*Bioinformatics*, 36(2), 2020, 393–399
doi: 10.1093/bioinformatics/btz569
Advance Access Publication Date: 22 July 2019
Original Paper



Sequence analysis

# QPARSE: searching for long-looped or multimeric G-quadruplexes potentially distinctive and druggable

**Michele Berselli[1], Enrico Lavezzo[1,*] and Stefano Toppo[1,2,*]**

[1]Department of Molecular Medicine and [2]CRIBI Biotech Centre, University of Padova, Padova I-35131, Italy

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 22, 2019; revised on June 4, 2019; editorial decision on July 15, 2019; accepted on July 16, 2019

## Abstract

**Motivation:** G-quadruplexes (G4s) are non-canonical nucleic acid conformations that are widespread in all kingdoms of life and are emerging as important regulators both in RNA and DNA. Recently, two new higher-order architectures have been reported: adjacent interacting G4s and G4s with stable long loops forming stem-loop structures. As there are no specialized tools to identify these conformations, we developed QPARSE.

**Results:** QPARSE can exhaustively search for degenerate potential quadruplex-forming sequences (PQSs) containing bulges and/or mismatches at genomic level, as well as either multimeric or long-looped PQS (MPQS and LLPQS, respectively). While its assessment versus known reference datasets is comparable with the state-of-the-art, what is more interesting is its performance in the identification of MPQS and LLPQS that present algorithms are not designed to search for. We report a comprehensive analysis of MPQS in human gene promoters and the analysis of LLPQS on three experimentally validated case studies from HIV-1, BCL2 and hTERT.

**Availability and implementation:** QPARSE is freely accessible on the web at http://www.medcomp.medicina.unipd.it/qparse/index or downloadable from github as a python 2.7 program https://github.com/B3rse/qparse

**Contact:** enrico.lavezzo@unipd.it or stefano.toppo@unipd.it

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

## 1 Introduction

DNA is known to form a wide range of local structures alternative to the canonical B-form, such as hairpins, cruciforms, triplexes, tetraplexes and others, collectively known as non-B DNAs (Bacolla and Wells, 2004; Svozil *et al.*, 2008). Among non-B DNAs, G-quadruplexes (G4s) currently represent a hot topic in research and their involvement has been demonstrated in numerous physiological functions in both eukaryotic and prokaryotic organisms, as well as in some viruses (Lavezzo *et al.*, 2018; Nicola *et al.*, 2016; Rhodes and Lipps, 2015). G4s consist in four-stranded structures that form in single-stranded guanine-rich nucleic acids. Four guanines (Gs) can arrange within a planar tetrad via Hoogsteen base-pairings and the stacking of at least two tetrads creates the G4 scaffold. The intervening sequences are extruded as single-stranded loops, which are usually short and can contain any of the four nucleotides (nt).

G4s are clustered in crucial genomic regions, particularly in nucleosome-depleted euchromatic regions (Hänsel-Hertsch *et al.*, 2016) such as gene promoters, recombination sites and telomeres, but also in mRNAs and non-coding RNAs. They are involved in several physiological and pathological processes in both cellular and non-cellular organisms, including DNA replication, gene expression and viral latency (Nicola *et al.*, 2016; Rhodes and Lipps, 2015).

Due to their presence in telomeres and in the promoter region of many oncogenes, G4s have become important targets for small-molecule drugs in cancer treatment, and they are emerging also as targets for antiviral therapies (Neidle, 2017; Ruggiero and Richter, 2018). Most of these ligands usually bind a range of G4s, which can

be either an advantage, since a multi-gene response can be achieved, or a disadvantage, resulting in off-targeting of unwanted G4s (Asamitsu *et al.*, 2019). This lack of specificity is indeed hampering the transition of G4 ligands beyond the animal model stage since these small molecules are often associated with toxic effects (Iachettini *et al.*, 2013; Rizzo *et al.*, 2014), although some notable exceptions exist (Drygin *et al.*, 2009; Xu *et al.*, 2017).

Several bioinformatics tools for the prediction of potential quadruplex-forming sequences (PQSs) are available: the older ones are based on simple pattern matching rules and were developed aiming at the detection of PQSs with perfect G-islands (Arora *et al.*, 2006; D'Antonio *et al.*, 2006), whereas the newest tools can also detect degenerate patterns (Bedrat *et al.*, 2016; Brázda *et al.*, 2019; Dhapola and Chowdhury, 2016; Garant *et al.*, 2017; Hon *et al.*, 2017; Varizhuk *et al.*, 2014). For a comprehensive summary of existing tools and their functionalities, please refer to the recent pqsfinder paper (Hon *et al.*, 2017) and to Kwok *et al.* (2018).

In this work, we focus on two new peculiar classes of G4s that recently emerged as interesting molecular targets for drug design. The first class does not properly represent a novel G4 architecture, but rather includes higher-order structures generated by the cross-talk of two or more independent G4s that are adjacent along the primary sequence (Palumbo *et al.*, 2009; Rigo and Sissi, 2017). The second class includes G4/hairpin hybrid conformations. These structures, besides the tetrads that are the core of the G4, are characterized by the presence of auxiliary stem-loop structures that can occur within long loops. While long loops usually destabilize the G4 scaffolds (Guédin *et al.*, 2010), it has been observed that in this scenario they can exert a stabilizing effect (Butovskaya *et al.*, 2018; Onel *et al.*, 2016; Palumbo *et al.*, 2009). Given that these additional stem-loops are more distinctive and can make the whole G4 structure unique, G4/hairpin hybrids offer extremely valuable molecular targets for the development of specific and more selective ligands.

