

# ONQUADRO: a database of experimentally determined quadruplex structures

Tomasz Zok<sup>1</sup>, Natalia Kraszewska<sup>1</sup>, Joanna Miskiewicz<sup>1</sup>, Paulina Pielacinska<sup>1</sup>,  
Michal Zurkowski<sup>1</sup> and Marta Szachniuk<sup>1,2,\*</sup>

<sup>1</sup>Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland and <sup>2</sup>Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland

Received August 11, 2021; Revised October 22, 2021; Editorial Decision October 22, 2021; Accepted October 25, 2021

## ABSTRACT

**ONQUADRO is an advanced database system that supports the study of the 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 visualise the data; scripts for statistical analysis; automatic updates and newsletter modules; and a web application that provides a user interface. The database is a self-updating resource, with new information arriving once a week. The preliminary data are downloaded from the 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/>. It can be used on both 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, the regulation of gene expression, DNA replication, etc. (2,3,5,6). A recent hypothesis, coined as the quadruplex

world, suggests that G4s may have been the molecules to initiate life on Earth (7). It takes its cue from the ability of guanines to form stable G-tetrads, the basic building units of a quadruplex. In a tetrad, four guanines arranged in the same plane connect via hydrogen bonds such that each acts as a donor of two hydrogen bonds at the Watson-Crick edge and an acceptor at the Hoogsteen edge (1,8,9). A complete G4 is assembled from at least two G-tetrads stacked one above another, and is stabilised by monovalent cations located in the ion channel (5,9). The stacking interactions of G-tetrads occurring independently of backbone connectivity define a G4-helix (10).

The general definition of what constitutes a quadruplex does not capture the complexity of its structure and the diversity of its features (2,8,9,11–14). Meanwhile, the latter is subjected to various studies, aiming - among others - to associate the motif's conformation with its function, find the relationship between its sequence and the higher-level structure, cluster and classify quadruplexes, learn about and fully describe their properties. We already know that tetrads can form from guanine as well as non-guanine nucleotides (15). Their spatial arrangement to their stacking neighbours 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 (16). 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 - cf. Webba da Silva formalism (17). The list of analysed attributes also includes the glycosidic bond angles, the groove width, the number of stacked tetrads, or G-tract continuity, and is probably not yet complete (11).

In the past decade, the unique structure of the quadruplex has focused the attention of many researchers, especially in medical sciences. G4s have become therapeutic targets, that is, for cancer and antiviral treatment (18–20). In the latter case, increased interest in targeting G-quadruplexes in viral genomes was prompted by the COVID-19 pandemic. The frequency and localisation of putative quadruplex sequences in different viral taxa, G4-binding viral domains,

\*To whom correspondence should be addressed. Tel: +48 616653030; Fax: +48 618771525; Email: [marta.szachniuk@cs.put.poznan.pl](mailto:marta.szachniuk@cs.put.poznan.pl)

and the potential of G4s as viral biosensors were investigated (20–25). These and other quadruplex studies provided a wealth of data for collection, organisation, and further analysis (26–28). It has initiated the development of computational methods and bioinformatics tools dedicated to G4s. Most of these deal with sequence data storage and processing (15,29–35). A few address higher-level structures (10,36–39), including databases that store G4-related data (36,40,41) - none, however, collects complete information about quadruplex structures at all levels of their organisation.

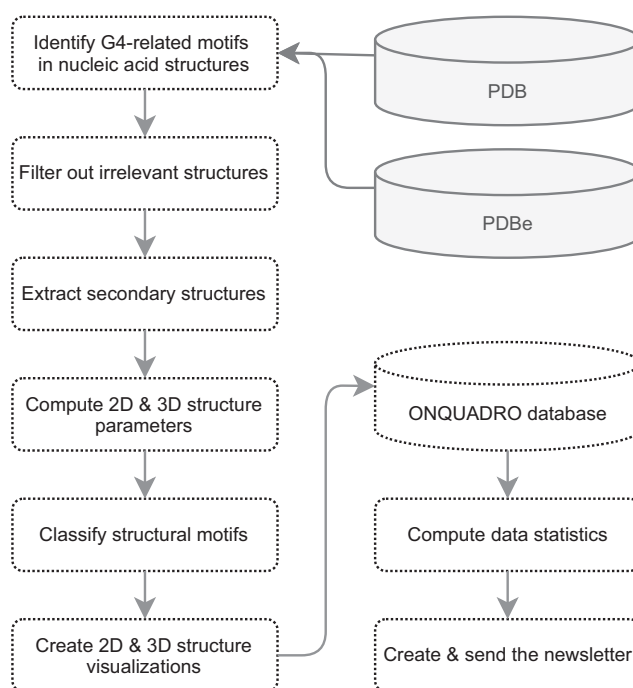
ONQUADRO is a new comprehensive database system that collects and shares data on tetrads, quadruplexes, and G4-helices, whose three-dimensional structures have been determined experimentally. Baseline data are regularly downloaded from the Protein Data Bank (42) and supplemented with parameters computed by specialised procedures of the system's engine. The incorporated programs prepare visualisations of the secondary and tertiary structure models of each motif. The analytical module generates statistics of the distribution of the structural parameters in the set of tetrads and quadruplexes. The system allows users to subscribe to a newsletter about all database newcomers. ONQUADRO, designed for 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 the 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 from the list of modified and deleted structures; entries for new motifs are created and added. The process of new data preparation takes place in several steps (Figure 1).

