DSSR-Enabled Automatic Identification and Annotation of G-quadruplexes in the PDB

Xiang-Jun Lu[1*] (xiangjun@x3dna.org), Wilma K. Olson[2], and Harmen J. Bussemaker[1]

[1]Department of Biological Sciences, Columbia University, New York, NY 10027, U.S.A.

[2]Department of Chemistry and Chemical Biology, Rutgers – The State University of New Jersey, Piscataway, NJ 08854, U.S.A.

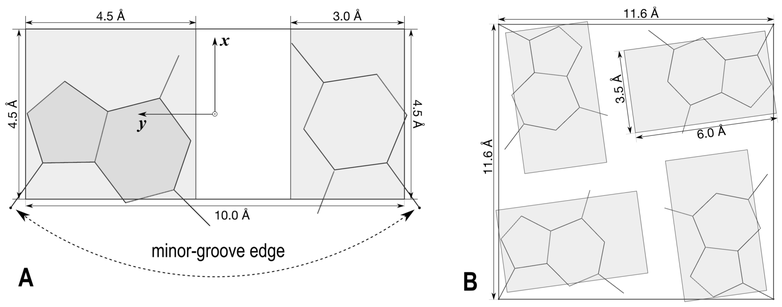
Abstract

G-quadruplexes (G4s) are a common type of higher-order nucleic acid structures formed from G-rich sequences, typically with the pattern $\mathbf{G_{3+}N_{1\text{-}7}G_{3+}N_{1\text{-}7}G_{3+}N_{1\text{-}7}G_{3+}}$. For example, the human telomere DNA has a repetitive sequence of $(GGGTTA)_n$. The building block of G4s is a tetrad of guanines (G-tetrad) arranged in a cyclic planar manner, held together by eight hydrogen-bonds via four consecutive G+G pairs. G4s are formed by the stacking of G-tetrads and stabilized by cations. They are known to play important roles in the regulation of gene expression, the maintenance of genome stability, and serve as potential therapeutic targets.

Experimentally solved 3D structures of G4s, deposited in the Protein Data Bank (PDB), provide important insights into their functions at the atomic level. However, limited annotations of G4s in the PDB unavoidably lead to false positive and false negative search results based on keywords. The lack of a widely accepted annotation tool also makes results reported in the literature difficult, if not impossible, to compare. Schematic representations that capture the essential features of G4s in a simple yet revealing manner are missing: it is easy to get lost in detailed descriptions and complicated graphics while reading publications on G4s. Clearly, the increasing number of G4s available in the PDB calls for a pragmatic software tool that can identify this important class of structures automatically, characterize them consistently, and visualize them intelligibly.

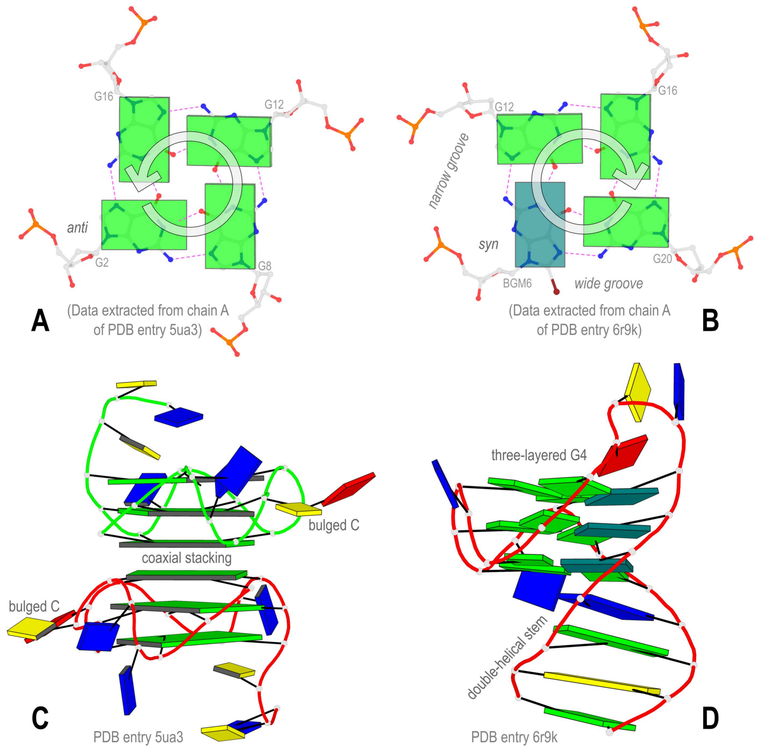
The DSSR (Dissecting the Spatial Structure of RNA) program detects and annotates G4s, starting from atomic coordinates in PDB or PDBx/mmCIF format. It identifies G-tetrads and arranges them into G4 helices, composed of G4 stems via coaxial stacking interactions. G4 stems are categorized in terms of loops connecting the four strands, by common names [e.g., chair, basket], or the *revised* Webba da Silva structural descriptors for G4 stems [e.g., **3(-P-Lw-Ln)** instead of **3(-p-$\mathbf{l_n}$-$\mathbf{l_w}$)** for PDB entry 2GKU]. DSSR accounts for bulges, identifies V-loops, characterizes G4 structures using rigid-body parameters, and quantifies stacking with overlapping areas. The program introduces innovative schematic representations, highlighting G-tetrads in PyMOL with unprecedented clarity. DSSR-annotated G4s from the PDB, easily searchable and regularly updated, are available at **http://G4.x3dna.org/.**

