV-1 LTR (PDB id: 6H1K) (Butovskaya *et al.*, 2018). Its three-dimensional structure has been determined in NMR experiment. The quadruplex is built of three tetrads (Fig. 6). The first one belongs to category Z- (lilac arcs in Fig. 6B), the second one is in group Z+ (purple arcs) and the third tetrad has been classified as O+ (blue arcs). Therefore, the motif has been assigned to category M that collects irregular structures.

5 Conclusion

Quadruplexes are one of the intensely studied structural motifs that form in nucleic acids. Their popularity results from quite wide potential applications of these structures in biomedical sciences. Therefore, the studies of their diverse architectures and functions are conducted by the research teams worldwide.

Until now, three different approaches existed to describe and classify quadruplex structures. One was based on sequences and G-tracts that contributed to building the G4 motif (Garant *et al.*, 2015, 2017). The other focused on the tertiary structure and took into account the conformation of loops between G-stems (Burge *et al.*, 2006; Dvorkin *et al.*, 2018). Finally, glycosidic bond angles were the reference for the third taxonomy that made possible the description of the relationship between type of loops and groove width of a quadruplex stem (Karsisiotis *et al.*, 2013; da Silva, 2007). In our approach, we have considered both tetrads and quadruplexes. We have analyzed the secondary structure topology of these motifs to propose new ONZ classification for tetrads. Further, our study has encompassed unimolecular quadruplexes, for which we have also defined ONZ-based categories. Our taxonomy has been accompanied by unique ways to encode the considered motifs in dot-bracket notation and represent them in arc diagrams. They reveal the diversity of tetrad and quadruplex topologies, even inside ONZ groups. We believe that the presented approach will enrich the knowledge about tetrads and quadruplexes, and allow for easier in-depth analysis of their characteristics.

Funding

This research was supported by the National Science Centre, Poland [2016/23/B/ST6/03931] and Mloda Kadra project [09/91/SBAD/0684] from Poznan University of Technology, and carried out in the European Centre for Bioinformatics and Genomics (Poland).

Conflict of Interest: none declared.

References

Antczak,M. *et al.* (2014) RNApdbee – a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Res.*, **42**, W368–W372.

Antczak,M. *et al.* (2018) New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics*, **34**, 1304–1312.

Bakalar,B. *et al.* (2019) A minimal sequence for left-handed G-quadruplex formation. *Angew. Chem. Int. Ed.*, **58**, 2331–2335.

Bedrat,A. *et al.* (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Rese.*, **44**, 1746–1759.

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhattacharya,S. *et al.* (2019) Going beyond base-pairs: topology-based characterization of base-multiplets in RNA. *RNA*, **25**, 573–589.

Burge,S. *et al.* (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.

Butovskaya,E. *et al.* (2018) Major G-quadruplex form of HIV-1 LTR reveals a (3 + 1) folding topology containing a stem-loop. *J. Am. Chem. Soc.*, **140**, 13654–13662.

Cammas,A. and Millevoi,S. (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res.*, **45**, 1584–1159.

Chambers,V.S. *et al.* (2015) High-throughput sequencing of DNA g-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.

Colasanti,A.V. *et al.* (2013) Analyzing and building nucleic acid structures with 3DNA. *J. Vis. Exp.*, **74**, e4401.

da Silva,M.W. (2007) Geometric formalism for DNA quadruplex folding. *Chem. Eur. J.*, **13**, 9738–9745.

Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.

Dvorkin,S.A. *et al.* (2018) Encoding canonical DNA quadruplex structure. *Sci. Adv.*, **4**, eaat3007.

Fay,M.M. *et al.* (2017) RNA G-quadruplexes in biology: principles and molecular mechanisms. *J. Mol. Biol.*, **429**, 2127–2147.

Garant,J.-M. *et al.* (2015) G4RNA: an RNA G-quadruplex database. *Database*, **2015**,

Garant,J.-M. *et al.* (2017) Motif independent identification of potential RNA g-quadruplexes by G4RNA screener. *Bioinformatics*, **33**, 3532–3537.

Gudanis,D. *et al.* (2016) Structural characterization of a dimer of RNA duplexes composed of 8-bromoguanosine modified CGG trinucleotide repeats: a novel architecture of RNA quadruplexes. *Nucleic Acids Res.*, **44**, 2409–2416.

Huppert,J. (2008) Hunting G-quadruplexes. *Biochimie*, **90**, 1140–1148.

Karsisiotis,A.I. *et al.* (2013) DNA quadruplex folding formalism – a tutorial on quadruplex topologies. *Methods*, **64**, 28–35.

Kikin,O. *et al.* (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.

Kikin,O. *et al*. (2007) GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. *Nucleic Acids Res*., **36**, D141–D148.

Kotar,A. *et al*. (2019) Two-quartet kit* g-quadruplex is formed via double-stranded pre-folded structure. *Nucleic Acids Res*., **47**, 2641–2653.

Kwok,C.K. and Merrick,C.J. (2017) G-quadruplexes: prediction, characterization, and biological application. *Trends Biotechnol*., **35**, 997–1013.

Lai,D. *et al*. (2012) R-chie: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res*., **40**, e95–e95.

Lorenz,R. *et al*. (2012) RNA folding algorithms with G-quadruplexes. In: *Advances in Bioinformatics and Computational Biology*. Springer, Berlin, Heidelberg, pp. 49–60.

Lorenz,R. *et al*. (2013) 2d meets 4G: G-quadruplexes in RNA secondary structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **10**, 832–844.

Lu,X.-J. *et al*. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res*., **43**, 142.

Malgowska,M. *et al*. (2016) Overview of RNA G-quadruplex structures. *Acta Biochimica Polonica*, **63**, 609–621.

Marsico,G. *et al*. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*., **47**, 3862–3874.

Marušič,M. and Plavec,J. (2015) The effect of DNA sequence directionality on G-quadruplex folding. *Angew. Chem. Int. Ed*., **54**, 11716–11719.

Mishra,S.K. *et al*. (2016) G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep*., **6**, 38144.

Pandey,S. *et al*. (2015) The RNA stem-loop to G-quadruplex equilibrium controls mature microRNA production inside the cell. *Biochemistry*, **54**, 7067–7078.

Parkinson,G.N. *et al*. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876–880.

Popenda,M. *et al*. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res*., **36**, D386–D391.

Reche,E.B. and Morales,J.C. (2019) G4-iM grinder: DNA and RNA G-quadruplex, i-Motif and higher order structure search and analyser tool. *bioRxiv*.

Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res*., **43**, 8627–8637.

Sahakyan,A.B. *et al*. (2017) Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep*., **7**, 14535.

Schrödinger,L.L.C. (2015) The PyMOL Molecular Graphics System, version 1.8.

Szachniuk,M. (2019) RNApolis: computational platform for RNA structure analysis. *Found. Comput. Decision Sci*., **44**, 241–257.

Tan,W. *et al*. (2016) Probing the G-quadruplex from hsa-miR-3620-5p and inhibition of its interaction with the target sequence. *Talanta*, **154**, 560–566.

Trachman,R.J. *et al*. (2017) Structural basis for high-affinity fluorophore binding and activation by RNA mango. *Nat. Chem. Biol*., **13**, 807–813.

Trajkovski,M. *et al*. (2012) Unique structural features of interconverting monomeric and dimeric G-quadruplexes adopted by a sequence from the intron of the N-myc gene. *J. Am. Chem. Soc*., **134**, 4132–4141.

Wang,Y. and Patel,D.J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*, **1**, 263–282.

Yadav,V. *et al*. (2017) G quadruplex in plants: a ubiquitous regulatory element and its biological relevance. *Front. Plant Sci*., **8**, 1163. doi: 10.3389/fpls.2017.01163.

Zok,T. *et al*. (2018) RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res*., **46**, W30–W35.

Zok,T. *et al*. (2019) ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *submitted for publication*.

SUPPLEMENTARY MATERIAL

Topology-based classification of tetrads and quadruplex structures

Mariusz Popenda[1], Joanna Miskiewicz[2], Joanna Sarzynska[1], Tomasz Zok[2,3] and Marta Szachniuk[1,2]*

[1]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
[2]Institute of Computing Science, and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[3]Poznan Supercomputing and Networking Center, Jana Pawla II 10, 61-139 Poznan, Poland

*corresponding author: marta.szachniuk@cs.put.poznan.pl

TABLE OF CONTENTS

Table S1. PDB-deposited structures containing unimolecular, *bimolecular* and tetramolecular quadruplexes (as of 18.04.2019). **3**

Figure S1. PDB structures containing quadruplexes. **4**

Table S2. Mixed N-quartets found in the dataset. **4**

Figure S2. Arc diagrams showing different topologies of regular 2-tetrad quadruplexes of (A) parallel, (B) antiparallel, and (C) hybrid type. Leftmost column presents diagrams for G4s with both tetrads in O+ class, middle column – N+ class, rightmost column – Z+ class. **5**

Table S1. PDB-deposited structures containing **unimolecular**, *bimolecular* and tetramolecular quadruplexes (as of 18.04.2019).

DNA					RNA	Other
139D	1PHJ	**2KZE**	*3EM2*	**5J4W**	1J6S	1S9L
143D	**1QDF**	**2L7V**	*3EQW*	**5J6U**	1J8G	2CHK
148D	**1QDH**	**2L88**	*3ERU*	**5LIG**	1MDG	4L0A
156D	1QDI	**2LBY**	*3ES0*	**5LQG**	**1MY9**	**4WB2**
186D	1QDK	**2LD8**	*3ET8*	**5LQH**	1N7A	**4WB3**
1A6H	**1RDE**	**2LE6**	*3EUI*	5LS8	1N7B	
1A8N	1S45	*2LED*	*3EUM*	*5M1L*	1RAU	
1A8W	1S47	**2LEE**	*3NYP*	5M2L	2AWE	
1AFF	1U64	**2LK7**	*3NZ7*	**5MBR**	2GRB	
1BUB	**1XAV**	**2LOD**	*3QCR*	**5MCR**	*2KBP*	
1C32	1XCE	**2LPW**	**3QLP**	**5MJX**	**2LA5**	
1C34	**1Y8D**	**2LXQ**	*3QSC*	**5MTA**	*2M18*	
1C35	**201D**	**2LXV**	*3QSF*	**5MTG**	**2RQJ**	
1C38	**230D**	**2LYG**	**3QXR**	**5MVB**	**2RSK**	
1D59	244D	*2M1G*	**3R6R**	**5NYS**	**2RU7**	
1D6D	**2A5P**	**2M27**	**3SC8**	**5NYT**	*3IBK*	
1EEG	**2A5R**	**2M4P**	**3T5E**	**5NYU**	**3JBV**	
1EMQ	2AKG	**2M53**	3TVB	**5O4D**	*3MIJ*	
1EVM	2AQY	**2M6V**	3UGO	**5OPH**	**4KZD**	
1EVN	2AVH	**2M6W**	3UGP	**5OV2**	**4KZE**	
1EVO	2AVJ	**2M8Z**	**3UYH**	**5UA3**	**4Q9Q**	
1F3S	2CHJ	**2M90**	**4DA3**	**5VHE**	**4Q9R**	
1FQP	**2E4I**	**2M91**	**4DAQ**	**5W77**	4RJ1	
1HAO	**2F8U**	**2M92**	**4DIH**	**5YEY**	4RKV	
1HAP	**2GKU**	**2M93**	**4DII**	**5Z80**	4RNE	
1HUT	2GW0	**2MAY**	**4FXM**	**5Z8F**	*4TS0*	
1I34	2GWE	**2MB2**	**4G0F**	**5ZEV**	*4TS2*	
1JB7	*2GWQ*	**2MB3**	*4H29*	*6A7Y*	4XK0	
1JPQ	*2HBN*	**2MB4**	**4LZ1**	6A85	**5BJO**	
1JRN	*2HRI*	**2MBJ**	**4LZ4**	**6AC7**	**5BJP**	
1JVC	**2HY9**	**2MCC**	**4NI7**	**6AU4**	**5DE5**	
1K4X	**2IDN**	**2MCO**	**4NI9**	**6CCW**	**5DE8**	
1K8P	**2JPZ**	**2MFT**	*4P1D*	**6EO6**	**5DEA**	
1KF1	**2JSK**	**2MFU**	4R44	**6EO7**	**5IWA**	
1L1H	**2JSL**	**2MGN**	*4R45*	**6ERL**	**5OB3**	
1LVS	**2JSM**	2MS6	*4R47*	**6EVV**	**5V3F**	
1MYQ	**2JSQ**	**2MS9**	**4U5M**	**6F4Z**	**6B14**	
1NP9	2JT7	**2MWZ**	4U92	**6FC9**	**6B3K**	
1NYD	2JWQ	**2N21**	**4WO2**	**6FFR**	**6C63**	
1NZM	*2KAZ*	**2N2D**	**4WO3**	**6FQ2**	**6C64**	
1O0K	**2KF7**	**2N3M**	**5CCW**	**6FTU**	**6C65**	
1OZ8	**2KF8**	**2N4Y**	*5CDB*	**6GH0**	**6E8S**	
1PA6	**2KKA**	**2N60**	**5CMX**	**6GN7**	**6E8T**	
1PH1	**2KM3**	**2N6C**	**5DWW**	**6GZN**	**6E8U**	
1PH2	**2KOW**	*2N9Q*	**5DWX**	**6H1K**	6GE1	
1PH3	**2KPR**	**2O3M**	**5EW1**	**6H5R**		
1PH4	**2KQG**	2O4F	**5EW2**	**6IA0**		
1PH5	**2KQH**	*2WCN*	5G35	**6IA4**		
1PH6	2KVY	352D	*5HIX*	**6JKN**		
1PH7	*2KYO*	*3CCO*	**5I2V**	**6NEB**		
1PH8	**2KYP**	**3CDM**	**5J05**			
1PH9	**2KZD**	*3CE5*	**5J4P**			

Figure S1. PDB structures containing quadruplexes.

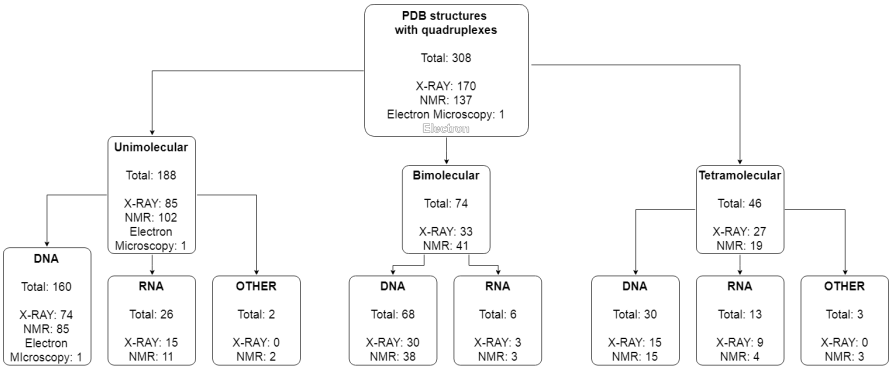
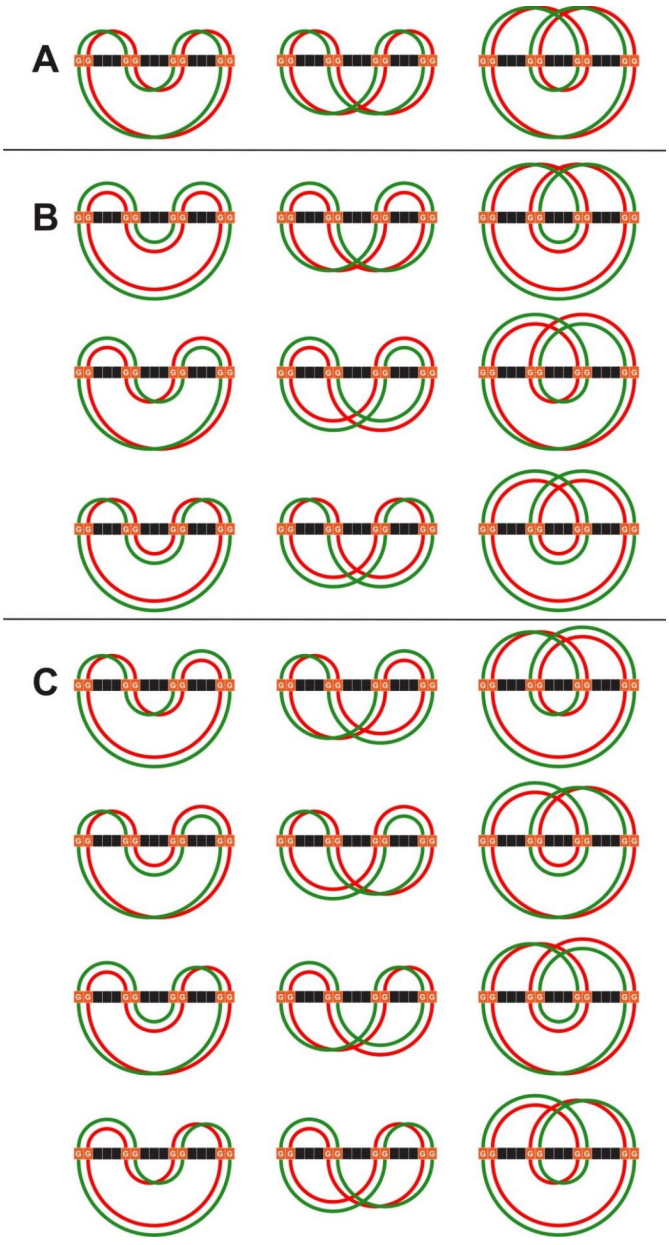


Table S2. Mixed N-quartets found in the dataset.

Tetrad	CGCG	AGAG	ATAT	CGAT	GGAT	AAGT	CGAU	GACC	GCCC
DNA	12	8	4	4	1	1	0	0	0
RNA	0	0	0	0	0	0	7	1	1

Figure S2. Arc diagrams showing different topologies of regular 2-tetrad quadruplexes of (A) parallel, (B) antiparallel, and (C) hybrid type. Leftmost column presents diagrams for G4s with both tetrads in O+ class, middle column – N+ class, rightmost column – Z+ class.



Briefings in Bioinformatics, 22(3), 2021, 1–14

doi: 10.1093/bib/bbaa201
Method Review

How bioinformatics resources work with G4 RNAs

Joanna Miskiewicz, Joanna Sarzynska and Marta Szachniuk

Corresponding author: Marta Szachniuk, Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965 Poznan, Poland. Tel.: +48 61 6653030; Fax: +48 61 8771525. E-mail: mszachniuk@cs.put.poznan.pl

Abstract

Quadruplexes (G4s) are of interest, which increases with the number of identified G4 structures and knowledge about their biomedical potential. These unique motifs form in many organisms, including humans, where their appearance correlates with various diseases. Scientists store and analyze quadruplexes using recently developed bioinformatic tools—many of them focused on DNA structures. With an expanding collection of G4 RNAs, we check how existing tools deal with them. We review all available bioinformatics resources dedicated to quadruplexes and examine their usefulness in G4 RNA analysis. We distinguish the following subsets of resources: databases, tools to predict putative quadruplex sequences, tools to predict secondary structure with quadruplexes and tools to analyze and visualize quadruplex structures. We share the results obtained from processing specially created RNA datasets with these tools.
Contact: mszachniuk@cs.put.poznan.pl
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

Key words: quadruplexes; RNA; databases; bioinformatics tools; structure analysis; PQS prediction

Introduction

G-quadruplexes (G4s) are non-canonical, four-stranded structures that form in guanine-rich nucleic acids. The basic structural unit of G4 is a G-tetrad: four guanines in the planar arrangement that interact with one another via hydrogen bonds. Quadruplexes can also contain non-G based quartets; about 20% of tetrads found in PDB-deposited quadruplex structures contain other nucleotides than guanine [1, 2]: U-tetrads, mixed ATAT, GCGC tetrads, etc. [2–6]. If n ($n \geq 2$) neighboring tetrads stack with one another, they create an n-layer quadruplex. Two or more G4s can associate through stacking interactions between their outer tetrads to form higher-order multimers, ranging from dimers to G-wires [7, 8]. All these structural motifs have several topological characteristics, the most important of which are strand orientation in the stem, the number of tetrads, loops' arrangement and groove dimension. One, two or four individual strands—linked by loops—participate in the quadruplex formation. They have parallel, antiparallel or hybrid orientation [9], which correlates to anti/syn conformation of guanines in the tetrad planes, base stacking geometry and loop types [10–12].

Scientists observed the very first biologically relevant G4s in eukaryotic chromosomal telomeric DNA. This discovery initiated the research into the role and distribution of quadruplexes in the genome. Chemical biology methods were developed to map G4 folding *in vitro*, like G4-seq [13], and *in vivo*, like G4ChIP-seq [14], in the genomes of various species, including humans [15,

Joanna Miskiewicz is a PhD student and a member of Laboratory of RNA Structural Bioinformatics, Poznan University of Technology. Her research interests include structural bioinformatics, motif identification and algorithms for RNA biology. She holds an MSc in bioinformatics (2015).
Joanna Sarzynska is a research associate at IBCh PAS. Her research interests include RNA structure, molecular dynamics and structural bioinformatics. She has authored papers published in top scientific journals on life sciences and holds a PhD in chemistry (1997).
Marta Szachniuk is a professor of technical sciences, vice-president of the Polish Bioinformatics Society and vice chair of EURO CBBM. Her research interests include algorithms for structural biology, operations research and AI. She is an author of highly cited papers and over 20 bioinformatics tools. She holds a PhD (2005) and a DSc (2015) in computing science, ProfTit (2020).
Submitted: 25 June 2020; **Received (in revised form):** 3 August 2020

16]. First genomic maps of DNA G4s appeared—a starting point for clarifying the purpose of quadruplex formation [17]. Many sequences with G4 potential turned out to be associated with cancer or neurodegenerative diseases and became an attractive target for molecular therapeutics [18]. Extensive research on the DNA quadruplexes generated interest in RNA G4s. RNA G-quadruplex sequencing protocols—rG4-seq [19, 20], G4RP-seq [21]—were developed for detecting RNA quadruplexes on a transcriptome-wide level. The studies revealed RNA G4s in coding and noncoding fragments of mRNA [22], mature human miRNA [23], other premature and mature noncoding RNAs (including telomeric RNA) and aptamers [24]. RNA quadruplexes showed to regulate pre-mRNA processing, control RNA localization and mechanisms like mRNA translation, miRNA biogenesis or protein binding [25]. G4 in 5'UTR of mRNAs proved the ability to repress translation *in vitro* [26]. Experiments indicated that under certain conditions—the presence of salts or ligands—pre-miRNA sequences could adopt the G4 structure avoiding the Dicer-mediated maturation and leading to subexpression of miRNA levels [27]. A structural study led to the determination of first quadruplex folds via X-ray [28] and nuclear magnetic resonance (NMR) [29]. Currently, Protein Data Bank collects about 300 G4-rich nucleic acid structures, 75% of them from DNAs [2, 30, 31]. Their analysis showed differences between DNA and RNA quadruplexes. DNA G4s are structurally polymorphic and thermodynamically less stable than their RNA equivalents [1, 32]. RNA quadruplexes are predominantly parallel; ribose favors anti-oriented guanines which results in the strands' parallelism.

The research into quadruplexes resulted in an increased demand for computer resources dedicated to these motifs. Bioinformatics responded with tools for DNA G4s that pretended to process quadruplexes regardless of the molecule type. They aimed to identify G4 forming sequences, model and analyze the secondary and tertiary structures, simulate molecular dynamics, calculate free energy or perform molecular docking [33]. Many resources for predicting G4 formation and stability, available before 2010, were reviewed in [34]. With the increasing number of experimental data on quadruplexes, including more G4-rich high-resolution structures and genomic maps of DNA G4s, new algorithms appeared [15, 16]. They go beyond canonical G4 predictions and apply non-trivial computational techniques, including machine learning (ML) [35–37]. Lombardi and Londoño-Vallejo [33] present a comprehensive overview of modern open-source software for G-quadruplex detection tested on a set of G4 DNAs verified experimentally *in vitro*.

In this paper, we focused on RNA quadruplexes. We described 35 bioinformatics resources that addressed quadruplexes and checked how they worked with RNA G4s. The set included 14 tools from [33] and 21 others. We grouped all programs into four categories: (i) databases, (ii) sequence-based tools predicting G4 location in the sequence, (iii) secondary structure prediction tools and (iv) secondary and tertiary structure-based tools analyzing and visualizing quadruplexes. We performed their tests on specially created datasets. We hope that our review will be helpful to scientists studying the specifics of RNA G4s.

Methods

In preparing this review, we searched global resources to create a potentially complete list of databases and analytical tools used in quadruplex research. As a result, the list includes 16 data repositories, 14 tools that predict quadruplex-forming sites in nucleic acid sequence, 1 tool that predicts the secondary structure and annotates potential location of quadruplexes and 4 tools that analyze the secondary and tertiary structure and visualize quadruplex topology (Figure 1). Twenty-one resources with web interfaces were tested by us mostly via a web browser. The remaining ones were downloaded, configured and run locally. In most cases, we applied the default input settings.

Databases with G4-related data

Currently, there exist 16 databases, which store information concerning quadruplexes. They fall into three categories: databases that collect primary or tertiary structures with experimentally verified G4s (DSSR-G4DB, G4IPDB, G4LDB, G4RNA, Lit392 and Lit638); databases storing data from high-throughput sequencing with mapped quadruplexes (GSE63874, GSE77282, GSE110582 and GSE129281); and databases of sequences with G4s identified *in silico* (Greglist, GRSDB2, G4-virus, Non-B DB v2.0, PlantGQ and QuadBase2). We describe them briefly in the following paragraphs and define the following features in Table 1: DNA and RNA indicate whether the database collects DNA and RNA sequences; G4 verification denotes whether quadruplexes are verified experimentally or predicted *in silico*; G4 sequence informs if the quadruplex sequence is available; the number of G4s gives the number of stored quadruplex sequences (as of 21 March 2020); DB records specifies the number and type of database entries (as of 21 March 2020); customized search shows whether the database has a search engine that allows to search its records with different criteria; web interface specifies if the database is web-interfaced; visual output indicates whether any visualization of the output data is available.

DSSR-G4DB [38] contains quadruplex nucleic acid structures found by DSSR in the Protein Data Bank [30], currently 354 entries. The data are annotated. Users can find information about G-tetrads, G4 helices and G4-stems and visualize the 3D models of G4 structures. Availability: webserver (http://g4.x3dna.org). Recent update: 5 June 2020.

G4IPDB [39] is a database of over 200 proteins interacting with DNA and RNA G-quadruplexes, based on the literature data. For each entry, it contains the G4 sequence, interacting protein name, and UniProt ID, the details of the interaction, PubMed ID of the paper being the source of information. Users browse the data and query the database by specifying G4IPDB interaction ID, DNA/RNA target name, gene name, etc. Availability: webserver (http://bsbe.iiti.ac.in/bsbe/ipdb). Updated: twice a year.

G4LDB [40] collects ligands (currently, over 800) that interact with G-quadruplexes. Each entry contains information about chemical structure, targeted G4 sequence, physical properties of a ligand and literature references. The 3D model of every ligand is also available. Users browse the database and search for ligands by defining their structure, ligand properties, ligand activity fields or bibliographic information. Availability: webserver (http://www.g4ldb.org/ci2/index.php). Updated: twice a year.