For each quadruplex, it derives the secondary structure and prepares its representation in two-line extended dot-bracket notation, computes the rise and twist parameters, identifies the number of contributing strands and tetrads, determines the strand direction, finds loops to calculate their lengths and types, classifies it according to Webba da Silva formalism based on its loop topology and tetrad combination (17), and assigns an ONZ class from the secondary structure topology (16). If the nucleic acid contains metal ions, the procedure determines their position relative to the quadruplex. Then, it describes 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. These 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.

Once the database is updated, the statistical analysis module generates graphs and tables of the data distribution.



**Figure 1.** Data flow during the weekly ONQUADRO update.

Statistics available for G4s are (i) the number of quadruplexes as a function of the number of constituent tetrads; (ii) the abundance of the set of uni-, bi-, and tetramolecular quadruplexes; (iii) ONZ class coverage by uni-, bi-, and tetramolecular quadruplexes; (iv) the geometric class distribution based on glycosidic bond angles and loop topology; (v) the loop length distribution in the subsets of lateral, propeller and diagonal loops and (vi) the distribution of the twist and rise values. Statistics prepared for tetrads include: (i) the distribution of tetrads concerning their sequence and molecule type; (ii) the coverage of ONZ classes by uni-, bi-, and tetramolecular tetrads; (iii) the chi angle value distribution in the ONZ classes; (iv) ONZ class coverage by ions and (v) the planarity value distribution in the tetrad set.

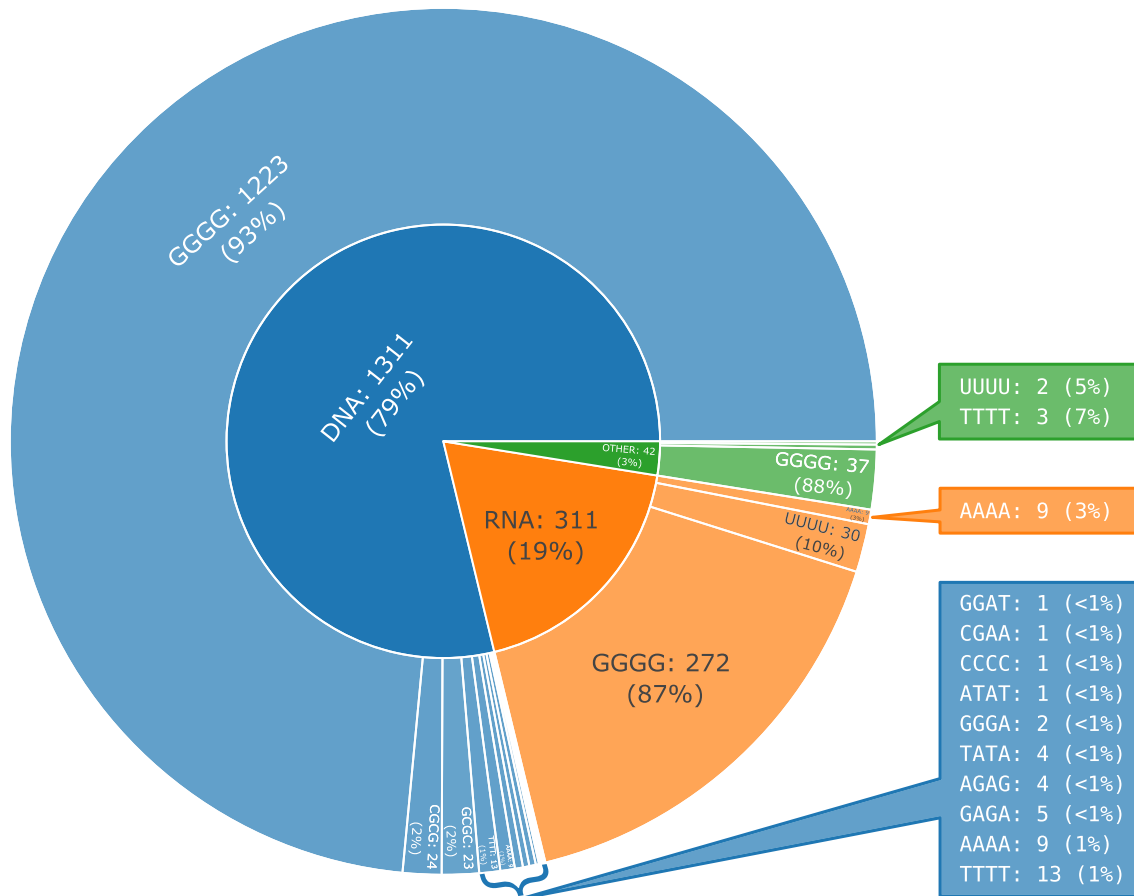
Finally, the system creates a hypertext newsletter listing all changes to the database and sends it to subscribers.