QPARSE is conceived to complement the common search of monomeric, monomolecular and possibly degenerate PQSs with the detection of more complex motifs characterized by additional sequence features. The tool aims to identify the following features: (i) intramolecular monomeric PQSs with perfect and (ii) degenerate G-islands; (iii) all possible redundant and overlapping PQSs resulting from the alternative usage of G-islands in G-rich regions; (iv) multimeric PQSs (MPQS, Fig. 1) that are adjacent in the linear sequence and could interact in the three-dimensional space; (v) PQSs with one or more long symmetric loops (long-looped PQS or LLPQS, Fig. 1) that could fold into hairpin-like structures, allowing the user to define his/her own symmetry rules. The tool is provided either as a Python program that can be easily downloaded and run on most computers or as a freely accessible web server available at http://www.medcomp.medicina.unipd.it/qparse/index.

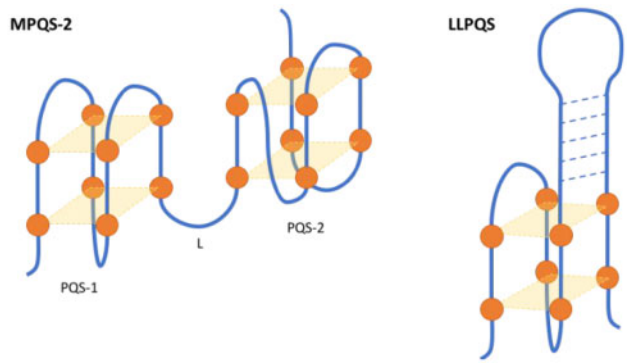


**Fig. 1.** Example of an MPQS of two G4s (left) and an LLPQS (right). Dots represent Gs. L indicates the maximum length allowed for the loops.

## 2 Materials and methods

### 2.1 Algorithm

QPARSE algorithm allows the user to set up multiple combinations of parameters (e.g. G-island degeneration, loop length and symmetry, number of consecutive PQSs, etc.) depending on specific needs. Since degeneration of G-tracts can result in ambiguous assignments between bulges, mismatches and loops (i.e. in the pattern GGAGG, the A is either a loop between two G-islands of length two, a bulge within a G-island of length four or a mismatch in a G-tract of length five), the tool does not take *a priori* decisions, but provides the exhaustive ensemble of all possible PQSs by building and traversing a direct acyclic graph (DAG). After this initial search, the results can be prioritized according to the assigned scores, or further refined by looking for internal symmetries within long loops. QPARSE algorithm can be summarized in the following steps: (i) detection of all possible G-islands, (ii) construction of the DAG, (iii) traversal of the DAG, (iv) scoring of the results and (v) output refinement.

#### G-islands detection

The input sequence is analyzed to identify all possible G-islands satisfying the input parameters, such as the minimum/maximum number of Gs required per island and the maximum number of mismatches/bulges allowed (Fig. 2, I). The islands are progressively detected using a finite-state machine that scans the sequence starting from each G, retrieving all the possible islands starting at that position.

#### DAG construction

The identified islands are modeled as nodes of a DAG, where non-overlapping islands within the maximum loop distance are connected through edges (Fig. 2, II).

#### DAG traversal

DAG traversal is performed by means of a Breadth-First Search approach coupled with a Depth-Limited Search (DLS) (see Supplementary Scheme S1). The algorithm accepts as parameters the number **N** of consecutive nodes **V** in the graph **D** that are required. Starting from the root, **D** is traversed in breadth and a DLS of depth **N** is performed for each node **V** that is retrieved. With this strategy, all potential paths of connected nodes **V** belonging to **D** and satisfying length **N** are exhaustively reported. Usually, **N** is four or a multiple of four, accounting for either canonical PQS or MPQS, respectively (Fig. 2, III).

#### Scoring the results

DAG navigation output can be highly redundant; hence, the results are ranked using a scoring system and only the highest scoring solutions are reported at each index (Fig. 2, IV). Depending on the analysis, the user can be interested in either classical PQSs/MPQSs, i.e. four/multiples of four proximal G-islands, or LLPQS. In the first case, corresponding to the default mode, the score only depends on the G-islands and is calculated as the sum of the scores of each individual island (Supplementary Scheme S2A). In the second case, the PQS score is calculated as the sum of the scores of both islands and loops participating in the motif, as described in Supplementary