G4RNA [36] stores published human RNA sequences, processed experimentally. This collection includes sequences with confirmed G-quadruplexes and sequences confirmed not to form G4s, currently, 567 entries in total. The system allows running a keyword-driven and position-driven search. Keywords include the G4 gene symbol or sequence (IUPAC-encoded or regular expression), the experiment name, reference paper and DOI. Position-driven search needs specifying at least one chromosome to contain G4. The results fall into four categories: sequences (nine options), experiments (four options), predictions (seven options) and QGRS Mapper (nine options). Users can choose to display the secondary structure

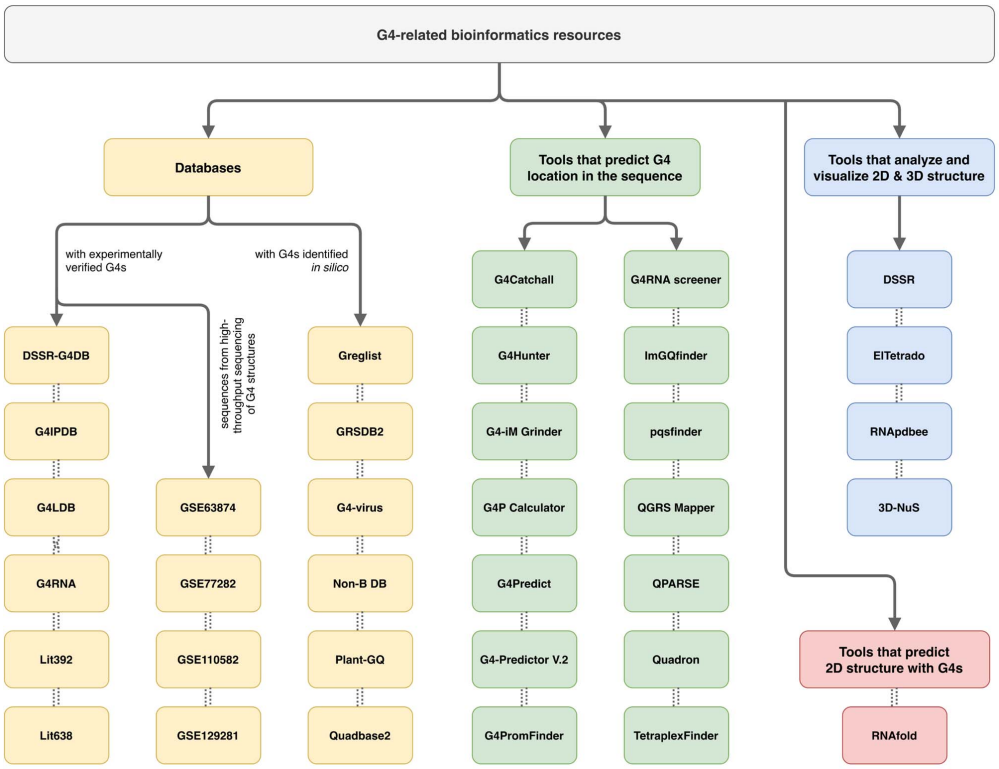


Figure 1. Bioinformatics resources for quadruplexes.

predicted by RNAfold with quadruplex annotated in the dot-bracket representation. The output data can be download in the xls file. Availability: webserver (http://scottgroup.med.usherbrooke.ca/G4RNA). Updated: monthly.

Lit392 [41] is a set of 392 DNA and RNA sequences for which the formation of G4s was experimentally confirmed (298 sequences) or disproven (94 sequences). The set mainly includes published sequences and several unpublished ones resulting from the experiments performed by the authors. The database was created to test the performance of G4Hunter. Availability: supplementary file to [41] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770238/bin/supp_gkw006_nar-02998-f-2015-File003.zip). Updated: no.

Lit638 [42] combines data from Lit392 and G4RNA databases. It contains 638 DNA and RNA sequences: 506 confirmed to form G4s and 132 non-forming quadruplexes. The database was created to test the performance of QPARSE method. Availability: upon request to authors. Updated: no.

GSE63874 [15] is a map of distinct canonical and non-canonical G4s, experimentally confirmed to form in the human genome. It consists of 716,310 DNA G4s, obtained from the high-throughput G4-seq method. The genomic DNA template was sequenced twice, first with Na^+ and second with K^+. Binding to K^+ cations enhances the structural stability of G4. Availability: BED files on GEO webserver (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63874). Updated: no.

GSE77282 [19] is a map of canonical and non-canonical G4s found in the human transcriptome. RNA G4s were identified *in vitro* by the rG4-seq method—over 3000 in the presence of K^+ cation or the PDS ligand. Availability: BED files on GEO webserver (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77282). Updated: no.

GSE110582 [16] is a map of DNA G4s, identified in the genomes of 12 species (model organisms—including human—and clinically important pathogens). The data—1,420,841 G-quadruplexes—come from high-throughput sequencing applying G4-seq2 (improved G4-seq method) with K^+, Li^+ and PDS ligand as the sequencing buffer. Availability: BED files on GEO webserver (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110582. Updated: no.

GSE129281 [43] is a map of G4 structures found in the genomes of *Pseudomonas aeruginosa* and *Escherichia coli*. The rG4-seq method, in the presence of K^+ or Li^+ cations, revealed 329 RNA G4 sites—168 in E. coli and 161 in P. aeruginosa. Availability: BED files on GEO webserver (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129281). Updated: no.

Greglist [44] is a database of potential G-quadruplex regulated genes in the genomes of various species, including humans. Users browse through the database or search for stored DNA putative quadruplex sequences (PQSs). Database records contain, among others, Ensembl ID, gene name, organism, number of PQS, G4 sequence and distance to TSS. PQSs were

Table 1. Selected features of G4-related databases

Database	DNA	RNA	G4 verification	G4 sequence	Number of G4s	DB records	Customized search	Web interface	Visual output
DSSR-G4DB	✓	✓	experimental	✓	354	354 (PDB structures)	✓	✓	✓
G4IPDB	✓	✓	experimental	✓	no data	216 (interactions)	✓	✓	✓
G4LDB	✓	✓	experimental	✓	no data	> 800 (ligands)	✓	✓	✓
G4RNA	✓	✓	experimental	✓	321	567 (human RNA sequences)	✓	✓	
Li392	✓	✓	experimental	✓	298	392 (DNA and RNA sequences)			
Li638	✓	✓	experimental	✓	506	638 (DNA and RNA sequences)			
GSE63874	✓		experimental		716,310	32 million (reads)			
GSE77282		✓	experimental		3383	1.15 billion (reads)			
GSE110582	✓		experimental		1,420,841	7675.39 Mb			
GSE129281		✓	experimental		329	3505 (hits)			
Greglist	✓		in silico	✓	no data	115442 (genes)	✓	✓	✓
GRSDB2		✓	in silico	✓	3,255,075	29,288 (genes)	✓	✓	✓
G4-virus	✓	✓	in silico		47	248 (viruses)		✓	✓
Non-B DB v2.0	✓		in silico	✓	3,864,596	12 (mammalian genomes)	✓	✓	✓
Plant-GQ	✓		in silico	✓	626,341,645	195 (plants)	✓	✓	✓
QuadBase2	✓		in silico	✓	no data	1897 (species)	✓	✓	✓

found using Quadparser [52] with $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ and $C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$ (to find quadruplexes on complementary strand). Availability: webserver (http://tubic.tju.edu.cn/greglist). Recent update: 19 October 2007.

GRSDB2 [45] stores PQS positions in computationally processed over 29,000 eukaryotic pre-mRNA sequences of eight organisms, including humans. QGRS Mapper, run to search for $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$ motif, identified 3,015,683 PQS for $x = 2$ and motif length ≤ 30 nts, and 239,392 PQS for $x = 3$ and motif length ≤ 45 nts. Availability: webserver (http://bioinformatics.ramapo.edu/GRSDB2/). Updated: no.

G4-virus [46] is a collection of PQS locations in the human viruses' genomes. Users search viruses by name or browse the database by virus class. Deposited data consist of PQS identified computationally on both strands, PQS positions in viral genomes, conservation degrees among different strains and statistical data for PQS. The system provides a graphical visualization of the PQS arrangement. Availability: webserver (http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus). Recent update: 30 July 2019.

Non-B DB v2.0 [47] contains non-B DNA structure predictions of 3,864,596 quadruplex forming motifs in 12 mammalian genomes, including the human genome. The database system allows searching by features or attributes. In plain feature search, users select species, classes, chromosome or gene type, start and stop position of the chromosome, query type (specifying how many features and where they should be found) and feature type (for example, quadruplex). Attribute search offers additional options: composition, sequence and tracts. Users can submit their data of non-B DNA motifs. The visualization of non-B DNA motifs for each organism is provided. Availability: webserver (https://nonb-abcc.ncifcrf.gov/apps/site/default). Recent update: 13 June 2012.

Plant-GQ [48] collects 626,341,645 DNA PQS from 195 plant species: 610,897,949 of them are two-tetrad G4s; 14,326,347 are three-tetrad G4s; 1,117,349 are four-tetrad G4s. The data were obtained by searching for motifs matching the following pattern $G_xN_yG_xN_yG_xN_yG_x$, where $x \in [2, 4]$, $y \in [1, 10]$. Database entries include PQS sequences and positions within the genome. Availability: webserver (http://biodb.sdau.edu.cn/plantgq/index.php). Updated: no.

QuadBase2 [49, 50] is a database of DNA PQS found in 178 species of eukaryotes (EuQuad module) and 1719 species of prokaryotes (ProQuad module). The data result from searching for $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$ pattern. Users introduce the query by selecting a group of organisms, the species, gene ID, region of interest (gene body, around TSS, CDS and UTRs, strand), PQS distance from TSS, search algorithm (greedy/non-greedy), bulge size, etc. Additional search parameters for G4s concern the stringency level, which can be low (two G-tetrads, loop size 1–12), medium (three G-tetrads, loop size 1–7) or high (three G-tetrads, loop size 1–3). Resulting PQSs are shown in circular histograms. Availability: webserver (http://quadbase.igib.res.in/). Updated: no.

Tools that predict G4 location in the sequence

The first bioinformatics tools, searching for the existence of PQS motifs, were published in 2004–2005 [51, 52]. To this day, several such programs appeared. They differ in the search algorithm, the searched pattern and the scoring function; for example, some tools apply an ML approach, others look for sequential motifs that match different regular expressions. In Table 2, we show these tools' options important from the user point of view.

Columns in the table: DNA and RNA indicate whether the tool accepts DNA and RNA sequences; multiple entry checked means that the program allows entering many sequences for a single run; allow mismatches denotes if the tool accepts mismatches and bulges in a G-tract/PQS; number of tetrads in G4 informs about limit for the number of G-tetrads per quadruplex; loop length specifies the limit imposed on every loop involved in quadruplex formation; max PQS' length is the program restriction for PQS length; accept overlapping indicates whether the tool searches for overlapping PQS; non-G PQS informs about the possibility to search for non-G-based quadruplexes; strands tells whether the tool provides the results only for the input sequence (+) or for both—the input and the complementary strand (+/−); show G4 position informs if the tool gives the PQS location within the sequence. Different search patterns with examples are presented in the Supplementary Materials (Table S5).

G4Catchall [53] looks for conservative G4s by fitting into the regular expression and user-defined parameters. Sequences that lack typical, uninterrupted G4s (pattern: Gn, where n denotes the number of guanines) are scanned with more complicated patterns: $G_xN_yG_{(n-x)}$ or $G_xN_yG_{(n-x-1)}$, where $x \in [1, n)$ (a detailed explanation of search patterns can be found in the Supplementary Materials). A definition of the G4 motif includes bulged G-tract, mismatched tract that contains non-guanine nucleotides and loops between tracts—one extreme loop, up to 50 nucleotides, is allowed. Users can provide minimum G-tract length (2 or 3), loop length (1–15), extreme loop length (1–50), the number of allowed bulges and mismatches (0–2), complementary strand search, overlapping G4s merge and flanking nucleotides inclusion. Input data formats: raw sequence, FASTA. Availability: Python script, web application (http://homes.ieu.edu.tr/odoluca/G4Catchall).

G4Hunter [41, 54] predicts PQS in DNA or RNA sequence based on its G-richness and G-skewness, meaning the fraction of Gs in the sequence and G/C ratio between the strands, respectively. The program allows users to define the size of a search window (web default: 25) and the threshold value (web default: 1.2). In the input sequence, each nucleotide and nucleotide aggregation has arbitrary assigned value: score(G) = 1, score(GG) = 2, score(GGG) = 3 and score(GGGG...) = 4. Cytosines have the opposite values: score(C) = −1, score(CC) = −2, score(CCC) = −3, and score(CCCC...) = −4. Other nucleotides are not counted. Input data format: FASTA. Availability: Python script, web application (http://bioinformatics.ibp.cz:8888/#/analyse/quadruplex). Web application stores users results in a relational database.

G4-iM Grinder [55] finds and characterizes quadruplexes and i-Motifs within a given DNA or RNA sequence. It uses a two-part search engine with 13 customizable functions (e.g. showing PQS on both strands, loop sequence and size). The first search method (M1A) identifies uninterrupted and bulged nucleotide runs. Users define the length of the run, the size of the bulge and the nucleotide the run is composed of (any nucleobase can be used to form a run). Discovered runs are passed to the second search method (M1B). The method finds a correlation between runs, starting with closest runs in a sequence and broadening the search if necessary. The distance between the runs is limited by the users. The search results are further analyzed by methods, one for overlapping and size-dependent PQS search (M2) and second for non-overlapping size-independent search (M3). M2 links the runs for final structure depending on user-defined limitations for the number of connected runs, the length of the final motif sequence and the number of bulges within the sequence. The linkage process is followed by the frequency count of the final structure within the input sequence. M3 searches for higher-order structures based on unlinked runs from the M1B

Table 2. Selected features of PQS prediction tools

Tool	DNA	RNA	Multiple entry	Allow mismatches	Number of tetrads in G4	Loop length	Max PQS'	Accept overlapping	Non-G PQS	Strands	Show G4 position
G4Catchall	✓	✓	✓	✓	≥ 2	1–50	99	✓		+/−	✓
G4Hunter	✓	✓		✓	≥ 0	≥ 0		✓		+/−	✓
G4-iM Grinder	✓	✓		✓	3	≥ 0		✓	✓	+/−	✓
G4P Calculator	✓		✓	✓	≥ 0	≥ 0		✓		+	
G4Predict	✓	✓		✓	≥ 0	≥ 0		✓		+/−	✓
G4-Predictor V.2	✓	✓			2–6	0–36	45			+/−	✓
G4PromFinder	✓				2–4	1–10	30			+/−	
G4RNA screener		✓	✓	✓	≥ 0	≥ 0		✓		+	✓
ImGQfinder	✓	✓		✓	2–10	≤ 25		✓	✓	+	✓
pqsfinder	✓	✓		✓	2–20	0–9		✓		+/−	✓
QGRS Mapper	✓	✓			2–6	0–36	45	✓		+	✓
QPARSE	✓	✓	✓	✓	≥ 2	≥ 0		✓	✓	+	✓
Quadron	✓				≥ 3	1–12		✓		+/−	✓
TetraplexFinder	✓	✓	✓	✓	2–5	1–50	170	✓		+/−	✓

method. After connecting the runs for higher-order structures, it calculates their frequencies in the input sequence. After the analysis, G4-iM Grinder counts predefined patterns provided by the users (both single and multiple nucleobase patterns are accepted) in all found PQS. Program can also compare found PQS with validated *in vitro* G4 structures. Evaluation of the results is prepared based on specified scoring methods, G4Hunter, PQSfinder (incorporated and modified using ML method) and/or cGcC. The final score is computed as a weighted average of the selected scoring systems. Input data format: FASTA. Availability: R package (https://github.com/EfresBR/G4iMGrinder).

G4P Calculator [56] finds PQS based on the density of guanine runs in a nucleic acid sequence. Although designed for DNA, it accepts also RNA sequences. The algorithm moves window frames along the sequence and counts frames that meet the specified criteria. Users can define window size, window shift, the minimum length of the G-run and the minimum number of G-runs per window. Default settings: window size = 100 nts, window shift = 20 nts, G-run $\geq$ 3 and G-runs per window $\geq$ 4. Input data format: FASTA. Availability: standalone software (http://depts.washington.edu/maizels9/G4calc.php).

G4Predict searches for intramolecular and intermolecular G4s based on a sequence motif. It extends the functionality of Quadparser [52]. The pattern to search for intramolecular G4s is defined as $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$. Users can determine if the overlapping sequences should be preserved or merged, they can define scores for the number of tetrads, loop lengths and the number of bulges. The bulge score factor is only available for the intramolecular G4 search. Users can also limit the loop size and guanines in the loops. In the intramolecular mode, users can determine the number of bulges ($\leq$ 1 per tetrad) and bulge length. In the intermolecular mode, users can limit the G-runs used to predict partial G4s. Input data format: FASTA. Availability: Python script (https://github.com/mparker2/g4predict).

G4-Predictor V.2 [39] locates non-overlapping G4s on sense and anti-sense strands of a given DNA or RNA sequence. Users can set the maximum length of G4 sequence (10–45), the minimum length of G-tract (2–6) and loop size (0–36). G4-Predictor V.2 is accessible from the Mishra group website along with G4IPDB, the G-quadruplex DNA/RNA Interacting Protein Database. Input data formats: raw sequence, FASTA. Availability: web application (http://bsbe.iiti.ac.in/bsbe/ipdb/pattern2.php).

G4PromFinder [57] identifies potential transcription promoters in bacterial genomes. It searches for promoters based on G4 DNA motifs and AT-richness. Designed for DNA sequences, the program can be easily modified to process RNA data. The identification of AT-rich fragments follows scanning for PQS in 50 bp upstream region from the 5' end of found AT-rich elements. Searched pattern is $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$, where $x \in [2,4]$, $y \in [1,10]$. The maximum length of the G4 sequence equals 30 nucleotides. Input data format: FASTA. Availability: Python script (https://github.com/MarcoDiSalvo90/G4PromFinder).

G4RNA screener [58, 59] aims to predict G4s in RNA sequences. Its ML algorithm is trained on experimentally validated G4s from sequences deposited in the G4RNA database [36]. The webserver version of the program allows submitting data of either 20,000 characters or 30 KB. Users can customize the output data, the program allows displaying a variety of optional features, for example, Ensembl gene ID, G4 strand and start and end position of G4. G4RNA screener incorporates consecutive G/C ratio threshold (cGcC), G4Hunter threshold (G4H) and G4 Neural Network threshold (G4NN). In the webserver version, these thresholds are set by default to 4.5, 0.9 and 0.5, respectively. The result table can be downloaded as XLSX or CSV file. Input data format: FASTA. Availability: Python script, web application (http://scottgroup.med.usherbrooke.ca/G4RNA_screener/).

ImGQfinder [60, 61] detects canonical and non-canonical PQS. Depending on users' preferences, it finds either guanine-based or cytosine-based quadruplexes. The search criteria rely on G-runs customized by users into a sequence pattern. Users can set the following parameters: the number of tetrads (2–10), the maximum loop length (2–25), canonical/non-canonical structure (0–1, where 1 implies one bulge or mismatch in G-run) and displaying overlapping/non-overlapping PQS. Non-canonical G4s that contain mismatch in G-run are represented by the search pattern $G_{i-1}NG_{n-i}$, bulged G-runs are defined as $G_{i-1}NG_{n-i+1}$, where $i \in (1,n)$, $n \geq 3$. n denotes the number of Gs in a G-run; i shows the position of bulge or mismatch in a G-run. Input data formats: raw sequence, FASTA. Availability: web application (http://imgqfinder.niifhm.ru). pqsfinder [35] searches for PQS in DNA or RNA sequences using regular expression for G-runs with length limitation. The searched motif defines as $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$, where $x \in [1,10]$, $y \in [0,9]$. Mismatches, bulges and long loops within the searched G-run motif are allowed. pqsfinder offers options within three categories: filters, scoring systems and advanced options. Filters allow selecting strands where PQS are searched; searching for overlapping G4s; limiting G-run length; and setting the minimum PQS score, maximum PQS length, loop size, the maximum number of bulges, mismatches and overall defects. The scoring scheme is based on the stability of the potential G4 structure. G-runs are evaluated individually, with bonus points for each G-tetrad stacking, and penalty points for each mismatch or bulge that occurs in the motif. In the scoring system, the users can set penalties and bonus points for complete G-tetrads. A regular expression, custom scoring and default scoring system can be set by users as advanced options. Input data format: FASTA. Availability: R package (Bioconductor) (https://bioconductor.org/packages/release/bioc/html/pqsfinder.html).

QGRS Mapper [62] finds putative G-quadruplexes within DNA or RNA sequences using a sequence motif. QGRS Mapper accepts A, C, T, G, U and N in the input sequence. It searches for pattern $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$, $x \geq 2$, fixed for all G-tracts. The users can set the maximum length of the quadruplex sequence (10–45), the minimum number of tetrads (2–6), loop size (0–36) and loop sequence, which makes program find at least one loop matching the defined character string. Loop sequence is provided as a regular expression, for example, *a*4, (a loop with 4 or more consecutive adenines). The sequence is scored and cleared of the overlapping PQS. The scoring function favors shorter, regular (equal in size) loops over the longer, irregular ones; and relies on the assumption that more stable quadruplexes have more G-tetrads. The results of sequence analysis are displayed as sequence view, data view, data view with overlapping quadruplexes and graphics view. The last one requires the Java Plugin installed on the local machine. The results can be downloaded as an Excel file. Input data formats: raw sequence, FASTA. Availability: web application (http://bioinformatics.ramapo.edu/QGRS/analyze.php).

QPARSE [42] is a graph-based algorithm to search for non-canonical G-quadruplexes. It constructs and then traverses a direct acyclic graph of discovered runs. QPARSE finds multimeric potential quadruplex-forming sequences, long-looped PQS and intramolecular monomeric quadruplexes. It applies the mfold-derived function to score the predicted loops based on their thermodynamic and conformation stability. The users can specify searched base in a run (default: G), run length, the maximum loop distance between runs within the same PQS, the number of consecutive runs in the same PQS, the minimum number of

uninterrupted runs and the maximum number of long loops (≥ 7 nt) per each PQS. Either one loop symmetry—mirror or palindrome—or both can be verified within input data. QPARSE webserver limits the input sequence up to 10,000 nucleotides or 15 KB of data. Input data format: FASTA. Availability: Python script, web application (https://github.com/B3rse/qparse).

Quadron [37] applies an ML model to predict PQS in DNA sequences. It was developed based on a tree gradient boosting machine. Human genome G4-seq sequences were divided into two sets: one was used as a training set (70% random sequences), and the other one served as a testing set (remaining 30% of the sequences). The general motif in Quadron search defines as $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$. Despite the model was dedicated to DNA sequences, it also handles RNAs. The users specify how many CPUs should algorithm use for calculation. Input data format: FASTA. Availability: R package (http://quadron.atgcdynamics.org/). The program requires an installation of an R and xgboost library. For users convenience, GUI is also available.

TetraplexFinder [50] searches for potential G-quadruplexes within DNA or RNA sequences. It is a partial module of QuadBase2 [49]—a webserver for PQS prediction within eukaryotic and prokaryotic sequences and user-provided sequences. The tool accepts 20 MB of data, which allows processing large datasets, differentiating whether the input file has a single or multiple entries. The users can set up G-tract length (2–5), loop size (1–50), strand where algorithms should search for PQS, bulge size (0–7) and the algorithm to be used (greedy/non-greedy). Bulges are searched only in GGG tetrads. The output data can be filtered on the website or downloaded as a BED file to the local machine. Input data format: FASTA. Availability: web application (http://quadbase.igib.res.in/TetraPlexFinder).

Tools that predict 2D structure with G4s

In the vast collection of programs that predict RNA secondary structure, only one refers to quadruplexes. RNAfold [63, 64] predicts the secondary structure of RNA or DNA and annotates potential quadruplexes in it. This core program of the ViennaRNA Package applies the thermodynamics-based function to optimize the structure. Additional option—Incorporate G-Quadruplex formation into the structure prediction algorithm—turned on makes the algorithm search for quadruplexes during the computational process. RNAfold outputs the secondary structure in a dot-bracket notation with '+' signs under guanines predicted to form the G-quadruplex. Input data formats: raw sequence, FASTA. Availability: standalone program, web application (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi).

Tools that analyze and visualize 2D and 3D structure

In 2017, the first computational tools appeared that could visualize the topology of quadruplexes and reveal their specificity based on the 3D structure data [65, 66]. Shortly afterward, bioinformatics developed methods to identify quadruplexes in nucleic acid structures and determine selected parameters characteristic of these motifs. Currently, four tools can analyze and visualize G4 structures. Their important features are summarized in Table 3: strand polarity indicates whether the tool outputs the information about strands' directions; G4 classification informs what is the basis for classifying the quadruplex topology; base-pair classification indicates if base pairs are classified according to known nomenclatures; area tells if the program calculates surface area of the tetrads; rise and twist denote that the program computes these parameters for each pair of neighboring tetrads in the quadruplex; planarity checked means that the program analyzes planarity deviation for every tetrad; torsion angles tell that the program outputs torsion angles for the structure; 2D view and 3D view indicate whether the tool visualizes the secondary and tertiary structure; moving camera means that the program allows rotating the visualized 3D model.

DSSR [38] processes the 3D structure of the RNA molecule and annotates its secondary structure. It is a part of the 3DNA suite [67] designed to work with the structures of nucleic acids. DSSR identifies, classifies and describes base pairs, multiplets and characteristic motifs of the secondary structure; helices, stems, hairpin loops, bulges, internal loops, junctions and others. It can also detect modules and tertiary structure patterns, including pseudoknots and kink-turns. The recent extension, DSSR-PyMOL [68], allows drawing cartoon-block schemes of the 3D structure and responds to the need for simplified visualization of quadruplexes. Input data formats: PDB, mmCIF and PDB ID. Availability: standalone program, web application (http://dssr.x3dna.org/, http://skmatic.x3dna.org/).

ElTetrado [31] specializes in identifying and describing tetrads and quadruplexes in the 3D structures of nucleic acids, by searching for G-based and non-G-based motifs. It classifies tetrads and quadruplexes into ONZ classes according to their secondary structure topology [2] and calculates strand direction, planarity deviation, rise and twist parameters. The program also outputs the graphical representation of the secondary structure (top-down arc diagram) and its dot-bracket encoding in a two-line format—both designed specially to handle quadruplexes. Input data formats: PDB, mmCIF. Availability: Python script (https://github.com/tzok/eltetrado).

RNApdbee [66, 69], a multifunctional tool from RNApolis suite [70], mainly aims to annotate secondary structures of knotted and unknotted RNAs based on the 3D structure data. Its usefulness in the study of quadruplexes lies in the appropriately matched visualization of the secondary structure, which facilitates the visual identification of these motifs on a diagram of the entire structure and highlights their topological features. A pictographic annotation of interactions within tetrads—according to Leontis–Westhof nomenclature—allows the immediate determination of the tetrad (and quadruplex) class in the recently developed ONZ taxonomy [2]. Input data formats: PDB, mmCIF, PDB ID, BPSEQ, CT, dot-bracket. Availability: web application (http://rnapdbee.cs.put.poznan.pl/).

3D-NuS [65] models and visualizes the 3D nucleic acid structures, including duplexes, triplexes and quadruplexes. It builds energy minimized 3D models of canonical and non-canonical G4 structures based on 17 classes defined for the intramolecular and intermolecular quadruplexes. The users provide strand orientation and type, the number of G-quartets, sequences of all G4-loops, to get the model visualized in JSmol along with selected structure data. Input data formats: G4 class, strand type, number of G-tetrads, loops' sequences. Availability: web application (http://iith.ac.in/3dnus/Quadruplex.html).

Results

Test sets

We created four sets of RNA sequences with and without the ability to form quadruplexes (Figure 2). We used them to test sequence-based tools for the prediction of G4 location in the sequence and prediction of RNA secondary structure

Table 3. Selected features of 2D and 3D structure analyzing tools

Tool	Strand polarity	G4 classification	Base-pair classification	Area	Rise	Twist	Planarity	Torsion angles	2D view	3D view	Moving camera
DSSR	✓	Loop-based	Saenger, Leontis–Westhof	✓	✓	✓	✓	✓	✓	✓	
ElTetrado	✓	ONZ	Leontis–Westhof		✓	✓	✓		✓		
RNApdbee			Saenger, Leontis–Westhof						✓		
3D-NuS		Q1–Q17			✓	✓		✓		✓	✓

with quadruplexes. To build the first test set (DP: dataset with positive examples), we searched the G4RNA database [36] for the sequences, for which the experiments confirmed that they formed G-quadruplexes. We found 321 examples, and after removing the redundant data (unnecessary duplicates), we obtained DP with 295 positive cases. The duplicates were identified and removed by using the MS Excel function *remove duplicates*. In the same database, we found 238 RNAs that did not tend to fold into quadruplexes. By selecting unique sequences, we ended up with 237 negative cases in the DN set (DN: dataset with negative examples). G4RNA is the only database of sequences experimentally tested for G-quadruplex folding, which contains both: G4 sequences and sequences confirmed not to form G4s. Therefore, we chose it as the test data source. Two other test sets were built based on miRBase [71]—the database collecting annotated pre-miRNA and mature miRNA sequences of various species; currently, about 40,000 pre-miRNAs from 271 organisms. We created DH (DH: dataset with human pre-miRNAs) containing 1864 non-redundant sequences selected from 1917 human pre-miRNAs and DV (DV: dataset with *Viridiplantae* pre-miRNAs) with 8354 unique sequences selected from 8615 *Viridiplantae* pre-miRNAs. DH and DV sets contain sequences with quadruplex forming propensity, although their formation was not confirmed experimentally. The data to create all test sets were collected from both repositories on 21 March 2020. More information on the datasets is available in the Supplementary Material (Table S1).

