Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

The expanding repertoire of G4 DNA structures

Anna Varizhuk [a, b], Dmitry Ischenko [a, c], Vladimir Tsvetkov [a, d], Roman Novikov [b, e], Nikolay Kulemin [a, c], Dmitry Kaluzhny [b], Maria Vlasenok [a, c], Vladimir Naumov [a], Igor Smirnov [a], Galina Pozmogova [a, *]

[a] *Research and Clinical Center for Physical Chemical Medicine, 119435 Moscow, Russia*
[b] *Engenlhardt Institute of Molecular Biology, 119991 Moscow, Russia*
[c] *Moscow Institute of Physics and Technology (State University), 117303 Moscow, Russia*
[d] *Department of Molecular Virology, FSBI Research Institute of Influenza, Ministry of Health of the Russian Federation, Saint Petersburg, Russia*
[e] *N.D. Zelinsky Institute of Organic Chemistry, Moscow 119991, Russia*

ARTICLE INFO

Article history:
Received 2 November 2016
Received in revised form
30 December 2016
Accepted 12 January 2017
Available online 19 January 2017

Keywords:
Noncanonical DNA structures
G-quadruplexes
Genomic analysis
G4 ligands

ABSTRACT

The definition of DNA and RNA G-quadruplexes (G4s) has recently been broadened to include structures with certain defects: bulges, G-vacancies or mismatches. Despite the striking progress in computational methods for assessing G4 folding propensity, predicting G4s with defects remains problematic, reflecting the enhanced sequential diversity of these motifs. "Imperfect" G4 motifs, i.e., those containing interrupted or truncated G-runs, are typically omitted from genomic analyses. We report here studies of G4s with defects and compare these structures with classical ("perfect") quadruplexes. Thermal stabilities and ligand interactions are also discussed. We exploited a simple in-house computational tool for mining putative G4s with defects in the human genome. The obtained profiles of the genomic distribution of imperfect G4 motifs were analyzed. Collectively, our findings suggest that, similar to classical G4s, imperfect G4s could be considered as potential regulatory elements, pathology biomarkers and therapeutic targets.

1. Introduction

G-quadruplexes (G4s), four-stranded DNA or RNA architectures stabilized by the stacking of square planar arrangements of guanine bases (G-tetrads), are impressively polymorphic [1], depending on the number of G-quartets; syn vs. anti Gua conformations; loop length, sequence and connectivities; and strand orientation and handedness [2]. Additional structural diversity arises from the accommodation of A-, U-, T- or C- homo-tetrads [3–5], mixed tetrads [6–9], bulges [10] and vacancies [11,12].

Despite the growing interest in the multifaceted biological role of G4 motifs [13–15] and the potential application of G4 aptamers [16,17], the determinants of G4 formation and function in vivo are not fully understood. Recent findings advocate short pyrimidine loops as the hallmarks of G4 motifs prone to triggering genomic instability [18]. This idea is consistent with the results of an in vitro sequence-stability analysis of intramolecular G4s [19] and G4 motifs in dsDNA [20]. Subsequent studies have revealed that the short minimum loop length and the presence of the thymine bases are two key composition properties that promote G4 formation.

Most G4-predicting software programs utilize the consensus sequence rule, according to S. Neidle and S. Balasubramanian, i.e., the $G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}N_{1\text{-}7}G_{3\text{-}5}$ formula. The limitations of this approach are apparent.

- First, the maximum loop length is a matter of debate. Restricting this length to 7 nt may be a reasonable compromise between the sensitivity and robustness of G4 motif mining. However, compelling evidence supports the biological significance of genomic G4s with longer loops.
- Second, the consensus formula does not apply to G4s with bulges, mismatches or vacancies.
- Third, the genomic context, relative G4 and duplex stabilities, and other factors that influence the thermodynamic probability of G4 formation are not considered.

New-generation G4 search tools employ more complex pattern-based rules. Computational methods for predicting G4 folding

* Corresponding author.
E-mail addresses: pozmogova@niifhm.ru, pozmge@gmail.com (G. Pozmogova).

http://dx.doi.org/10.1016/j.biochi.2017.01.003

propensity based on regression models have been reported. J. Huppert and coworkers proposed a general regression model that facilitates the flexible incorporation of various stability-related sequence features [21]. Subsequently, J.P. Perreault reported a somewhat analogous algorithm that considers the impact of flanking sequences [22]. Recently, J.L. Mergny and coworkers developed a search tool (G4Hunter) that considered the relative likelihood of canonical and non-canonical structures [23]. The underlying algorithm relies on two basic factors: the presence of G-blocks and G/C skewness. Kim et al. reported an alternative algorithm that relies on dsDNA-specific pattern rules [20] but has serious limitations: the current version is inapplicable to mixed-sequence loops, particularly those containing G, as apparently this algorithm cannot handle G4s with bulges, mismatches or vacancies (interrupted G-runs are not tolerated).

The classical algorithm yields approximately 376,000 PQS sites in the human genome, whereas the estimation based on the high-resolution sequencing method [24] revealed 716,310 distinct G4 structures. The re-evaluation of G4 abundance in the human genome using G4Hunter yielded a nearly identical number with an appropriate window size and score threshold [23]. A large subset of the G4s, which were not predicted using first-generation search tools, includes noncanonical structures with “defects,” such as long loops or bulges, highlighting the necessity of improving computational methods and suggesting the potential in vivo function of the previously omitted G4s.

In the present study, we report the investigation of G4s containing interrupted or truncated G-runs (“imperfect” quadruplexes), i.e., G4s with bulges, vacancies or mismatches, and compare these structures with classical (“perfect”) quadruplexes. Thermal stabilities, ligand interactions and genomic distributions are also discussed.

2. Materials and methods

2.1. Oligonucleotide synthesis, purification and secondary structure verification using optical methods

Oligonucleotide (ON) synthesis, purification, MALDI TOF MS analysis and ethidium bromide (EtBr) rotational relaxation time measurements were performed as previously described [25]. The purity of all ONs was ≥95% based on HPLC analysis. UV spectra, CD spectra and CD melting curves were recorded using a Chirascan spectrophotometer (Applied Photophysics, UK) equipped with a thermoregulated cuvette holder. Molar CD per nucleotide residue was calculated as follows: $\Delta\varepsilon = \theta/(32.982 \times C \times l \times n)$, where θ is ellipticity (degree), C is ON concentration (M); l is optical pathlength (cm) and n is the number of nucleotide residues in the ON. All ON solutions were denatured at 95 °C for 5 min and snap cooled on ice prior to measurement to ensure intramolecular folding.

2.2. NMR studies

NMR samples were prepared at a concentration of ~0.1 mM in 0.6 ml of $H_2O + D_2O$ (10%) buffer solution containing 20 mM Tris-HCl (pH 7.5) and 100 mM KCl and annealed (heated to 90 °C for 3 min, and subsequently cooled on ice prior to spectral measurements to ensure unimolecular quadruplex folding). The 1H NMR spectra were recorded using Bruker AMX III (400.1 MHz) and Bruker AVANCE II 600 (600.1 MHz) spectrometers. The 1H chemical shifts were referenced relative to an external standard, sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS). The spectra were recorded using presaturation or pulsed-field gradient WATERGATE W5 pulse sequences (zgprsp and zggpw5 from the Bruker library, respectively) for H_2O suppression.