## IMPLEMENTATION

The ONQUADRO system consists of the database, web application, and computational engine. It runs on a quad-core machine with 8GB RAM in a Ubuntu GNU/Linux environment, hosted and maintained by the Institute of Computing Science, Poznan University of Technology.

### Database

ONQUADRO has been developed as a 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



**Figure 2.** Example statistics generated by ONQUADRO: quadruplexes by the number of constituent tetrads.

dates, molecule name, a secondary structure diagram, and a 3D 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, and glycosidic bond for nucleotides.

All sequences in ONQUADRO are coded in a one-letter format, in the 5'–3' direction. The secondary structures of tetrads, quadruplexes, and G4-helices are represented using dot-bracket notation - an unpaired nucleotide corresponds to a dot and a base pair to a pair of opening and closing brackets. Since in the considered motifs, each nucleotide pairs with two others, basic dot-bracket notation is not sufficient to unambiguously encode the secondary structure of a tetrad, so for a quadruplex or G4-helix. Therefore, in (16), we introduced a two-line dot-bracket and used an extended set of brackets to label paired nucleotides. This includes parentheses ( ), square brackets [ ], curly brackets { } and angle brackets < >. An arc diagram representing the secondary structure is also adjusted to unambiguously reflect all pairings. It is associated with dot-bracket notation - the top of the diagram corresponds to the first line of the dot-bracket notation, and the bottom part corresponds to the second line (16). The secondary structure of every motif is

also visualised in a classical diagram. The 3D structure is represented by a layer diagram and three molecular models (ball-and-stick, surface, vdw-balls).

### Computational engine

The computational engine is composed of scripts utilising in-house and third-party procedures, responsible for data collection, quadruplex identification, computation of structure parameters, secondary structure annotation, visualisation of the secondary and tertiary structure models, database queries, generation of statistics, and newsletter preparation. DSSR (`--pair-only` mode) (36) and ElTetrado (39) functionalities are applied to identify quadruplexes, tetrads, and G4-helices in nucleic acid structures. Procedures from ElTetrado (39) 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 a top-down arc diagram. The embedded LiteMol (47) module generates models (ball-and-stick, surface, vdw-balls) of the three-dimensional structure. The Python script draws a layer diagram of the quadruplex based on the data in JSON format obtained from ElTetrado. Every nucleotide in the diagram is colour-coded - yellow indicates the anti conformation, orange is for syn. The script optimises the quadruplex position in three-dimensional space to get a