Schematic Blocks



Definition of various types of DSSR blocks, illustrated using idealized, planar Watson-Crick (WC) base-pair and G-tetrad geometries. (A) Purine (A, G) and pyrimidine (C, T, U) base blocks and the WC-pair block in default dimensions. (B) The slim purine block and the square G-tetrad block for simplified visualizations of G-quadruplexes.

Simplified Representations



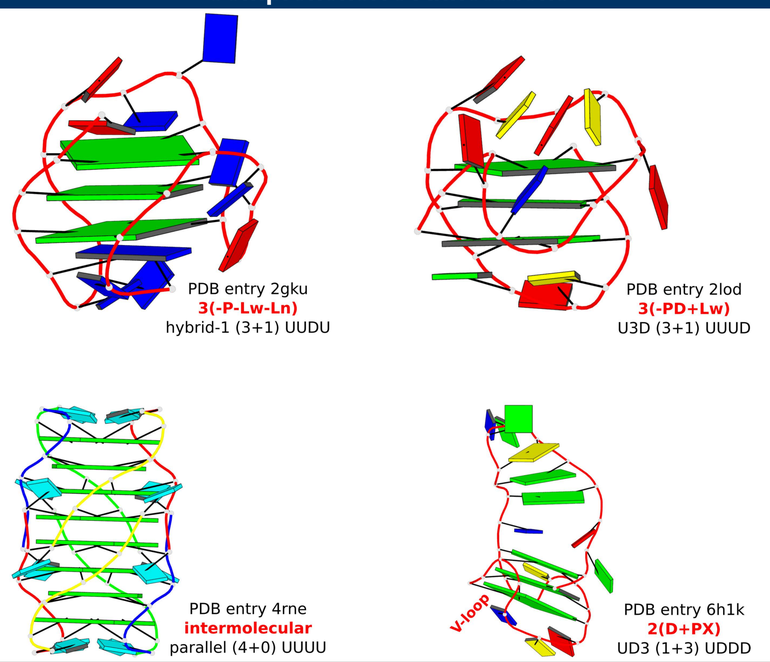
Innovative block representations introduced in DSSR to simplify the visualization of G-quadruplexes. (A) The *anti* guanine (lower-left) leads to a counter-clockwise orientation of H-bonding interactions (dashed lines in magenta) from the WC edge to the major-groove (Hoogsteen) edge. (B) The *syn* guanine (colored teal) reverses the direction of the H-bonds to clockwise, and it also creates a wide groove and a narrow groove. (C) A parallel G-quadruplex illustrated with square blocks for G-tetrads, highlighting a six-layered G4-helix composed of two three-layered G4-stems via coaxial stacking interactions. The two bulged cytosines are marked. (D) Another G4 with a (1+3) hybrid conformation, emphasizing the duplex-quadruplex transition, and the three guanines in *syn* conformation.