2.3. Molecular modeling

2.3.1. Model preparation and molecular dynamics simulations

Ct1 GQ models 1 and 2 were created in the following manner. The starting positions of the GQ core atoms were obtained from the PDB (139D and 2KQH). The core of every GQ was created using Swiss-PDB Viewer. Subsequently, the loops were added step by step, utilizing the SYBYL 8.0 molecular modeling package. To remove unfavorable van der Waals interactions, the models were re-optimized after attaching each loop using SYBYL 8.0 and the Powell method with the following parameters: Gasteiger-Hückel charges, TRIPOS force field, non-bonded cut-off distance of 8 Å, distance-dependent dielectric function, 1000 iterations, the simplex method in an initial optimization and the energy gradient convergence criterion with a threshold of 0.05 kcal*mol^{-1}*Å^{-1}. The GQ core was frozen during the above re-optimizations.

The molecular dynamics simulations (MD) were performed using the Amber 10 suite with ff99SB and parmbsc0 force fields as previously described [25]. The trajectory length was 35 ns. Snapshot visualization and hydrogen bond analysis were performed using VMD (http://www.ks.uiuc.edu/Research/vmd/) with a donor-acceptor distance of 3 Å and an angle cutoff of 20°. Snapshots were taken every 0.1 ns.

2.3.2. Free energy calculations

The MM-GBSA method was used to calculate the Ct1 GQ free energies for models 1 and 2. In this approach, the free energy is calculated according to the formula $G = E_{MM} + G_{sol}$ -TS, where E_{MM}, G_{sol} and TS are the total mechanical energy of the molecule in gas phase, the free energy of hydration and the entropic contribution, respectively. E_{MM} was calculated as the sum of the electrostatic energies, van der Waals energies and the energies of internal strain (bonds, angles and dihedrals) using a molecular-mechanics approach. G_{sol} was calculated as the sum of the polar (G_{polar}) and nonpolar ($G_{nonpolar}$) terms. The electrostatic contribution to the hydration energy G_{polar} was computed using the Generalized Born (GB) method [26], which utilizes the algorithm developed by Onufriev et al. [27,28] to calculate the effective Born radii. The nonpolar component of hydration energy $G_{nonpolar}$, which includes solute-solvent van der Waals interactions and the free energy of cavity formation in the solvent, was calculated using the formula $G_{nonpolar} = \alpha{*}SASA$, where $SASA$ is the solvent accessible surface area. $SASA$ was computed using the LCPO method [29] with $\alpha = 0.00542$ kcal/mol^{-1} Å^{-2}. The entropic term was not explicitly calculated, but was implicitly accounted for via the GQ conformational mobility. Snapshots taken from a single trajectory of the MD simulation of the complex were used to calculate the binding free energy.

2.4. G4 ligands

Pyridostatin (PDS) was obtained from ApexBio, Thyoflavin T (ThT) was obtained from Abcam, and N-methyl mesoporphyrin IX (NMM) was obtained from Frontier Scientific. G4 melting with PDS or NMM was performed as described in the previous section. PDS or NMM were added to pre-annealed ON solutions. ThT fluorescence was registered using a Chirascan spectrophotometer (Applied Photophysics) upon excitation at 425 nm at 20 °C.

2.5. G4 motifs in the genome

RefSeq genomic sequences (http://www.ncbi.nlm.nih.gov/refseq/rsg/about/) were used for the analysis of GQ/imGQ abundance and distribution in the human genome. The sequences were obtained from the UCSC database (table hg19.knownGene) and referenced on zero-based GRCh37.p13 (UCSC version, hg19). The frequencies of non-overlapping GQ/ImGQ motifs in the genomic regions were calculated using option 2 in ImGQfinder (http://imgqfinder.niifhm.ru). Each G-rich fragment containing 4 to 7 G-blocks (uninterrupted or interrupted by non-G-rich fragments not exceeding the maximum loop length) was counted as one putative GQ/ImGQ, each fragment containing 8 to 11 G-blocks was counted as two putative GQs/ImGQs, etc. Fragments with both GQ and ImGQ-forming potential were counted as putative GQs because the latter are generally more stable.

To account for the impact of the G/C content, non-G4 fragments with 70–80% G/C-content (a number equal to the total number of putative GQs/imGQs in the whole genome) were randomly selected from 10 million 33-nt random chromosome fragments. This process was repeated 10 times. The profiles of the non-G4 G-rich fragment distribution relative to TSS/TTS sites and exon/intron boundaries were obtained and superimposed onto the respective profiles of putative GQ/imGQ motifs. Ten 'negative control' profiles correspond to ten repeats.

2.6. Functional enrichment analysis

Genes containing 'imperfect' G4 motifs at the exon/intron boundaries were analyzed for the common molecular function, related biological process and cellular localization of the respective proteins. The enrichment analysis was performed using the Bingo application of Cytoscape 3.2.0 software with Bonferroni's family-wise error rate (FWER) correction and the GeneOnthology database as a reference. The P-value threshold (after the FWER correction) was 0.05.

3. Results and discussion

3.1. Imperfect G4 structures: general description, examples and thermal stability studies

Hereafter, we will exclusively use the abbreviation 'GQ' for perfect G4 structures. Intramolecular GQ motifs contain four uninterrupted (G_{3+}) G-runs and comply with the classical formula (except the maximum loop length may exceed the canonical limit of 7 nucleotides). The abbreviation 'imGQ' stands for imperfect G4s, i.e., those containing at least one interrupted or truncated G-run. ImGQ motifs defy the classical formula. Intramolecular imGQs with single defects (mismatches, bulges or vacancies) comply with the formulas presented in Table 1.

Apparently, some imGQ-prone sequences can also fold into canonical GQs with fewer tetrads, e.g., a putative 3-tetrad imGQ with a mismatch or vacancy in the external tetrad may actually adopt a 2-tetrad GQ conformation. Moreover, the distinction between the sequence rules for various types of defects is illusive. ImGQs with bulges and imGQs with mismatches are defined by substantially overlapping motif sets.

Any G-run-interrupting nucleotide (N) can either occupy a position in a respective tetrad as a G-substitute (a mismatch) or project from the G4 core between the tetrads (a bulge). Theoretically, a G-run-interrupting nucleotide could also project from the tetrad plane, leaving one unoccupied position in the core (a vacancy). However, the latter case has not been experimentally confirmed thus far. G-vacancies have only been observed in external (terminal) tetrads [11]. The molecular modeling experiments conducted herein, which will be discussed below in detail, also suggest that imGQs with internal-tetrad vacancies are prone to rearrangements. Unlike vacancies, mismatches in internal tetrads may be tolerated because the mismatching nucleic base can participate in stacking with the neighboring tetrads or even form H-bonds within its own tetrad [6].

Examples of previously unreported putative imGQs from the human genome (NCBI Reference Sequence: NC_000018.9) are shown in Fig. 1. They were chosen randomly from the fragments of chromosome 18 that comply with the formulas in Table 1 ($n = 4$). Sequences PSTP and Ct1 are located in the intron of the CTIF gene (chr18: +46379322 to +46379344) and at the PSTPIP2 intron/exon boundary (chr18: −43572049 to −43572072), respectively. According to the conventional G4 motif description, these motifs can only form 2-tetrad GQs and would be skipped upon genome mining using the classical G4-predicting software. However, CD and 1H NMR spectra of these ONs bear the specific signatures of G4 structures.