To test the tools that analyze and visualize 2D and 3D structures, we selected two PDB-deposited RNA structures that formed G-quadruplexes: an RNA aptamer (PDB id: 2RQJ) [72] with canonical G4 topology and r(GGAGGAGGAGGA) sequence and a synthetic construct of r(UGGUGGU)4 structure (PDB id: 6GE1) [3] containing U-tetrads.

Computational experiments with sequence-based tools

In this part of our study, we processed four test sets with 14 tools aimed to predict G4 locations in RNA sequences and one for RNA secondary structure prediction with G4s. Most programs performed with the default settings, apart from G4Catchall tested in two modes, G4RNA screener in four and TetraplexFinder in three. *2 tetrads* and *3 tetrads* modes of G4Catchall mean that the tool searches for two- or three-tetrad motifs, respectively. G4Predict in the *Intra* mode looks for intramolecular G-quadruplexes. *cGcC*, *G4H* and *G4NN* are thresholds used in G4RNA screener; all mode aggregates the results obtained for all of them. *G2 L1-12*, *G3 L1-7* and *G3 L1-3* modes of TetraplexFinder correspond to three stringency levels in PQS search: low (two-tetrad G4s, loop size 1–12), medium (three-tetrad G4s, loop size 1–7) and high (three-tetrad G4s, loop size 1–3). G4-iM Grinder was launched with parameters: Name = LmajorESTs, Sequence = Sequence, DNA = FALSE, RunComposition=G, MinRunSize = 1, MinNRuns = 1, MinPQSSize = 1, Complementary = FALSE. RNAfold was executed with the advanced option *Incorporate G-Quadruplex formation into the structure prediction algorithm*, which set it to the G4 search mode. We post-processed the results using in-house scripts [73, 74] in Python, bash and R.

We executed each tool for each sequence in the DP, DN, DH and DV sets, and we noted whether the tool predicted quadruplexes in it or not. Within every test set, we calculated the number of G4-positive and G4-negative sequences found by each tool (Supplementary Material, Tables S2 and S3, Figure S1), where a sequence with predicted quadruplex is counted as G4-positive, and a sequence without quadruplex is G4-negative.

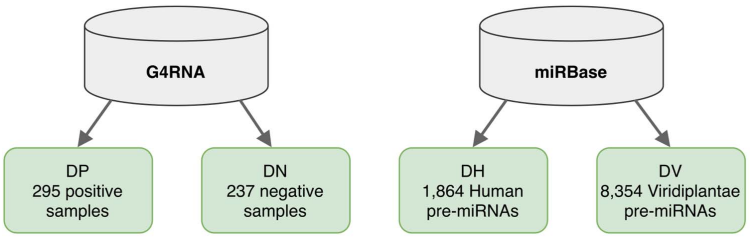


Figure 2. Test sets created for the analysis of sequence-based tools.

Table 4. Test set coverage [%] with PQS predictions: positive (+) for all sets, negative (–) for DP and DN

Tool	Mode	DP+	DP–	DN+	DN–	DH+	DV+
G4Catchall	2 tetrads	91.5	8.5	40.5	59.5	15.7	4.2
	3 tetrads	86.4	13.6	30.8	69.2	9.0	1.4
G4Hunter	n/a	97.6	2.4	67.5	32.5	53.9	44.4
G4-iM Grinder	n/a	100.0	0.0	97.5	2.5	93.2	92.1
G4P Calculator	n/a	78.0	22.0	38.8	61.2	10.2	3.2
G4Predict	Intra	14.6	85.4	2.5	97.5	0.6	0.1
G4-Predictor V.2	n/a	99.0	1.0	66.7	33.3	44.2	30.9
G4PromFinder	n/a	28.1	71.9	30.8	69.2	14.8	19.6
G4RNA screener	cGcC	85.1	14.9	40.1	59.9	8.3	21.0
	G4H	81.0	19.0	25.7	74.3	3.1	1.1
	G4NN	84.4	15.6	30.0	70.0	6.0	3.3
	all	95.6	4.4	46.8	53.2	10.9	22.2
ImGQfinder	n/a	16.3	83.7	5.5	94.5	0.8	0.1
pqsfinder	n/a	86.8	13.2	36.7	63.3	10.6	1.4
QGRS Mapper	n/a	99.0	1.0	66.2	33.8	44.2	30.9
QPARSE	n/a	97.3	2.7	65.0	35.0	43.0	29.8
Quadron	n/a	70.5	29.5	21.1	78.9	5.8	0.9
TetraplexFinder	G2 L1-12	39.3	60.7	29.1	70.9	5.4	1.7
	G3 L1-7	14.6	85.4	2.1	97.9	0.4	0.0
	G3 L1-3	9.2	90.8	1.3	98.7	0.4	0.0
RNAfold	Advanced	73.2	26.8	21.9	78.1	2.5	0.8

In Table 4, we show the coverage of all sets with positive (+) predictions and, additionally, the coverage of DP and DN with the negative (–) ones. Let us recall that the DP set contains sequences for which experiments confirmed the formation of quadruplexes, sequences from DN do not form quadruplexes, DH and DV include sequences with quadruplex forming propensity. We expected the best tools to show high coverage of DP with positive PQS predictions and DN with negative predictions. As shown in Figure 3, several programs meet these conditions. The best results achieved G4RNA screener, G4Catchall and RNAfold. G4RNA screener (in all modes) identified PQS in >80% sequences in DP and classified >50% of DN sequences as non-PQS; however, note that this algorithm was trained on data from G4RNA database (as of 2017) and this result was expected. G4Catchall generated around 90% of positive predictions for DP and 60–70% negative ones for DN. RNAfold showed over 70% of correct predictions for both sets. Just behind these three programs are Quadron and pqsfinder—both cover >60% of DP and DN with correct predictions. Relatively few PQS were found in the DV and DH datasets, which contain sequences potentially forming quadruplexes. In the vast majority of cases, the coverage of these sets with positive predictions does not exceed 10%.

A separate group of tools maximize the number of predicted PQS. Among them, G4-iM Grinder stands out in the foreground—it found quadruplexes in all sequences of DP and 97.5% sequences of DN. The opposite strategy is adopted by G4Predict, ImGQfinder and TetraplexFinder, which in all data sets found few sequences with the potential to create quadruplexes. In most cases, these programs give at most 15% coverage with correct predictions.

Finally, G4PromFinder surprisingly recognizes more PQS in DN than DP set. This program addresses large sequences (bacterial genomes) where it searches for potential promoters. Therefore, the input sequence length should exceed 50 nucleotides, with >40% of adenines and uracils, and motif length ≤ 30 nucleotides. In the DP test set, 147 of the 295 sequences (49.8%) consist of ≥ 50 nucleotides, while in DN, 177 of the 237 (74.7%). G4PromFinder predicted 83 PQS in DP set (28.1%) and 73 PQS in the DN set (30.8%). Such predictions are the result of the program's prerequisites.

Based on the results obtained for DP and DN sets, we evaluated the quality of prediction (Table 5) by computing the following:

- accuracy: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$
- sensitivity (true positive rate): $TPR = \frac{TP}{TP+FN}$

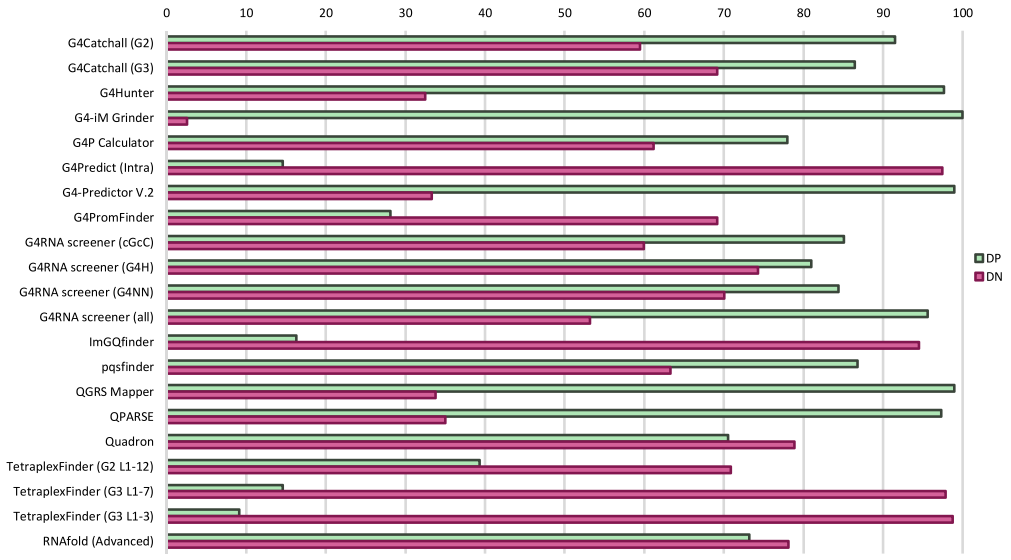


Figure 3. Coverage of DP and DN datasets with correct predictions: positive in DP and negative in DN [%].

- specificity (true negative rate): $TNR = \frac{TN}{TN+FP}$
- precision (positive predictive value): $PPV = \frac{TP}{TP+FP}$
- negative predictive value: $NPV = \frac{TN}{TN+FN}$
- false discovery rate: $FDR = \frac{FP}{TP+FP}$
- F-score: $F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV+TPR}$

They all use four basic measures: true positives (TP)—PQS predicted for DP sequences, true negatives (TN)—negative predictions in DN set, false positives (FP)—PQS predicted for DN sequences and false negatives (FN)—negative predictions in DP set (Supplementary Material, Table S4).

Accuracy (ACC) is the ratio of correct predictions to the total number of input sequences. G4Catchall, G4RNA screener and pqsfinder have the best (the highest) accuracy, G4PromFinder—the worst (the lowest) one, which confirms our observations from the previous paragraphs. Sensitivity (TPR) indicates what part of the actual PQS has been predicted by the program. The highest TPR (i.e. the best one) belongs to G4-iM Grinder, G4-Predictor V.2 and QGRS Mapper, but we already know that these tools aim to maximize the prediction of PQS. Such a strategy also causes poor (the lowest) specificity (TNR). Low TNR value indicates a small fraction of correctly predicted PQS-negative sequences. TetraplexFinder, ImGQfinder and G4Predict are the leaders of specificity, with its value exceeding 0.95. TetraplexFinder and G4Predict have also the highest precision (PPV). PPV shows a fraction of positive predictions, which are natively positive. In turn, NPV determines which part of the negative predictions is actually negative. A reliable tool takes high values of all mentioned factors, but especially PPV and NPV should be close to 1. False discovery rate (FDR) is the only measure from Table 5 to be minimized. It points the fraction of incorrectly predicted PQS among all positive predictions. The lowest FDR belongs to TetraplexFinder and G4Predict. Finally, the F-score, a weighted harmonic mean of precision and sensitivity, is computed to find the balance between these two measures. It aims to assess the accuracy when the distribution between classes is uneven—especially if there is a large number of true negatives. The best (highest) F-score has G4Catchall and G4RNA screener, with pqsfinder right behind, while the worst F-score belongs to TetraplexFinder and G4Predict—two tools showing the best precision. In Table 5, the best value of each computed measure is highlighted in bold.

Computational experiments with structure-based tools

In this part of the study, we tested four tools that analyze and visualize the secondary and the tertiary structures with quadruplexes. We executed them for two quadruplex RNA structures—an RNA aptamer (PDB id: 2RQJ) [72] and a synthetic construct of r(UGGUGGU)4 structure (PDB id: 6GE1) [3]—to find out what details of the structure we can obtain and in what form, numerical and graphical, they are presented. Let us note that 2RQJ is a dimer with two unimolecular structures obtained via NMR; each comprising one G-quadruplex with two tetrads of O+ type in the ONZ classification [2]. 6GE1 is an NMR-determined, tetramolecular structure with unusual topology. It is composed of seven tetrads—four G-tetrads and three U-tetrads—six of them classified as O+ and one as O– according to the ONZ taxonomy. ElTetrado and RNApdbee were run with the default input settings, DSSR with the PDB identifier at the input only, 3D-NuS required to select the quadruplex class.

DSSR and ElTetrado identified quadruplexes in the input PDB files. Both programs focused on structural aspects of the input molecule, explicitly informing about quadruplexes and tetrads within the structure. DSSR provided an extensive analysis of 3D structures and output the data about G-tetrads, G-helices and G4-stems. It computed planarity for each G-tetrad and gave the sections area, rise and twist parameters for G4-helix and G4-stems. The program automatically assigned loop topologies

Table 5. Quality of PQS prediction based on DP and DN set processing

Tool	Mode	ACC	TPR	TNR	PPV	NPV	FDR	F1
G4Catchall	G2	0.77	0.92	0.59	0.74	0.85	0.26	**0.82**
	G3	**0.79**	0.86	0.69	0.78	0.80	0.22	**0.82**
G4Hunter	n/a	0.69	0.98	0.32	0.64	0.92	0.36	0.78
G4-iM Grinder	n/a	0.57	**1.00**	0.03	0.56	**1.00**	0.44	0.72
G4P Calculator	n/a	0.70	0.78	0.61	0.71	0.69	0.29	0.75
G4Predict	Intra	0.52	0.15	0.97	0.88	0.48	0.12	0.25
G4-Predictor V.2	n/a	0.70	0.99	0.33	0.65	0.96	0.35	0.78
G4PromFinder	n/a	0.46	0.28	0.69	0.53	0.44	0.47	0.37
G4RNA screener	cGcC	0.74	0.85	0.60	0.73	0.76	0.27	0.78
	G4H	0.78	0.81	0.74	0.80	0.76	0.20	0.80
	G4NN	0.78	0.84	0.70	0.78	0.78	0.22	0.81
	all	0.77	0.96	0.53	0.72	0.91	0.28	**0.82**
ImGQfinder	n/a	0.51	0.16	0.95	0.79	0.48	0.21	0.27
pqsfinder	n/a	0.76	0.87	0.63	0.75	0.79	0.25	0.80
QGRSMapper	n/a	0.70	0.99	0.34	0.65	0.96	0.35	0.78
QPARSE	n/a	0.70	0.97	0.35	0.65	0.91	0.35	0.78
Quadron	n/a	0.74	0.71	0.79	0.81	0.68	0.19	0.75
TetraplexFinder	G2 L1-12	0.53	0.39	0.71	0.63	0.48	0.37	0.48
	G3 L1-7	0.52	0.15	0.98	**0.90**	0.48	**0.10**	0.25
	G3 L1-3	0.49	0.09	**0.99**	**0.90**	0.47	**0.10**	0.17
RNAfold	Advanced	0.75	0.73	0.78	0.81	0.70	0.19	0.77

according to the predefined types (P—parallel, D—diagonal and L—lateral) and their orientation (+/−). DSSR-PyMOL generated block schemes of both quadruplexes (Figure 4A3 and B3). ElTetrado also calculated planarity, rise and twist parameters and identified strand directions for both quadruplexes. It classified the quadruplexes and their component tetrads to ONZ classes. Finally, it generated the arc diagram (Figure 4A1 and B1) and two-line dot-bracket encoding of every quadruplex.

RNApdbee, as opposed to the previous programs, does not explicitly inform that it has identified tetrads and quadruplexes in the input data. Its purpose is to annotate and visualize the secondary structure and determine its parameters, focusing on pseudoknots [66, 69]. For the analyzed structures, RNApdbee generated an extensive report on the secondary structure, including information on canonical and non-canonical interactions, their classification in Saenger and Leontis–Westhof nomenclatures, base-phosphate interactions, stacking interactions, base-ribose interactions, structure motifs, dot-bracket and the secondary structure diagram—drawn by VARNA-based procedure—with base pairs annotated according to Leontis–Westhof [75] (Figure 4A2 and B2). Note that only one, VARNA-based, drawing procedure of RNApdbee can visualize quadruplexes. The other two, PseudoViewer-based procedure and R-chie, are based on canonical interactions and their visualizations of quadruplex structures are incomplete.

3D-NuS aims to generate the 3D models of G4s based on 17 classes of G-quadruplex folds. Thus, its input and output data differ from the other tools in this section. The program requires input information about quadruplex topology: quadruplex class, subclass and the number of tetrads and sequences of loops. It outputs the tertiary structure in PDB format and provides its visualization. We tried 3D-NuS for different quadruplexes, including the ones selected for the analysis. Tests have shown that 3D-NuS is not a suitable tool for modeling RNA quadruplexes. It does not form non-G tetrads, it has problems with modeling short loops, uni- and bimolecular guadruplexes; no such observations appeared when it modeled DNA quadruplexes with similar sequence and topology. Provided input data and output data generated by 3D-NuS for exemplary structures are presented in the Supplementary Material.

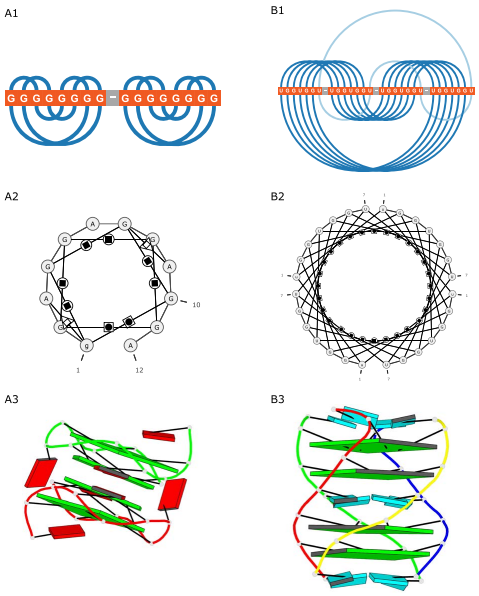


Figure 4. Visualization of (A) 2RQJ and (B) 6GE1 structures generated by (1) ElTetrado, (2) RNApdbee and (3) DSSR-PyMOL.

Conclusion

With the growing interest in quadruplexes, computer programs for their analysis began to appear. Most of them rely solely

on a sequence and parse it to find a predefined G4 motif. This goes hand in hand with creating G4-related databases that primarily collect information about sequences with the ability to form quadruplexes. Our experiments with sequence-based tools applied for RNA sequences showed a very good performance of G4Catchall (motif-based algorithm), whose flexibility certainly contributed to this result. Right behind was RNAfold, the tool for secondary structure prediction enriched with the quadruplex annotation option. Four existing structure-based tools addressing G4s focus on different structural aspects. DSSR comprehensively examines the G4 structure, determines a variety of its parameters and provides the schematic 3D view. ElTetrado identifies tetrads and quadruplexes in the structure, computes their basic parameters, classifies according to ONZ taxonomy and gives the secondary structure in the arc diagram and dot-bracket notation. RNApdbee draws secondary structure diagrams and classifies base-pairs. 3D-NuS builds the 3D model of the quadruplex based on user-defined topology if the quadruplex topology fits one of the classes supported by the tool. These tools complement each other in revealing the full picture of quadruplex space, although they do not deal equally well with all quadruplex types, e.g. 3D-NuS is limited to 17 G4 classes and can reliably model DNA quadruplexes only.

Despite the already significant number of bioinformatics programs that can be used to study DNA and RNA quadruplexes, there are still issues that lack *in silico* solutions. One of them is the modeling of the secondary structure. Among the huge number of programs to predict the RNA 2D structure, only RNAfold touched the problem of quadruplexes. It annotates the places of G4 formation in the dot-bracket representation of the structure. However, this notation does not reflect the quadruplex topology and cannot be easily transformed into secondary structure visualization. Prediction and modeling of the quadruplex 3D structure is also a challenge. First reported attempts of blind, template-free prediction of the 3D G4 structure were made within RNA-Puzzles challenge 23, with seven human and one webserver participant. From the assessment table (available online), it can be seen that this structure was one of the most difficult ones in RNA-Puzzles history. A reliable prediction of the 2D and 3D structure of quadruplexes requires experimental data inclusion, e.g. thermodynamic parameters. A resource collecting such data for G4s could be very supportive. A database that would integrate various data from existing archives would also be a helpful tool or a specialized search engine, browsing the existing databases for related information on a given quadruplex.

Key points

- G4-related computational tools concentrate on discovering, analyzing, and visualizing quadruplexes, mostly on the sequence level.
- In this work, we analyzed 35 bioinformatics resources: 10 dedicated solely to DNA G4s; 4 for RNA G4s; 21 for any nucleic acid.
- Tests of existing G4-related sequence-based tools against four RNA datasets identified G4Catchall, a motif-based method, as the best tool for finding reliable RNA PQS.
- Only 3 tools analyze the 2D and 3D structures of RNA quadruplexes. Their functions are complementary: each considers the other set of structural features and generates a different view of the G4 topology.
- The sequence-based modeling of the G4 structure is challenging. Only one program for 2D structure prediction—RNAfold—reliably indicates G4 location in the sequence but does not give its topology. The 3D structure prediction of RNA G4s is still a challenge.

Funding

National Science Centre, Poland (2016/23/B/ST6/03931, 2019/35/B/ST6/03074); Poznan University of Technology.

References

1. Lightfoot HL, Hagen T, Tatum NJ, *et al.* The diverse structural landscape of quadruplexes. *FEBS Lett* 2019;**593**(16):2083–102. doi: 10.1002/1873-3468.13547.
2. Popenda M, Miskiewicz J, Sarzynska J, *et al.* Topology-based classification of tetrads and quadruplex structures. *Bioinformatics* 2020;**36**(4):1129–34. doi: 10.1093/bioinformatics/btz738.
3. Andrałojć W, Małgowska M, Sarzyńska J, *et al.* Unraveling the structural basis for the exceptional stability of RNA G-quadruplexes capped by a uridine tetrad at the 3′ terminus. *RNA* 2018;**25**(1):121–34. doi: 10.1261/rna.068163.118.
4. Zhang N, Gorin A, Majumdar A, *et al.* Dimeric DNA quadruplex containing major groove-aligned A·T·A·T and G·C·G·C tetrads stabilized by inter-subunit Watson-crick a·T and G·C pairs. *J Mol Biol* 2001;**312**(5):1073–88. doi: 10.1006/jmbi.2001.5002.
5. Heddi B, Martín-Pintado N, Serimbetov Z, *et al.* G-quadruplexes with (4n - 1) guanines in the G-tetrad core: formation of a G-triad·water complex and implication for small-molecule binding. *Nucleic Acids Res* 2015;**44**(2):910–6. doi: 10.1093/nar/gkv1357.
6. Kettani A, Gorin A, Majumdar A, *et al.* A dimeric DNA interface stabilized by stacked A· (G · G · G · G)· A hexads and coordinated monovalent cations. *J Mol Biol* 2000;**297**(3):627–44. doi: 10.1006/jmbi.2000.3524.
7. Kogut M, Kleist C, Czub J. Why do G-quadruplexes dimerize through the 5'-ends? Driving forces for G4 DNA dimerization examined in atomic detail. *PLoS Comput Biol* 2019 e1007383; **15**(9). doi: 10.1371/journal.pcbi.1007383.
8. Kolesnikova S, Curtis EA. Structure and function of multimeric G-quadruplexes. *Molecules* 2019;**24**(17):3074. doi: 10.3390/molecules24173074.
9. Webba da Silva M. Geometric formalism for DNA quadruplex folding. *Chem A Eur J* 2007;**13**(35):9738–45. doi: 10.1002/chem.200701255.
10. Lech CJ, Heddi B, Phan AT. Guanine base stacking in G-quadruplex nucleic acids. *Nucleic Acids Res* 2012;**41**(3):2034–46. doi: 10.1093/nar/gks1110.
11. Webba da Silva M, Trajkovski M, Sannohe Y, *et al.* Design of a G-quadruplex topology through glycosidic bond angles. *Angewandte Chemie* 2009;**121**(48):9331–4. doi: 10.1002/ange.200902454.
12. Dvorkin SA, Karsisiotis AI, da Silva MW. Encoding canonical DNA quadruplex structure. *Sci Adv* 2018;**4**(8): eaat3007. doi: 10.1126/sciadv.aat3007.
13. Ravichandran S, Ahn J-H, Kim KK. Unraveling the regulatory G-quadruplex puzzle: lessons from genome and

transcriptome-wide studies. *Front Genet* 2019;**10**:1002. doi: 10.3389/fgene.2019.01002.
14. Hänsel-Hertsch R, Spiegel J, Marsico G, *et al.* Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* 2018;**13**(3):551–64. doi: 10.1038/nprot.2017.150.
15. Chambers VS, Marsico G, Boutell JM, *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* 2015;**33**(8):877–81. doi: 10.1038/nbt.3295.
16. Marsico G, Chambers VS, Sahakyan AB, *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res* 2019;**47**(8):3862–74. doi: 10.1093/nar/gkz179.
17. Raguseo F, Chowdhury S, Minard A, *et al.* Chemical-biology approaches to probe DNA and RNA G-quadruplex structures in the genome. *Chem Commun* 2020;**56**(9):1317–24. doi: 10.1039/c9cc09107f.
18. Che T, Wang Y-Q, Huang Z-L, *et al.* Natural alkaloids and heterocycles as G-quadruplex ligands and potential anticancer agents. *Molecules* 2018;**23**(2):493. doi: 10.3390/molecules23020493.
19. Kwok CK, Marsico G, Sahakyan AB, *et al.* rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods* 2016;**13**(10):841–4. doi: 10.1038/nmeth.3965.
20. Yeung PY, Zhao J, Chow EY-C, *et al.* Systematic evaluation and optimization of the experimental steps in RNA G-quadruplex structure sequencing. *Sci Rep* 2019;**9**(1):8091. doi: 10.1038/s41598-019-44541-4.
21. Yang SY, Lejault P, Chevrier S, *et al.* Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun* 2018;**9**(1):4730. doi: 10.1038/s41467-018-07224-8.
22. Lee DSM, Ghanem LR, Barash Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat Commun* 2020;**11**(1):527. doi: 10.1038/s41467-020-14404-y.
23. Chan KL, Peng B, Umar MI, *et al.* Structural analysis reveals the formation and role of RNA G-quadruplex structures in human mature microRNAs. *Chem Commun* 2018;**54**(77):10878–81. doi: 10.1039/c8cc04635b.
24. Roxo C, Kotkowiak W, Pasternak A. G-quadruplex-forming aptamers - characteristics, applications, and perspectives. *Molecules* 2019;**24**(20):3781. doi: 10.3390/molecules24203781.
25. Brázda V, Hároníková L, Liao J, *et al.* DNA and RNA quadruplex-binding proteins. *Int J Mol Sci* 2014;**15**(10):17493–517. doi: 10.3390/ijms151017493.
26. Serikawa T, Spanos C, von Hacht N, *et al.* Comprehensive identification of proteins binding to RNA G-quadruplex motifs in the 5′ UTR of tumor-associated mRNAs. *Biochimie* 2018;**144**:169–84. doi: 10.1016/j.biochi.2017.11.003.
27. Rouleau SG, Garant J-M, Bolduc F, *et al.* G-quadruplexes influence pri-microRNA processing. *RNA Biol* 2017;**15**(2):198–206. doi: 10.1080/15476286.2017.1405211.
28. Kang C, Zhang X, Ratliff R, *et al.* Crystal structure of four-stranded Oxytricha telomeric DNA. *Nature* 1992;**356**:126–31. doi: 10.1038/356126a0.
29. Cheong C, Moore PB. Solution structure of an unusually stable RNA tetraplex containing G- and U-quartet structures. *Biochemistry* 1992;**31**:8406–14. doi: 10.1021/bi00151a003.
30. Berman HM. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**(1):235–42. doi: 10.1093/nar/28.1.235.
31. Zok T, Popenda M, Szachniuk M. ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *BMC Bioinformatics* 2020;**21**:40. doi: 10.1186/s12859-020-3385-1.
32. Joachimi A, Benz A, Hartig JS. A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem* 2009;**17**(19):6811–5. doi: 10.1016/j.bmc.2009.08.043.
33. Lombardi EP, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res* 2020;**48**(3):1603. doi: 10.1093/nar/gkaa033.
34. Wong HM, Stegle O, Rodgers S, *et al.* A toolbox for predicting G-quadruplex formation and stability. *J Nucleic Acids* 2010;**2010**:1–6. doi: 10.4061/2010/564946.
35. Hon J, Martínek T, Zendulka J, *et al.* PQSfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* 2017;**33**(21):3373–9. doi: 10.1093/bioinformatics/btx413.
36. Garant J-M, Luce MJ, Scott MS, *et al.* G4RNA: an RNA G-quadruplex database. *Database* 2015. doi: 10.1093/database/bav059.
37. Sahakyan AB, Chambers VS, Marsico G, *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* 2017;**7**(1):14535. doi: 10.1038/s41598-017-14017-4.
38. Lu X-J, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* 2015;**43**(21):e142. doi: 10.1093/nar/gkv716.
39. Mishra SK, Tawani A, Mishra A, *et al.* G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* 2016;**6**:38144. doi: 10.1038/srep38144.
40. Li Q, Xiang J-F, Yang Q-F, *et al.* G4ldb: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res* 2012;**41**(D1):D1115–23. doi: 10.1093/nar/gks1101.
41. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* 2016;**44**(4):1746–59. doi: 10.1093/nar/gkw006.
42. Berselli M, Lavezzo E, Toppo S. QPARSE: searching for long-looped or multimeric G-quadruplexes potentially distinctive and druggable. *Bioinformatics* 2019;**36**(2):393–9. doi: 10.1093/bioinformatics/btz569.
43. Shao X, Zhang W, Umar MI, *et al.* RNA G-quadruplex structures mediate gene regulation in bacteria. *MBio* 2020;**11**(1):e02926-19. doi: 10.1128/mbio.02926-19.
44. Zhang R, Lin Y, Zhang C-T. Greglist: a database listing potential G-quadruplex regulated genes. *Nucleic Acids Res* 2007;**36**(D1):D372–6. doi: 10.1093/nar/gkm787.
45. Kikin O, Zappala Z, D'Antonio L, *et al.* GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. *Nucleic Acids Res* 2007;**36**(D1):D141–8. doi: 10.1093/nar/gkm982.
46. Lavezzo E, Berselli M, Frasson I, *et al.* G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide. *PLoS Comput Biol* 2018;**14**(12):e1006675. doi: 10.1371/journal.pcbi.1006675.
47. Cer RZ, Donohue DE, Mudunuri US, *et al.* Non-b DB v2.0: a database of predicted non-b DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2012;**41**(D1):D94–D100. doi: 10.1093/nar/gks955.
48. Ge F, Wang Y, Li H, *et al.* Plant-GQ: an integrative database of G-quadruplex in plant. *J Comput Biol* 2019;**26**(9):1013–9. doi: 10.1089/cmb.2019.0010.
49. Yadav VK, Abraham JK, Mani P, *et al.* QuadBase: genome-wide database of G4 DNA occurrence and conservation in human, chimpanzee, mouse and rat promoters and