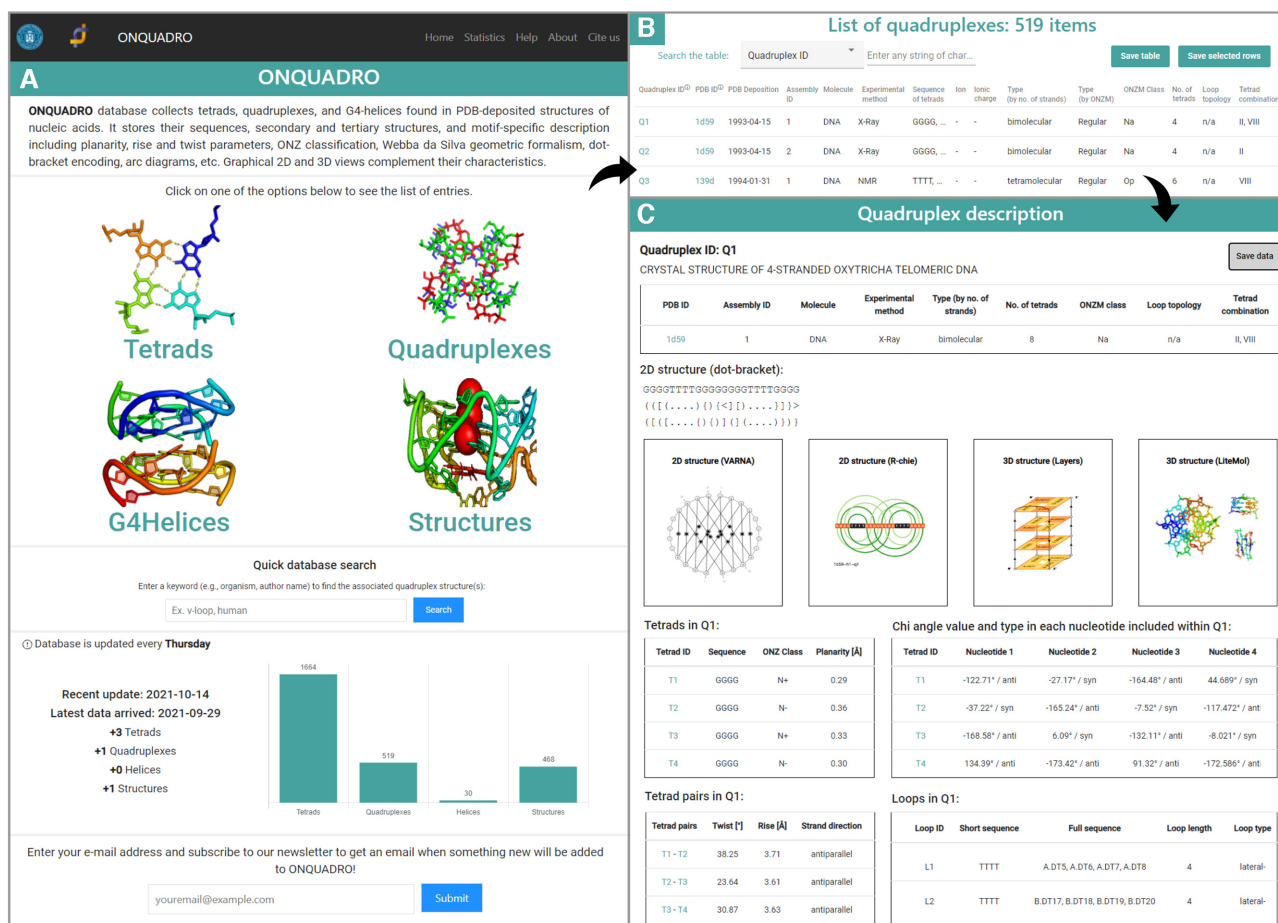


Figure 3. ONQUADRO interface: (A) main page, (B) quadruplex table and (C) quadruplex details based on 1D59 structure.

clear view of the G4, with the least number of crossing strands. The optimisation 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 (Figure 2).

### Web application

The web application provides an interface for 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 brief information about the database resources and the latest update. From this page, users can access the sets 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 detailed structural information to be viewed about the selected element or link to the corresponding page in the Protein Data Bank. The table can be sorted (ascending or descending) for 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. Search-

ing runs in real-time. The element counter at the 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 a selected part (*Save selected rows* button after clicking on the check-boxes in the rightmost column). Tabular data (from any subpage) are downloaded in a CSV file, structure visualisations can be saved in 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 graphical (pie chart, tree map, or bar plot) and tabular form. Plots can be saved in HTML format. They are interactive - upon clicking the plot, users can enlarge fragments and see selected parts of the 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 them according to geometric formalism (17) and topological ONZ nomenclature (16), represent the secondary structure in dot-bracket notation and a specially adjusted top-down arc diagram,



draw a 3D model in a schematic layer diagram, and trigger statistics. Some are G4-adapted routines applied in our previously released tools; others are brand new and have not yet been published (e.g. automatic creation of layer diagrams - a much-needed function in the research community). The user-friendly interface allows browsing of the database contents divided into four subsets (tetrads, quadruplexes, G4-helices, PDB structures), searching and sorting the data by various parameters and keywords, 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 collecting and sharing quadruplex information. We hope it will facilitate the study of G4 structures and their modelling *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