Additional examples of imGQ-forming ONs and their GQ-forming analogs are presented in Table 2. Ct2-Ct4, CtA, CtC and CtG are Ct1 mutants. Bcl is a fragment of the BCL2 promoter region (chr18: −60985942 to −60985966). BclT, BclA and BclG are Bcl mutants. G3, G4, and their derivatives are model ONs. For UV-melting curves, thermal difference spectra (TDS, [30]), CD spectra and inter- *vs.* intra-molecular folding analysis of the GQ/imGQ ONs, see Figs. 1 and 2 in Ref. [31]. The MALDI-TOF MS data are provided in Table 1 in Ref. [31]. As evident from Table 2, all the tested intramolecular imGQs are thermodynamically stable under physiological conditions.

ON Bcl is referred to as a putative GQ in Table 2 because its sequence complies with the classical formula for 3-tetrad G4s. However, Bcl could also adopt a 4-tetrad imGQ structure. We were unable to clarify Bcl folding based on NMR data as a result of signal overlapping (the G4 signature in the imino region is present, but conformational polymorphism or oligomerization at high concentrations hamper spectra interpretation, Fig. S1).

14 Imino proton signals are present in the NMR spectra of Ct1 and PSTP (Fig. 1), indicating the formation of imGQs with three perfect G-tetrads and one mismatched tetrad or G-triad. Ts that interrupt the 5'-terminal G-runs either bulge or stack with the neighboring Gs. The 12 signals in the imino proton region of the ribo-Ct1 spectrum suggest a 3-tetrad G4 structure (T is bulging). The strand orientations in the schematic representations were ascribed based on CD data: large positive bands at 265 nm and smaller negative bands at 240 nm suggest parallel G4-folding. A minor band at 295 nm in the spectrum of deoxy-Ct1 can be attributed to the presence of antiparallel or hybrid G4 admixtures [32]. Another possible explanation is an unusual parallel G4 structure comprising three similar-polarity tetrads and one external tetrad of an opposite polarity (The CD spectrum of Ct1 resembles that of a known unusual structure from the RET promoter, which also defies the "all-anti" conformational rule for parallel G4s [33]).

3.2. ImGQ core dynamics (molecular modeling studies)

To obtain some insight into imGQ core dynamics and the favorable positions of the G-run-interrupting nucleotides, we performed molecular modeling of Ct1. A schematic representation of the 50-ns molecular dynamics (MD) simulation results is shown in Fig. 2. Smoothed MD trajectories are provided as supplementary movies SM1 and SM2. Two initial models of the imGQ with a

Table 1
ImGQ motifs with single defects. L denotes any nucleotide in a loop; N denotes any mismatching or bulging nucleotide (if N = G, the formula describes a GQ motif); n is the number of tetrads (n ≥ 3); Xj is the number of nucleotides in a loop; j is a loop index; and i is the position of a defect in the G-run (1 ≤ i ≤ n for bulges; 1 < i < n for mismatches, and i = 1 or n for vacancies).

Defect position	Defect type	
	Bulge	Mismatch or vacancy
5′-terminal G-run	$G_{i-1}NG_{n-i+1}(L_{Xj}G_n)_3$, j = 1,2,3	$G_{i-1}NG_{n-i}(L_{Xj}G_n)_3$, j = 1,2,3
Middle-strand G-runs	$G_nL_{X1}G_{i-1}NG_{n-i+1}(L_{Xj}G_n)_2$, j = 2,3 $(G_nL_{X1})_2G_{i-1}NG_{n-i+1}L_{X3}G_n$, j = 1,2	$G_nL_{X1}G_{i-1}NG_{n-i}(L_{Xj}G_n)_2$, j = 2,3 $(G_nL_{X1})_2G_{i-1}NG_{n-i}L_{X3}G_n$, j = 1,2
3′-terminal G-run	$(G_nL_{Xj})_3G_{i-1}NG_{n-i+1}$, j = 1,2,3	$(G_nL_{Xj})_3G_{i-1}NG_{n-i}$, j = 1,2,3

mismatch (top scheme in Fig. 1) were obtained using available XDR data for well-characterized structurally relevant GQs [34,35]. Model 1 is a typical parallel G4 with all core Gs in the anti-conformation. Model 2 is a parallel G4 with syn-Gs in the external tetrad. These quadruplex structures have slightly different twist angles. The G-run-interrupting T was initially stacked with Gs from the neighboring tetrads in both models but did not participate in H-bonding.

Supplementary video related to this article can be found at http://dx.doi.org/10.1016/j.biochi.2017.01.003.

The core of the model 1 ImGQ (Fig. 2, left panel) was stable throughout the simulation (although moderate fluctuations in the external tetrad resulted in the loss of one potassium ion at approximately 1 ns), and T participated in H-bonding after approximately 2 ns (see Fig. S2A for H-bonding time plots). As evident from the 2.5-ns snapshot (Fig. 2, left panel, top view of the green tetrad; other tetrads are hidden for clear representation), O4 and H3 of Thy-3 formed H-bonds with H1 of Gua-22 and O6 of Gua-10, respectively, yielding 4 imino protons involved in base pairing in the mismatched tetrad and 16 imino protons in the whole imGQ. The modeling results contradict the 1H NMR data (14 signals in the imino region of the Ct1 spectrum, Fig. 1).

Better consistency with the NMR data was achieved using Ct1 model 2 (Fig. 2, right panel). A slightly different initial geometry generated certain freedom of Thy-3, and instead of forming H-bonds within the tetrad this nucleobase deviated from the tetrad plane. Interestingly, the obtained vacancy was almost immediately filled with the neighboring base from the external tetrad (Gua-4). The latter process is illustrated as a time plot of Gua-4 H-bonding with Gua-22 and Gua-10 (Fig. S2B). Free energy estimations for models 1 and 2 produced comparable results (Fig. S2C).

Thus, the MD data suggest that a mismatching base in the internal tetrad of a single-defect imGQ either forms H-bonds within the tetrad or bulges out to initiate "shifting" of the defect to the external tetrad.