146 microbes. *Nucleic Acids Res* 2007;**36**(D1):D381–5. doi: 10.1093/nar/gkm781.
50. Dhapola P, Chowdhury S. QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res* 2016;**44**(W1):W277–83. doi: 10.1093/nar/gkw425.
51. D'Antonio L, Bagga P. Computational methods for predicting intramolecular G-quadruplexes in nucleotide sequences. In: *Proceedings of the IEEE Computational Systems Bioinformatics Conference* 2004. 590–1. Stanford, CA: IEEE. doi: 10.1109/CSB.2004.1332508.
52. Huppert JL. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* 2005;**33**(9):2908–16. doi: 10.1093/nar/gki609.
53. Doluca O. G4Catchall: a G-quadruplex prediction approach considering atypical features. *J Theor Biol* 2019;**463**:92–8. doi: 10.1016/j.jtbi.2018.12.007.
54. Brázda V, Kolomazník J, Lýsek J, *et al*. G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* 2019;**35**(18):3493–5. doi: 10.1093/bioinformatics/btz087.
55. Belmonte-Reche E, Morales JC. G4-iM grinder: DNA and RNA G-Quadruplex, i-motif and higher order structure search and analyser tool. *NAR Genom Bioinfor* 2020;**2**(1):lqz005. doi: 10.1093/nargab/lqz005.
56. Eddy J, Maizels N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res* 2006;**34**(14):3887–96. doi: 10.1093/nar/gkl529.
57. Salvo MD, Pinatel E, Talà A, *et al*. G4PromFinder: an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs. *BMC Bioinformatics* 2018;**19**(1):36. doi: 10.1186/s12859-018-2049-x.
58. Garant J-M, Perreault J-P, Scott MS. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics* 2017;**33**(22):3532–7. doi: 10.1093/bioinformatics/btx498.
59. Garant J-M, Perreault J-P, Scott MS. G4RNA screener web server: user focused interface for RNA G-quadruplex prediction. *Biochimie* 2018;**151**:115–8. doi: 10.1016/j.biochi.2018.06.002.
60. Varizhuk A, Ischenko D, Tsvetkov V, *et al*. The expanding repertoire of G4 DNA structures. *Biochimie* 2017;**135**:54–62. doi: 10.1016/j.biochi.2017.01.003.
61. Varizhuk A, Ischenko D, Smirnov I, *et al*. An improved search algorithm to find G-quadruplexes in genome sequences. *bioRxiv* 2014. doi: 10.1101/001990.
62. Kikin O, D'Antonio L, Bagga PS. QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* 2006;**34**(W1):W676–82. doi: 10.1093/nar/gkl253.
63. Gruber AR, Lorenz R, Bernhart SH, *et al*. The Vienna RNA Websuite. *Nucleic Acids Res* 2008;**36**(W1):W70–4. doi: 10.1093/nar/gkn188.
64. Lorenz R, Bernhart SH, zu Biederdissen CH, *et al*. ViennaRNA Package 2.0. *Algorithm Mol Biol* 2011;**6**(1):26. doi: 10.1186/1748-7188-6-26.
65. Patro LPP, Kumar A, Kolimi N, *et al*. 3d-NuS: a web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. *J Mol Biol* 2017;**429**(16):2438–48. doi: 10.1016/j.jmb.2017.06.013.
66. Zok T, Antczak M, Zurkowski M, *et al*. RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res* 2018;**46**(W1):W30–5. doi: 10.1093/nar/gky314.
67. Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* 2008;**3**(7):1213–27. doi: 10.1038/nprot.2008.104.
68. Lu X-J. DSSR-enabled innovative schematics of 3d nucleic acid structures with PyMOL. *Nucleic Acids Res* 2020;**48**:e77. doi: 10.1093/nar/gkaa426.
69. Antczak M, Popenda M, Zok T, *et al*. New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics* 2018;**34**(8):1304–12. doi: 10.1093/bioinformatics/btx783.
70. Szachniuk M. RNApolis: computational platform for RNA structure analysis. *Found Comput Decis Sci* 2019;**44**(2):241–57. doi: 10.2478/fcds-2019-0012.
71. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2018;**47**(D1):D155–62. doi: 10.1093/nar/gky1141.
72. Mashima T, Matsugami A, Nishikawa F, *et al*. Unique quadruplex structure and interaction of an RNA aptamer against bovine prion protein. *Nucleic Acids Res* 2009;**37**(18):6249–58. doi: 10.1093/nar/gkp647.
73. Miskiewicz J, Tomczyk K, Mickiewicz A, *et al*. Bioinformatics study of structural patterns in plant microRNA precursors. *Biomed Res Int* 2017;6783010. doi: 10.1155/2017/6783010.
74. Miskiewicz J, Szachniuk M. Discovering structural motifs in miRNA precursors from Viridiplantae kingdom. *Molecules* 2018;**23**(6):1367. doi: 10.3390/molecules23061367.
75. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA* 2001;**7**(4):499–512. doi: 10.1017/s1355838201002515.

SUPPLEMENTARY MATERIAL

How bioinformatics resources work with G4 RNAs

Joanna Miskiewicz[1], Joanna Sarzynska[2], Marta Szachniuk[1,2,*]

[1]Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

*corresponding author: marta.szachniuk@cs.put.poznan.pl

Table S1. Description of the test sets.

	DP	DN	DH	DV
Number of nonredundant sequences in the test set	295	237	1864	8354
Number of sequences with at least 8 Guanines	295	231	1859	8348
% of sequences with at least 8 Guanines	100%	97.47%	99.73%	99.93%
Number of sequences with at least 12 Guanines	279	214	1803	8260
% of sequences with at least 12 Guanines	94.58%	90.30%	96.73%	98.87%
Number of sequences with at least 50 nucleotides	147	177	1852	8344
% of sequences with at least 50 nucleotides	49.83%	74.68%	99.36%	99.88%
Average sequence length [nts]	130	152	82	149

Table S2. The number of PQS-positive sequences predicted by the sequence-based tools.

Tool	Mode	DP	DN	DH	DV
G4Catchall	2 tetrads	270	96	293	353
	3 tetrads	255	73	167	117
G4Hunter	n/a	288	160	1004	3708
G4-iM Grinder	n/a	295	231	1737	7690
G4P Calculator	n/a	230	92	191	271
G4Predict	Intra	43	6	11	9
G4-Predictor V.2	n/a	292	158	823	2570
G4PromFinder	n/a	83	73	275	1640
G4RNA screener	cGcC	251	95	154	1757
	G4H	239	61	57	94
	G4NN	249	71	111	279
	all	282	111	203	1853
ImGQfinder	n/a	48	13	15	5
pqsfinder	n/a	256	87	197	119
QGRS Mapper	n/a	292	157	823	2582
QPARSE	n/a	287	154	801	2492
Quadron	n/a	208	50	109	78
TetraplexFinder	G2 L1-12	116	69	100	141
	G3 L1-7	43	5	8	4
	G3 L1-3	27	3	7	3
RNAfold	Advanced	216	52	46	70

Table S3. The number of PQS-negative samples resulting from test set processing by the sequence-based tools.

Tool	Mode	DP	DN	DH	DV
G4Catchall	2 tetrads	25	141	1571	8001
	3 tetrads	40	164	1697	8237
G4Hunter	n/a	7	77	860	4646
G4-iM Grinder	n/a	0	6	127	664
G4P Calculator	n/a	65	145	1673	8083
G4Predict	Intra	252	231	1853	8345
G4-Predictor V.2	n/a	3	79	1041	5784
G4PromFinder	n/a	212	164	1589	6714
G4RNA screener	cGcC	44	142	1710	6597
	G4H	56	176	1807	8260
	G4NN	46	166	1753	8075
	all	13	126	1661	6501
ImGQfinder	n/a	247	224	1849	8349
pqsfinder	n/a	39	150	1667	8235
QGRS Mapper	n/a	3	80	1041	5772
QPARSE	n/a	8	83	1063	5862
Quadron	n/a	87	187	1755	8276
TetraplexFinder	G2 L1-12	179	168	1764	8213
	G3 L1-7	252	232	1856	8350
	G3 L1-3	268	234	1857	8351
RNAfold	Advanced	79	185	1818	8284

Table S4. Confusion matrix computed based on DP and DN sets processing (population: 532 sequences).

Tool	Mode	TP	TN	FP	FN	ACC	TPR	TNR	PPV	NPV	FDR	F1	FNR	FPR	FOR
G4Catchall	G2	270	141	96	25	0.77	0.92	0.59	0.74	0.85	0.26	0.82	0.08	0.41	0.15
	G3	255	164	73	40	0.79	0.86	0.69	0.78	0.80	0.22	0.82	0.14	0.31	0.20
G4Hunter	n/a	288	77	160	7	0.69	0.98	0.32	0.64	0.92	0.36	0.78	0.02	0.68	0.08
G4-iM Grinder	n/a	295	6	231	0	0.57	1.00	0.03	0.56	1.00	0.44	0.72	0.00	0.97	0.00
G4P Calculator	n/a	230	145	92	65	0.70	0.78	0.61	0.71	0.69	0.29	0.75	0.22	0.39	0.31
G4Predict	Intra	43	231	6	252	0.52	0.15	0.97	0.88	0.48	0.12	0.25	0.85	0.03	0.52
G4-Predictor V.2	n/a	292	79	158	3	0.70	0.99	0.33	0.65	0.96	0.35	0.78	0.01	0.67	0.04
G4PromFinder	n/a	83	164	73	212	0.46	0.28	0.69	0.53	0.44	0.47	0.37	0.72	0.31	0.56
G4RNA screener	cGcC	251	142	95	44	0.74	0.85	0.60	0.73	0.76	0.27	0.78	0.15	0.40	0.24
	G4H	239	176	61	56	0.78	0.81	0.74	0.80	0.76	0.20	0.80	0.19	0.26	0.24
	G4NN	249	166	71	46	0.78	0.84	0.70	0.78	0.78	0.22	0.81	0.16	0.30	0.22
	all	282	126	111	13	0.77	0.96	0.53	0.72	0.91	0.28	0.82	0.04	0.47	0.09
ImGQfinder	n/a	48	224	13	247	0.51	0.16	0.95	0.79	0.48	0.21	0.27	0.84	0.05	0.52
pqsfinder	n/a	256	150	87	39	0.76	0.87	0.63	0.75	0.79	0.25	0.80	0.13	0.37	0.21
QGRS Mapper	n/a	292	80	157	3	0.70	0.99	0.34	0.65	0.96	0.35	0.78	0.01	0.66	0.04
QPARSE	n/a	287	83	154	8	0.70	0.97	0.35	0.65	0.91	0.35	0.78	0.03	0.65	0.09
Quadron	n/a	208	187	50	87	0.74	0.71	0.79	0.81	0.68	0.19	0.75	0.29	0.21	0.32
TetraplexFinder	G2 L1-12	116	168	69	179	0.53	0.39	0.71	0.63	0.48	0.37	0.48	0.61	0.29	0.52
	G3 L1-7	43	232	5	252	0.52	0.15	0.98	0.90	0.48	0.10	0.25	0.85	0.02	0.52
	G3 L1-3	27	234	3	268	0.49	0.09	0.99	0.90	0.47	0.10	0.17	0.91	0.01	0.53
RNAfold	Advanced	216	185	52	79	0.75	0.73	0.78	0.81	0.70	0.19	0.77	0.27	0.22	0.30

Figure S1. Coverage of DP and DN test sets with positive PQS predictions [%].

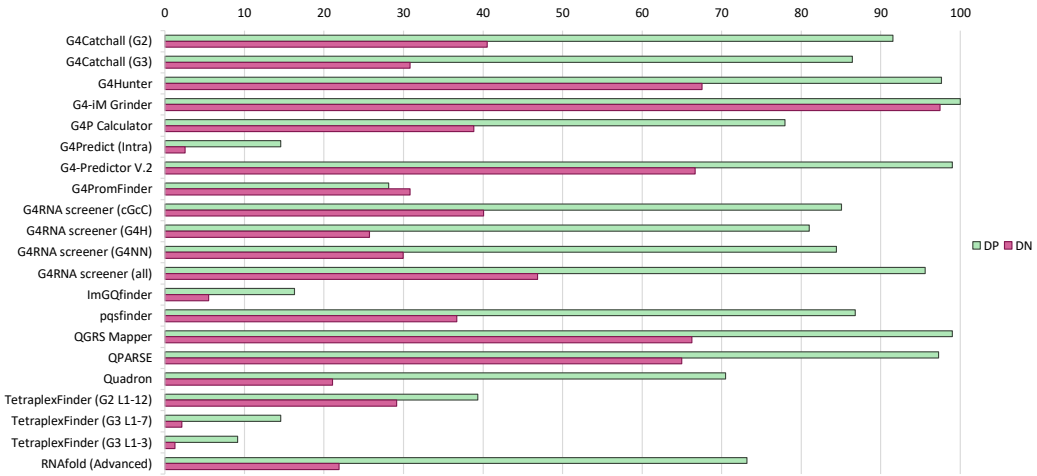


Table S5. Patterns used to search for quadruplex motifs and exemplary sequence motifs.

Tool	Pattern	Pattern description	Exemplary motif	
G4Catchall	G_n	Perfect G-tract	GGG	n = 3
	$G_xN_yG_{(n-x)}$	Bulged G-tract	GGAG	n = 3; x = 2; y = 1
	$G_xN_yG_{(n-x-1)}$	Mismatched G-tract	GGAG	n = 4; x = 2; y = 1
G4Hunter	n/a	n/a	n/a	
G4-iM Grinder	n/a	n/a	n/a	
G4P Calculator	G_n	Perfect G-tract	GGG	n = 3
G4Predict (Intra)	$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for intramolecular PQS	GGAAGGAAAGGAAGG	x1= x2= x3= x4= 2; y1= y3= 2; y2= 3
G4-Predictor V.2	$G_{2\text{-}6}N_{0\text{-}36}G_{2\text{-}6}N_{0\text{-}36}G_{2\text{-}6}N_{0\text{-}36}G_{2\text{-}6}$	Patterns depend on users' settings	GGAGGGAAGGAAGGG	
G4PromFinder	$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for PQS	GGAGGGAAGGAAGGG	x1= x3= 2; x2= x4= 3; y1= 1; y2= y3= 2
G4RNA screener	n/a	n/a	n/a	
ImGQfinder	G_n	Perfect G-tract	GGG	n = 3
	$G_{i-1}NG_{n-i+1}$	Bulged G-tract	GGAGG	n = 4; i = 3
	$G_{i-1}NG_{n-i}$	Mismatched G-tract	GGAG	n = 4; i = 3
pqsfinder	$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for PQS	GGAGGAAGGAAGGG	x1= x2= x3= 2; x4= 3; y1= 1; y2= y3= 2
QGRS Mapper	$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for PQS	GGAGGAAGGAAGG	x1= x2= x3= x4= 2; y1= 1; y2= y3= 2
QPARSE	n/a	Patterns depend on users' settings	n/a	
Quadron	*$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for PQS	GGAGGAAGGAAGG	x1= x2= x3= x4= 2; y1= 1; y2= y3= 2
TetraplexFinder	$G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$	General pattern for PQS	GGAAAGGAAGGAAGG	x1= x2= x3= x4= 2; y1= 3; y2= y3= 2
RNAfold	n/a	n/a	n/a	

*Quadron is a machine-learning method, pattern was only used to find PQS in human genome

DSSR output for 2RQJ structure

```
****************************************************************************
File name: 2rqj.pdb
    no. of DNA/RNA chains: 2 [A=12,B=12]
    no. of nucleotides:    24
    no. of atoms:          806
    no. of waters:         0
    no. of metals:         0

****************************************************************************
List of 20 base pairs
     nt1            nt2            bp  name        Saenger   LW   DSSR
   1 A.G1           A.A3           G-A Sheared     11-XI     tSH  tm-M
   2 A.G1           A.G4           G+G --          06-VI     cWH  cW+M
   3 A.G1           A.G10          G+G --          06-VI     cHW  cM+W
   4 A.G2           A.G5           G+G --          06-VI     cWH  cW+M
   5 A.G2           A.G11          G+G --          06-VI     cHW  cM+W
   6 A.G4           A.G7           G+G --          06-VI     cWH  cW+M
   7 A.G5           A.G8           G+G --          06-VI     cWH  cW+M
   8 A.G7           A.A9           G-A Sheared     11-XI     tSH  tm-M
   9 A.G7           A.G10          G+G --          06-VI     cWH  cW+M
  10 A.G8           A.G11          G+G --          06-VI     cWH  cW+M
  11 B.G13          B.A15          G-A Sheared     11-XI     tSH  tm-M
  12 B.G13          B.G16          G+G --          06-VI     cWH  cW+M
  13 B.G13          B.G22          G+G --          06-VI     cHW  cM+W
  14 B.G14          B.G17          G+G --          06-VI     cWH  cW+M
  15 B.G14          B.G23          G+G --          06-VI     cHW  cM+W
  16 B.G16          B.G19          G+G --          06-VI     cWH  cW+M
  17 B.G17          B.G20          G+G --          06-VI     cWH  cW+M
  18 B.G19          B.A21          G-A Sheared     11-XI     tSH  tm-M
  19 B.G19          B.G22          G+G --          06-VI     cWH  cW+M
  20 B.G20          B.G23          G+G --          06-VI     cWH  cW+M

****************************************************************************
List of 4 multiplets
   1 nts=4 GGGG A.G2,A.G5,A.G8,A.G11
   2 nts=4 GGGG B.G14,B.G17,B.G20,B.G23
   3 nts=6 GAGGAG A.G1,A.A3,A.G4,A.G7,A.A9,A.G10
   4 nts=6 GAGGAG B.G13,B.A15,B.G16,B.G19,B.A21,B.G22

****************************************************************************
List of 2 helices
  Note: a helix is defined by base-stacking interactions, regardless of bp
        type and backbone connectivity, and may contain more than one stem.
      helix#number[stems-contained] bps=number-of-base-pairs in the helix
      bp-type: '|' for a canonical WC/wobble pair, '.' otherwise
      helix-form: classification of a dinucleotide step comprising the bp
        above the given designation and the bp that follows it. Types
        include 'A', 'B' or 'Z' for the common A-, B- and Z-form helices,
        '.' for an unclassified step, and 'x' for a step without a
        continuous backbone.
      ------------------------------------------------------------------
  helix#1[0] bps=4
      strand-1 5'-GGGG-3'
       bp-type    ....
      strand-2 5'-GGGG-3'
      helix-form  xx.
   1 A.G2           A.G5           G+G --          06-VI     cWH  cW+M
   2 A.G1           A.G4           G+G --          06-VI     cWH  cW+M
   3 B.G13          B.G16          G+G --          06-VI     cWH  cW+M
   4 B.G14          B.G17          G+G --          06-VI     cWH  cW+M
  ----------------------------------------------------------------------------
  helix#2[0] bps=4
      strand-1 5'-GGGG-3'
       bp-type    ....
      strand-2 5'-GGGG-3'
      helix-form  .x.
   1 A.G8           A.G11          G+G --          06-VI     cWH  cW+M
```

```
   2 A.G7           A.G10          G+G --           06-VI     cWH  cW+M
   3 B.G19          B.G22          G+G --           06-VI     cWH  cW+M
   4 B.G20          B.G23          G+G --           06-VI     cWH  cW+M

*****************************************************************************
List of 6 stacks
  Note: a stack is an ordered list of nucleotides assembled together via
        base-stacking interactions, regardless of backbone connectivity.
        Stacking interactions within a stem are *not* included.
   1 nts=2 AA A.A3,B.A15
   2 nts=2 AA A.A9,B.A21
   3 nts=4 GGGG A.G2,A.G1,B.G16,B.G17
   4 nts=4 GGGG A.G5,A.G4,B.G13,B.G14
   5 nts=5 GGGGA A.G8,A.G7,B.G22,B.G23,B.A24
   6 nts=5 AGGGG A.A12,A.G11,A.G10,B.G19,B.G20

*****************************************************************************
Nucleotides not involved in stacking interactions
     nts=2 AA A.A6,B.A18

*****************************************************************************
List of 2 non-loop single-stranded segments
   1 nts=12 GGAGGAGGAGGA
A.G1,A.G2,A.A3,A.G4,A.G5,A.A6,A.G7,A.G8,A.A9,A.G10,A.G11,A.A12
   2 nts=12 GGAGGAGGAGGA
B.G13,B.G14,B.A15,B.G16,B.G17,B.A18,B.G19,B.G20,B.A21,B.G22,B.G23,B.A24

*****************************************************************************
List of 4 G-tetrads
   1 glyco-bond=---- groove=---- planarity=0.092 type=planar nts=4 GGGG
A.G1,A.G4,A.G7,A.G10
   2 glyco-bond=---- groove=---- planarity=0.076 type=planar nts=4 GGGG
A.G2,A.G5,A.G8,A.G11
   3 glyco-bond=---- groove=---- planarity=0.112 type=planar nts=4 GGGG
B.G13,B.G16,B.G19,B.G22
   4 glyco-bond=---- groove=---- planarity=0.075 type=planar nts=4 GGGG
B.G14,B.G17,B.G20,B.G23

*****************************************************************************
List of 1 G4-helix
  Note: a G4-helix is defined by stacking interactions of G4-tetrads, regardless
        of backbone connectivity, and may contain more than one G4-stem.
  helix#1[2] stems=[#1,#2] layers=4 inter-molecular
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G2,A.G5,A.G8,A.G11
   2* glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G1,A.G4,A.G7,A.G10
   3  glyco-bond=---- groove=---- Major-->WC nts=4 GGGG B.G16,B.G13,B.G22,B.G19
   4  glyco-bond=---- groove=---- Major-->WC nts=4 GGGG B.G17,B.G14,B.G23,B.G20
    step#1  mp(<<,backward) area=11.25 rise=3.18 twist=35.2
    step#2  mm(<>,outward)  area=19.27 rise=3.39 twist=5.3
    step#3  pm(>>,forward)  area=11.51 rise=3.17 twist=35.6
    strand#1 RNA glyco-bond=---- nts=4 GGGG A.G2,A.G1,B.G16,B.G17
    strand#2 RNA glyco-bond=---- nts=4 GGGG A.G5,A.G4,B.G13,B.G14
    strand#3 RNA glyco-bond=---- nts=4 GGGG A.G8,A.G7,B.G22,B.G23
    strand#4 RNA glyco-bond=---- nts=4 GGGG A.G11,A.G10,B.G19,B.G20

*****************************************************************************
List of 2 G4-stems
  Note: a G4-stem is defined as a G4-helix with backbone connectivity.
        Bulges are also allowed along each of the four strands.
  stem#1[#1] layers=2 INTRA-molecular loops=3 descriptor=2(-P-P-P)
note=parallel(4+0) UUUU parallel
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G1,A.G4,A.G7,A.G10
   2  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G2,A.G5,A.G8,A.G11
    step#1  pm(>>,forward)  area=11.25 rise=3.18 twist=35.2
    strand#1  U RNA glyco-bond=-- nts=2 GG A.G1,A.G2
    strand#2  U RNA glyco-bond=-- nts=2 GG A.G4,A.G5
    strand#3  U RNA glyco-bond=-- nts=2 GG A.G7,A.G8
    strand#4  U RNA glyco-bond=-- nts=2 GG A.G10,A.G11
```

```
    loop#1 type=propeller strands=[#1,#2] nts=1 A A.A3
    loop#2 type=propeller strands=[#2,#3] nts=1 A A.A6
    loop#3 type=propeller strands=[#3,#4] nts=1 A A.A9
  ------------------------------------------------------------------------
  stem#2[#1] layers=2 INTRA-molecular loops=3 descriptor=2(-P-P-P)
note=parallel(4+0) UUUU parallel
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG B.G13,B.G16,B.G19,B.G22
   2  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG B.G14,B.G17,B.G20,B.G23
    step#1  pm(>>,forward)  area=11.51 rise=3.17 twist=35.6
    strand#1  U RNA glyco-bond=-- nts=2 GG B.G13,B.G14
    strand#2  U RNA glyco-bond=-- nts=2 GG B.G16,B.G17
    strand#3  U RNA glyco-bond=-- nts=2 GG B.G19,B.G20
    strand#4  U RNA glyco-bond=-- nts=2 GG B.G22,B.G23
    loop#1 type=propeller strands=[#1,#2] nts=1 A B.A15
    loop#2 type=propeller strands=[#2,#3] nts=1 A B.A18
    loop#3 type=propeller strands=[#3,#4] nts=1 A B.A21

****************************************************************************
List of 1 G4 coaxial stack
   1 G4 helix#1 contains 2 G4 stems: [#1,#2]  [5'/5']

****************************************************************************
Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>2rqj nts=24 [whole]
GGAGGAGGAGGA&GGAGGAGGAGGA
............&............
>2rqj-A #1 nts=12 0.58(2.82) [chain] RNA
GGAGGAGGAGGA
............
>2rqj-B #2 nts=12 0.49(2.84) [chain] RNA
GGAGGAGGAGGA
............

****************************************************************************
Summary of structural features of 24 nucleotides
  Note: the first five columns are: (1) serial number, (2) one-letter
    shorthand name, (3) dbn, (4) id string, (5) rmsd (~zero) of base
    ring atoms fitted against those in a standard base reference
    frame. The sixth (last) column contains a comma-separated list of
    features whose meanings are mostly self-explanatory, except for:
      turn: angle C1'(i-1)--C1'(i)--C1'(i+1) < 90 degrees
      break: no backbone linkage between O3'(i-1) and P(i)
   1  G . A.G1  0.010  anti,~C2'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
   2  G . A.G2  0.011  anti,~C2'-endo,BII,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
   3  A . A.A3  0.017  turn,syn,~C3'-endo,non-canonical,non-pair-
contact,multiplet,ss-non-loop
   4  G . A.G4  0.011  anti,~C3'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
   5  G . A.G5  0.010  anti,~C2'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad
   6  A . A.A6  0.014  turn,syn,~C2'-endo,non-stack,non-pair-contact,ss-non-loop
   7  G . A.G7  0.010  anti,~C2'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
   8  G . A.G8  0.012  anti,~C2'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
   9  A . A.A9  0.013  turn,anti,~C3'-endo,non-canonical,non-pair-
contact,multiplet,ss-non-loop
  10  G . A.G10 0.010  anti,~C3'-endo,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
  11  G . A.G11 0.010  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad
  12  A . A.A12 0.013  anti,non-pair-contact,ss-non-loop
  13  G . B.G13 0.010  anti,~C2'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
  14  G . B.G14 0.010  anti,~C2'-endo,BII,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
```

```
  15  A . B.A15 0.017  turn,syn,~C3'-endo,non-canonical,non-pair-
contact,multiplet,ss-non-loop
  16  G . B.G16 0.011  anti,~C3'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad,phosphate
  17  G . B.G17 0.010  anti,~C2'-endo,BII,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad
  18  A . B.A18 0.013  turn,syn,~C2'-endo,non-stack,non-pair-contact,ss-non-loop
  19  G . B.G19 0.009  anti,~C2'-endo,BI,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
  20  G . B.G20 0.012  anti,~C2'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
  21  A . B.A21 0.013  turn,anti,~C3'-endo,non-canonical,non-pair-
contact,multiplet,ss-non-loop
  22  G . B.G22 0.010  anti,~C3'-endo,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
  23  G . B.G23 0.010  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad
  24  A . B.A24 0.013  anti,non-pair-contact,ss-non-loop

*****************************************************************************
List of 8 additional files
   1 dssr-pairs.pdb -- an ensemble of base pairs
   2 dssr-multiplets.pdb -- an ensemble of multiplets
   3 dssr-helices.pdb -- an ensemble of helices (coaxial stacking)
   4 dssr-2ndstrs.bpseq -- secondary structure in bpseq format
   5 dssr-2ndstrs.ct -- secondary structure in connectivity table format
   6 dssr-2ndstrs.dbn -- secondary structure in dot-bracket notation
   7 dssr-torsions.txt -- backbone torsion angles and suite names
   8 dssr-stacks.pdb -- an ensemble of stacks
```

