


Structural bioinformatics

# Topology-based classification of tetrads and quadruplex structures

Mariusz Popenda<sup>1,†</sup>, Joanna Miskiewicz<sup>2,†</sup>, Joanna Sarzynska<sup>1</sup>, Tomasz Zok<sup>2,3</sup> and Marta Szachniuk <sup>1,2,\*</sup>

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

\*To whom correspondence should be addressed.

<sup>†</sup>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](mailto: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 (Cammass 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 (Cammass 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 (Cammass and Millevoi, 2017; Malgowska *et al.*, 2016). Thus, one quadruplex, also referred to as N<sub>4</sub> 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 G<sub>4</sub>. 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. GRSDDB2 (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 canonicity, 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 RNAPdbec 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/eltradrado/>. 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 RNAPdbec 2.0 (Antczak *et al.*, 2018, 2014; Zok *et al.*, 2018). This webserver is a part of RNAPdbec 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.

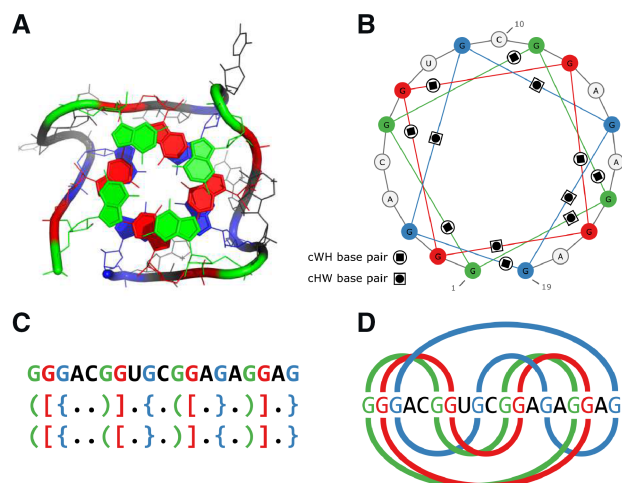


Fig. 1. Quadruplex formed in RNA Mango (PDB id: 5V3F chain B) (Trachman *et al.*, 2017) in (A) 3D and (B) secondary structure view, (C) dot-bracket encoding, (D) arc diagram. For clarity, each tetrad has a different color

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 hour-glass (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

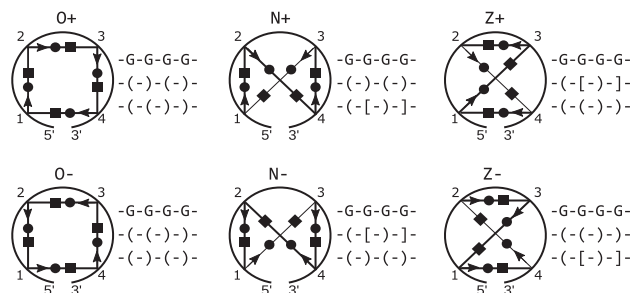


Fig. 2. ONZ classes of tetrads in diagrams of unimolecular structures and their dot-bracket representations. Arrows ease discerning between + and -. Here, black filled circle denotes Watson-Crick edge, square—Hoogsteen edge

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 subclasses: O+, O-, N+, N-, Z+, Z-. Each of these subclasses has a unique dot-bracket representation (Fig. 2).

### 3.3 Classification of quadruplexes

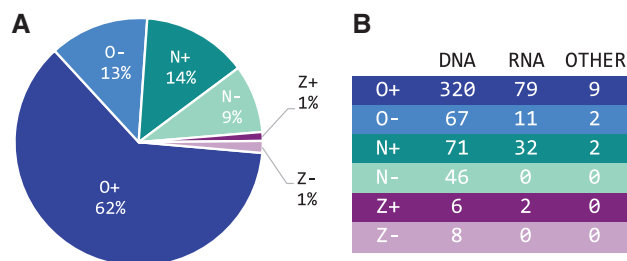
Making use of the new tetrad classification, we have proposed ONZ-based taxonomy for quadruplexes. We have assumed that the classes of component tetrads automatically determine the category of the whole quadruplex. From this it follows that a quadruplex consisting of O-type tetrads belongs to class O, 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](#)).

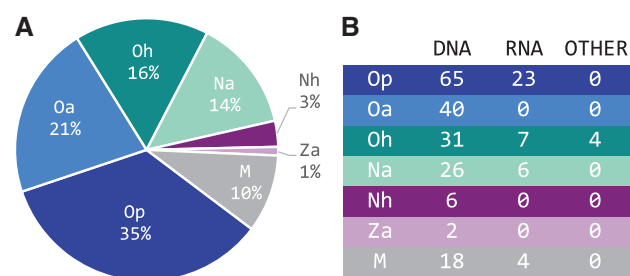
Next, we have investigated how many strands contributed to building quadruplexes ([Table 3](#)). Our research has shown that more than half of the collection is made up of unimolecular motifs. These 242 quadruplexes could be next studied with respect to ONZ-based taxonomy. The remaining 181 motifs belong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

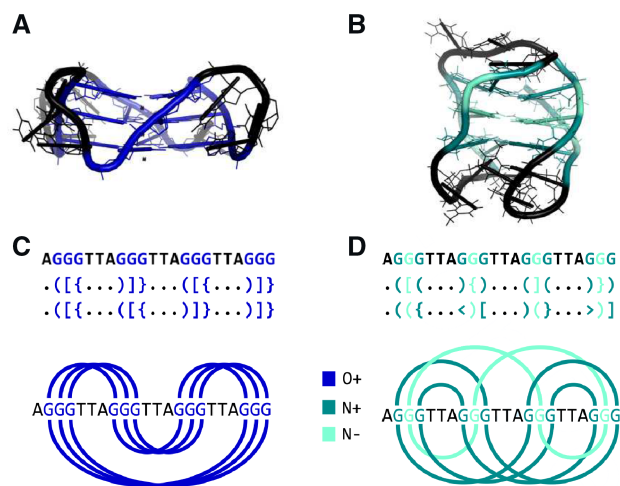


Fig. 5. 3D structure of human telomeric DNA (A) 1Kf1 (Parkinson *et al.*, 2002) and (B) 143D (Wang and Patel, 1993) with colored tetrads and (C, D) their secondary structure topologies in dot-bracket and arc diagram representations. The colors in the secondary structure representations indicate the tetrad types

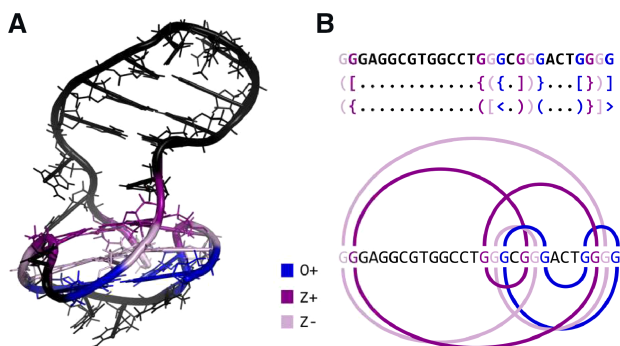


Fig. 6. (A) 3D structure of major G-quadruplex form in HIV-1 LTR (PDB id: 6H1K) (Butovskaya *et al.*, 2018) with colored tetrads, and (B) its secondary structure topology in dot-bracket and arc diagram representation. The colors in the secondary structure representations indicate the tetrad types

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 Młoda 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 Res.*, **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-multiplies 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.
- Cammass, 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.
- Dhapol, 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) RNAPdb 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*.