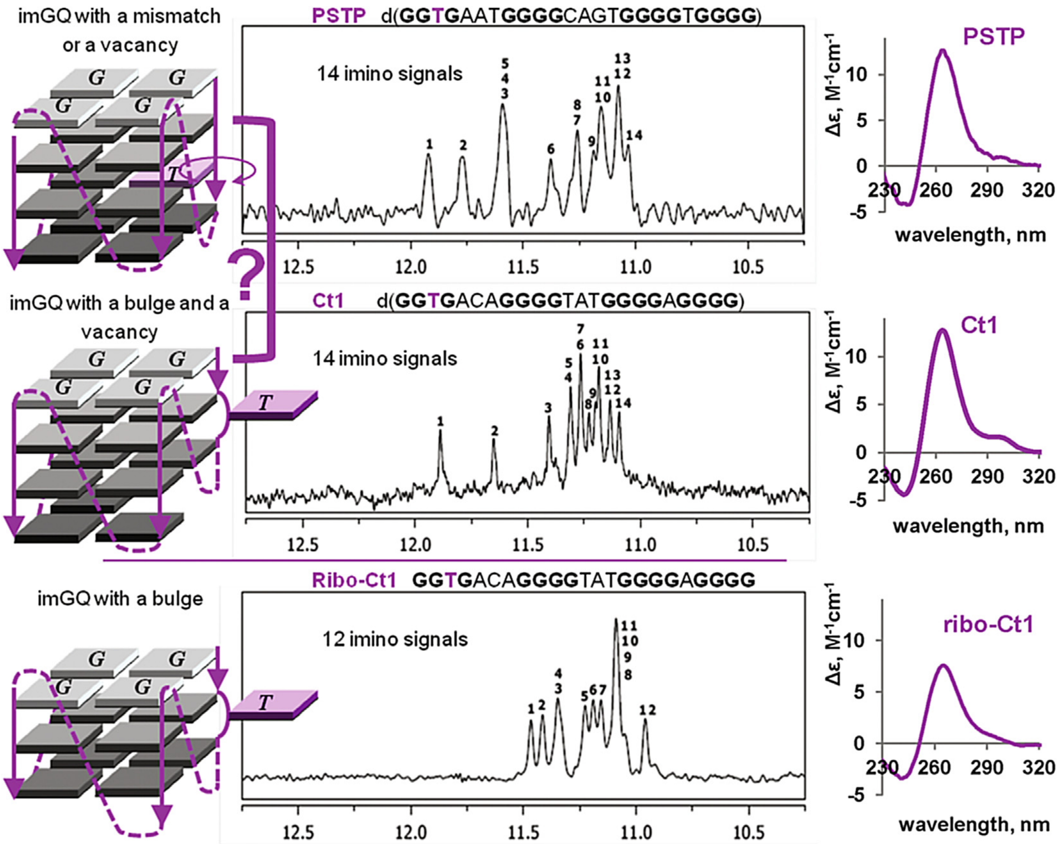

Fig. 1. Examples of putative imGQs from the human genome. Schematic representations (left panel), fragments of 1H NMR spectra (middle panel) and CD spectra (right panel). Buffer conditions: 25 mM Tris-HCl (pH 7.5) and 100 mM KCl. The ON concentration was 5 μM in CD experiments and ~0.1 mM in the NMR experiments.

Table 2
Sequences of G4-forming ONs and characteristics of their solution structures.

ON code	Sequence, 5′→3′	G4 motif type	G4 T_m^{295} values, °C±1			G4 type[a]
			100 mM KCl	10 mM KCl	100 mM LiCl	
PSTP	***GGTG***AATGGGGCAGTGGGGTGGGG	ImGQ (1defect)	75	61	no intramol.G4s	p
Ct1	***GGTG***ACAGGGGTATGGGGAGGGG		74	57		p
Ct2	GGGGACA***GGTG***TATGGGGAGGGG		56	40		h
Ct3	GGGGACAGGGGTAT***GGTG***AGGGG		62	44		h
Ct4	GGGGACAGGGGTATGGGGA***GGTG***		72	44		h
CtA	***GGAG***ACAGGGGTATGGGGAGGGG		70	56		p
CtC	***GGCG***ACAGGGGTATGGGGAGGGG		intermolecular G4			p
CtG	GGGGACAGGGGTATGGGGAGGGG	GQ	>85	69		h
Bcl	GGGGGCCGTGGGGT***GGGAG***CTGGGG		77	62		p
BclT	GGGGGCCGTGGGGT***GTGAG***CTGGGG	imGQ (1 or 2 defects)	52	46		h
BclA	GGGGGCCGTGGGGT***GAGAG***CTGGGG		intermolecular G4			h
BclG	GGGGGCCGTGGGGTGGGGGCTGGGG	GQ	>85	>80	44	h
G4	GGGGTGGGGTGGGGTGGGG		>85	79	57	p
G3	GGGTGGGTGGGTGGG		>90	>85	45	p
G3A	GGGT***GAG***TGGGTGGG	ImGQ (1defect)	intermolecular G4			p
G4A	GGGGT***GAGG***TGGGGTGGGG		>85	75	50	p
G4AA	GGGGT***GAGA***TGGGGTGGGG	imGQ (1 or 2 defects)	>80	57	48	p

Interrupted G3 and G4 runs are in Italics and bold.

[a] p = parallel GQ/imGQ (the CD spectrum contains a positive band at about 265 nm and a negative band at about 245 nm), h = hybrid GQ/imGQ (the CD spectrum contains a positive band at about 290 nm in addition to the bands at 265 nm and 240 nm).

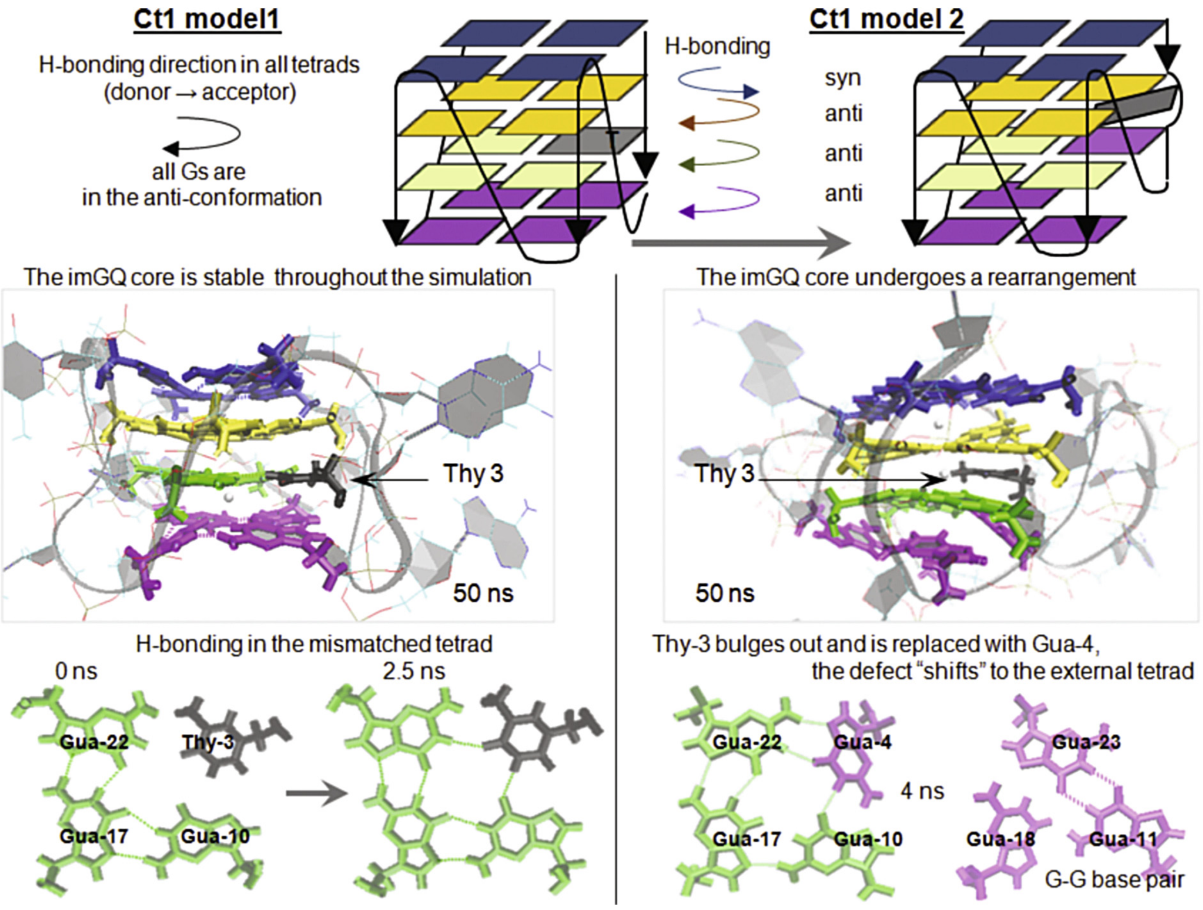
Ct1 model1
H-bonding direction in all tetrads
(donor → acceptor)
all Gs are
in the anti-conformation
H-bonding
syn
anti
anti
anti
Ct1 model 2
The imGQ core is stable throughout the simulation
Thy 3
50 ns
The imGQ core undergoes a rearrangement
Thy 3
50 ns
H-bonding in the mismatched tetrad
0 ns
2.5 ns
Gua-22
Thy-3
Gua-17
Gua-10
Thy-3 bulges out and is replaced with Gua-4,
the defect "shifts" to the external tetrad
Gua-22
Gua-4
4 ns
Gua-17
Gua-10
Gua-23
Gua-18
Gua-11
G-G base pair

Fig. 2. Ct1 modeling and MD simulations.