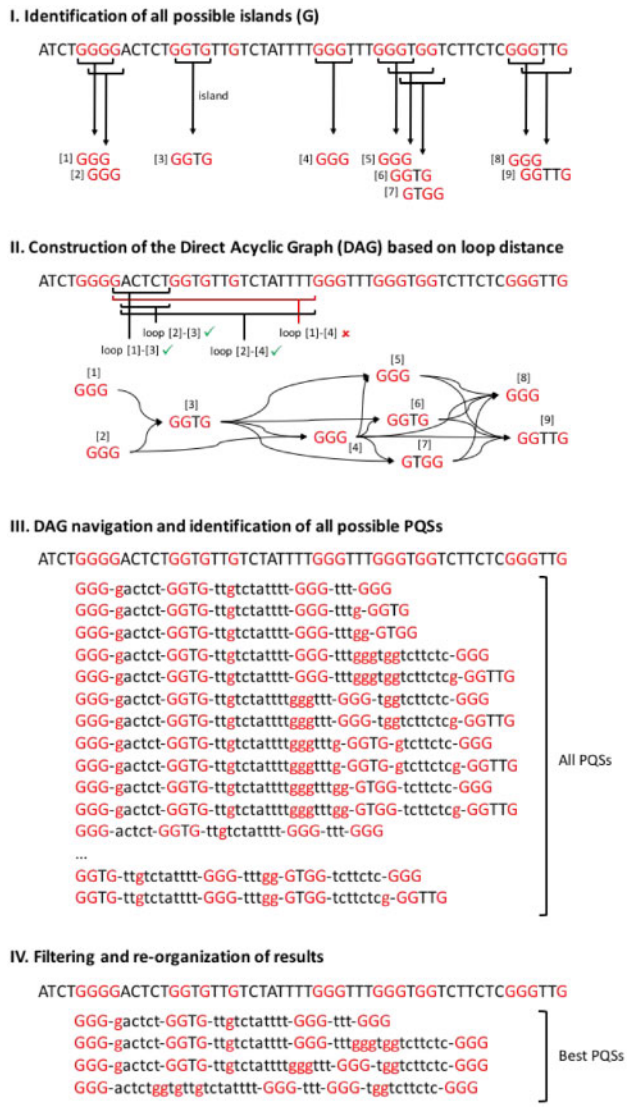


**Fig. 2.** Example of QPARSE algorithmic steps: (**I**) detection of G-islands of length three, with maximum one bulge and a maximum bulge-length of two nt [−m 3 –g 1 –l 2]; (**II**) DAG construction using a maximum loop length of 20 nt [−L 20]; (**III**) DAG traversal looking for monomeric PQSs, represented by all paths connecting four nodes. All possible results are reported in this step; (**IV**) filtering of the results based on the highest scores. In III and IV, uppercase Gs contribute to G-islands that participate in the PQSs.

Scheme S2B. The score assigned to loops is based on loop length and linearly decreases with the increase in length, accounting for the experimental observation that longer loops tend to destabilize the G4. However, longer loops can contain self-complementary regions that form stable hairpin structures that stabilize the G4. To evaluate this possibility, loops longer than 6 nt are analyzed for symmetric patterns, i.e. palindromes and mirrors, using a dynamic programming approach as described in Berselli *et al.* (2018). The minimum length of long loops is set to 7 to allow at least two base pairings in the stem, and at least three nucleotides in the hairpin loop. The nucleotides involved in the hairpin loop do not contribute to the final score. Presently, the implemented substitution matrices reward both palindrome and mirror symmetries (i.e. GG, CC, TT, AA, CG and AT pairings are allowed) (Supplementary Scheme S2C), as suggested by currently validated structures where both Watson-Crick and Hoogsteen hydrogen bonds contribute to the hairpin formation (Butovskaya *et al.*, 2018; Gajarský *et al.*, 2017; Onel *et al.*, 2016; Palumbo *et al.*, 2009). For each loop, the best self-alignment contributes to the final score, even negatively in the absence of symmetries. Three scoring matrices are currently implemented in the tool for the evaluation of different properties: palindrome (−sP), mirror (−sM), mixed palindrome-mirror (−sX). Alternatively, the user can specify a custom matrix to be used (−sC).

### mfold implementation

To extend QPARSE functionality and to evaluate energy stabilization of hairpin loops connecting G-islands, we integrated a wrapper around mfold (Zuker, 2003) algorithm for predicting DNA/RNA secondary structures. mfold can be run on long loops detected by QPARSE and calculates both energy stability and conformation by using thermodynamic methods. mfold can be effective only when palindromes (−sP option) are evaluated because it does not consider non-canonical base pairings.

### Output refinement

To better organize the results and reduce the output, an additional parser is provided: (i) to merge overlapping PQSs into one sequence, (ii) to filter the results and return only the maximum number of non-overlapping PQSs, (iii) to rank the results based on the final score (ranking is reported by index as default) and (iv) to convert the results into useful text formats like GFF or TSV for an easier visualization. Finally, if loop symmetry is considered, the parser allows to explicitly visualize the best loop self-alignment for each detected LLPQS, presented in a blast-like format.

## 2.2 Implementation

QPARSE is a Python program (v. 2.7), easily portable in different operating systems, which only requires the *numpy* library (http://www.numpy.org/). A detailed user guide is provided in the installation package. The command line version of QPARSE is fast and can scale up to the analysis of genome-size datasets to be performed in most desktop computers. When applied to human chromosome 1 (∼250 million nucleotides) the time and memory requirements for searching the G-patterns are reported in Supplementary Table S1. Additionally, a web server is available at http://www.medcomp.medicina.unipd.it/qparse/index, where an interactive graphical interface guides the user through the analysis. The web server is implemented using the python framework Flask and allows the analysis of small datasets (up to 10 000 nt).