DSSR output for 6GE1 structure

```
*****************************************************************************
File name: 6ge1.pdb
    no. of DNA/RNA chains: 4 [A=7,B=7,C=7,D=7]
    no. of nucleotides:    28
    no. of atoms:          900
    no. of waters:         0
    no. of metals:         0
*****************************************************************************
List of 28 base pairs
     nt1            nt2            bp  name        Saenger   LW   DSSR
   1 A.U1           B.U1           U+U --          n/a       cHW  cM+W
   2 A.U1           D.U1           U+U --          n/a       cWH  cW+M
   3 A.G2           B.G2           G+G --          06-VI     cHW  cM+W
   4 A.G2           D.G2           G+G --          06-VI     cWH  cW+M
   5 A.G3           B.G3           G+G --          06-VI     cHW  cM+W
   6 A.G3           D.G3           G+G --          06-VI     cWH  cW+M
   7 A.U4           B.U4           U+U --          n/a       cHW  cM+W
   8 A.U4           D.U4           U+U --          n/a       cWH  cW+M
   9 A.G5           B.G5           G+G --          06-VI     cHW  cM+W
  10 A.G5           D.G5           G+G --          06-VI     cWH  cW+M
  11 A.G6           B.G6           G+G --          06-VI     cHW  cM+W
  12 A.G6           D.G6           G+G --          06-VI     cWH  cW+M
  13 A.U7           B.U7           U+U --          n/a       cWH  cW+M
  14 A.U7           D.U7           U+U --          n/a       cHW  cM+W
  15 B.U1           C.U1           U+U --          n/a       cHW  cM+W
  16 B.G2           C.G2           G+G --          06-VI     cHW  cM+W
  17 B.G3           C.G3           G+G --          06-VI     cHW  cM+W
  18 B.U4           C.U4           U+U --          n/a       cHW  cM+W
  19 B.G5           C.G5           G+G --          06-VI     cHW  cM+W
  20 B.G6           C.G6           G+G --          06-VI     cHW  cM+W
  21 B.U7           C.U7           U+U --          n/a       cWH  cW+M
  22 C.U1           D.U1           U+U --          n/a       cHW  cM+W
  23 C.G2           D.G2           G+G --          06-VI     cHW  cM+W
  24 C.G3           D.G3           G+G --          06-VI     cHW  cM+W
  25 C.U4           D.U4           U+U --          n/a       cHW  cM+W
```

```
  26 C.G5           D.G5           G+G --          06-VI     cHW  cM+W
  27 C.G6           D.G6           G+G --          06-VI     cHW  cM+W
  28 C.U7           D.U7           U+U --          n/a       cWH  cW+M
*************************************************************************
List of 7 multiplets
   1 nts=4 UUUU A.U1,B.U1,C.U1,D.U1
   2 nts=4 GGGG A.G2,B.G2,C.G2,D.G2
   3 nts=4 GGGG A.G3,B.G3,C.G3,D.G3
   4 nts=4 UUUU A.U4,B.U4,C.U4,D.U4
   5 nts=4 GGGG A.G5,B.G5,C.G5,D.G5
   6 nts=4 GGGG A.G6,B.G6,C.G6,D.G6
   7 nts=4 UUUU A.U7,B.U7,C.U7,D.U7
*************************************************************************
List of 2 helices
  Note: a helix is defined by base-stacking interactions, regardless of bp
        type and backbone connectivity, and may contain more than one stem.
      helix#number[stems-contained] bps=number-of-base-pairs in the helix
      bp-type: '|' for a canonical WC/wobble pair, '.' otherwise
      helix-form: classification of a dinucleotide step comprising the bp
        above the given designation and the bp that follows it. Types
        include 'A', 'B' or 'Z' for the common A-, B- and Z-form helices,
        '.' for an unclassified step, and 'x' for a step without a
        continuous backbone.
      ---------------------------------------------------------------------
  helix#1[0] bps=6 parallel
      strand-1 5'-UGGUGG-3'
       bp-type    ......
      strand-2 5'-UGGUGG-3'
      helix-form  .....
   1 A.U1           D.U1           U+U --          n/a       cWH  cW+M
   2 A.G2           D.G2           G+G --          06-VI     cWH  cW+M
   3 A.G3           D.G3           G+G --          06-VI     cWH  cW+M
   4 A.U4           D.U4           U+U --          n/a       cWH  cW+M
   5 A.G5           D.G5           G+G --          06-VI     cWH  cW+M
   6 A.G6           D.G6           G+G --          06-VI     cWH  cW+M
  ------------------------------------------------------------------------
  helix#2[0] bps=7
      strand-1 5'-UGGUGGU-3'
       bp-type    .......
      strand-2 5'-UGGUGGU-3'
      helix-form  .....x
   1 B.U1           C.U1           U+U --          n/a       cHW  cM+W
   2 B.G2           C.G2           G+G --          06-VI     cHW  cM+W
   3 B.G3           C.G3           G+G --          06-VI     cHW  cM+W
   4 B.U4           C.U4           U+U --          n/a       cHW  cM+W
   5 B.G5           C.G5           G+G --          06-VI     cHW  cM+W
   6 B.G6           C.G6           G+G --          06-VI     cHW  cM+W
   7 B.U7           A.U7           U+U --          n/a       cHW  cM+W
*************************************************************************
List of 4 stacks
  Note: a stack is an ordered list of nucleotides assembled together via
        base-stacking interactions, regardless of backbone connectivity.
        Stacking interactions within a stem are *not* included.
   1 nts=6 UGGUGG D.U1,D.G2,D.G3,D.U4,D.G5,D.G6
   2 nts=7 UGGUGGU A.U1,A.G2,A.G3,A.U4,A.G5,A.G6,D.U7
   3 nts=7 UGGUGGU B.U1,B.G2,B.G3,B.U4,B.G5,B.G6,B.U7
   4 nts=7 UGGUGGU C.U1,C.G2,C.G3,C.U4,C.G5,C.G6,C.U7
*************************************************************************
List of 4 non-loop single-stranded segments
   1 nts=7 UGGUGGU A.U1,A.G2,A.G3,A.U4,A.G5,A.G6,A.U7
   2 nts=7 UGGUGGU B.U1,B.G2,B.G3,B.U4,B.G5,B.G6,B.U7
   3 nts=7 UGGUGGU C.U1,C.G2,C.G3,C.U4,C.G5,C.G6,C.U7
   4 nts=7 UGGUGGU D.U1,D.G2,D.G3,D.U4,D.G5,D.G6,D.U7
*************************************************************************
List of 4 G-tetrads
   1 glyco-bond=---- groove=---- planarity=0.167 type=other  nts=4 GGGG
A.G2,D.G2,C.G2,B.G2
```

```
   2 glyco-bond=---- groove=---- planarity=0.262 type=other  nts=4 GGGG
A.G3,D.G3,C.G3,B.G3
   3 glyco-bond=---- groove=---- planarity=0.111 type=planar nts=4 GGGG
A.G5,D.G5,C.G5,B.G5
   4 glyco-bond=---- groove=---- planarity=0.368 type=bowl   nts=4 GGGG
A.G6,D.G6,C.G6,B.G6
*************************************************************************
List of 1 G4-helix
  Note: a G4-helix is defined by stacking interactions of G4-tetrads, regardless
        of backbone connectivity, and may contain more than one G4-stem.
  helix#1[2] stems=[#1,#2] layers=4 inter-molecular
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G2,D.G2,C.G2,B.G2
   2* glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G3,D.G3,C.G3,B.G3
   3  glyco-bond=---- groove=---- Major-->WC nts=4 GGGG A.G5,B.G5,C.G5,D.G5
   4  glyco-bond=---- groove=---- Major-->WC nts=4 GGGG A.G6,B.G6,C.G6,D.G6
    step#1  pm(>>,forward)  area=10.68 rise=3.18 twist=34.6
    step#2  pm(>>,forward)  area=0.00  rise=7.06 twist=-27.4
    step#3  pm(>>,forward)  area=17.62 rise=3.66 twist=19.8
    strand#1 RNA glyco-bond=---- nts=4 GGGG A.G2,A.G3,A.G5,A.G6
    strand#2 RNA glyco-bond=---- nts=4 GGGG D.G2,D.G3,B.G5,B.G6
    strand#3 RNA glyco-bond=---- nts=4 GGGG C.G2,C.G3,C.G5,C.G6
    strand#4 RNA glyco-bond=---- nts=4 GGGG B.G2,B.G3,D.G5,D.G6
*************************************************************************
List of 2 G4-stems
  Note: a G4-stem is defined as a G4-helix with backbone connectivity.
        Bulges are also allowed along each of the four strands.
  stem#1[#1] layers=2 inter-molecular loops=0 note=parallel(4+0) UUUU parallel
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G2,D.G2,C.G2,B.G2
   2  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G3,D.G3,C.G3,B.G3
    step#1  pm(>>,forward)  area=10.68 rise=3.18 twist=34.6
    strand#1  U RNA glyco-bond=-- nts=2 GG A.G2,A.G3
    strand#2  U RNA glyco-bond=-- nts=2 GG D.G2,D.G3
    strand#3  U RNA glyco-bond=-- nts=2 GG C.G2,C.G3
    strand#4  U RNA glyco-bond=-- nts=2 GG B.G2,B.G3
  ------------------------------------------------------------------------
  stem#2[#1] layers=2 inter-molecular loops=0 note=parallel(4+0) UUUU parallel
   1  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G5,D.G5,C.G5,B.G5
   2  glyco-bond=---- groove=---- WC-->Major nts=4 GGGG A.G6,D.G6,C.G6,B.G6
    step#1  pm(>>,forward)  area=17.62 rise=3.66 twist=19.8
    strand#1  U RNA glyco-bond=-- nts=2 GG A.G5,A.G6
    strand#2  U RNA glyco-bond=-- nts=2 GG D.G5,D.G6
    strand#3  U RNA glyco-bond=-- nts=2 GG C.G5,C.G6
    strand#4  U RNA glyco-bond=-- nts=2 GG B.G5,B.G6
*************************************************************************
List of 1 G4 coaxial stack
   1 G4 helix#1 contains 2 G4 stems: [#1,#2]  [3'/5']
*************************************************************************
Secondary structures in dot-bracket notation (dbn) as a whole and per chain
>6ge1 nts=28 [whole]
UGGUGGU&UGGUGGU&UGGUGGU&UGGUGGU
.......&.......&.......&.......
>6ge1-A #1 nts=7 3.84(0.57) [chain] RNA
UGGUGGU
.......
>6ge1-B #2 nts=7 3.85(0.50) [chain] RNA
UGGUGGU
.......
>6ge1-C #3 nts=7 3.75(0.58) [chain] RNA
UGGUGGU
.......
>6ge1-D #4 nts=7 3.76(0.68) [chain] RNA
UGGUGGU
.......
*************************************************************************
Summary of structural features of 28 nucleotides
  Note: the first five columns are: (1) serial number, (2) one-letter
    shorthand name, (3) dbn, (4) id string, (5) rmsd (~zero) of base
    ring atoms fitted against those in a standard base reference
```

```
    frame. The sixth (last) column contains a comma-separated list of
    features whose meanings are mostly self-explanatory, except for:
      turn: angle C1'(i-1)--C1'(i)--C1'(i+1) < 90 degrees
      break: no backbone linkage between O3'(i-1) and P(i)
   1  U . A.U1 0.025  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop
   2  G . A.G2 0.047  anti,~C2'-endo,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
   3  G . A.G3 0.028  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
   4  U . A.U4 0.024  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop
   5  G . A.G5 0.035  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
   6  G . A.G6 0.029  anti,~C3'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
   7  U . A.U7 0.034  anti,~C2'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,phosphate
   8  U . B.U1 0.026  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop
   9  G . B.G2 0.045  anti,~C2'-endo,non-canonical,non-pair-
contact,helix,multiplet,ss-non-loop,G-tetrad
  10  G . B.G3 0.027  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
  11  U . B.U4 0.024  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop
  12  G . B.G5 0.041  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
  13  G . B.G6 0.023  anti,~C3'-endo,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
  14  U . B.U7 0.028  anti,~C2'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop
  15  U . C.U1 0.023  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop
  16  G . C.G2 0.046  anti,~C2'-endo,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
  17  G . C.G3 0.032  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
  18  U . C.U4 0.028  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop
  19  G . C.G5 0.038  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
  20  G . C.G6 0.022  anti,~C3'-endo,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
  21  U . C.U7 0.023  anti,~C2'-endo,non-canonical,non-pair-contact,multiplet,ss-
non-loop
  22  U . D.U1 0.026  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop
  23  G . D.G2 0.055  anti,~C2'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
  24  G . D.G3 0.028  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad,phosphate
  25  U . D.U4 0.026  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop
  26  G . D.G5 0.032  anti,~C3'-endo,BI,non-canonical,non-pair-contact,helix,
multiplet,ss-non-loop,G-tetrad
  27  G . D.G6 0.031  turn,anti,~C3'-endo,non-canonical,non-pair-contact,helix-
end,multiplet,ss-non-loop,G-tetrad,phosphate
  28  U . D.U7 0.033  anti,~C2'-endo,non-canonical,non-pair-contact,multiplet,ss-
non-loop,phosphate
*******************************************************************************
List of 8 additional files
   1 dssr-pairs.pdb -- an ensemble of base pairs
   2 dssr-multiplets.pdb -- an ensemble of multiplets
   3 dssr-helices.pdb -- an ensemble of helices (coaxial stacking)
   4 dssr-2ndstrs.bpseq -- secondary structure in bpseq format
   5 dssr-2ndstrs.ct -- secondary structure in connectivity table format
   6 dssr-2ndstrs.dbn -- secondary structure in dot-bracket notation
```

```
   7 dssr-torsions.txt -- backbone torsion angles and suite names
   8 dssr-stacks.pdb -- an ensemble of stacks
****************************************************************************
```

ElTetrado output for 2RQJ structure

```
Chain order: A, B
n4-helix with 4 tetrads
  Op+ quadruplex with 2 tetrads
    A.G1 A.G4 A.G7 A.G10 cWH-cWH-cWH-cWH O+ planarity=0.28
      direction=parallel rise=3.35 twist=32.98
    A.G2 A.G5 A.G8 A.G11 cWH-cWH-cWH-cWH O+ planarity=0.21
  Op+ quadruplex with 2 tetrads
    B.G13 B.G16 B.G19 B.G22 cWH-cWH-cWH-cWH O+ planarity=0.16
      direction=parallel rise=3.25 twist=37.79
    B.G14 B.G17 B.G20 B.G23 cWH-cWH-cWH-cWH O+ planarity=0.11

GGAGGAGGAGGA-GGAGGAGGAGGA
([.)].([.)].-([.)].([.)].
([.([.)].)].-([.([.)].)].
```

ElTetrado output for 6GE1 structure

```
Chain order: A, D, C, B
n4-helix with 7 tetrads
  Op* quadruplex with 7 tetrads
    A.U1 D.U1 C.U1 B.U1 cWH-cWH-cWH-cWH O+ planarity=1.21
      direction=parallel rise=3.4 twist=19.87
    A.G2 D.G2 C.G2 B.G2 cWH-cWH-cWH-cWH O+ planarity=0.09
      direction=parallel rise=3.08 twist=35.23
    A.G3 D.G3 C.G3 B.G3 cWH-cWH-cWH-cWH O+ planarity=0.5
      direction=parallel rise=3.57 twist=33.57
    A.U4 D.U4 C.U4 B.U4 cWH-cWH-cWH-cWH O+ planarity=0.59
      direction=parallel rise=3.17 twist=29.96
    A.G5 D.G5 C.G5 B.G5 cWH-cWH-cWH-cWH O+ planarity=0.15
      direction=parallel rise=3.54 twist=21.28
    A.G6 D.G6 C.G6 B.G6 cWH-cWH-cWH-cWH O+ planarity=0.75
      direction=parallel rise=3.28 twist=38.1
    A.U7 B.U7 C.U7 D.U7 cWH-cWH-cWH-cWH O- planarity=1.28

UGGUGGU-UGGUGGU-UGGUGGU-UGGUGGU
([{<ABC-)]}>abC-([{<ABc-)]}>abc
([{<ABC-([{<ABc-)]}>abC-)]}>abc

Chain order: A, D, C, B
n4-helix with 7 tetrads
  Op* quadruplex with 7 tetrads
    A.U1 D.U1 C.U1 B.U1 cWH-cWH-cWH-cWH O+ planarity=1.21
      direction=parallel rise=3.4 twist=23.29
    A.G2 D.G2 C.G2 B.G2 cWH-cWH-cWH-cWH O+ planarity=0.09
      direction=parallel rise=3.08 twist=35.23
    A.G3 D.G3 C.G3 B.G3 cWH-cWH-cWH-cWH O+ planarity=0.5
      direction=parallel rise=3.57 twist=33.57
    A.U4 D.U4 C.U4 B.U4 cWH-cWH-cWH-cWH O+ planarity=0.59
      direction=parallel rise=3.17 twist=29.96
    A.G5 D.G5 C.G5 B.G5 cWH-cWH-cWH-cWH O+ planarity=0.15
      direction=parallel rise=3.54 twist=21.28
    A.G6 D.G6 C.G6 B.G6 cWH-cWH-cWH-cWH O+ planarity=0.75
      direction=parallel rise=3.28 twist=8.01
    A.U7 B.U7 C.U7 D.U7 cWH-cWH-cWH-cWH O- planarity=1.28

UGGUGGU-UGGUGGU-UGGUGGU-UGGUGGU
([{<ABC-)]}>abC-([{<ABc-)]}>abc
([{<ABC-([{<ABc-)]}>abC-)]}>abc
```

RNApdbee output for 2RQJ structure

```
Dot-bracket notation
>strand_A
gGAGGAGGAGGA
............
```

Noncanonical interactions

Base-pair	Interaction type	Canonical	Saenger	Leontis-Westhof
A.g1 - A.A3	base - base	N	XI	S/H trans
A.g1 - A.G4	base - base	N	VI	W/H cis
A.g1 - A.G10	base - base	N	VI	H/W cis
A.G2 - A.G5	base - base	N	VI	W/H cis
A.G2 - A.G11	base - base	N	VI	H/W cis
A.G4 - A.G7	base - base	N	VI	W/H cis
A.G5 - A.G8	base - base	N	VI	W/H cis
A.G7 - A.A9	base - base	N	XI	S/H trans
A.G7 - A.G10	base - base	N	VI	W/H cis
A.G8 - A.G11	base - base	N	VI	W/H cis

RNApdbee output for 6GE1 structure

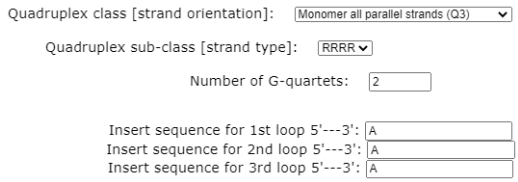
```
Dot-bracket notation
>strand_A
uGGUGGU
.......
>strand_B
uGGUGGU
.......
>strand_C
uGGUGGU
.......
>strand_D
uGGUGGU
.......
```

Noncanonical interactions

Base-pair	Interaction type	Canonical	Saenger	Leontis-Westhof
A.u1 - B.u1	base - base	N	n/a	H/W cis
A.u1 - D.u1	base - base	N	n/a	W/H cis
A.G2 - B.G2	base - base	N	VI	H/W cis
A.G2 - D.G2	base - base	N	VI	W/H cis
A.G3 - B.G3	base - base	N	VI	H/W cis
A.G3 - D.G3	base - base	N	VI	W/H cis
A.U4 - B.U4	base - base	N	n/a	H/W cis
A.U4 - D.U4	base - base	N	n/a	W/H cis
A.G5 - B.G5	base - base	N	VI	H/W cis
A.G5 - D.G5	base - base	N	VI	W/H cis
A.G6 - B.G6	base - base	N	VI	H/W cis
A.G6 - D.G6	base - base	N	VI	W/H cis
A.U7 - B.U7	base - base	N	n/a	W/H cis
A.U7 - D.U7	base - base	N	n/a	H/W cis
B.u1 - C.u1	base - base	N	n/a	H/W cis
B.G2 - C.G2	base - base	N	VI	H/W cis
B.G3 - C.G3	base - base	N	VI	H/W cis
B.U4 - C.U4	base - base	N	n/a	H/W cis
B.G5 - C.G5	base - base	N	VI	H/W cis
B.G6 - C.G6	base - base	N	VI	H/W cis
B.U7 - C.U7	base - base	N	n/a	W/H cis
C.u1 - D.u1	base - base	N	n/a	H/W cis
C.G2 - D.G2	base - base	N	VI	H/W cis
C.G3 - D.G3	base - base	N	VI	H/W cis
C.U4 - D.U4	base - base	N	n/a	H/W cis
C.G5 - D.G5	base - base	N	VI	H/W cis
C.G6 - D.G6	base - base	N	VI	H/W cis
C.U7 - D.U7	base - base	N	n/a	W/H cis

3D-NuS input for 2RQJ structure

Figure S2. Input parameters for 2RQJ structure.



3D-NuS output for 2RQJ structure

Figure S3. Generated 3D model.

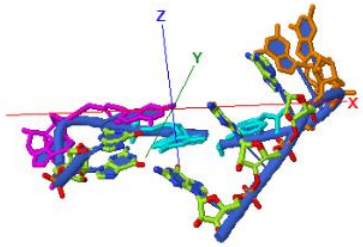


Figure S4. Phase angles of pseudorotation of sugar ring of nucleotide bases for strand I, II, III, IV of 2RQJ structure.

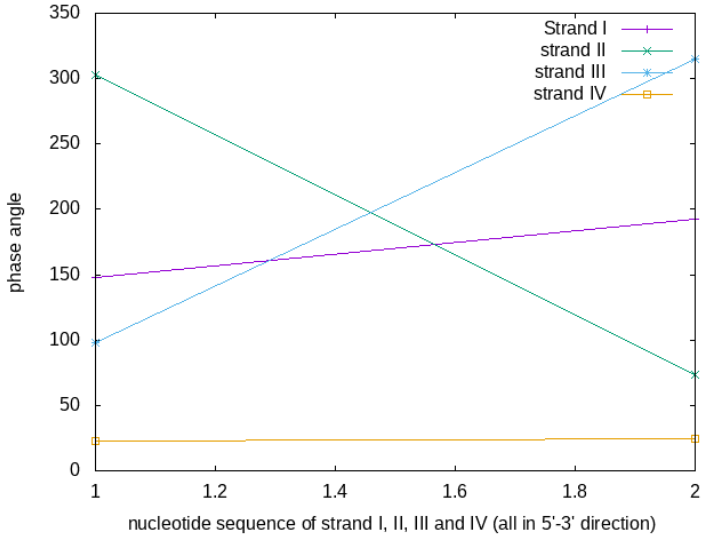


Figure S5. Backbone torsion angles for strand I of 2RQJ structure.

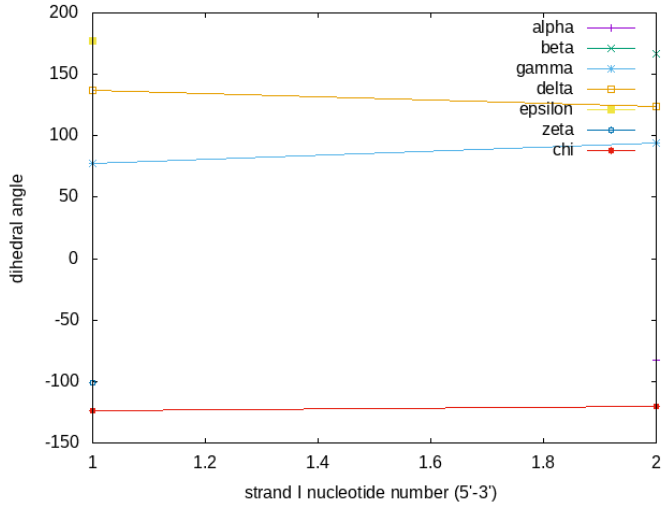


Figure S6. Backbone torsion angles for strand II of 2RQJ structure.

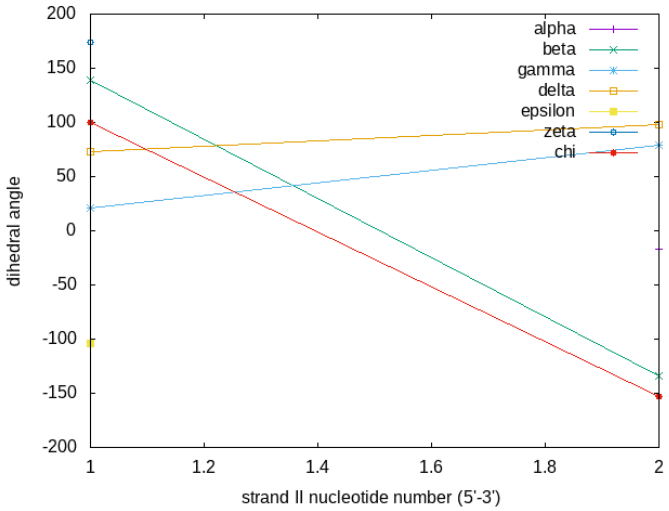


Figure S7. Backbone torsion angles for strand III of 2RQJ structure.