3.3. ImGQ interactions with ligands

We analyzed imGQ binding with several well-characterized G4-recognizing small molecules to clarify whether imGQs can be stabilized and visualized using known ligands and fluorescent probes, respectively. The studies of GQ and imGQ interactions with pyridostatin (PDS), N-methylmesoporphyrin IX (NMM) and thioflavin T (ThT) are summarized in Fig. 3. For more data, see Figs. 3–5 in Ref. [31].

PDS is a well-known ligand that stabilizes G4s with good

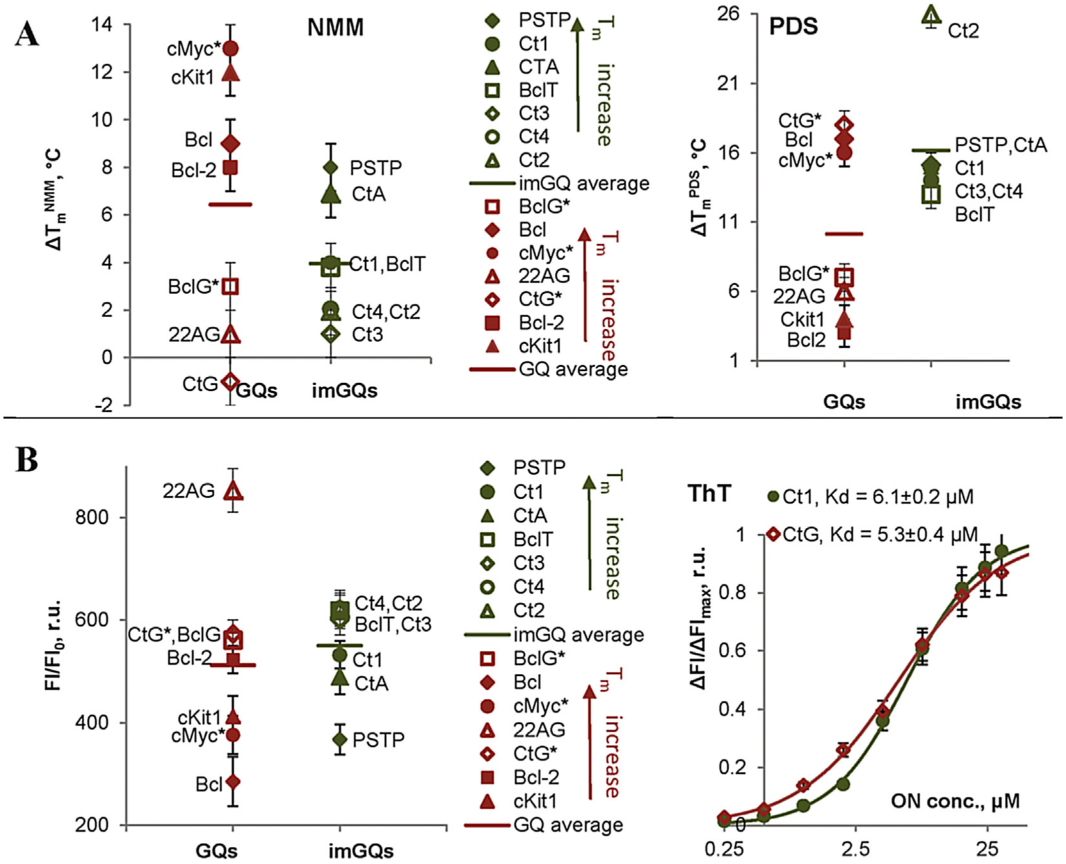

Fig. 3. GQ and imGQ interactions with ligands. A: Effects of NMM and PDS on G4 melting temperatures, $\Delta Tm^{ligand} = Tm_{G4}^{with\ ligand} - Tm_{G4}^{without\ ligand}$. Conditions: 1.5 μM G4 and 3 μM NMM/PDS in buffer 1 (20 mM Tris-HCl (pH 7.6), 10 mM KCl) unless otherwise specified (* = 1 mM KCl instead of 10 mM KCl). imGQs, red; GQs, green; parallel or mostly parallel G4s, filled markers; hybrid G4s, hollow markers. B: ThT interaction with G4s. Left graph: relative fluorescence intensities (FI) at the ThT emission maxima (490 nm) in the presence of G4s (425 nm excitation) at 20 °C. FI_0 = ThT fluorescence intensity in the absence of ONs. Conditions: 1.5 μM G4 and 1.5 μM ThT in buffer 1. Right graph: Ct1 (imGQ) and Ctg (GQ) binding with ThT in buffer 1 was monitored by ThT fluorescence at 490 nm. The ThT concentration was 1.25 μM. All florescence measurements and melting experiments were performed in duplicate.

selectivity relative to double-stranded DNA [36]. This ligand has been used as a platform for developing matrices for selective G4 isolation [37]. Its ability to promote DNA double strand breaks in vivo (presumably, by causing polymerase stalling during transcription and replication [38]) has enabled G4 detection and mapping in genomic DNA. PDS-induced stalling of a reverse transcriptase has been exploited to detect and map RNA G4s in cellular transcripts [39]. NMM, another popular ligand with remarkable specificity for G4s [40], stabilizes predominantly parallel structures. NMM fluorescence intensity substantially increases in the presence of parallel G4s, whereas antiparallel G4s cause only a moderate increase. Thus, NMM could be used as a reporter of strand orientation [41]. However, its ability to induce the rearrangement of antiparallel quadruplexes into parallel quadruplexes [40] raises concerns for the reliability of this strand orientation assignment. ThT is a fluorogenic dye that exhibits striking fluorescence upon binding with the G4 DNA [42,43] and G4 RNA [44], representing a convenient fluorescent sensor for G4s.

The imGQs evaluated in the present study are described in Table 2. A set of GQs from Table 2 was complemented with ‘control’ GQs from oncogene promoters (Bcl-2, cMyc and cKit) and the telomeric GQ 22AG (see Table 1 in Ref. [31] for sequences and MS data). We aimed to obtain statistically comparable GQ and imGQ sets, each including G4s of different topologies and thermal stabilities (40 °C ≤ $T_m^{GQ/imGQ}$ ≤ 70 °C; T_m dispersions: $\sigma^2(T_m^{GQ}) = 65$, $\sigma^2(T_m^{imGQ}) = 94$, $T_m^{GQ}{}_{average} = 55$ °C, $T_m^{imGQ}{}_{average} = 50$ °C).