# 3 Results

## 3.1 Benchmark

The assessment of QPARSE was performed at two distinct levels. First, we evaluated QPARSE in the context of state-of-the-art G4-datasets and tools for PQS prediction. The aim of this benchmark was to assess QPARSE ability to correctly detect experimentally validated G4s and be competitive with the best tools available in literature (Bedrat *et al.*, 2016; Hon *et al.*, 2017; Sahakyan *et al.*, 2017). We used two independent datasets: one was created by merging the Lit392 (Bedrat *et al.*, 2016) with the G4RNA database (Garant *et al.*, 2015), obtaining a collection of 506 positive and 132 negative samples (Lit638); the other is the G4-seq dataset published by Chambers *et al.* (2015). Second, we aimed at assessing the novel features implemented in QPARSE, namely the possibility of looking for MPQSs and LLPQSs. Since the discovery of such structures is quite recent and no benchmark test sets are available, we evaluated

QPARSE on case studies taken from recent literature. Among all available tools for PQS prediction, we selected pqsfinder (Hon *et al.*, 2017) and G4Hunter (Bedrat *et al.*, 2016) for comparative purposes based on the ability to scale up to genome wide analyses and search for degenerate patterns. Another recent tool, Quadron (Sahakyan *et al.*, 2017), was excluded from the comparison because, despite relying on a promising machine learning approach, it can currently identify only non-degenerate PQSs.

On Lit638 dataset, QPARSE and pqsfinder achieved a similar accuracy of 0.9 outperforming G4Hunter that reached 0.75 (Supplementary Table S2). On the G4-seq dataset, QPARSE still proved to be competitive with pqsfinder as shown in the precision–recall and *F*-measure plots (Supplementary Fig. S1).

## 3.2 Detection of MPQS

Rigo and Sissi (2017) recently demonstrated that two G4s, adjacent in the linear sequence of the c-KIT promoter, interact in the three-dimensional space into a higher-order structure that might be involved in regulating the gene expression. Similarly, another cross-talk between consecutive G4s was reported in the promoter region of the hTERT gene (Palumbo *et al.*, 2009). Although *in vivo* validation and functional characterization are still needed, the G4–G4 interaction surface may provide a new and attractive target for drug design, representing a totally unexplored field for the development of new therapies. Considering these recent evidences, QPARSE was designed to search for MPQSs, intended as consecutive PQS that are at loop distance range. This feature is activated by the -n option, which defines the number of consecutive islands to search for. If a multiple of four is selected, the tool will detect MPQSs. As a proof of principle, we performed a search for MPQSs in the promoter regions of all human genes (see Supplementary Materials). For each gene annotated in gencode (v28) (Harrow *et al.*, 2012), we extracted 15 000 nt both upstream and downstream the transcription start site (TSS), and searched for monomeric (PQS), dimeric (MPQS-2) and trimeric (MPQS-3) PQSs. As shown by the peaks in Fig. 3 (line 1), a large fraction of the genomic regions across the TSS display at least one PQS (∼49% of the total regions) if we consider the range −200/+600 nt around TSS. Moreover, there is a consistent fraction of genes having at least one MPQS-2 (Fig. 3, line 2, ∼15%) in the same region or one MPQS-3 (Fig. 3, line 3, ∼4%). Since the G-C content increases in the same region (Fig. 3, dashed line 4), we performed additional investigations to assess whether the number of observed PQSs was something unexpected or it was merely a consequence of this bias in base frequency. By comparing the results obtained in real and reshuffled sequences, we confirmed that the observed amount of PQSs is significantly higher than the expected, especially for MPQSs, even in extremely G/C rich sequences (Supplementary Figs S2–S4), suggesting for a potential role of MPQSs that would be worth investigating.

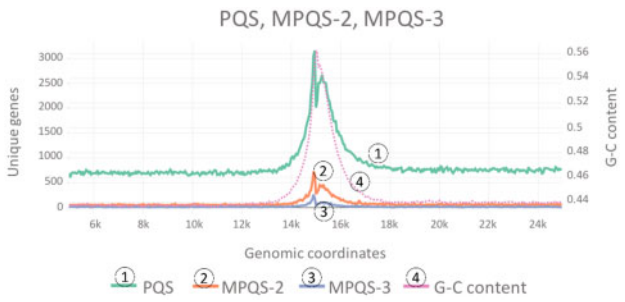


**Fig. 3.** Distribution of PQS, MPQS-2 and MPQS-3 in the promoter regions of all annotated human genes. TSS is located at 15k on the *x*-axis. The average G-C content is reported alongside.