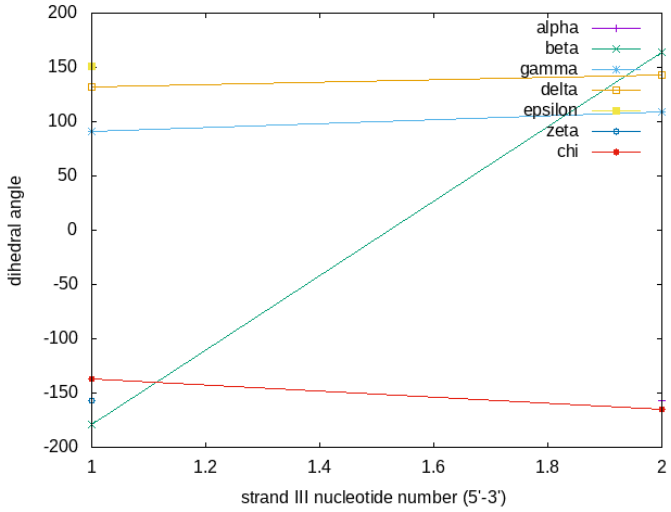
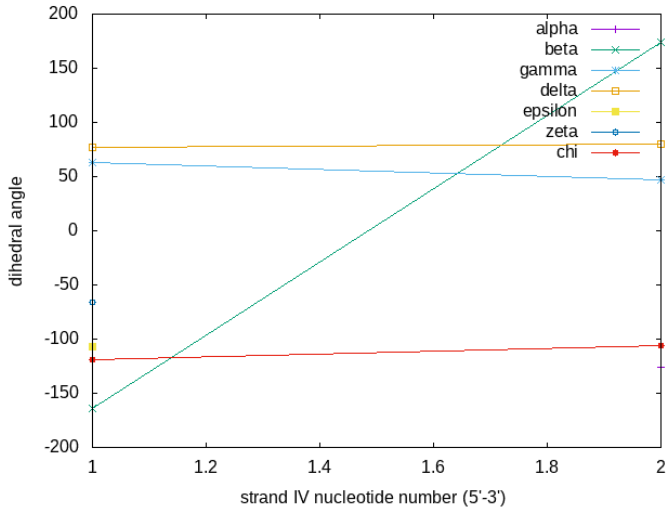


Figure S8. Backbone torsion angles for strand IV of 2RQJ structure.



Output file

```
****************************************************************************
    3DNA v2.3.2-2017dec26, created and maintained by Xiang-Jun Lu (PhD)
****************************************************************************
1. The list of the parameters given below correspond to the 5' to 3' direction
   of strand I and 3' to 5' direction of strand II.

2. All angular parameters, except for the phase angle of sugar pseudo-
   rotation, are measured in degrees in the range of [-180, +180], and all
   displacements are measured in Angstrom units.
****************************************************************************
File name: quadruplex_3dnus.pdb.temp
Date and time: Thu Jul 16 22:05:53 2020

Number of base-pairs: 2
Number of atoms: 370
****************************************************************************
****************************************************************************
RMSD of the bases (----- for WC bp, + for isolated bp, x for helix change)

            Strand I                    Strand II          Helix
   1   (0.025) ....>-:...1_:[GUA]G-**+-G[GUA]:...4_:-<.... (0.087)     |
   2   (0.056) ....>-:...2_:[GUA]G-**+-G[GUA]:...5_:-<.... (0.063)     |

Note: This structure contains 2[2] non-Watson-Crick base-pairs.
****************************************************************************
Detailed H-bond information: atom-name pair and length [ O N]
   1 G-**+-G  [2]  N2 - N7  2.75  O6 * O6  2.49
   2 G-**+-G  [2]  N2 - N7  2.85  O6 * O6  2.46
****************************************************************************
Overlap area in Angstrom^2 between polygons defined by atoms on successive
bases. Polygons projected in the mean plane of the designed base-pair step.

Values in parentheses measure the overlap of base ring atoms only. Those
outside parentheses include exocyclic atoms on the ring. Intra- and
inter-strand overlap is designated according to the following diagram:

                    i2  3'      5' j2
                       /|\       |
                        |        |
                Strand I |       | II
                         |       |
                         |       |
                         |       |
                         |      \|/
                    i1  5'      3' j1

     step      i1-i2        i1-j2        j1-i2        j1-j2        sum
   1 GG/GG  2.34( 0.61)  0.00( 0.00)  0.00( 0.00)  3.63( 1.99)  5.97( 2.60)
****************************************************************************
Origin (Ox, Oy, Oz) and mean normal vector (Nx, Ny, Nz) of each base-pair in
   the coordinate system of the given structure

      bp         Ox         Oy         Oz         Nx         Ny         Nz
    1 G+G     -1.235      1.673      6.703      0.026      0.058     -0.998
    2 G+G     -0.348      1.739      3.605      0.122      0.014     -0.992
****************************************************************************
Local base-pair parameters
     bp        Shear    Stretch   Stagger    Buckle  Propeller  Opening
    1 G+G       0.48       3.93     -0.06     -5.14       0.86    -86.89
    2 G+G       0.71       3.74      0.35     -8.62     -13.14    -89.57
****************************************************************************
Local base-pair step parameters
    step       Shift     Slide      Rise      Tilt      Roll     Twist
   1 GG/GG      0.21     -0.62      3.16      4.73      3.82     40.50
****************************************************************************
Local base-pair helical parameters
    step       X-disp    Y-disp   h-Rise     Incl.       Tip   h-Twist
```

```
   1 GG/GG     -1.29      0.20      3.09      5.48     -6.78     40.93
********************************************************************************
The 'simple' base-pair and step parameters were introduced into 3DNA as of
v2.3-2016jan01. This list of 'simple' parameters is reported by default,
but can be turned off by specifying 'analyze -simple=false'.

The simple parameters are 'intuitive' for non-Watson-Crick base pairs and
associated base-pair steps, where the above corresponding 3DNA parameters
may appear cryptic. Note that the following sets of simple parameters are
for structural description only, not to be fed into the 'rebuild' program.
Overall, they complement the rigorous characterization of base-pair
geometry, as exemplified by the original 3DNA analyze/rebuild programs.

In short, the 'simple' base-pair parameters employ the RC8--YC6 (default,
or RN9--YN1) vector as the (long) y-axis of the pair. As before, the
z-axis is the average of two base normals, taking consideration of the M-N
vs M+N base-pair classification. In essence, the 'simple' parameters make
geometrical sense through the introduction of an ad hoc base-pair
reference frame in each case. See x3dna.org for more details. The same
idea applies to the 'simple' inter-base-pair step parameters, which use
consecutive C1'--C1' vectors.

Here, 'angle' refers to the inter-base angle of each pair, with values in
the range of [0, 90] degrees corresponding to the net non-planarity, i.e.,
sqrt(buckle^2 + propeller^2).

This structure contains 2 non-Watson-Crick (with leading *) base pair(s)
--------------------------------------------------------------------------------
Simple base-pair parameters based on RC8--YC6 vectors
      bp         Shear    Stretch   Stagger    Buckle  Propeller  Opening  angle
*    1 G+G       -3.25      2.26     -0.06     -3.16     -4.14    -86.94    5.2
*    2 G+G       -3.04      2.28      0.35      8.01    -13.51    -89.36   15.7
--------------------------------------------------------------------------------
Simple base-pair step parameters based on consecutive C1'-C1' vectors
      step       Shift     Slide      Rise      Tilt      Roll     Twist
*    1 GG/GG      0.60     -0.27      3.16      0.36      6.07     43.77
********************************************************************************
Structure classification:

This is a parallel duplex structure
********************************************************************************
lambda: virtual angle between C1'-YN1 or C1'-RN9 glycosidic bonds and the
        base-pair C1'-C1' line

C1'-C1': distance between C1' atoms for each base-pair
RN9-YN1: distance between RN9-YN1 atoms for each base-pair
RC8-YC6: distance between RC8-YC6 atoms for each base-pair

    bp     lambda(I) lambda(II)  C1'-C1'   RN9-YN1   RC8-YC6
   1 G+G      60.9      41.2      10.6       8.8       8.3
   2 G+G      61.0      22.5      11.1       9.1       8.6
********************************************************************************
Classification of each dinucleotide step in a right-handed nucleic acid
structure: A-like; B-like; TA-like; intermediate of A and B, or other cases.

    step       Xp      Yp      Zp     XpH     YpH     ZpH    Form
   1 GG/GG   -1.74    5.96    1.26   -3.10    6.02    1.58
********************************************************************************
Minor and major groove widths: direct P-P distances and refined P-P distances
   which take into account the directions of the sugar-phosphate backbones

   (Subtract 5.8 Angstrom from the values to take account of the vdw radii
    of the phosphate groups, and for comparison with FreeHelix and Curves.)

Ref: M. A. El Hassan and C. R. Calladine (1998). ``Two Distinct Modes of
     Protein-induced Bending in DNA.'' J. Mol. Biol., v282, pp331-343.

                  Minor Groove        Major Groove
```

```
                P-P     Refined     P-P     Refined
   1 GG/GG      ---       ---       ---       ---
********************************************************************************
Global linear helical axis defined by equivalent C1' and RN9/YN1 atom pairs
Deviation from regular linear helix: 4.02(0.03)
Helix:     0.0981    0.1157   -0.9884
HETATM 9998  XS    X X 999      -1.554   0.751   6.840
HETATM 9999  XE    X X 999      -1.159   1.216   2.862
Average and standard deviation of helix radius:
      P: 10.05(1.51), O4': 7.46(0.10),  C1': 6.68(0.25)

Global parameters based on C1'-C1' vectors:

disp.: displacement of the middle C1'-C1' point from the helix
angle: inclination between C1'-C1' vector and helix (subtracted from 90)
twist: helical twist angle between consecutive C1'-C1' vectors
rise:  helical rise by projection of the vector connecting consecutive
       C1'-C1' middle points onto the helical axis

     bp       disp.    angle     twist      rise
   1 G+G      4.11     -6.03     44.23      4.02
   2 G+G      3.74     -5.99      ---       ---
********************************************************************************
Main chain and chi torsion angles:

Note: alpha:   O3'(i-1)-P-O5'-C5'
      beta:    P-O5'-C5'-C4'
      gamma:   O5'-C5'-C4'-C3'
      delta:   C5'-C4'-C3'-O3'
      epsilon: C4'-C3'-O3'-P(i+1)
      zeta:    C3'-O3'-P(i+1)-O5'(i+1)

      chi for pyrimidines(Y): O4'-C1'-N1-C2
          chi for purines(R): O4'-C1'-N9-C4

Strand I
  base    alpha    beta   gamma   delta  epsilon   zeta    chi
   1 G     ---     ---     77.8   137.2   177.5  -100.7  -123.5
   2 G    -82.4   167.1    94.3   124.2  -179.1  -166.5  -120.2

Strand II
  base    alpha    beta   gamma   delta  epsilon   zeta    chi
   1 G    -17.2  -133.9    78.5    98.1  -177.5   -93.9  -152.7
   2 G   -111.7  -178.8    91.4   132.4   151.0  -157.3  -137.0
********************************************************************************
Sugar conformational parameters:

Note: v0: C4'-O4'-C1'-C2'
      v1: O4'-C1'-C2'-C3'
      v2: C1'-C2'-C3'-C4'
      v3: C2'-C3'-C4'-O4'
      v4: C3'-C4'-O4'-C1'

      tm: the amplitude of pucker
      P:  the phase angle of pseudorotation

Strand I
 base       v0      v1      v2      v3      v4      tm       P    Puckering
   1 G    -23.8    30.3   -24.8    13.0     6.0    29.2   148.3    C2'-endo
   2 G     -2.9    20.3   -26.1    26.4   -15.2    26.8   192.9    C3'-exo

Strand II
 base       v0      v1      v2      v3      v4      tm       P    Puckering
   1 G    -24.3     9.1     7.6   -20.3    25.9    27.0    73.6    O4'-endo
   2 G     -7.2     5.0    -1.0    -2.9     6.1     6.9    98.0    O4'-endo
********************************************************************************
Same strand P--P and C1'--C1' virtual bond distances
```

```
                Strand I                          Strand II
    step      P--P     C1'--C1'       step      P--P     C1'--C1'
   1 G/G       ---      6.22         1 G/G      7.92      6.65
*****************************************************************************
Helix radius (radial displacement of P, O4', and C1' atoms in local helix
   frame of each dimer)

                       Strand I                          Strand II
     step          P        O4'       C1'        P        O4'        C1'
   1 GG/GG     10.58      8.47      7.73     12.65      8.14      7.17
*****************************************************************************
Position (Px, Py, Pz) and local helical axis vector (Hx, Hy, Hz)
         for each dinucleotide step

     step       Px        Py        Pz        Hx        Hy        Hz
   1 GG/GG     -0.32      0.58      5.27      0.01     -0.09     -1.00
```

3D-NuS input for 6GE1 structure

Figure S9. Input parameters for 6GE1 structure.

Quadruplex class [strand orientation]: Tetramer all parallel strands (Q10)

Quadruplex sub-class [strand type]: RRRR

Number of G-quartets: 4

Insert sequence for 1st loop 5'---3':

Insert sequence for 2nd loop 5'---3':

Insert sequence for 3rd loop 5'---3':

3D-NuS output for 6GE1 structure

Figure S10. Generated 3D model.

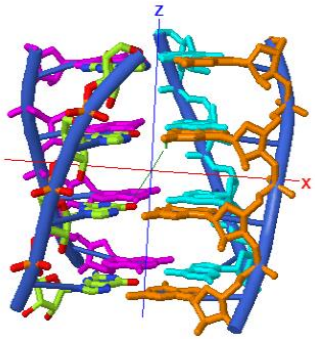


Figure S11. Phase angles of pseudorotation of sugar ring of nucleotide bases for strand I, II, III, IV of 6GE1 structure.

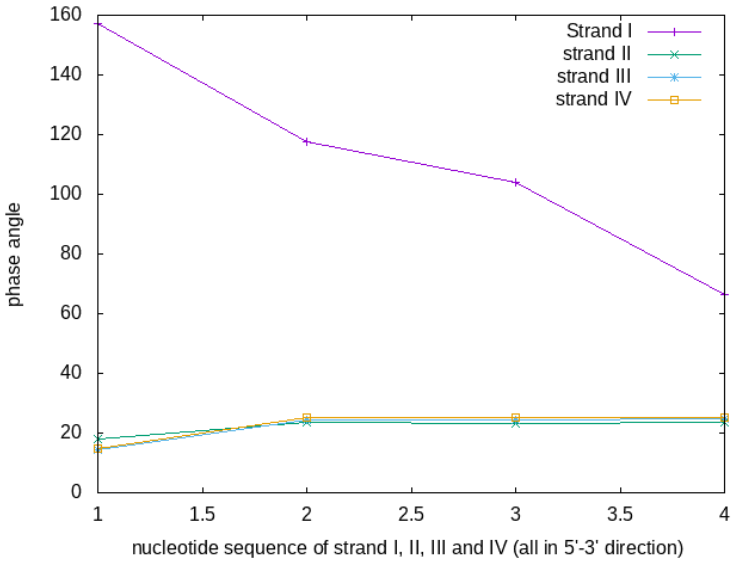


Figure S12. Backbone torsion angles for strand I of 6GE1 structure.

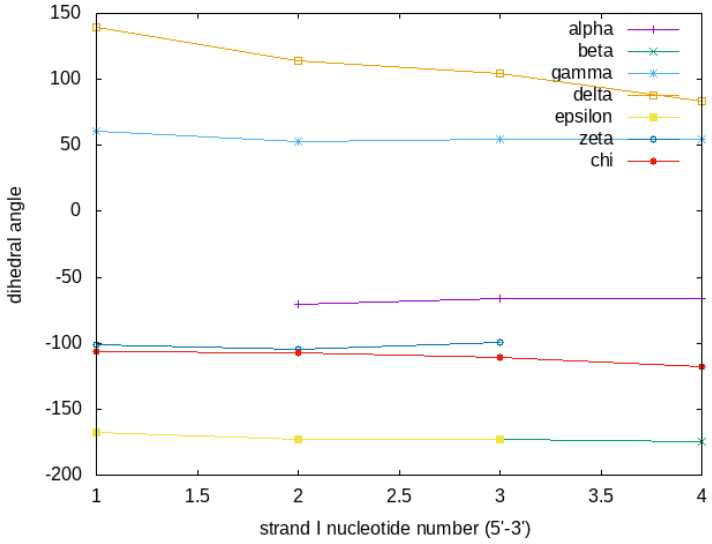


Figure S13. Backbone torsion angles for strand II of 6GE1 structure.

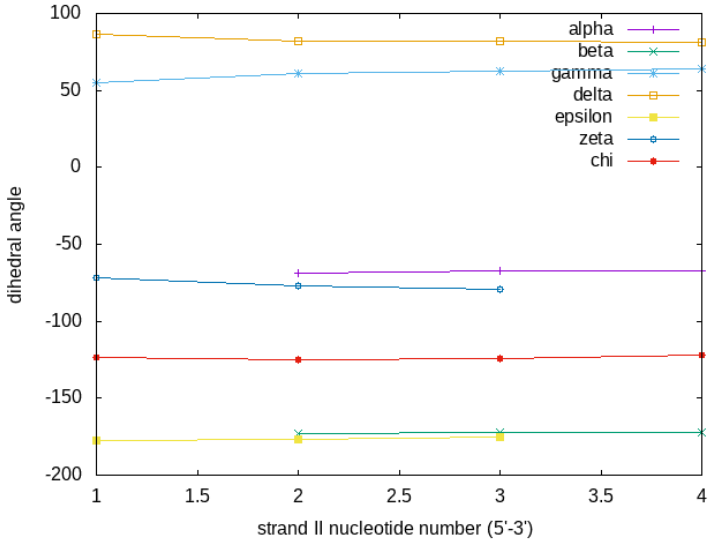


Figure S14. Backbone torsion angles for strand III of 6GE1 structure.

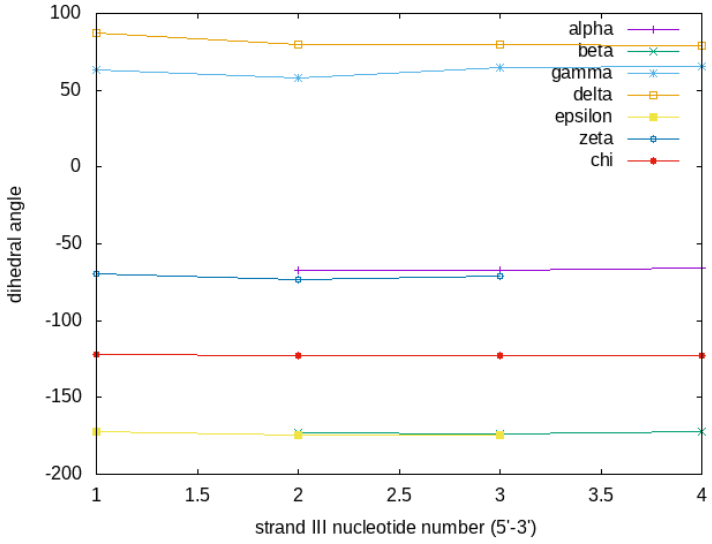
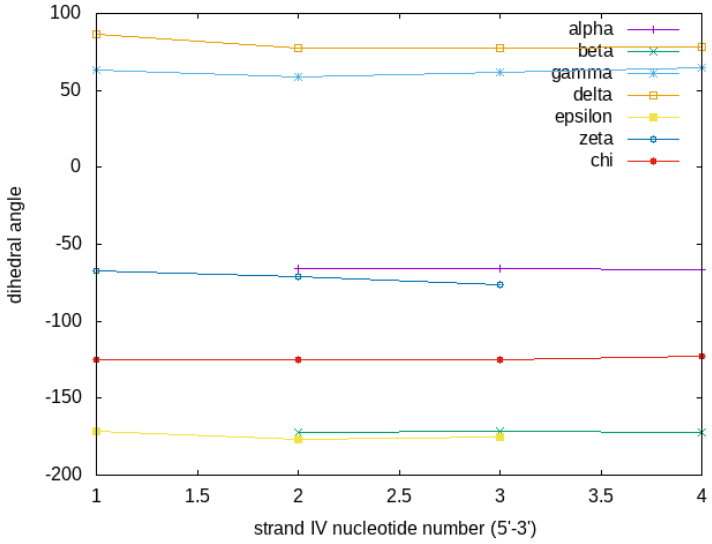


Figure S15. Backbone torsion angles for strand IV of 6GE1 structure.



Output file

```
This structure has broken O3'[i] to P[i+1] linkages
********************************************************************************
    3DNA v2.3.2-2017dec26, created and maintained by Xiang-Jun Lu (PhD)
********************************************************************************
1. The list of the parameters given below correspond to the 5' to 3' direction
   of strand I and 3' to 5' direction of strand II.

2. All angular parameters, except for the phase angle of sugar pseudo-
   rotation, are measured in degrees in the range of [-180, +180], and all
   displacements are measured in Angstrom units.
********************************************************************************
File name: quadruplex_3dnus.pdb.temp
Date and time: Thu Jul 16 22:20:36 2020

Number of base-pairs: 8
Number of atoms: 540
********************************************************************************
********************************************************************************
RMSD of the bases (----- for WC bp, + for isolated bp, x for helix change)

            Strand I                    Strand II          Helix
   1   (0.004) ....>-:...1_:[GUA]G-**+-G[GUA]:...5_:-<.... (0.004)     |
   2   (0.006) ....>-:..13_:[GUA]G-**+-G[GUA]:...9_:-<.... (0.003)     |
   3   (0.005) ....>-:...2_:[GUA]G-**+-G[GUA]:...6_:-<.... (0.005)     x
   4   (0.005) ....>-:..10_:[GUA]G-**+-G[GUA]:..14_:-<.... (0.008)     +
   5   (0.005) ....>-:...4_:[GUA]G-**+-G[GUA]:...8_:-<.... (0.003)     |
   6   (0.005) ....>-:..11_:[GUA]G-**+-G[GUA]:...7_:-<.... (0.004)     |
   7   (0.005) ....>-:..12_:[GUA]G-**+-G[GUA]:..16_:-<.... (0.004)     |
   8   (0.007) ....>-:...3_:[GUA]G-**+-G[GUA]:..15_:-<.... (0.007)     |

Note: This structure contains 8[8] non-Watson-Crick base-pairs.
********************************************************************************
Detailed H-bond information: atom-name pair and length [ O N]
```

```
   1 G-**+-G  [2]  N2 - N7  2.94  O6 * O6  2.71
   2 G-**+-G  [2]  O6 * O6  2.71  N7 - N2  2.89
   3 G-**+-G  [2]  N2 - N7  2.91  O6 * O6  2.78
   4 G-**+-G  [2]  N2 - N7  2.92  O6 * O6  2.74
   5 G-**+-G  [2]  N2 - N7  2.96  N1 - O6  3.15
   6 G-**+-G  [2]  O6 * O6  2.71  N7 - N2  2.93
   7 G-**+-G  [2]  N2 - N7  2.99  N1 - O6  3.15
   8 G-**+-G  [2]  O6 * O6  2.73  N7 - N2  2.93
*******************************************************************************
Overlap area in Angstrom^2 between polygons defined by atoms on successive
bases. Polygons projected in the mean plane of the designed base-pair step.

Values in parentheses measure the overlap of base ring atoms only. Those
outside parentheses include exocyclic atoms on the ring. Intra- and
inter-strand overlap is designated according to the following diagram:

                    i2  3'      5' j2
                       /|\      |
                        |       |
                Strand I |      | II
                         |      |
                         |      |
                         |     \|/
                    i1  5'      3' j1

     step      i1-i2        i1-j2        j1-i2        j1-j2        sum
   1 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   2 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   3 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   4 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   5 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   6 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
   7 GG/GG  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)  0.00( 0.00)
*******************************************************************************
Origin (Ox, Oy, Oz) and mean normal vector (Nx, Ny, Nz) of each base-pair in
   the coordinate system of the given structure

      bp        Ox        Oy        Oz        Nx        Ny        Nz
    1 G+G     -1.645    -0.538    16.454     0.120     0.033    -0.992
    2 G+G      2.284    -1.330    16.860     0.081     0.011    -0.997
    3 G+G     -1.344     0.394    11.232     0.117     0.022    -0.993
    4 G+G      2.332    -1.252    11.555     0.053     0.004    -0.999
    5 G+G      0.082     2.194    -0.120     0.090     0.020    -0.996
    6 G+G      1.908     1.761     5.866     0.079     0.005    -0.997
    7 G+G      1.949    -1.194    -0.026     0.054     0.015    -0.998
    8 G+G     -0.463    -1.531     5.674     0.079     0.017    -0.997
*******************************************************************************
Local base-pair parameters
     bp        Shear    Stretch   Stagger    Buckle  Propeller  Opening
    1 G+G       0.74      3.92      0.02     -0.98      0.44    -91.82
    2 G+G      -0.73     -3.94     -0.03      0.86      2.11     90.36
    3 G+G       0.84      3.93     -0.03     -1.09     -0.16    -90.19
    4 G+G       0.79      3.92      0.01      0.16     -2.85    -91.00
    5 G+G       2.52      3.33     -0.05      0.57      0.63    -82.43
    6 G+G      -0.76     -3.92      0.03     -0.31      3.49     91.36
    7 G+G       2.51      3.32     -0.08      0.39     -1.14    -83.13
    8 G+G      -0.79     -3.92      0.07     -3.47      3.19     91.00
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
      ave.       0.64      0.83     -0.01     -0.48      0.72    -20.73
      s.d.       1.36      3.95      0.05      1.40      2.16     92.51
*******************************************************************************
Local base-pair step parameters
    step       Shift     Slide      Rise      Tilt      Roll     Twist
   1 GG/GG     -0.81     -3.95     -0.03     -1.58      2.06    179.85
   2 GG/GG      0.72      4.42      5.27      1.45     -1.63   -164.28
   3 GG/GG      ----      ----      ----      ----      ----      ----
   4 GG/GG      ----      ----      ----      ----      ----      ----
   5 GG/GG     -0.30     -2.34     -5.81     -0.36      1.02     73.79
```

```
   6 GG/GG     -0.51     -2.99      5.85     -0.16      1.58    106.21
   7 GG/GG     -0.81     -1.88     -5.85     -1.35      0.49     74.08
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
      ave.     -0.34     -1.35     -0.11     -0.40      0.70     53.93
      s.d.      0.63      3.31      5.70      1.20      1.43    129.43
*******************************************************************************
Local base-pair helical parameters
    step       X-disp    Y-disp   h-Rise     Incl.       Tip   h-Twist
   1 GG/GG     -1.97      0.41     -0.03      1.03      0.79    179.85
   2 GG/GG     -2.22      0.37      5.28      0.82      0.73   -164.28
   3 GG/GG      ----      ----      ----      ----      ----      ----
   4 GG/GG      ----      ----      ----      ----      ----      ----
   5 GG/GG     -1.89      0.27     -5.84      0.85      0.30     73.80
   6 GG/GG     -1.91      0.31      5.82      0.98      0.10    106.22
   7 GG/GG     -1.53      0.75     -5.85      0.41      1.12     74.09
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
      ave.     -1.90      0.42     -0.12      0.82      0.61     53.93
      s.d.      0.25      0.19      5.70      0.25      0.41    129.43
*******************************************************************************
The 'simple' base-pair and step parameters were introduced into 3DNA as of
v2.3-2016jan01. This list of 'simple' parameters is reported by default,
but can be turned off by specifying 'analyze -simple=false'.

The simple parameters are 'intuitive' for non-Watson-Crick base pairs and
associated base-pair steps, where the above corresponding 3DNA parameters
may appear cryptic. Note that the following sets of simple parameters are
for structural description only, not to be fed into the 'rebuild' program.
Overall, they complement the rigorous characterization of base-pair
geometry, as exemplified by the original 3DNA analyze/rebuild programs.

In short, the 'simple' base-pair parameters employ the RC8--YC6 (default,
or RN9--YN1) vector as the (long) y-axis of the pair. As before, the
z-axis is the average of two base normals, taking consideration of the M-N
vs M+N base-pair classification. In essence, the 'simple' parameters make
geometrical sense through the introduction of an ad hoc base-pair
reference frame in each case. See x3dna.org for more details. The same
idea applies to the 'simple' inter-base-pair step parameters, which use
consecutive C1'--C1' vectors.

Here, 'angle' refers to the inter-base angle of each pair, with values in
the range of [0, 90] degrees corresponding to the net non-planarity, i.e.,
sqrt(buckle^2 + propeller^2).

This structure contains 8 non-Watson-Crick (with leading *) base pair(s)
-------------------------------------------------------------------------------
Simple base-pair parameters based on RC8--YC6 vectors
      bp        Shear    Stretch   Stagger    Buckle  Propeller  Opening  angle
*    1 G+G      -3.19      2.40      0.02     -0.82     -0.68    -91.82    1.1
*    2 G+G      -3.19      2.42     -0.03      1.50     -1.72     90.35    2.3
*    3 G+G      -3.15      2.50     -0.03     -0.34     -1.05    -90.20    1.1
*    4 G+G      -3.16      2.45      0.01      2.62     -1.11    -90.98    2.9
*    5 G+G      -2.26      3.50     -0.05     -0.40      0.75    -82.43    0.9
*    6 G+G      -3.18      2.42      0.03      3.27     -1.27     91.34    3.5
*    7 G+G      -2.27      3.49     -0.08      1.21     -0.02    -83.13    1.2
*    8 G+G      -3.16      2.45      0.07      4.40      1.70     91.00    4.7
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
       ave.     -2.94      2.70     -0.01      1.43     -0.43    -20.73
       s.d.      0.42      0.49      0.05      1.90      1.16     92.50
-------------------------------------------------------------------------------
Simple base-pair step parameters based on consecutive C1'-C1' vectors
      step       Shift     Slide      Rise      Tilt      Roll     Twist
*    1 GG/GG      2.63     -0.02      3.05    177.41      0.17    171.95
*    2 GG/GG     -3.47     -2.83      5.27     -0.01      2.18     15.35
*    3 GG/GG      ----      ----      ----      ----      ----      ----
*    4 GG/GG      ----      ----      ----      ----      ----      ----
*    5 GG/GG      1.67     -1.66      5.81     -0.34      1.02     99.21
*    6 GG/GG     -2.24     -2.05      5.85      0.84      1.34    -80.83
*    7 GG/GG      1.79     -0.99      5.85      0.76      1.21     98.89
```

```
           ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ave.      0.08     -1.51      5.17     35.73      1.19     60.92
        s.d.      2.74      1.07      1.21     79.20      0.72     96.70
****************************************************************************
Structure classification:

****************************************************************************
lambda: virtual angle between C1'-YN1 or C1'-RN9 glycosidic bonds and the
        base-pair C1'-C1' line

C1'-C1': distance between C1' atoms for each base-pair
RN9-YN1: distance between RN9-YN1 atoms for each base-pair
RC8-YC6: distance between RC8-YC6 atoms for each base-pair

    bp     lambda(I) lambda(II)  C1'-C1'   RN9-YN1   RC8-YC6
   1 G+G      56.7      30.6      11.3       9.3       8.8
   2 G+G      31.8      57.5      11.2       9.2       8.7
   3 G+G      58.9      31.5      11.3       9.3       8.8
   4 G+G      57.9      31.0      11.3       9.3       8.8
   5 G+G      69.5      28.8      11.7       9.9       9.6
   6 G+G      31.0      57.5      11.3       9.3       8.8
   7 G+G      69.1      28.1      11.7       9.9       9.7
   8 G+G      31.3      58.4      11.3       9.3       8.8
****************************************************************************
Classification of each dinucleotide step in a right-handed nucleic acid
structure: A-like; B-like; TA-like; intermediate of A and B, or other cases.

    step       Xp      Yp      Zp     XpH     YpH     ZpH    Form
   1 GG/GG    ---     ---     ---     ---     ---     ---     ---
   2 GG/GG    ---     ---     ---     ---     ---     ---     ---
   3 GG/GG    ---     ---     ---     ---     ---     ---     ---
   4 GG/GG    ---     ---     ---     ---     ---     ---     ---
   5 GG/GG  -5.17   -5.57   -2.79   -6.69   -5.53   -2.84
   6 GG/GG   4.35    4.65    2.62    3.21    4.62    2.68
   7 GG/GG  -5.90   -6.62   -3.07   -7.16   -6.61   -3.05
****************************************************************************
Minor and major groove widths: direct P-P distances and refined P-P distances
   which take into account the directions of the sugar-phosphate backbones

   (Subtract 5.8 Angstrom from the values to take account of the vdw radii
    of the phosphate groups, and for comparison with FreeHelix and Curves.)

Ref: M. A. El Hassan and C. R. Calladine (1998). ``Two Distinct Modes of
     Protein-induced Bending in DNA.'' J. Mol. Biol., v282, pp331-343.

                Minor Groove        Major Groove
               P-P     Refined     P-P     Refined
   1 GG/GG     ---       ---       ---       ---
   2 GG/GG     ---       ---       ---       ---
   3 GG/GG     ---       ---       ---       ---
   4 GG/GG     ---       ---       ---       ---
   5 GG/GG     ---       ---       ---       ---
   6 GG/GG     ---       ---       ---       ---
   7 GG/GG     ---       ---       ---       ---
****************************************************************************
Global linear helical axis defined by equivalent C1' and RN9/YN1 atom pairs
Deviation from regular linear helix: 1.66(5.18)
****************************************************************************
Main chain and chi torsion angles:

Note: alpha:   O3'(i-1)-P-O5'-C5'
      beta:    P-O5'-C5'-C4'
      gamma:   O5'-C5'-C4'-C3'
      delta:   C5'-C4'-C3'-O3'
      epsilon: C4'-C3'-O3'-P(i+1)
      zeta:    C3'-O3'-P(i+1)-O5'(i+1)

      chi for pyrimidines(Y): O4'-C1'-N1-C2
```