NMM stabilized parallel imGQs (filled markers in Fig. 3) more efficiently than the hybrid imGQs (hollow markers in Fig. 3), consistent with the previously reported data for GQs [40]. Average NMM-induced change of GQ T_m ($\Delta T_m^{NMM} = 6 \pm 1$ °C) was slightly higher than that of imGQ T_m ($\Delta T_m^{NMM} = 4 \pm 1$ °C; Fig. 3A, left graph). PDS more efficiently stabilizes imGQs (Fig. 3A, right graph); however, ΔT_m^{PDS} dispersions in the GQ and imGQ sets apparently differ too much, thus Student's t-test cannot be performed. The average ThT fluorescence intensities in complexes with GQs and imGQ are relatively close (Fig. 3B, left graph; $FI/FI_0^{imGQ} > FI/FI_0^{GQ}$, but the differences are insignificant according to Student's t-test, assuming normal distributions). The dissociation constants (Kd) were estimated only for Ct1-ThT and CtG-ThT complexes (Fig. 3B, right graph). Both GQ and imGQ bound to ThT in the low-micromolar concentration range; the Kd values were almost equal.

Thus, these data indicate that imGQs can be efficiently visualized using ThT and are generally at least as responsive to NMM and PDS as classical G4s. Should ImGQs exist in vivo and behave similarly to GQs, their interactions with GQ-targeted therapeutics might account for some of the side effects of these therapies. This hypothesis prompted us to develop ImGQ-predicting software for the analysis of genomic sequences.

3.4. ImGQ motifs in the human genome

To evaluate the genomic abundance and investigate the

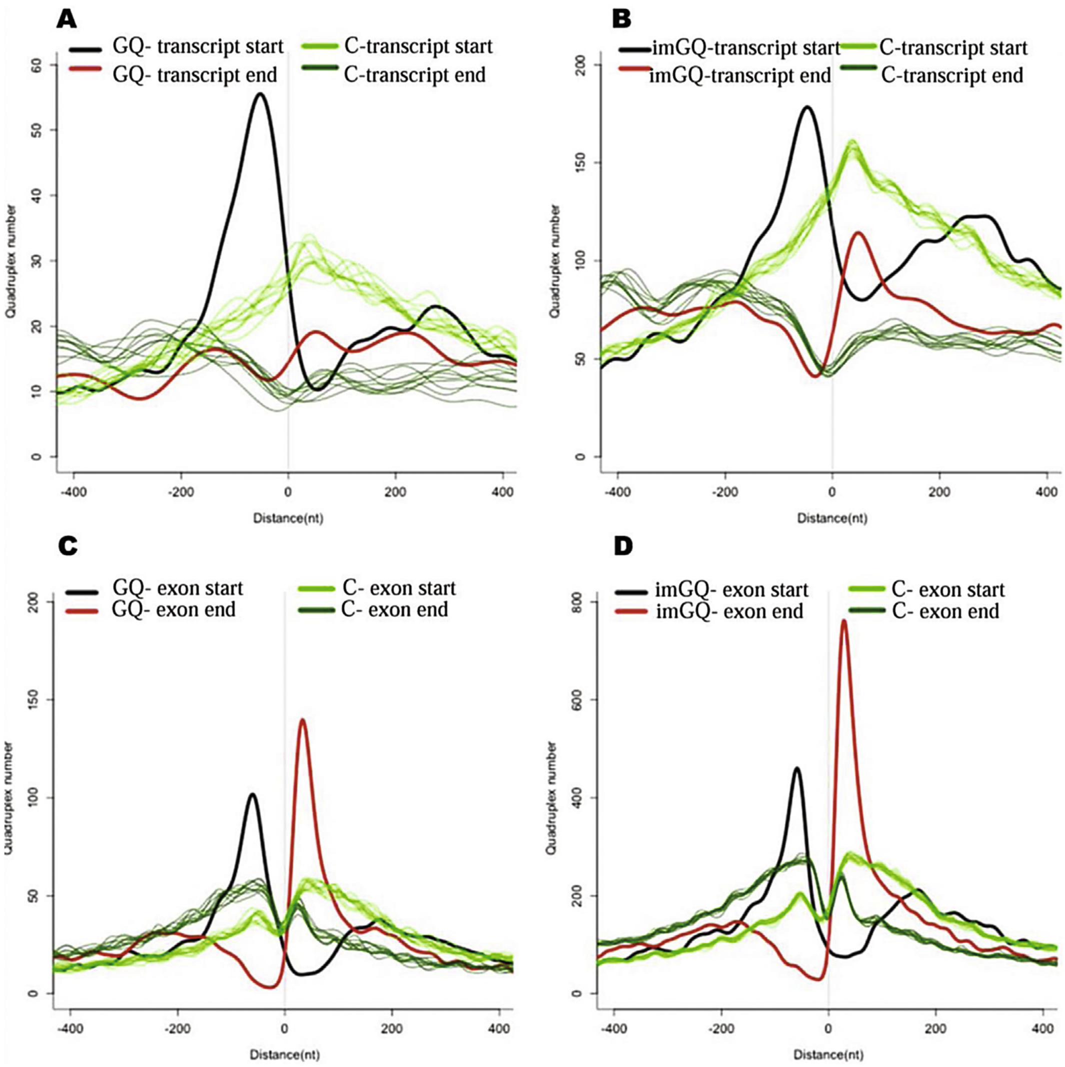

Fig. 4. GQ/imGQ motif distribution within genes. A and B: GQ and imGQ motif frequencies, respectively, in the proximity of RefSeq TSS/TTS. C and D: GQ and imGQ motif frequencies, respectively, in the proximity of RefSeq exon/intron boundaries. C^- is a negative control (normalized frequencies of G-rich non-G4 sequences). G-rich non-G4 fragments are significantly more frequent in the genome than G4 motifs. Therefore, the data were normalized to equalize the whole-genome integral of the negative control with that of GQs or imGQs, respectively.

genomic localization of imGQ motifs, we developed a simple online search tool, ImGQfinder, which is freely accessible at the URL http://imgqfinder.niifhm.ru. The program is implemented in Perl. The underlying algorithm is based on the set of formulas presented in Table 1. In this algorithm, we searched for G-runs, determined the distance between them and selected fragments that complied with the predetermined conditions for the maximum loop length and the minimum number of nucleotides in a G-run (i.e., the number of tetrads). ImGQfinder searches for all GQ and imGQ motifs in a given sequence. The input parameters include the queried nucleotide sequence in raw or FASTA format, the number of tetrads and defects and the maximum loop length. The hits are displayed in a table. Two options are available:

1) Identify all putative GQs or imGQs, including overlapping ones;
2) Calculate the maximum number of non-overlapping GQs/ImGQs.

Option 1 is recommended for detailed conformational analyses of relatively short sequences. The results include a list of all putative GQs/imGQs with coordinates for each G-run start, GQ/imGQ sequence (G-runs are highlighted in yellow) and position of the defects.