## 3.3 PQS characterized by long symmetric loops (LLPQS)

The most innovative feature of QPARSE is the possibility to identify peculiar PQSs containing long loops with symmetric properties. Long loops are known to destabilize the overall structure of G4s (Guédin *et al.*, 2010); however, the presence of long loops that can fold into auxiliary secondary structures, i.e. hairpins or G-hairpins, has been shown to contribute to the G4 stability. So far, three independent studies demonstrated that LLPQSs can be stable *in vitro*: two in the promoter region of human genes, hTERT (Palumbo *et al.*, 2009) and BCL2 (Onel *et al.*, 2016), and one in the long terminal repeat (LTR) of the human immunodeficiency virus 1 (HIV-1) (Butovskaya *et al.*, 2018). QPARSE options can be tweaked to allow the detection of LLPQS and to discriminate even sets of mutants that are known not to fold. In all three cases, QPARSE was run with the following combination of parameters: (i) minimum island length [−m] of 3 Gs; (ii) maximum loop length [−L] of 30 nt and (iii) mixed scoring matrix to evaluate long-loop symmetry [−sX].

Finally, the results were processed with QPARSE_parser script to sort them by score [−s] and exclude all PQSs with more than one long loop (>6 nt, [−mL 1]). pqsfinder was used on the same case studies as a comparison and the 'overlapping' option was set to 'TRUE' to extend the standard output.

In addition, the long loops predicted by QPARSE were evaluated with mfold [−mfold_s] to calculate their thermodynamic and conformation stability. Detailed results are reported in Supplementary Materials S8–S10.

### 3.3.1 HIV-1

HIV-1 is a lentivirus responsible for the acquired immunodeficiency syndrome that can integrate into the host chromosomes. In this integrated form, the 5′-LTR serves as the unique promoter for viral transcription and is involved in reactivation from latency (Pereira *et al.*, 2000). Several G-tracts are present in the U3 region of the 5′-LTR, and they can fold into distinct G4s, which have a role in the regulation of HIV-1 activity. Among them, the G4 present in the *LTR-III* region is the most stable and its structure has been determined by nuclear magnetic resonance (NMR) (Butovskaya *et al.*, 2018), revealing two peculiar features: (i) a duplex–quadruplex structure combination, due to the presence of a long (12 nt) palindromic loop and (ii) a broken strand structure, resulting in a V-shaped loop and a non-canonical usage of G-islands in the construction of the stacked tetrads. QPARSE was run on the complete *LTR-II-III-IV* region of the proviral genome (from −100 to −47 genomic coordinates) and the results are summarized in Table 1. The highest scoring PQSs (results #1 and #2) are found in the *LTR-III* region and correspond to the experimental results. In both cases, the predicted loop slightly differs from that determined by NMR: G3 is assigned to the first G-island instead of being part of the first loop, and one additional base pairing is predicted, but not demonstrated experimentally (Watson-Crick pairing between A4 and T14). pqsfinder detected the same hits in the sequence region even though not ranked with the highest scores (Table 1 and Supplementary Fig. S5). QPARSE was also run on a set of *LTR-III* mutated sequences that were originally tested by the authors to assess the importance of different nucleotides in loop folding (Supplementary Table S3) (Butovskaya *et al.*, 2018). In most cases, mutants able to maintain the fold (G10A, G6A-C12T) receive higher scores than the others (ΔA4, ΔT14), whereas the exceptions are ΔG3 (completely missed because