chi for purines(R): O4'-C1'-N9-C4

Strand I

base	alpha	beta	gamma	delta	epsilon	zeta	chi
1 G	---	---	60.5	139.6	-167.9	-100.8	-106.2
2 G	---	---	63.2	86.6	-171.5	-67.1	-124.9
3 G	-70.6	-173.0	52.7	114.1	-173.0	-104.9	-107.2
4 G	-67.2	-173.0	58.3	79.8	-174.8	-73.4	-122.4
5 G	-66.5	-174.6	54.9	83.4	---	---	-117.4
6 G	-67.6	-173.5	64.6	80.1	-174.4	-71.2	-122.9
7 G	-65.9	-172.6	65.5	79.3	---	---	-122.5
8 G	-66.2	-173.0	55.0	104.9	-172.5	-99.8	-110.6

Strand II

base	alpha	beta	gamma	delta	epsilon	zeta	chi
1 G	---	---	55.2	86.5	-177.8	-71.7	-123.4
2 G	---	---	63.2	87.4	-172.4	-69.7	-121.8
3 G	-68.4	-172.7	60.8	82.1	-176.8	-77.1	-125.1
4 G	-66.1	-172.4	58.7	77.6	-176.7	-71.0	-125.2
5 G	-67.5	-172.5	64.2	81.6	---	---	-122.3
6 G	-67.4	-172.0	62.4	82.0	-175.1	-79.1	-124.1
7 G	-66.2	-171.9	65.1	78.5	---	---	-122.8
8 G	-65.8	-171.3	62.1	77.8	-175.6	-76.4	-124.7

****

Sugar conformational parameters:

Note: v0: C4'-O4'-C1'-C2'
v1: O4'-C1'-C2'-C3'
v2: C1'-C2'-C3'-C4'
v3: C2'-C3'-C4'-O4'
v4: C3'-C4'-O4'-C1'

tm: the amplitude of pucker
P: the phase angle of pseudorotation

Strand I

base	v0	v1	v2	v3	v4	tm	P	Puckering
1 G	-24.0	35.1	-32.3	19.5	2.6	35.1	157.1	C2'-endo
2 G	2.4	-24.0	35.1	-34.8	20.4	36.3	14.9	C3'-endo
3 G	-43.3	38.3	-19.8	-4.3	29.8	42.5	117.7	C1'-exo
4 G	-3.9	-20.5	35.8	-39.1	27.0	39.2	24.2	C3'-endo
5 G	-32.2	8.2	16.9	-35.5	42.6	42.0	66.3	C4'-exo
6 G	-3.9	-20.5	35.7	-39.0	27.0	39.2	24.2	C3'-endo
7 G	-4.2	-20.2	35.5	-39.0	27.2	39.1	24.7	C3'-endo
8 G	-42.6	31.6	-10.1	-13.7	35.2	41.3	104.1	O4'-endo

Strand II

base	v0	v1	v2	v3	v4	tm	P	Puckering
1 G	0.6	-21.9	33.6	-34.4	21.3	35.3	17.8	C3'-endo
2 G	2.6	-23.0	33.5	-33.0	19.2	34.6	14.5	C3'-endo
3 G	-3.3	-20.1	34.5	-37.5	25.6	37.7	23.6	C3'-endo
4 G	-4.8	-20.8	37.0	-40.8	28.5	40.9	25.1	C3'-endo
5 G	-3.3	-20.1	34.5	-37.4	25.6	37.6	23.5	C3'-endo
6 G	-3.1	-20.4	34.7	-37.5	25.5	37.8	23.2	C3'-endo
7 G	-4.6	-20.0	35.6	-39.2	27.5	39.3	25.1	C3'-endo
8 G	-4.8	-20.6	36.7	-40.6	28.4	40.6	25.2	C3'-endo

****

Same strand P--P and C1'--C1' virtual bond distances

Strand I			Strand II		
step	P--P	C1'--C1'	step	P--P	C1'--C1'
1 G/G	---	11.15	1 G/G	---	11.18
2 G/G	---	13.90	2 G/G	---	11.25
3 G/G	---	---	3 G/G	---	---
4 G/G	---	---	4 G/G	---	---
5 G/G	21.80	16.52	5 G/G	6.14	6.09
6 G/G	6.17	5.92	6 G/G	21.26	17.37
7 G/G	23.11	16.63	7 G/G	6.09	6.11

****
Helix radius (radial displacement of P, O4', and C1' atoms in local helix frame of each dimer)

	Strand I			Strand II		
step	P	O4'	C1'	P	O4'	C1'
1 GG/GG	----	8.55	7.92	----	8.59	7.95
2 GG/GG	11.53	8.68	8.03	----	8.71	8.07
3 GG/GG	----	----	----	----	----	----
4 GG/GG	----	----	----	----	----	----
5 GG/GG	10.02	8.49	7.78	10.44	8.86	8.22
6 GG/GG	10.42	8.53	7.81	9.96	8.92	8.25
7 GG/GG	11.46	8.61	7.88	10.79	9.07	8.33

****
Position (Px, Py, Pz) and local helical axis vector (Hx, Hy, Hz) for each dinucleotide step

step	Px	Py	Pz	Hx	Hy	Hz
1 GG/GG	0.32	-0.94	16.66	0.10	0.02	-0.99
2 GG/GG	0.36	-0.75	14.03	0.10	0.02	-1.00
3 GG/GG	----	----	----	----	----	----
4 GG/GG	----	----	----	----	----	----
5 GG/GG	0.78	0.47	2.83	0.07	0.02	-1.00
6 GG/GG	0.78	0.42	2.84	0.07	0.02	-1.00
7 GG/GG	0.52	-0.02	2.81	0.07	-0.00	-1.00

Published online DD MM 2021

Nucleic Acids Research, 2021, Vol. xx, No. xx ***1–5***
doi:10.1093/nar/gkn000

ONQUADRO: a database of experimentally determined quadruplex structures

Tomasz Zok [1], Natalia Kraszewska [1], Joanna Miskiewicz [1], Paulina Pielacinska [1], Michal Zurkowski [1], and Marta Szachniuk [1,2*]

[1]Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland, and
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland

Received 2021-08-06; Revised 2021-MM-DD; Accepted 2021-MM-DD

ABSTRACT

ONQUADRO is an advanced database system that supports studying structures of canonical and non-canonical quadruplexes. It combines a relational database that collects comprehensive information on tetrads, quadruplexes, and G4-helices; programs to compute structure parameters and visualize the data; scripts for statistical analysis; automatic update and newsletter modules; a web application that provides a user interface. The database is a self-updating resource, with new arrivals coming once a week. The preliminary data are downloaded from Protein Data Bank, processed, annotated, and completed. As of August 2021, ONQUADRO contains 1,661 tetrads, 518 quadruplexes, and 30 G4-helices found in 467 experimentally determined 3D structures of nucleic acids. Users can view and download their description: sequence, secondary structure (dot-bracket, classical diagram, arc diagram), tertiary structure (ball-and-stick, surface or vdw-ball model, layer diagram), planarity, twist, rise, chi angle (value and type), loop characteristics, strand directionality, metal ions, ONZ, and Webba da Silva classification (the latter by loop topology and tetrad combination), origin structure ID, assembly ID, experimental method, and molecule type. The database is freely available at https://onquadro.cs.put.poznan.pl/. One can use it both from desktop computers and mobile devices.

INTRODUCTION

G-quadruplexes (G4s) are unique structures folded in G-rich nucleic acids (1, 2), found in eukaryotic, prokaryotic, and viral genomes (1, 3, 4). In biological processes, they play crucial regulatory roles by participating in telomere maintenance, regulation of gene expression, DNA replication, etc. (2, 3, 5, 6). Recent hypothesis, coined as the quadruplex world, suggests that G4s may have been the first simple molecules to appear on Earth (7). It takes the cue from the ability of guanines to form stable G-tetrads, the basic building units of a quadruplex. In the tetrad, four guanines arranged in the plane connect via hydrogen bonds such that each of them acts as a donor of two hydrogen bonds at the Watson-Crick edge and an acceptor at the Hoogsteen edge (1, 8, 9). Complete G4 assembles of at least two G-tetrads stacked one above another and stabilized by monovalent cations located in the ion channel (5, 9).

The general definition of what constitutes a quadruplex does not capture the complexity of its structure and feature diversity (2, 8, 9, 10, 11, 12, 13). Meanwhile, the latter is subjected to various studies, aiming - among others - to associate the motif's conformation with its function, find the relationship between the sequence and higher-level structure, cluster and classify quadruplexes, learn and fully describe their properties. We already know that tetrads can form from guanine as well as non-guanine nucleotides (14). Their spatial arrangement to the stacking neighbors is diverse, defined by the rise and twist parameters. The topologies of secondary structures differ in both the tetrad and the quadruplex set, as reflected in the ONZ classification (15). They are influenced by the number of strands contributing to the motif, their lengths, and directionality. Strands may form loops, which can be a part of the quadruplex - c.f. Webba da Silva formalism (16). The list of analyzed attributes also includes glycosidic bond angles, groove width, number of stacked tetrads, or G-tract continuity and is probably not yet complete (10).

In the recent decade, the unique structure of the quadruplex has focused the attention of many researchers, especially in the medical sciences. G4s have become therapeutic targets, i.a. for cancer and antiviral treatment (17, 18, 19). In the latter case, increased interest in targeting G-quadruplexes in viral genomes was prompted by the Covid-19 pandemic. The frequency and localization of putative quadruplex sequences in different viral taxa, G4-binding viral domains, and the G4 potential as viral biosensors were investigated (19, 20, 21, 22, 23, 24). These and other quadruplex studies provided a wealth of data for collection, organization, and further analysis (25, 26, 27). It has initiated the development of computational methods and bioinformatics tools dedicated to G4s. Most of them deal with sequence data storage and processing (14, 28, 29, 30, 31, 32, 33, 34). A few address higher-level structures (35, 36, 37, 38, 39), including databases that store

*To whom correspondence should be addressed. Tel: +48 616653030; Fax: +48 618771525; Email: mszachniuk@cs.put.poznan.pl

G4-related data (35, 40, 41) - none of them, however, collects complete information about quadruplex structures at all levels of their organization.

ONQUADRO is a new comprehensive database system that collects and shares data on tetrads, quadruplexes, and G4-helices, whose three-dimensional structures were determined experimentally. Baseline data are regularly downloaded from the Protein Data Bank (42) and supplemented with parameters computed by specialized procedures of the system's engine. Incorporated programs prepare visualizations of the secondary and tertiary structure models of each motif. The analytical module generates statistics of the distribution of structural parameters in the set of tetrads and quadruplexes. The system allows users to subscribe to a newsletter that informs of all database newcomers. ONQUADRO, designed to use on desktop computers and mobile devices, is freely available at https://onquadro.cs.put.poznan.pl/.

METHOD OUTLINE

Every Thursday, the update module of ONQUADRO connects to the PDB FTP site and searches for new information about nucleic acids (including protein-nucleic acid complexes). Next, it queries PDBe (43) for biological assemblies to associate them with items found. The module creates a list with identifiers of newly added, modified, or deleted structures containing tetrads, quadruplexes, and G4-helices. Changes to the ONQUADRO database are made upon the list with modified and deleted structures; the entries for new motifs are created and added. The process of new data preparation takes place in several steps (Figure 1).

For each quadruplex, we derive the secondary structure and prepare its representation in the two-line extended dot-bracket notation, compute the rise and twist parameters, identify the number of contributing strands and tetrads, determine strand direction, find loops to calculate their lengths and types, classify due to Webba da Silva formalism based on loop topology and tetrad combination (16), assign ONZ class upon secondary structure topology (15). If the nucleic acid contains metal ions, the procedure determines their position relative to the quadruplex. Then, we describe every tetrad by planarity, chi angle value and type, ONZ class, and tetrad combination. In the next step, graphical models of the secondary and tertiary structure of every motif are prepared. They include a classical diagram, arc diagram, layer diagram, and 3D molecule models. After calculating all the parameters and preparing graphical models, the system populates the database and maintains the relationships between the entries.

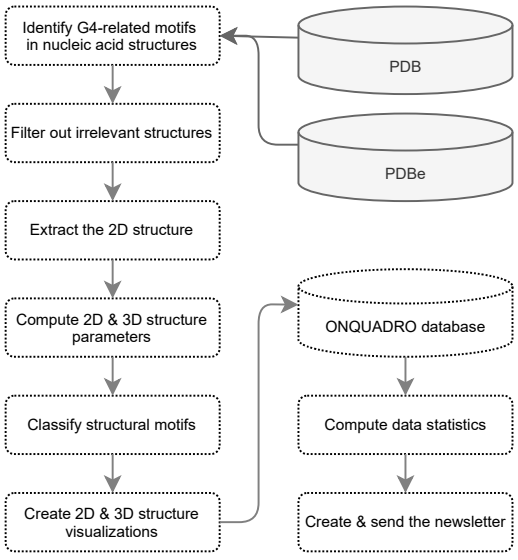


Figure 1. Data flow during the weekly update of ONQUADRO.

Once the database is updated, the statistical analysis module generates graphs and tables of the data distribution. Statistics available for G4s are (1) the number of quadruplexes as a function of the number of constituent tetrads; (2) the abundance of the set of uni-, bi-, and tetramolecular quadruplexes; (3) ONZ class coverage by uni-, bi-, and tetramolecular quadruplexes; (4) geometric class distribution based on glycosidic bond angles and loop topology; (5) loop length distribution in the subsets of lateral, propeller and diagonal loops; (6) twist and rise value distribution. Statistics prepared for tetrads include (1) distribution of tetrads concerning their sequence and molecule type; (2) ONZ class coverage by uni-, bi-, and tetramolecular tetrads; (3) chi angle value distribution in ONZ classes; (4) ONZ class coverage by ions; (5) planarity value distribution in the tetrad set.

Finally, the system creates a hypertext newsletter listing all changes to the database and sends it to the subscribers.

IMPLEMENTATION

The ONQUADRO system consists of the database, web application, and computational engine. It runs on a quad-core machine with 8 GB RAM in the Ubuntu GNU/Linux environment, hosted and maintained by the Institute of Computing Science, Poznan University of Technology.

Database

ONQUADRO has been developed as the relational database in PostgreSQL. It is composed of tables that correspond to PDB structures, G4-helices, quadruplexes, tetrads, tetrad pairs, base pairs, nucleotides, tracts, loops, and ions. The database stores the following information about every nucleic acid structure: PDB id, assembly id, experimental method, resolution, deposition, release and revision dates, molecule name, a secondary structure diagram, three-dimensional structure. The secondary and tertiary structures are collected separately for tetrads, quadruplexes, and G4-helices. Additionally, the database contains data on the ONZ class, type (uni-, bi-, or tetramolecular), loops, and ions for quadruplexes, ONZ class and planarity for tetrads, strand direction, rise and twist for tetrad pairs, stericity and edge names for base pairs, model, chain, glycosidic bond for nucleotides.

All sequences in ONQUADRO are coded in the one-letter format, in the 5'–3' direction. The secondary structures of tetrads, quadruplexes, and G4-helices are represented using

the dot-bracket notation - an unpaired nucleotide corresponds to a dot and a base pair to a pair of opening and closing brackets. Since in considered motifs, each nucleotide pairs with two others, basic dot-bracket notation is not sufficient to unambiguously encode the secondary structure of a tetrad, as well as a quadruplex or G4-helix. Therefore, in (15) we introduced a two-line dot-bracket and used an extended set of brackets to label paired nucleotides. It includes parentheses ( ), square brackets [ ], curly brackets { }, and angle brackets ⟨ ⟩. Arc diagram representing the secondary structure is also adjusted to reflect all pairings unambiguously. It is associated with dot-bracket notation - the top of the diagram corresponds to the first line of dot-bracket notation, and the bottom part corresponds to the second line (15). The secondary structure of every motif is also visualized in the classical diagram. The 3D structure is represented by a layer diagram and three molecular models (balls-and-sticks, surface, vdw-balls).

Computational engine

The computational engine is composed of scripts utilizing in-house and third-party procedures, responsible for data collection, quadruplex identification, computation of structure parameters, secondary structure annotation, visualization of the secondary and tertiary structure models, database queries, generation of statistics, and newsletter preparation. DSSR (`--pair-only` mode) (35) and ElTetrado (38) functionalities are applied to identify quadruplexes, tetrads, and G4-helices in nucleic acid structures. Procedures from in ElTetrado (38) and the BioCommons library (44) compute a variety of structure parameters. The VARNA-based routine (45) creates a classical diagram of the secondary structure. The R-Chie-driven function (46) produces the top-down arc diagram. The embedded LiteMol (47) module generates models (balls-and-sticks, surface, vdw-balls) of the three-dimensional structure. The Python script draws the layer diagram of the quadruplex based on data in JSON format obtained from ElTetrado. Every nucleotide in the diagram is color-coded – yellow indicates the anti conformation, orange is for syn. The script optimizes the quadruplex position in three-dimensional space to get a clear view of G4, with the least number of crossing strands. The optimization algorithm has been implemented in C++. Statistics are generated using the script in R and the plotly library. They self-update whenever new entries appear in the database.

Web application

The web application provides an interface to the ONQUADRO system. The client application has been created using the Angular framework and Bootstrap styling sheet; the server is implemented in C#. The client and the server communicate via REST API.

Figure 3 presents screenshots of the ONQUADRO system. On the homepage (https://onquadro.cs.put.poznan.pl/), users can see a brief information about the database resources and the latest update. From this page, one can enter the set of tetrads, quadruplexes, G4-helices, or structures. The selected subpage displays a list of items with a basic description. Some data are clickable - they allow to see detailed structural information about the selected element or link to the corresponding page in the Protein Data Bank. The

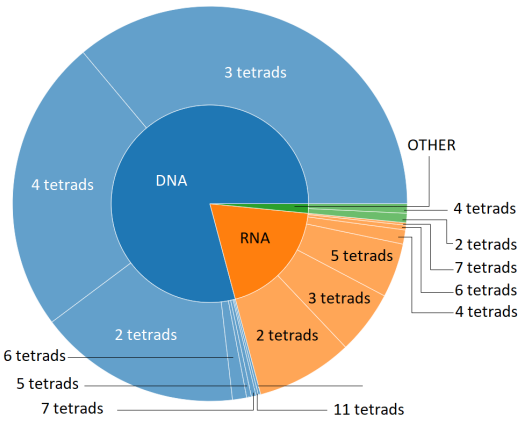


Figure 2. Example statistics generated by ONQUADRO: quadruplexes by the number of constituent tetrads.

table can be sorted (ascending or descending) to the contents of any column by clicking on the column header. Users can search the list of items by using the *Search the table* option and typing the string of interest. Searching runs in real-time. Element counter at top of the page shows how many elements contain the queried string. These elements are displayed in the table. Users can save the content of each table as a whole (*Save table* button) or in a selected part (*Save selected rows* button after clicking on check-boxes in the rightmost column). Tabular data (from any subpage) are downloaded in a CSV file, structure visualizations can be saved in the SVG format.

The *Statistics* option in the menu bar leads to a page listing statistics for tetrads and quadruplexes. User-selected stats are displayed in the graphical (pie chart, tree map, or bar plot) and tabular form. Plots can be saved in the HTML format. They are interactive - upon clicking the plot, one can enlarge its fragment and see selected part of data.

CONCLUSIONS

ONQUADRO gathers information about all tetrads, quadruplexes, and G4-helices found in experimentally determined nucleic acid structures deposited in the Protein Data Bank (42). The system's computational engine combines self-developed procedures to annotate these motifs, derive their secondary structures, classify according to the geometric formalism (16) and the topological ONZ nomenclature (15), represent the secondary structure in the dot-bracket notation and specially adjusted top-down arc diagram, draw the 3D model in the schematic layer diagram, and trigger statistics. Some of them are G4-adapted routines applied in our previously released tools; the others are brand new and have not been published yet (e.g., automatic creation of layer diagrams – a much-needed function in the research community). The user-friendly interface allows browsing the database contents divided into four subsets (tetrads, quadruplexes, G4-helices, PDB structures), searching and sorting the data by various parameters and keywords,

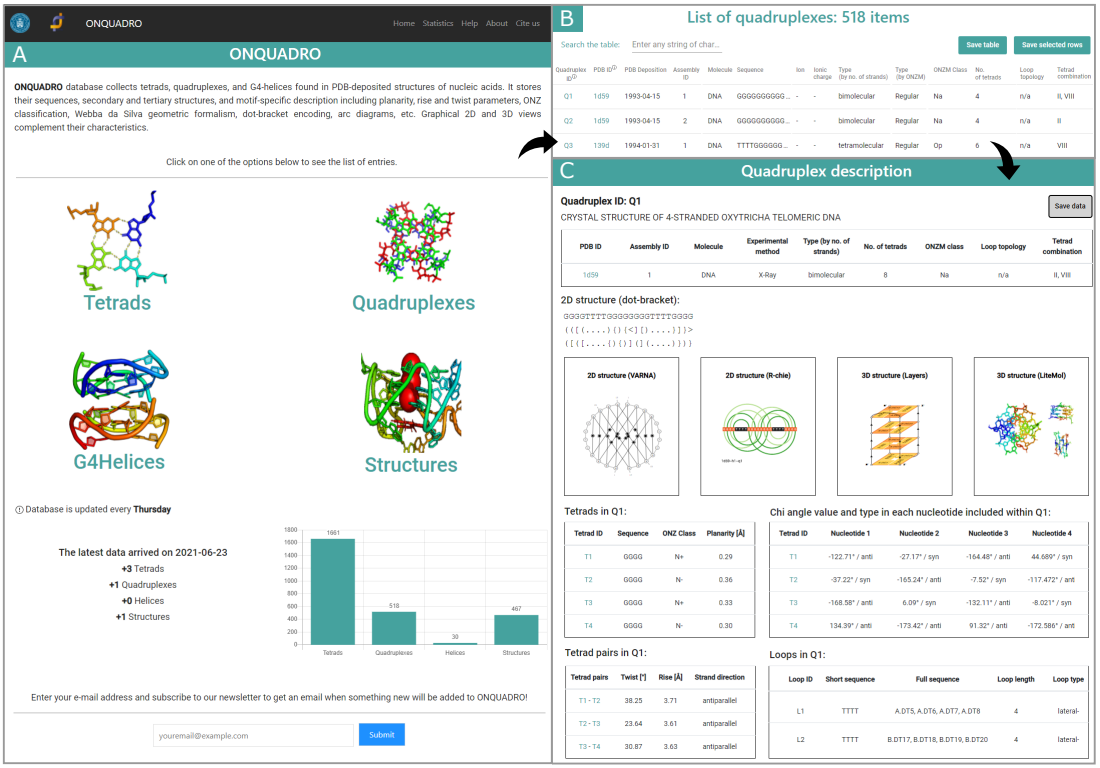


Figure 3. ONQUADRO interface: (A) main page, (B) quadruplex table, and (C) quadruplex details based on 2HY9 structure.

displaying and downloading detailed structural information on selected motifs, viewing and downloading statistics in graphical and textual form. ONQUADRO is a unique online resource that takes a comprehensive approach to collect and share quadruplex information. We hope it will facilitate the study of G4 structures and their modeling *in silico* - a great challenge for modern structural bioinformatics.