National Science Center, Poland [2019/35/B/ST6/03074 to M.S.]; statutory funds of Poznan University of Technology; Institute of Bioorganic Chemistry, PAS. Funding for open access charge: Poznan University of Technology.  
*Conflict of interest statement.* None declared.

## REFERENCES

- Malgowska, M., Czajczynska, K., Gudanis, D., Tworak, A. and Gdaniec, Z. (2016) Overview of RNA G-quadruplex structures. *Acta Biochim. Pol.*, **63**, 609–621.
- Kwok, C. and Merrick, C. (2017) G-Quadruplexes: prediction, characterization, and biological application. *Trends Biotechnol.*, **35**, 997–1013.
- Joachimi, A., Benz, A. and Hartig, J. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorg. Med. Chem.*, **17**, 6811–6815.
- Antonio, M., Ponjavic, A., Radzevicius, A., Ranasinghe, R., Catalano, M., Zhang, X., Shen, J., Needham, L., Lee, S., Klenerman, D. *et al.* (2020) Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat. Chem.*, **12**, 832–837.
- Webba da Silva, M., Trajkovski, M., Sannohe, Y., Ma'ni Hessari, N., Sugiyama, H. and Plavec, J. (2009) Design of a G-quadruplex topology through glycosidic bond angles. *Angew. Chem. Int. Ed. Engl.*, **48**, 9167–9170.
- Kolesnikova, S. and Curtis, E. (2019) Structure and Function of Multimeric G-Quadruplexes. *Molecules*, **24**, 3074.
- Kankia, B. (2021) Quadruplex world. *Orig. Life Evol. Biosophys.*, **51**, 273–286.
- Dvorkin, S., Karsisiotis, A. and Webba da Silva, M. (2018) Encoding canonical DNA quadruplex structure. *Sci. Adv.*, **4**, eaat3007.
- Spiegel, J., Adhikari, S. and Balasubramanian, S. (2020) The structure and function of DNA G-quadruplexes. *Trends Chem.*, **2**, 123–136.
- Lu, X. (2020) DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. *Nucleic Acids Res.*, **48**, e74.
- Burge, S., Parkinson, G., Hazel, P., Todd, A. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
- Cang, X., Spomer, J. and Cheatham, T. (2011) Explaining the varied glycosidic conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic Acids Res.*, **39**, 4499–4512.
- Mukundan, V. and Phan, A. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017–5028.
- Lightfoot, H., Hagen, T., Tatum, N. and Hall, J. (2019) The diverse structural landscape of quadruplexes. *FEBS Lett.*, **593**, 2083–2102.
- Miskiewicz, J., Sarzynska, J. and Szachniuk, M. (2021) How bioinformatics resources work with G4 RNAs. *Brief Bioinform.*, **22**, bbaa201.
- Popenda, M., Miskiewicz, J., Sarzynska, J., Zok, T. and Szachniuk, M. (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics*, **36**, 1129–1134.
- Webba da Silva, M. (2007) Geometric formalism for DNA quadruplex folding. *Chemistry*, **13**, 9738–9745.
- Carvalho, J., Mergny, J., Salgado, G., Queiroz, J. and Cruz, C. (2020) G-quadruplex, friend or foe: the role of the G-quartet in anticancer strategies. *Trends Mol. Med.*, **26**, 848–861.
- Hansel-Hertsch, R., Simeone, A., Shea, A., Hui, W., Zyner, K., Marsico, G., Rueda, O., Bruna, A., Martin, A., Zhang, X. *et al.* (2020) Landscape of G-quadruplex DNA structural regions in breast cancer. *Nat. Genet.*, **52**, 878–883.
- Panera, N., Tozzi, A. and Alisi, A. (2020) The G-quadruplex/helicase world as a potential antiviral approach against COVID-19. *Drugs*, **80**, 941–946.
- Lavezzo, E., Berselli, M., Frasson, I., Perrone, R., Palu, G., Brazzale, A., Richter, S. and Toppo, S. (2018) G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLoS Comput. Biol.*, **14**, e1006675.
- Ji, D., Juhas, M., Tsang, C., Kwok, C., Li, Y. and Zhang, Y. (2020) Discovery of G-quadruplex-forming sequences in SARS-CoV-2. *Brief Bioinform.*, **22**, 1150–1160.
- Wang, S., Min, Y., Wang, J., Liu, C., Fu, B., Wu, F., Wu, L., Qiao, Z., Song, Y., Xu, G. *et al.* (2016) A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new antihepatitis C target. *Sci. Adv.*, **2**, e1501535.
- Tan, J., Vornrhein, C., Smart, O., Bricogne, G., Bollati, M., Kusov, Y., Hansen, G., Mesters, J., Schmidt, C. and Hilgenfeld, R. (2009) The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog.*, **5**, e1000428.
- Xi, H., Juhas, M. and Zhang, Y. (2020) G-quadruplex based biosensor: a potential tool for SARS-CoV-2 detection. *Biosens Bioelectron.*, **167**, 112494.
- Gudanis, D., Popenda, L., Szpotkowski, K., Kierzek, R. and Gdaniec, Z. (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.
- Andralojc, W., Malgowska, M., Sarzynska, J., Pasternak, K., Szpotkowski, K., Kierzek, R. and Gdaniec, Z. (2018) 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.
- Frelih, T., Wang, B., Plavec, J. and Sket, P. (2020) Pre-folded structures govern folding pathways of human telomeric G-quadruplexes. *Nucleic Acids Res.*, **48**, 2189–2197.
- Wong, H., Stegle, O., Rodgers, S. and Huppert, J. (2010) A toolbox for predicting G-quadruplex formation and stability. *J. Nucleic Acids*, **2010**, 564946.
- Garant, J., Luce, M., Scott, M. and Perreault, J. (2015) G4RNA: an RNA G-quadruplex database. *Database*, **2015**, bav059.
- Bedrat, A., Lacroix, L. and Mergny, J. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **4**, 1746–1759.
- Sahakyan, A., Chambers, V., Marsico, G., Santner, T., Di Antonio, M. and Balasubramanian, S. (2017) Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.*, **7**, 14535.
- Ge, F., Wang, Y., Li, H., Zhang, R., Wang, X., Li, Q., Liang, Z. and Yang, L. (2019) Plant-GQ: an integrative database of G-quadruplex in plant. *J. Comput. Biol.*, **26**, 1013–1019.
- Lombardi, E. and Londono-Vallejo, A. (2020) A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res.*, **48**, 1603–1603.
- Kudla, M., Gutowska, K., Synak, J., Weber, M., Bohnsack, K., Lukasiak, P., Villmann, T., Blazewicz, J. and Szachniuk, M. (2020) Virxicon: a lexicon of viral sequences. *Bioinformatics*, **36**, 5507–5513.