**Table 1.** Results of the analyses of QPARSE and pqsfinder on HIV-1 (A), hTERT (B) and BCL2 (C)

| **(A)** | **Sequence** | **Long loop alignment** | **QPARSE score** |
|---|---|---|---|
| **HIV-1 LTR-II-III-IV** | 1 5 10 15 20 25 30<br>tttccgctggggactttccagggaggcgtggcctgggcgggactggggagtggt | | |
| **Experimentally validated G4** | GGgaggcgtggcctGGGcGGGactGGGG* | gaggc**g**<br>\|\|\| **t**<br>-tccg**g** | - |
| **LLPQS predicted by QPARSE** | (1) GGGaggcgtggcctGGGcGGGactGGG | aggc**g**<br>\|\|\|\| **t**<br>tccg**g** | 90<br>(**8**\|6\|4) |
| | (2) GGGaggcgtggcctGGGcGGGactgGGG | aggc**g**<br>\|\|\|\| **t**<br>tccg**g** | 89<br>(**8**\|6\|3) |
| **PQS predicted by pqsfinder** | (1) GGGGactttccaGGGAGGCGTGGCCTGGGCGGGactGGGG | NA | 59 |
| | (.) GGGaggcgtggcctGGGcGGGactgGGG | NA | 55 |
| **(B)** | **Sequence** | **Long loop alignment** | **QPARSE score** |
| **hTERT core promoter (−22 to −90 versus TSS)** | 1 2 3 4 5 6 7 8<br>ggggagggctgggagggcccggaggggctgggccggggacccggga<br>9 10 11 12<br>ggggtcgggacggggcgggg | | |
| **Experimentally validated G4** | 5 6 11 12<br>gGGGgctGGGccggggacccgggaggggtcgggacgGGGcGGGg | −ccggggacccg**g**<br>:::\|\|\|\|\| **g**<br>gcagggctggg**a** | - |
| **LLPQS predicted by QPARSE** | 5 6 11 12<br>(1) GGGctGGGccggggacccgggaggggtcgggacgGGGcGGG | cc−ggggacccg**g**<br>\|: :::\|\|\|\|\|: **g**<br>gcagggctggg**a** | 97<br>(5\|**14**\|6) |
| | 1 2 7 8<br>(2) GGGaGGGgctgggagggcccggaggggctgggccgGGGacccGGG | gctgggagggccc**g**<br>:: :::\|\|::\|\|\| **g**<br>gccgggtcggggg**a** | 96<br>(6\|**15**\|3) |
| **PQS predicted by pqsfinder** | 7 9 11 12<br>(1) GGGGacccgggaGGGGtcgggacGGGGcGGGG | NA | 95 |
| | 5 7 9 10−11−12<br>(.) GGGGctgggccGGGGacccgggaGGGGtcggGACGGGGCGGG | NA | 70 |
| **(C)** | **Sequence** | **Long loop alignment** | **QPARSE score** |
| **BCL2 region overlapping P1 promoter** | ccccggcaccttcgctggcagcggcggcggcggcagcgcggcggggccacggag<br>agcggc<br>1 2 3 4 5<br>gggcgggagcgcggcgggcgggcgggcaggcggcgcggaggggcgggcgcggga<br>ggaagg<br>gggcgggagcggggctgtggtgcctg | | |
| **Experimentally validated G4 (P1G4)** | 1 2 4 5<br>cGGGcGGGagcgcggcgggcGGGcGGGc | −agc**g**<br>\|\| **c**<br>ggcg**g** | - |
| **LLPQS predicted by QPARSE** | 3 4 5<br>(1) GGGgccacggagagcggcgggcgggagcgcggcGGGcGGGcGGG | gccacggagagcgg**c**<br>\|\|\| ::\| :::\|:\| **gg**<br>cgg−cgc−gagggc**g** | 96<br>(**12**\|6\|6) |
| | 3 4 5<br>(2) GGGcGGGcGGGcaggcggcgcggaggggcgggcgcGGG | caggcggcgc**g**<br>: \|:\|:\|\|:\| **ga**<br>cgcgggcggg**g** | 96<br>(6\|6\|**12**) |
| | … | … | … |
| | 1 2 4 5<br>(26) GGGcGGGagcgcggcgggcGGGcGGG | agcg**c**<br>:\|: **gg**<br>cggg**c** | 86<br>(6\|**2**\|6) |
| **PQS predicted by pqsfinder** | 1 2 3 4<br>(1) GGGcGGGagcgcggcGGGcGGG | NA | 90 |
| | 1 2 3 5<br>(.) GGGcGGGagcgcggcGGGcgggcGGG | NA | 57 |

*Note*: For each case-study the table reports: (i) the genomic sequence; (ii) the experimentally validated G4 with the observed hairpin structure; (iii) the top two PQSs predicted by QPARSE with the predicted hairpin structure; (iv) the top PQS predicted by pqsfinder and the predicted PQS more similar to the validated G4. Upper case Gs are those involved in G4 tetrads. In the long-loop alignment column, vertical dashes represent the putative Watson-Crick pairing, colons stand for putative Hoogsteen pairing, nucleotides highlighted in bold fall into hairpin loops; hence, they are not involved in base pairing. In the score column, numbers within round brackets represent the score of each individual loop of the PQS. (A) The coordinates reported on the HIV-1 LTR-II-III-IV sequence refer to the experimentally validated G4 and all other reported sequences are aligned to them. (B) The 12G-tracts present in the core promoter of hTERT are underlined and numbered. (C) The five G-tracts present in the P1G4 G4 sequence are underlined and numbered.

[a]Underlined Gs are those constituting the same (atypical) G-island.

of the loss of the first G-island) and T14A (highly scored because it maintains most of the symmetry, but the mutation is probably affecting the duplex–quadruplex junction).

#### 3.3.2 hTERT

The catalytic domain of telomerase (hTERT) is responsible for the addition of the nucleotide stretch TTAGGG to the end of chromosomes telomeres, thus preventing their degradation following multiple replication cycles. Since its activation is critical for cells immortalization, hTERT is expressed in almost 90% of cancers and represents one of the major targets in cancer therapy. hTERT core promoter spans the 180 nt upstream the TSS and contains multiple binding sites for transcription factors, and several G-tracts that are involved in the formation of different G4s. The more stable G4 involves the G-tracts 5-6-11-12 (Table 1), and is characterized by a 26-nt loop, which was demonstrated to form a hairpin structure crucial for the G4 stability (Palumbo *et al.*, 2009). QPARSE was run on a 400-nt sequence extracted from the hTERT gene (200 nt upstream and downstream the TSS), as well as on a set of mutated sequences that were tested to evaluate the hairpin stability. Despite the presence of multiple short PQS in the region, QPARSE correctly retrieved the one with the 26-nt loop as the best hit, thanks to the score contribution given by the symmetry detected in the loop. Some additional base pairings were predicted but not observed experimentally (Table 1), resulting in a slightly different loop conformation; nonetheless, the central pairing core is perfectly detected and confirms the reported structure. Interestingly, another PQS received a very high score (result #2); it involves G-tracts 1-2-7-8 and a slightly longer loop (29 nt), with almost complete symmetry. The other high scoring results are isomers of hits #1 and #2, due to the redundant number of guanines in several G-islands, and most of them were confirmed by dimethyl sulfate protection assays (Palumbo *et al.*, 2009). pqsfinder was not able to identify the right G4 with the correct topology in the region (Table 1 and Supplementary Fig. S6). The analyzed mutated sequences include different substitutions of guanines in G-tracts 7 and 9, which should impact the stability of the hairpin structure. The replacement of either one or both G-tracts with thymines results in the disruption of the hairpin stem. QPARSE assigns slightly lower scores to these sequences, and interestingly identifies alternative shorter loops that still maintain a high degree of symmetry (Supplementary Table S4). Finally, when swapping a G-tract with a C-tract in the stem, QPARSE correctly detects the 26-nt loop with the same score as in the original sequence.