DATA AVAILABILITY

ONQUADRO is a continuously maintained, weekly self-updating resource available at https://onquadro.cs.put.poznan.pl. No registration or login is required to access the data and take full advantage of the system's functionality.

FUNDING

This work was supported by the National Science Center, Poland [2019/35/B/ST6/03074 to MS]; the statutory funds of Poznan University of Technology; and the Institute of Bioorganic Chemistry, PAS.

Conflict of interest statement. None declared.

REFERENCES

1. M. Malgowska, K. Czajczynska, D. Gudanis, A. Tworak, and Z. Gdaniec. Overview of RNA G-quadruplex structures. *Acta Biochim Pol*, 63:609–621, 2016.
2. C.K. Kwok and C.J. Merrick. G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol*, 35:997–1013, 2017.
3. A. Joachimi, A. Benz, and J.S. Hartig. A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg Med Chem*, 17:6811–6815, 2009.
4. M.D. Antonio, A. Ponjavic, A. Radzevicius, R.T. Ranasinghe, M. Catalano, X. Zhang, J. Shen, L.M. Needham, S.F. Lee, D. Klenerman, and S. Balasubramanian. Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat Chem*, 12:832–837, 2020.
5. M. Webba da Silva, M. Trajkovski, Y. Sannohe, N. Ma'ni Hessari, H. Sugiyama, and J. Plavec. Design of a G-Quadruplex Topology through Glycosidic Bond Angles. *Angew Chem Int Ed Engl*, 48:9167–9170, 2009.
6. S. Kolesnikova and E.A. Curtis. Structure and Function of Multimeric G-Quadruplexes. *Molecules*, 24:3074, 2019.
7. B. Kankia. Quadruplex World. *Orig Life Evol Biosph*, 2021.
8. S.A. Dvorkin, A.I. Karsisiotis, and M. Webba da Silva. Encoding canonical DNA quadruplex structure. *Sci Adv*, 4:eaat3007, 2018.
9. J. Spiegel, S. Adhikari, and S. Balasubramanian. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry*, 2:123–136, 2020.
10. S. Burge, G.N. Parkinson, P. Hazel, A.K. Todd, and S. Neidle. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res*, 34:5402–5415, 2006.
11. X. Cang, J. Sponer, and T.E. Cheatham. Explaining the varied glycosidic

conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic Acids Res*, 39:4499–4512, 2011.
12. V.T. Mukundan and A.T. Phan. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *J Am Chem Soc*, 135:5017–5028, 2013.
13. H.L. Lightfoot, T. Hagen, N.J. Tatum, and J. Hall. The diverse structural landscape of quadruplexes. *FEBS Letters*, 593:2083–2102, 2019.
14. J. Miskiewicz, J. Sarzynska, and M. Szachniuk. How bioinformatics resources work with G4 RNAs. *Brief Bioinform*, 22:bbaa201, 2021.
15. M. Popenda, J. Miskiewicz, J. Sarzynska, T. Zok, and M. Szachniuk. Topology-based classification of tetrads and quadruplex structures. *Bioinformatics*, 36:1129–1134, 2020.
16. M. Webba da Silva. Geometric Formalism for DNA Quadruplex Folding. *Chemistry*, 13:9738–9745, 2007.
17. J. Carvalho, J.L. Mergny, G.F. Salgado, J.A. Queiroz, and C. Cruz. G-quadruplex, Friend or Foe: The Role of the G-quartet in Anticancer Strategies. *Trends Mol Med*, 26:848–861, 2020.
18. R. Hansel-Hertsch, A. Simeone, A. Shea, W.W.I. Hui, K.G. Zyner, G. Marsico, O.M. Rueda, A. Bruna, A. Martin, X. Zhang, S. Adhikari, D. Tannahill, C. Caldas, and S. Balasubramanian. Landscape of G-quadruplex DNA structural regions in breast cancer. *Nat Genet*, 52:878–883, 2020.
19. N. Panera and A. Tozzi, A.and Alisi. The G-quadruplex/helicase world as a potential antiviral approach against COVID-19. *Drugs*, 80:941–946, 2020.
20. E. Lavezzo, M. Berselli, I. Frasson, R. Perrone, G. Palu, A.R. Brazzale, S.N. Richter, and S. Toppo. G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLoS Comput Biol*, 14:e1006675, 2018.
21. D. Ji, M. Juhas, C.M. Tsang, C.K. Kwok, Y. Li, and Y. Zhang. Discovery of G-quadruplex-forming sequences in SARS-CoV-2. *Brief Bioinform*, 22:1150–1160, 2020.
22. S.R. Wang, Y.Q. Min, J.Q. Wang, C.X. Liu, B.S. Fu, F. Wu, L.Y. Wu, Z.X. Qiao, Y.Y. Song, G.H. Xu, Z.G. Wu, G. Huang, N.F. Peng, R. Huang, W.X. Mao, S. Peng, Y.Q. Chen, Y. Zhu, T. Tian, X.L. Zhang, and X. Zhou. A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new antihepatitis C target. *Sci Adv*, 2:e1501535, 2016.
23. J. Tan, C. Vonrhein, O.S. Smart, G. Bricogne, M. Bollati, Y. Kusov, G. Hansen, J.R. Mesters, C.L. Schmidt, and R. Hilgenfeld. The SARS-Unique Domain (SUD) of SARS Coronavirus Contains Two Macrodomains That Bind G-Quadruplexes. *PLoS Pathog*, 5:e1000428, 2009.
24. H. Xi, M. Juhas, and Y. Zhang. G-quadruplex based biosensor: A potential tool for SARS-CoV-2 detection. *Biosens Bioelectron*, 167:112494, 2020.
25. D. Gudanis, L. Popenda, K. Szpotkowski, R. Kierzek, and Z. Gdaniec. Structural characterization of a dimer of RNA duplexes composed of 8-bromoguanosine modified CGG trinucleotide repeats: a novel architecture of RNA quadruplexes. *Nucleic Acids Res*, 44:2409–2416, 2016.
26. W. Andralojc, M. Malgowska, J. Sarzynska, K. Pasternak, K. Szpotkowski, R. Kierzek, and Z. Gdaniec. Unraveling the structural basis for the exceptional stability of RNA G-quadruplexes capped by a uridine tetrad at the 3' terminus. *RNA*, 25:121–134, 2018.
27. T. Frelih, B. Wang, J. Plavec, and P. Sket. Pre-folded structures govern folding pathways of human telomeric G-quadruplexes. *Nucleic Acids Res*, 48:2189–2197, 2020.
28. H.M. Wong, O. Stegle, S. Rodgers, and J.L. Huppert. A toolbox for predicting G-quadruplex formation and stability. *J Nucleic Acids*, 2010:564946, 2010.
29. J.M. Garant, M.J. Luce, M.S. Scott, and J.P. Perreault. G4RNA: an RNA G-quadruplex database. *Database*, page bav059, 2015.
30. A. Bedrat, L. Lacroix, and J.L. Mergny. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res*, 4:1746–1759, 2016.
31. A.B. Sahakyan, V.S. Chambers, G. Marsico, T. Santner, M.D. Antonio, and S. Balasubramanian. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep*, 7:14535, 2017.
32. F. Ge, Y. Wang, H. Li, R. Zhang, X. Wang, Q. Li, Z. Liang, and L. Yang. Plant-GQ: An Integrative Database of G-Quadruplex in Plant. *J Comput Biol*, 26:1013–1019, 2019.
33. E.P. Lombardi and A. Londono-Vallejo. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res*, 48:1603–1603, 2020.
34. M. Kudla, K. Gutowska, J. Synak, M. Weber, K.S. Bohnsack, P. Lukasiak, T. Villmann, J. Blazewicz, and M. Szachniuk. Virxicon: a lexicon of viral sequences. *Bioinformatics*, 36:5507–5513, 2020.
35. X.J. Lu, H.J. Bussemaker, and W.K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res*, 43:e142, 2015.
36. A. Rybarczyk, N. Szostak, M. Antczak, T. Zok, M. Popenda, R.W. Adamiak, J. Blazewicz, and M. Szachniuk. New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinformatics*, 16:276, 2015.
37. L.P.P. Patro, A. Kumar, N. Kolimi, and T. Rathinavelan. 3d-NuS: a web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. *J Mol Biol*, 429:2438–2448, 2017.
38. T. Zok, M. Popenda, and M. Szachniuk. ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *BMC Bioinformatics*, 21:40, 2020.
39. X.J. Lu. DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. *Nucleic Acids Res*, 48:e74, 2020.
40. Q. Li, J.F. Xiang, Q.F. Yang, H.X. Sun, A.J. Guan, and Y.L. Tang. G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res*, 41:D1115–D1123, 2012.
41. S.K. Mishra, A. Tawani, A. Mishra, and A. Kumar. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep*, 6:38144, 2016.
42. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28:235–242, 2000.
43. S. Velankar, Y. Alhroub, C. Best, S. Caboche, M.J. Conroy, J.M. Dana, M.A.F. Montecelo, G. van Ginkel, A. Golovin, S.P. Gore, A. Gutmanas, P. Haslam, P.M.S. Hendrickx, E. Heuson, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L.E. Newman, T.J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-Garcia, S. Sen, R. Slowley, A. Suarez-Uruena, G.J. Swaminathan, M.F. Symmons, W.F. Vranken, M. Wainwright, and G.J. Kleywegt. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*, 42:D445–D452, 2011.
44. Tomasz Zok. BioCommons: A robust Java library for RNA structural bioinformatics. *Bioinformatics*, 2021.
45. K. Darty, A. Denise, and Y. Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25:1974–1975, 2009.
46. D. Lai, J.R. Proctor, J.Y.A. Zhu, and I.M. Meyer. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res*, 40:e95, 2012.
47. D. Sehnal, M. Deshpande, R.S. Varekova, S. Mir, K. Berka, A. Midlik, L. Pravda, S. Velankar, and J. Koca. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat Methods*, 14:1121–1122, 2017.

Co-author declarations

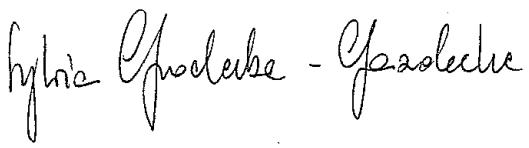
Sylwia Grodecka-Gazdecka
Department of Surgery, Chair and Clinic of Oncology
Poznan University of Medical Sciences
Szamarzewskiego 84, 60-569 Poznan, Poland

Poznan, 20 September, 2021

Declaration

I declare that in the following publication:

Tomasz P. Lehmann, Joanna Miskiewicz, Natalia Szostak, Marta Szachniuk, Sylwia Grodecka-Gazdecka, Pawel P. Jagodzinski. In Vitro and in Silico Analysis of miR-125a with rs12976445 Polymorphism in Breast Cancer Patients. Applied Sciences 10(20), 2020, pp. 7275
my contribution consisted in consulting the results and preparing a draft of the publication.



September 18, 2021

Natalia Kraszewska
ul. Słoneczna 11
78-200 Białogard

Declaration

Hereby, I declare that as a co-author of the paper

Tomasz Zok, Natalia Kraszewska, Joanna Miskiewicz, Paulina Pielacinska, Michal Zurkowski, Marta Szachniuk, *ONQUADRO: a database of experimentally determined quadruplex structure* (submitted)

I participated in the research carried under the supervision of prof. Marta Szachniuk described there. My task was to implement the web application of the ONQUADRO system.

Natalia Kraszewska

Poznan, 19 September, 2021

Tomasz P. Lehmann
Paweł P. Jagodziński
Department of Surgery, Chair and Clinic of Oncology
Poznan University of Medical Sciences
ul. Święcickiego 6
60-781 Poznan, Poland

Declaration of co-authors

We declare the following contributions to the publication

Tomasz P. Lehmann, Joanna Miskiewicz, Natalia Szostak, Marta Szachniuk, Sylwia Grodecka-Gazdecka, Pawel P. Jagodzinski. In Vitro and in Silico Analysis ofmiR-125a with rs12976445 Polymorphism in Breast Cancer Patients. Applied Sciences 10(20), 2020, pp. 7275

- Dr Tomasz P. Lehmann: conceptualization, methodology, data curation, supervision, and formal analysis;

- Prof. Paweł P. Jagodziński: project administration and original draft preparation.

prof. Paweł. P. Jagodziński

dr Tomasz P. Lehmann

Poznań, September 18, 2021

Agnieszka Mickiewicz-Piotrowska
Os. Stefana Batorego 59D/83
60-687 Poznań

Declaration

In connection with doctoral thesis of Joanna Miśkiewicz, I declare the following contribution to scientific publication:

J. Miskiewicz, K. Tomczyk, A. Mickiewicz, J. Sarzynska, M. Szachniuk. Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *BioMed Research International* 2017, 6783010

- participation in the team discussions,
- proposing thermodynamic profiles for selected miRNA:miRNA* duplexes for additional analysis,
- contributing to the preliminary version of the manuscript.

I performed all the tasks under the supervision of dr J. Sarzyńska and prof. M. Szachniuk.

A. Mickiewicz-Piotrowska

Paulina Pielacińska
Os. Piastowskie 120/24
61-166 Poznań

Poznań, 17.09.2021

Declaration of a co-author

This declaration concerns the paper “ONQUADRO: a database of experimentally determined quadruplex structure” by T. Zok, N. Kraszewska, J. Miskiewicz, P. Pielacinska, M. Zurkowski, and M. Szachniuk, submitted for publication in Nucleic Acids Research.

I declare the following contribution as an author of this paper: implementation of a prototype version of the interface for the ONQUADRO database.

Paulina Pielacińska

Poznan, 20 September 2021

Mariusz Popenda
Department of Structural Bioinformatics
Institute of Bioorganic Chemistry
Polish Academy of Sciences

Declaration

I declare that Joanna Miskiewicz and myself share the first authorship of the following paper:

Mariusz Popenda, Joanna Miskiewicz, Joanna Sarzynska, Tomasz Zok, and Marta Szachniuk (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics* 36(4): 1129–1134.

My contribution to this paper consisted in participating in the team discussions, analyzing the secondary structure topologies of nucleic acids, proposing the ONZ classification for quadruplexes, generating structure visualizations, selecting examples for publication.

Mariusz Popenda

Poznan, 18 September 2021

Joanna Sarzyńska
Department of Structural Bioinformatics
Institute of Bioorganic Chemistry
Polish Academy of Sciences

Declaration of a co-author

I declare the following contributions to publications co-authored with Joanna Miśkiewicz:

(1) Miskiewicz J, Tomczyk K, Mickiewicz A, Sarzynska J, Szachniuk M (2017) Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *BioMed Research International* 2017: 6783010

- co-supervising Agnieszka Mickiewicz, assistance in interpreting the results and contribution to manuscript preparation;

(2) Popenda M, Miskiewicz J, Sarzynska J, Zok T, Szachniuk M (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics* 36(4): 1129–1134

- participation in the team discussions, literature overview, analysing of the 3D structures of quadruplexes, selecting examples for the publication;

(3) Miskiewicz J, Sarzynska J, Szachniuk M (2021) How bioinformatics resources work with G4 RNAs. *Briefings in Bioinformatics* 22(3): bbaa201

- supervising computational experiments carried by Joanna Miskiewicz, involvement in the results' compilation and manuscript preparation.

Poznan, September 19, 2021

Natalia Szóstak
Laboratory of Bioinformatics
Institute of Bioorganic Chemistry
Polish Academy of Sciences

Declaration of a co-author

Hereby, I declare the following contribution to the paper –

Tomasz P. Lehmann, Joanna Miskiewicz, Natalia Szostak, Marta Szachniuk, Sylwia Grodecka-Gazdecka, Pawel P. Jagodzinski. In Vitro and in Silico Analysis of miR-125a with rs12976445 Polymorphism in Breast Cancer Patients. Applied Sciences 10(20), 2020, pp. 7275,

– which I co-authored: methodology of bioinformatics analysis, *in silico* investigation of miR-125a rs12976445 polymorphism and binding motifs, visualizations, and manuscript writing and editing.

.....Natalia Szóstak.....

September 18, 2021

Katarzyna Tomczyk
Kunickiego 9, 61-418 Poznań

Declaration of a co-author

I declare that my contribution to the publication:

Miskiewicz J, Tomczyk K, Mickiewicz A, Sarzynska J, Szachniuk M (2017) Bioinformatics Study of Structural Patterns in Plant MicroRNA Precursors. *BioMed Research International* 2017: 6783010

was to participate in the team discussions, prepare the literature overview, and contribute to the implementation of script for loop searching within pre-miRNA secondary structures. All the works were performed under the supervision of prof. Marta Szachniuk.

Katarzyna Tomczyk

Poznan, 19 September, 2021

Tomasz Żok
Faculty of Computing and Telecommunications
Poznan University of Technology
Poznan, Poland

Declaration of a co-author

I declare the following contributions to publications co-authored with Joanna Miśkiewicz:

(1) M. Popenda, J. Miskiewicz, J. Sarzynska, T. Zok, M. Szachniuk (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics* 36(4), pp. 1129–1134

- participating in the team discussions,
- developing the ElTetrado program for quadruplex annotation,
- contributing to the manuscript preparation;

(2) T. Zok, N. Kraszewska, J. Miskiewicz, P. Pielacinska, M. Zurkowski, M. Szachniuk (2021) ONQUADRO: a database of experimentally determined quadruplex structure. *submitted for publication*

- participating in the team discussions,
- designing the ONQUADRO database system in cooperation with Joanna Miśkiewicz and prof. Marta Szachniuk,
- creating the database scheme,
- implementation of the backend (json-to-database layer),
- preparation of the auto-update functionality.

Tomasz Żok

Poznan, 17 September 2021

Michal Żurkowski
Faculty of Computing and Telecommunications
Poznan University of Technology
Poznan, Poland

Declaration

As the co-author of the following paper:

Tomasz Zok, Natalia Kraszewska, Joanna Miskiewicz, Paulina Pielacinska, Michal Zurkowski, Marta Szachniuk "ONQUADRO: a database of experimentally determined quadruplex structure", *submitted for publication.*

I declare that my contribution was to develop and implement an algorithm to create layer diagrams for visualization of the 3D structure of quadruplexes.

Extended abstract in Polish

W literaturze, sztuce, czy muzyce spotykamy się z pewnymi powtarzalnymi schematami, po których rozpoznajemy ich twórców lub epokę, z której dzieła pochodzą. Te powtarzające się wzory nazywane są motywami i występują również w naukach o życiu – w sieciach metabolicznych, procesach regulacyjnych komórki, czy też w strukturach kwasów nukleinowych. Każdy biologiczny motyw ma nie tylko określoną formę, ale też specyficzną rolę do odegrania w organizmie. Odnajdując dany motyw w cząsteczce naukowcy są w stanie powiązać z nim funkcję jaką pełni w systemie. Trudniejszym zadaniem jest jednak poszukiwanie motywu, który odpowiada za konkretne działania cząsteczki.

Niniejsza praca doktorska poświęcona jest badaniom motywów strukturalnych w cząsteczkach RNA pochodzących z różnych organizmów. Prace wykonane podczas doktoratu polegały na wyszukiwaniu i analizie motywów w sekwencjach oraz strukturach drugo- i trzeciorzędowych. Pierwsze badania skupione były na poszukiwaniu motywów strukturalnych w zbiorze roślinnych mikroRNA na przykładzie organizmu modelowego – *Arabidopsis thaliana*. Zaobserwowano schemat powtarzania się małych pętli wewnętrznych w okolicach dupleksu miRNA:miRNA*, co może wskazywać na obecność motywu rozpoznawalnego przez enzym wycinający dupleks z cząsteczki. Uzyskane wyniki były inspiracją do rozszerzenia badań na pre-miRNA z całego królestwa roślin zielonych – *Viridiplantae*. W anal-

izowanych strukturach wykryto podobny motyw jak przy analizach pre-miRNA w *Arabidopsis thaliana*. Kolejne badania dotyczyły struktury pierwotnego transkryptu miR-125a w dwóch wariantach sekwencyjnych (zmiana pojedynczego nukleotydu, SNP). Bioinformatyczna analiza wskazywała na zależność rodzaju wiązanych białek do traksryptu od wybranego typu wariantu sekwencyjnego. Ponadto, predykcja struktury drugorzędowej wskazywała na różnice strukturalne wynikające ze zmiany pojedynczego nukleotydu w transkrypcie. Najnowsze badania koncentrowały się na motywach kwadrupleksów, ich topologii oraz analizie parametrycznej z użyciem narzędzi bioinformatycznych. Zaowocowały one opracowaniem nowej klasyfikacji kwadrupleksów w oparciu o ich strukturę drugorzędową oraz stworzeniem nowych reprezentacji umożliwiających zapisywanie informacji o strukturze drugorzędowej w dwuliniowej notacji kropkowo-nawiasowej i w postaci dwuczęściowego diagramu łukowego. Przebadaliśmy wszystkie dostępne zasoby bioinformatyczne pod kątem ich wykorzystania do badań kwadrupleksów RNA oraz utworzyliśmy bazę danych ONQUADRO gromadzącą i przetwarzającą dane o strukturach kwadrupleksów otrzymanych drogą eksperymentalną. Przeanalizowaliśmy ludzkie sekwencje mikroRNA pod kątem ich potencjału do formowania motywów kwadrupleksów. W tym celu wykorzystaliśmy algorytm bazujący na dopasowaniu wyrażeń regularnych. Sekwencje zostały również zbadane pod kątem nasycenia guaninami, w celu sprawdzenia wielkości zbioru, który spełnia minimalny wymóg do posiadania motywu kwadrupleksu (8G i 12G kolejno dla dwu- i trójtetradowych kwadrupleksów).

W badaniach do pracy doktorskiej wykorzystywane były dostępne narzędzia bioinformatyczne, jak również nowo stworzone metody do analizy zbiorów danych strukturalnych. Wszystkie analizowane dane pochodzą z publicznie dostępnych repozytoriów.

Appendices

APPENDIX A

Participation in research projects

- Theme of the project: RNApolis - methods and algorithms to model and analyze the RNA structure (RNApolis - metody i algorytmy do modelowania i analizy struktury RNA)
  Grant number: 2016/23/B/ST6/03931
  Participation period: 01.09.2017 – 19.11.2020
  Principal investigator: prof. Marta Szachniuk

- Theme of the project: Classifier and database of quadruplex motifs (Klasyfikator i baza danych motywów kwadrupleksowych)
  Grant number: 09/91/SBAD/0684 (Młoda Kadra)
  Participation period: 01.06.2019 – 30.09.2019
  Principal investigator: dr Tomasz Żok

- Theme of the project: Feature exploration and modelling of quadruplex structures (Eksploracja cech i modelowanie struktury kwadrupleksów)
  Grant number: 2019/35/B/ST6/03074
  Participation period: since 14.07.2020
  Principal investigator: prof. Marta Szachniuk

APPENDIX B

Conference presentations

During my Ph.D. study, I gave 21 presentations (talks and posters) at national and international scientific conferences and seminars:

- *Bioinformatic Analysis of Motifs in Plant MicroRNA*, BIT'15: Bioinformatics in Torun, June 2015, Torun, Poland.
- *Bioinformatic Analysis of the Neighbourhood of Plant MicroRNA*, RNA Structure and Function Conference, organized by International Centre for Genetic Engineering and Biotechnology (ICGEB), March 2016, Trieste, Italy.
- *Bioinformatic Analysis of Motifs in the Vicinity of Plant MicroRNA*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, May 2016, Poznan, Poland.
- *Searching for Structural Patterns in the Vicinity of MicroRNA in Plants*, BIT'16: Bioinformatics in Torun, June 2016, Torun, Poland.
- *Searching for Structural Patterns in the Vicinity of MicroRNA in Plants*, IX Convention of the Polish Bioinformatics Society, September 2016, Bialystok, Poland.

- *Bioinformatics Study of Structural Patterns in Plant MicroRNA*, BIT'17: Bioinformatics in Torun, June 2017, Torun, Poland.
- *Bioinformatics Study of Structural Patterns in Plant MicroRNA*, X Convention of the Polish Bioinformatics Society, September 2017, Uniejow, Poland.
- *Structural Patterns in Plant MicroRNA Recognition*, RECOMB'18: The 22nd Annual International Conference on Research in Computational Molecular Biology, May 2018, Paris, France.
- *G-quadruplex Structures in Human MicroRNA*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, June 2018, Poznan, Poland.
- *Computational Analysis and Visualization of G-quadruplex Structures*, ECCO 2018: 31st Conference of the European Chapter on Combinatorial Optimization, June 2018, Fribourg, Switzerland.
- *Bioinformatic Approach to Visualization and Analysis of G-quadruplex Structures*, BIT'18: Bioinformatics in Torun, June 2018, Torun, Poland.
- *Computational Analysis and Visualization of G-quadruplex Structures*, 29th European Conference On Operational Research (EURO), July 2018, Valencia, Spain.
- *G-quadruplex Topology from Bioinformatics Perspective*, XI Convention of the Polish Bioinformatics Society, September 2018, Wroclaw, Poland.
- *Computational Approach to G-quadruplex Topology*, International Workshop on Scheduling and Sequencing (ICOLE), September 2018, Lessach, Austria.

- *New computational solutions in the quadruplex world*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, May 2019, Poznan, Poland.

- *Novel quadruplex classification – ONZ approach*, BIT'19: Bioinformatics in Torun, June 2019, Torun, Poland.

- *ONZ as novel quadruplex classification*, 7th International Meeting on Quadruplex Nucleic Acids, September 2019, Changchun, China.

- *Secondary structure-based classification of nucleic acid quadruplexes*, XII Convention of the Polish Bioinformatics Society, September 2019, Cracow, Poland.

- *In silico exploration of quadruplex structures*, Seminar of the Laboratory of Algorithm Design and Data Structures, Institute of Computing Science, Poznan University of Technology, November 2020, Poznan, Poland.

- *In silico exploration of quadruplex structures*, Autumn Workshop of the Polish Bioinformatics Society, November 2020, online event.

- *Novel database for quadruplexes... and more – ONQUADRO*, Autumn Workshop of the Polish Bioinformatics Society, September 2021, online event.

APPENDIX C

Awards and distinctions

- ICGEB scholarship for young researchers to participate in RNA Structure and Function Conference in Trieste, Italy (2016).
- Travel Award granted by Polish Bioinformatics Society for young scientists' participation in BIT'18: Bioinformatics in Torun (2018).
- Best Presentation Award granted by the Polish Bioinformatics Society for the talk presented at the XII Convention of the Polish Bioinformatics Society (2019).
- Pro-quality scholarship for the best Ph.D. students granted by the Rector of Poznan University of Technology (2018/2019, 2019/2020, 2020/2021).
- Scholarship for best Ph.D. students at the Faculty of Computing granted by the Rector of Poznan University of Technology (2016/2017, 2017/2018, 2018/2019, 2019/2020, 2020/2021).
- Best Poster Presentation Award granted by the Polish Bioinformatics Society for the poster presented at Autumn Workshop of the Polish Bioinformatics Society (2020).
- Laureate of the 240+ incentive program in the Institute of Computer Science, Poznan University of Technology (2021).