Option 2 is recommended for statistical analyses of lengthy sequences. The results include a list of all non-overlapping G-rich fragments with GQ/imGQ-forming potential (the fragments may contain more than four G-runs, which implies various folding modes) separated by non-G-rich fragments that exceed the maximum loop length. The coordinates of the fragments, their sequences (possible GQs/imGQs are highlighted) and the maximum number of non-overlapping GQs/imGQs in each uninterrupted G-rich fragment are shown. The maximum number of non-overlapping GQ/imGQs that can exist simultaneously in the queried sequence is also shown.

ImGQfinder was utilized to reassess the number of G4-prone sites in the human genome. First, to validate the algorithm, we calculated the number of all non-overlapping 3-tetrad GQ motifs (option 2 in ImGQfinder, number of defects = 0) and compared these results to the data in the literature. The obtained value (359 k) is consistent with previous estimations [45]. Next, we considered 4-

tetrad GQs and imGQs. As expected, imGQ motifs are substantially more abundant than GQ motifs (approximately 5 times in the case of 4-tetrad structures, Table S1). If we assume that only 3^+-tetrad GQs and 4^+-tetrad imGQs are generally sufficiently stable to exist in vivo, then accounting for imGQs adds approximately 30% to the total number of putative G4 structures.

GQ and imGQ motifs have basically similar distributions within RefSeq genes. Exons generally tend to be depleted of both GQs and imGQs, except for certain enrichment in the first exons (Fig. S3). Large clusters of GQ/imGQ motifs were detected in promoters 50–150 nt downstream of the transcription start site (TSS), in the 3′-UTRs 50–100 bp upstream of the transcription termination sites (TTS) and in the introns near exon/intron boundaries (Fig. 4). These clusters do not exclusively reflect G-richness (the 'negative control' profiles, obtained for non-G4 G-rich sequences do not coincide with the GQ/imGQ profiles). These findings are consistent with the previously reported GQ motif distribution studies [46,47] and suggest imGQ participation in transcription, translation or splicing regulation.

The ratio of G4 and negative control amplitudes at the GQ/imGQ profile extreme provides a rough estimation of the impact of the putative secondary structure. In the promoters, the G4:control ratio is apparently higher for GQs than for imGQs (the imGQ motif clusterization is less pronounced). Near the exon/intron boundaries, the imGQ:control and GQ:control ratios are similar.

It has previously been hypothesized that certain types of imGQs (G4 with vacancies) may have a distinct biological role [11], and a correlation between gene function and the potential for G4 formation has been reported [48]. We performed a functional enrichment analysis of the genes containing 4^+-tetrad imGQ motifs near exon/intron boundaries (the areas where imGQ motif distribution is clearly non-random). The analysis revealed a statistically significant enrichment of genes encoding regulators of signal transduction, ion transport and neurogenesis (Table S2). Although further studies are required to support the biological significance of imGQ sites, it is tempting to speculate that these sites may contribute to the fine-tuning of the expression of signaling factors and other regulatory proteins.

Thus, a genome-wide analysis of imGQ motifs was performed using the imGQfinder program. We propose that this simple imGQ search tool could be helpful in fundamental studies and could be used in combination with the existing second-generation GQ-predicting tools.

3.5. Conclusions

We studied the thermal stabilities and conformational polymorphisms of imGQs under physiological conditions. We showed that quadruplexes with single defects are generally thermodynamically stable, interact with the commonly used GQ-targeted ligands and can be efficiently visualized using the popular fluorescent probe ThT. We developed a search tool that accounts for several major types of defects in G4 structures and analyzed the distribution of the respective motifs in the human genome. Consistent with the results of G4-sequencing experiments, imperfect G4 motifs were abundant and clustered in the same regulatory regions as classical putative quadruplex sites.

Funding

This work was supported by the Russian Science Foundation [14-25-00013] (the computational part and secondary structure verification by optical methods]; Russian Foundation for Basic Research [14-04-01244] (molecular modeling) and the Dynasty Foundation (interaction with ligands).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.biochi.2017.01.003.

References

[1] S. Burge, G.N. Parkinson, P. Hazel, A.K. Todd, S. Neidle, Quadruplex DNA: sequence, topology and structure, Nucleic Acids Res. 34 (2006) 5402–5415.

[2] W.J. Chung, B. Heddi, E. Schmitt, K.W. Lim, Y. Mechulam, A.T. Phan, Structure of a left-handed DNA G-quadruplex, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 2729–2733.

[3] G. Oliviero, J. Amato, N. Borbone, A. Galeone, M. Varra, G. Piccialli, L. Mayol, Synthesis and characterization of DNA quadruplexes containing T-tetrads formed by bunch-oligonucleotides, Biopolymers 81 (2006) 194–201.

[4] M.S. Searle, H.E. Williams, C.T. Gallagher, R.J. Grant, M.F. Stevens, Structure and K+ ion-dependent stability of a parallel-stranded DNA quadruplex containing a core A-tetrad, Org. Biomol. Chem. 2 (2004) 810–812.

[5] N.S. Bhavesh, P.K. Patel, S. Karthikeyan, R.V. Hosur, Distinctive features in the structure and dynamics of the DNA repeat sequence GGCGGG, Biochem. Biophys. Res. Commun. 317 (2004) 625–633.

[6] J.D. Wen, D.M. Gray, The Ff gene 5 single-stranded DNA-binding protein binds to the transiently folded form of an intramolecular G-quadruplex, Biochemistry 41 (2002) 11438–11448.

[7] M. Webba da Silva, Association of DNA quadruplexes through G: C:G: C tetrads. Solution structure of d(GCGGTGGAT), Biochemistry 42 (2003) 14356–14365.

[8] J. Viladoms, N. Escaja, M. Frieden, I. Gomez-Pinto, E. Pedroso, C. Gonzalez, Self-association of short DNA loops through minor groove C: G:G: C tetrads, Nucleic Acids Res. 37 (2009) 3264–3275.

[9] P.L. Tran, A. De Cian, J. Gros, R. Moriyama, J.L. Mergny, Tetramolecular quadruplex stability and assembly, Top. Curr. Chem. 330 (2013) 243–273.

[10] V.T. Mukundan, A.T. Phan, Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences, J. Am. Chem. Soc. 135 (2013) 5017–5028.

[11] X.M. Li, K.W. Zheng, J.Y. Zhang, H.H. Liu, Y.D. He, B.F. Yuan, Y.H. Hao, Z. Tan, Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 14581–14586.

[12] B. Heddi, N. Martin-Pintado, Z. Serimbetov, T.M. Kari, A.T. Phan, G-quadruplexes with (4n - 1) guanines in the G-tetrad core: formation of a G-triad.water complex and implication for small-molecule binding, Nucleic Acids Res. 44 (2016) 910–916.

[13] M.L. Bochman, K. Paeschke, V.A. Zakian, DNA secondary structures: stability and function of G-quadruplex structures, Nat. Rev. Genet. 13 (2012) 770–780.

[14] P. Murat, S. Balasubramanian, Existence and consequences of G-quadruplex structures in DNA, Curr. Opin. Genet. Dev. 25 (2014) 22–29.

[15] R. Simone, P. Fratta, S. Neidle, G.N. Parkinson, A.M. Isaacs, G-quadruplexes: emerging roles in neurodegenerative diseases and the non-coding transcriptome, FEBS Lett. 589 (2015) 1653–1668.

[16] A. Varizhuk, N. Ilyinsky, I. Smirnov, G. Pozmogova, G4 aptamers: trends in structural design, Mini Rev. Med. Chem. (2016) 1321–1329.