#### 3.3.3 BCL2

The B-cell lymphoma-2 (BCL2) is a protein located in the mitochondrial outer membrane, which plays a crucial role in promoting cell survival and inhibiting apoptosis (Chipuk *et al.*, 2010). Its expression is frequently high in several tumors, representing a popular target for many anti-cancer strategies (Montero and Letai, 2018). The BCL2 promoter region is extremely GC-rich and contains distinct and interchangeable G4s. In this context, a G4 named P1G4 was recently reported to have a predominant role in BCL2 transcription repression; it contains five G-tracts and a central 12-nt loop that assumes a stem-loop conformation and can represent a potential target for specific small molecules (Onel *et al.*, 2016). Furthermore, P1G4 was shown to be in a dynamic equilibrium between two structures, one regular and one broken strand G4. QPARSE was run on a 400-nt sequence centered around the TSS of the BCL2 gene and results are summarized in Table 1. Interestingly, there are several PQSs detected by QPARSE that contain extremely long symmetric loops, thus obtaining very high scores; as a consequence, the experimentally validated P1G4 is only at the 26th position in the rank list. Nonetheless, the first QPARSE hit is a super-string of P1G4, whose first G-tract is completely embedded in a 30 nt highly symmetric loop. While we do not have experimental data to support this prediction, the folding of one or more of these high scoring PQSs cannot be excluded and was never tested. Another interesting point is the structure of P1G4 hairpin loop: in NMR experiments, only two Watson-Crick base pairings are detected, although the stability of the stem-loop itself is surprisingly high (melting temperature of 45°C), as stated by the authors (Onel *et al.*, 2016). QPARSE predicts a different point of symmetry for the loop, resulting in five potential base pairings. A similar conformation, or even partially similar, could explain the higher than expected stability of the loop. Even in this case, pqsfinder was not able to detect the right G4 with the correct topology in the region (Table 1 and Supplementary Fig. S7). To further validate the results, the authors also tested two P1G4 mutated sequences: (i) in P1G4_runIII_G/T the third G-island, the one contained in the stem-loop structure, was replaced by thymines, thus maintaining the tetrads scaffold of the G4; (ii) P1G4_KO is a complete knock-out of the G4, since the middle guanine of each G-tract was replaced by a thymine. In both cases, QPARSE results are correct, with a PQS predicted in P1G4_runIII_G/T and no results obtained for the P1G4_KO.

## 4 Discussion

G4s are promising therapeutic targets for several pathologic conditions, including cancer and infectious diseases (Neidle, 2017; Ruggiero and Richter, 2018). However, G4 targeting with small molecules is challenged by the lack of specificity due to their intrinsic structural features: an overall conserved core of stacked G-quartets connected by short loops, which provide little discrimination between the different G4s. Recently, new peculiar G4s conformations have been identified adding structural complexity to the basic G4 scaffold: (i) multimeric G4s, which are independent G4s adjacent in the linear nucleotide sequence that interact in the three-dimensional space (Palumbo *et al.*, 2009; Rigo and Sissi, 2017) and (ii) hybrids G4/hairpins, which are G4s with one long loop that can fold into a stable stem-loop structure (Butovskaya *et al.*, 2018; Onel *et al.*, 2016; Palumbo *et al.*, 2009).

The detection of PQSs in nucleic acids represents a field of research that attracts a lot of interest, boosting the development of several bioinformatics tools based on different algorithms and searching strategies (Hon *et al.*, 2017). However, none of them is currently designed to specifically detect either MPQSs, or hybrids G4/hairpins (LLPQS).

QPARSE is a tool able to address these needs relying on a novel graph-based algorithm and a dynamic programming approach to search for MPQSs and LLPQSs. It can detect 'standard' intramolecular PQSs, i.e. characterized by four G-islands (either perfect or degenerated) interspaced by short loops. To this basic usage mode, which we proved to be competitive with the best state-of-the-art tools on two different benchmark test sets (Bedrat *et al.*, 2016; Hon *et al.*, 2017), we added additional features for the detection of higher-order structures. We demonstrate that QPARSE can detect both MPQS and LLPQS with long symmetric loops. In the first case, we show that MPQSs are present in the promoter regions of several human genes and their presence is not entirely ascribable to the high G-C content that characterizes these regions, claiming for some biological relevance. Although these are just *in silico* predictions, the potential presence of novel specific molecular targets in some genes

is intriguing and deserves additional consideration. Regarding G4/hairpin hybrids, we demonstrate that QPARSE performs well on three independent and experimentally validated case studies (Butovskaya *et al.*, 2018; Onel *et al.*, 2016; Palumbo *et al.*, 2009), thanks to the implementation of a customizable scoring matrix for the detection of symmetries within loops. In such cases, atypical G4s characterized by one long loop were shown to be able not only to fold, but also to be extremely stable, mainly because of the symmetric loops which lead to the formation of auxiliary hairpin structures. Nonetheless, descriptions of such structures are still rare and it is difficult to generalize properties and understand how widespread they are. As an example, different structure combinations might exist, such as multiple hairpins in different loops, or even the presence of secondary structures other than hairpins.