Cover Images of the *RNA* Journal

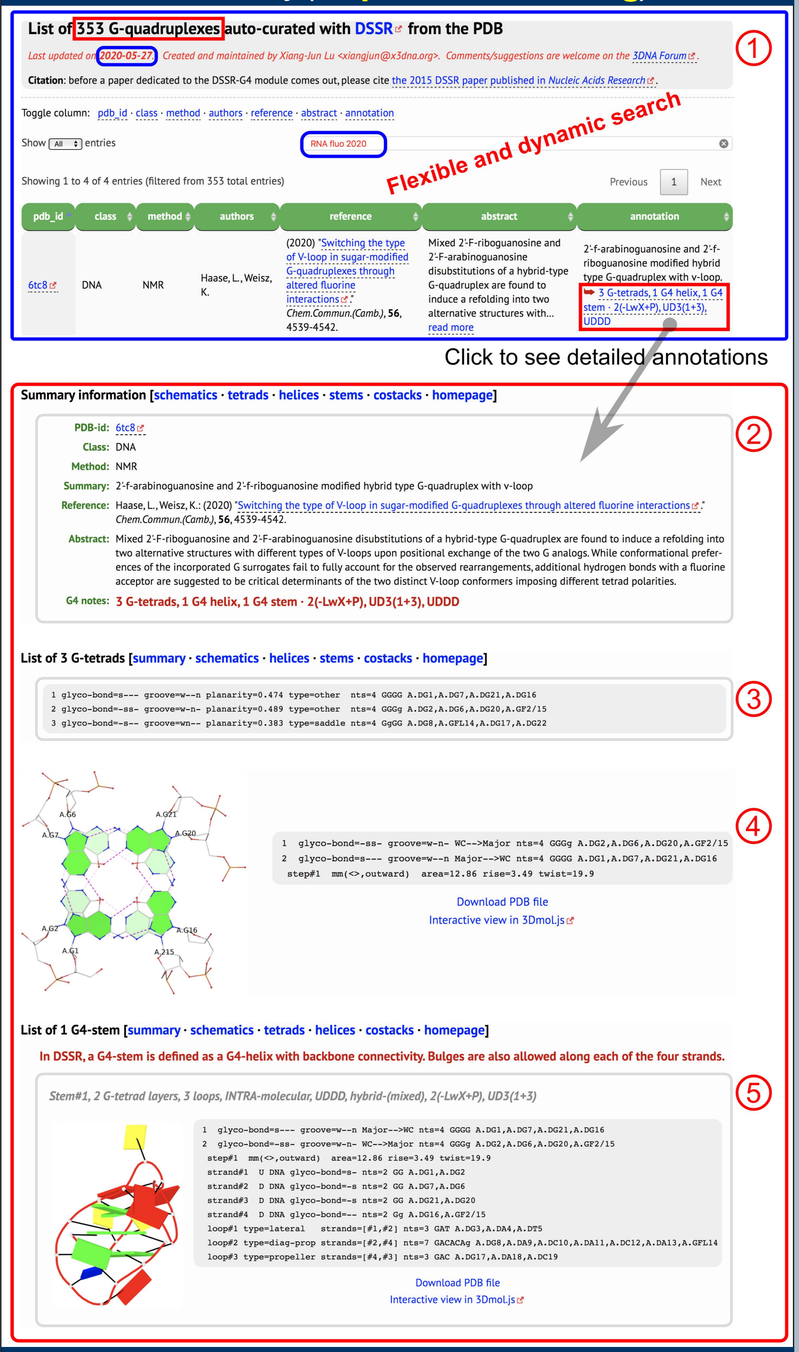
Nine out of 12 in 2019 and Jan-Jun in 2020 cover images of the RNA Journal, contributed by the NDB, have been generated via 3DNA-blocview and PyMOL.

The DSSR-PyMOL integration supersedes that approach; it is easier to use, and generates better schematic images.

Comprehensive Annotations



PDB Survey (http://G4.x3dna.org)



Summary

Starting simply from 3D atomic coordinates, such as those from the PDB or MD simulations, the G4 module of DSSR can:

- Identify G4 structures automatically
- Characterize G4 structures comprehensively
- Visualize G4 structures lucidly in PyMOL

Two typical use cases:

- Analyze one G4 structure to elucidate specific features
- Compare many G4 structures to draw general principles

References

Lu,X.J., Bussemaker,H.J., and Olson,W.K. (2015) "DSSR: an integrated software tool for dissecting the spatial structure of RNA." *Nucleic Acids Res*, **43**, e142.

Hanson,R.M. and **Lu,X.J.** (2017) "DSSR-enhanced visualization of nucleic acid structures in Jmol." *Nucleic Acids Res*, **45**, W528–W533.

Lu,X.J. (2020) "DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL." *Nucleic Acids Res* (online, May 22), DOI: 10.1093/nar/gkaa426.

This work has been made possible by the NIH grant R01GM096889 (XJL)