[17] M. Prokofjeva, V. Tsvetkov, D. Basmanov, A. Varizhuk, M. Lagarkova, I. Smirnov, K. Prusakov, D. Klinov, V. Prassolov, G. Pozmogova, S.N. Mikhailov, Anti-hiv activities of intramolecular G4 and non-G4 oligonucleotides, Nucleic Acid. Ther. (2016), http://dx.doi.org/10.1089/nat.2016.0624.

[18] A. Piazza, M. Adrian, F. Samazan, B. Heddi, F. Hamon, A. Serero, J. Lopes, M.P. Teulade-Fichou, A.T. Phan, A. Nicolas, Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites, EMBO J. 34 (2015) 1718–1734.

[19] A. Guedin, A. De Cian, J. Gros, L. Lacroix, J.L. Mergny, Sequence effects in single-base loops for quadruplexes, Biochimie 90 (2008) 686–696.

[20] M. Kim, A. Kreig, C.Y. Lee, H.T. Rube, J. Calvert, J.S. Song, S. Myong, Quantitative analysis and prediction of G-quadruplex forming sequences in double-stranded DNA, Nucleic Acids Res. 44 (2016) 4807–4817.

[21] O. Stegle, L. Payet, J.L. Mergny, D.J. MacKay, J.H. Leon, Predicting and understanding the stability of G-quadruplexes, Bioinformatics 25 (2009) i374–382.

[22] J.D. Beaudoin, R. Jodoin, J.P. Perreault, New scoring system to identify RNA G-quadruplex folding, Nucleic Acids Res. 42 (2014) 1209–1223.

[23] A. Bedrat, L. Lacroix, J.L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, Nucleic Acids Res. 44 (2016) 1746–1759.

[24] V.S. Chambers, G. Marsico, J.M. Boutell, M. Di Antonio, G.P. Smith, S. Balasubramanian, High-throughput sequencing of DNA G-quadruplex structures in the human genome, Nat. Biotechnol. 33 (2015) 877–881.

[25] A.M. Varizhuk, V.B. Tsvetkov, O.N. Tatarinova, D.N. Kaluzhny, V.L. Florentiev, E.N. Timofeev, A.K. Shchyolkina, O.F. Borisova, I.P. Smirnov, S.L. Grokhovsky, A.V. Aseychev, G.E. Pozmogova, Synthesis, characterization and in vitro activity of thrombin-binding DNA aptamers with triazole internucleotide linkages, Eur. J. Med. Chem. 67C (2013) 90–97.

[26] A. Onufriev, D. Bashford, D.A. Case, Modification of the generalized Born model suitable for macromolecules, J. Phys. Chem. B 104 (2000) 3712–3720.

[27] A. Onufriev, D.A. Case, D. Bashford, Effective Born radii in the generalized Born approximation: the importance of being perfect, J. Comput. Chem. 23 (2002) 1297–1304.

[28] J. Weiser, P.S. Shenkin, W.C. Still, Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO), J. Comput. Chem. 20 (1999) 217–230.

[29] J. Srinivasan, J. Miller, P.A. Kollman, D.A. Case, Continuum solvent studies of the stability of RNA hairpin loops and helices, J. Biomol. Struct. Dyn. 16 (1998) 671–682.

[30] J.L. Mergny, J. Li, L. Lacroix, S. Amrane, J.B. Chaires, Thermal difference spectra: a specific signature for nucleic acid structures, Nucleic Acids Res. 33 (2005) e138.

[31] A. Varizhuk, M. Vlasenok, D. Kaluzhny, I. Smirnov, G. Pozmogova, Data on secondary structures and ligand interactions of G-rich oligonucleotides that defy the classical formula for G4 motifs, Data in Brief, 2017 submitted.

[32] D.M. Gray, J.D. Wen, C.W. Gray, R. Repges, C. Repges, G. Raabe, J. Fleischhauer, Measured and calculated CD spectra of G-quartets stacked with the same or opposite polarities, Chirality 20 (2008) 431–440.

[33] X. Tong, W. Lan, X. Zhang, H. Wu, M. Liu, C. Cao, Solution structure of all parallel G-quadruplex formed by the oncogene RET promoter sequence, Nucleic Acids Res. 39 (2011) 6753–6763.

[34] Y. Wang, D.J. Patel, Solution structure of a parallel-stranded G-quadruplex DNA, J. Mol. Biol. 234 (1993) 1171–1183.

[35] S.T. Hsu, P. Varnai, A. Bugaut, A.P. Reszka, S. Neidle, S. Balasubramanian, A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics, J. Am. Chem. Soc. 131 (2009) 13399–13409.

[36] R. Rodriguez, S. Muller, J.A. Yeoman, C. Trentesaux, J.F. Riou, S. Balasubramanian, A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres, J. Am. Chem. Soc. 130 (2008) 15758–15759.

[37] S. Muller, S. Kumari, R. Rodriguez, S. Balasubramanian, Small-molecule-mediated G-quadruplex isolation from human cells, Nat. Chem. 2 (2010) 1095–1098.

[38] M. Di Antonio, R. Rodriguez, S. Balasubramanian, Experimental approaches to identify cellular G-quadruplex structures and functions, Methods 57 (2012) 84–92.

[39] C.K. Kwok, S. Balasubramanian, Targeted detection of g-quadruplexes in cellular RNAs, Angew. Chem. Int. Ed. Engl. 54 (2015) 6751–6754.

[40] J.M. Nicoludis, S.P. Barrett, J.L. Mergny, L.A. Yatsunyk, Interaction of human telomeric DNA with N-methyl mesoporphyrin IX, Nucleic Acids Res. 40 (2012) 5432–5447.

[41] N.C. Sabharwal, V. Savikhin, J.R. Turek-Herman, J.M. Nicoludis, V.A. Szalai, L.A. Yatsunyk, N-methylmesoporphyrin IX fluorescence as a reporter of strand orientation in guanine quadruplexes, FEBS J. 281 (2014) 1726–1737.

[42] V. Gabelica, R. Maeda, T. Fujimoto, H. Yaku, T. Murashima, N. Sugimoto, D. Miyoshi, Multiple and cooperative binding of fluorescence light-up probe thioflavin T with human telomere DNA G-quadruplex, Biochemistry 52 (2013) 5620–5628.

[43] A. Renaud de la Faverie, A. Guedin, A. Bedrat, L.A. Yatsunyk, J.L. Mergny, Thioflavin T as a fluorescence light-up probe for G4 formation, Nucleic Acids Res. 42 (2014) e65.

[44] S. Xu, Q. Li, J. Xiang, Q. Yang, H. Sun, A. Guan, L. Wang, Y. Liu, L. Yu, Y. Shi, H. Chen, Y. Tang, Thioflavin T as an efficient fluorescence sensor for selective recognition of RNA G-quadruplexes, Sci. Rep. 6 (2016) 24793.

[45] J.L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome, Nucleic Acids Res. 33 (2005) 2908–2916.

[46] J.L. Huppert, A. Bugaut, S. Kumari, S. Balasubramanian, G-quadruplexes: the beginning and end of UTRs, Nucleic Acids Res. 36 (2008) 6260–6268.

[47] J. Eddy, N. Maizels, Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes, Nucleic Acids Res. 36 (2008) 1321–1333.

[48] J. Eddy, N. Maizels, Gene function correlates with potential for G4 DNA formation in the human genome, Nucleic Acids Res. 34 (2006) 3887–3896.