Interestingly, QPARSE, when challenged on the G4-seq dataset with the long symmetric loop option activated, reached almost a recall of 90% containing high scoring hits. This suggests that palindromic/mirror symmetries are potentially widespread in the long loops of the experimental data from G4-seq dataset.

QPARSE is computationally efficient, able to run on millions of nucleotides in few minutes and requires a small amount of memory. It can provide results in several formats, including GFF that can be easily parsed and uploaded to most genome browsers. The tool can be downloaded and run locally or accessed through the web server available at http://www.medcomp.medicina.unipd.it/qparse/index.

In summary, we believe that QPARSE fills in the gap of present PQS predictors because it offers a solid starting point for subsequent experimental validation of MPQSs and LLPQSs that are emerging as promising candidates for drug design.

## Funding

This work was supported by University of Padova [grant number TOPP_SID19_01] to ST.

*Conflict of Interest*: none declared.

## References

Arora,A. *et al.* (2006) Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res.*, **34**, W683–W685.

Asamitsu,S. *et al.* (2019) Recent progress of targeted G-quadruplex-preferred ligands toward cancer therapy. *Molecules*, **24**, 429.

Bacolla,A. and Wells,R.D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*, **279**, 47411–47414.

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

Berselli,M. *et al.* (2018) NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics*, **34**, 2503–2505.

Brázda,V. *et al.* (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*

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.

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

Chipuk,J.E. *et al.* (2010) The BCL-2 family reunion. *Mol. Cell*, **37**, 299–310.

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

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

Drygin,D. *et al.* (2009) Anticancer activity of CX-3543: a direct Inhibitor of rRNA biogenesis. *Cancer Res.*, **69**, 7653–7661.

Gajarský,M. *et al.* (2017) Structure of a stable G-hairpin. *J. Am. Chem. Soc.*, **139**, 3591–3594.

Garant,J.M. *et al.* (2015) G4RNA: an RNA G-quadruplex database. *Database*, **2015**, bav059.

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

Guédin,A. *et al.* (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.

Hänsel-Hertsch,R. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

Hon,J. *et al.* (2017) pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*, **33**, 3373–3379.

Iachettini,S. *et al.* (2013) On and off-target effects of telomere uncapping G-quadruplex selective ligands based on pentacyclic acridinium salts. *J. Exp. Clin. Cancer Res.*, **32**, 68.

Kwok,C.K. *et al.* (2018) Detecting RNA G-quadruplexes (rG4s) in the transcriptome. *Cold Spring Harb. Perspect. Biol.*, **10**, a032284.

Lavezzo,E. *et al.* (2018) G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide. *PLoS Comput. Biol.*, **14**, e1006675.

Montero,J. and Letai,A. (2018) Why do BCL-2 inhibitors work and where should we use them in the clinic?. *Cell Death Differ.*, **25**, 56.

Neidle,S. (2017) Quadruplex nucleic acids as targets for anticancer therapeutics. *Nat. Rev. Chem.*, **1**, 0041.

Nicola,B.D. *et al.* (2016) Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome. *Nucleic Acids Res.*, **44**, 6442–6451.

Onel,B. *et al.* (2016) A new G-quadruplex with hairpin loop immediately upstream of the human BCL2 P1 promoter modulates transcription. *J. Am. Chem. Soc.*, **138**, 2563–2570.

Palumbo,S.L. *et al.* (2009) Formation of a unique end-to-end stacked pair of G-Quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. *J. Am. Chem. Soc.*, **131**, 10878–10891.

Pereira,L.A. *et al.* (2000) A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res.*, **28**, 663–668.

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

Rigo,R. and Sissi,C. (2017) Characterization of G4–G4 crosstalk in the c-KIT promoter region. *J. Am. Chem. Soc.*, **56**, 4309–4312.

Rizzo,A. *et al.* (2014) Identification of novel RHPS4-derivative ligands with improved toxicological profiles and telomere-targeting activities. *J. Exp. Clin. Cancer Res.*, **33**, 81.

Ruggiero,E. and Richter,S.N. (2018) G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. *Nucleic Acids Res.*, **46**, 3270–3283.

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

Svozil,D. *et al.* (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.

Varizhuk,A. *et al.* (2014). An improved search algorithm to find G-quadruplexes in genome sequences. *bioRxiv*, doi:10.1101/001990.

Xu,H. *et al.* (2017) CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat. Commun.*, **8**, 1443.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.