36. Lu, X., Bussemaker, H. and Olson, W. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
37. Rybarczyk, A., Szostak, N., Antczak, M., Zok, T., Popena, M., Adamiak, R., Blazewicz, J. and Szachniuk, M. (2015) New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinformatics*, **16**, 276.
38. Patro, L., Kumar, A., Kolimi, N. and Rathinavelan, T. (2017) 3d-NuS: a web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. *J. Mol. Biol.*, **429**, 2438–2448.
39. Zok, T., Popena, M. and Szachniuk, M. (2020) ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *BMC Bioinformatics*, **21**, 40.
40. Li, Q., Xiang, J., Yang, Q., Sun, H., Guan, A. and Tang, Y. (2012) G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res.*, **41**, D1115–D1123.
41. Mishra, S., Tawani, A., Mishra, A. and Kumar, A. (2016) G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.*, **6**, 38144.
42. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
43. Velankar, S., Alhroub, Y., Best, C., Caboche, S., Conroy, M., Dana, J., Montecelo, M., van Ginkel, G., Golovin, A., Gore, S. *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D445–D452.
44. Zok, T. (2021) BioCommons: a robust Java library for RNA structural bioinformatics. *Bioinformatics*, **37**, 2766–2767.
45. Darty, K., Denise, A. and Ponty, Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
46. Lai, D., Proctor, J., Zhu, J. and Meyer, I. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
47. Sehnal, D., Deshpande, M., Varekova, R., Mir, S., Berka, K., Midlik, A., Pravda, L., Velankar, S. and Koca, J. (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods*, **14**, 1121–1122.