

Flexibility and structural conservation in a c-KIT G-quadruplex

Dengguo Wei[†], Jarmila Husby[†] and Stephen Neidle^{*}

UCL School of Pharmacy, University College London, London WC1N 1AX, UK

Received October 14, 2014; Revised November 21, 2014; Accepted November 21, 2014

ABSTRACT

A quadruplex sequence from the promoter region of the c-KIT gene forms a stable quadruplex, as characterized by crystallographic and NMR methods. Two new crystal structures are reported here, together with molecular dynamics simulation studies on these quadruplex crystal structures and an NMR structure. The new crystal structures, each in a distinct space group and lattice packing arrangement, together with the existing structures, demonstrate that the c-KIT quadruplex fold does not change with differing environments, suggesting that quadruplex topological dynamism is not a general phenomenon. The single and dinucleotide loops in these structures show a high degree of conformational flexibility within the three crystal forms and the NMR ensemble, with no evidence of clustering to particular conformers. This is in accord with the findings of high loop flexibility from the molecular dynamics studies. It is suggested that intramolecular quadruplexes can be grouped into two broad classes (i) those with at least one single-nucleotide loop, often showing singular topologies even though loops are highly flexible, and (ii) with all loops comprising at least two nucleotides, leading to topological dynamism. The loops can have more stable and less dynamic base-stacked secondary structures.

INTRODUCTION

Quadruplex DNA and RNA structures are formed from repeats of guanine tracts that are interspersed by mixed sequences ('loops'). These serve to connect together the basic building-blocks of these four-stranded arrangements, the G-quartet (1,2). Sequences encoding for quadruplex nucleic acids have been found in a variety of locations in the human (3,4) and other genomes (5), principally in eukaryotic

telomeres (6,7), in promoter regions (8–10), in untranslated regions especially 5' and 3'-UTRs (11,12) and in a number of breakpoint regions (13–16). Their biological functions remain to be fully elucidated, although the demonstration of the existence of quadruplexes in cancer cells (17,18) and of enhanced levels in human cancer tissues (19) is strongly supportive of the concept that the presence of quadruplexes may be deleterious when they cannot be fully unwound, as in cancer cells having for example deficiencies in the FANCD1 helicase.

The concept that sequences within a promoter comprising repeats of short G-tracts could form higher-order quadruplex structures was first suggested for the c-MYC gene (9,20). Such sequences have subsequently found to be widely prevalent in the human genome, with a particular over-representation in genes involved in proliferation, notably in oncogenes such as k-RAS (21), b-RAF (22), SRC (23), BCL-2 (24,25) and c-KIT (26–32). Promoter quadruplexes can be stabilized by the binding of appropriate small molecules and it has been suggested that this can be an effective approach to transcriptional inhibition at the single-gene level (9,10). A large number of studies have reported down-regulation of expression of one or other these genes in cell-based systems when quadruplex-binding small molecules have been used (see (10,33,34) for recent reviews). Correlations have been suggested in a number of these studies between *in vitro* small molecule quadruplex affinity to a particular promoter quadruplex and expression changes in that particular gene, although this is not always the case (35). Even though there is no definitive evidence to date of cause and effect in cells for a particular promoter quadruplex gene target, the attractiveness of the concept has led to its current emphasis as a novel drug discovery strategy.

The c-KIT gene codes for the KIT tyrosine kinase gene product, which is implicated in a number of human cancers, notably gastro-intestinal cancer (GIST), where down-regulation of c-KIT expression is the primary causative event of this disease (36–38). Two quadruplexes have been identified in the promoter of this gene, one positioned between –87 and –109 bp (termed here c-KIT G4) (26) and

^{*}To whom correspondence should be addressed. Tel: +44 207 753 5969; Fax: +44 207 753 5970; Email: s.neidle@ucl.ac.uk

Present addresses:

Dengguo Wei, Department of Chemistry, Umeå University, Umeå SE-901 87, Sweden.

Jarmila Husby, Istituto Italiano di Tecnologia, Genova, Italy.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

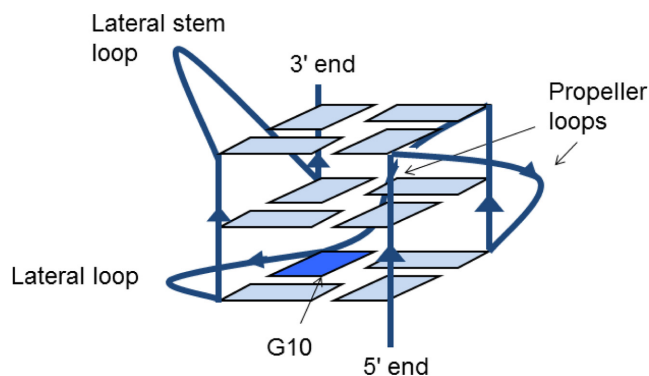


Figure 1. Schematic representation of the c-KIT quadruplex topology.

the other –140 and –160 bp (termed here c-KIT1 G4) (27), upstream of the transcription initiation site. A number of studies with low molecular-weight compounds have shown correlations between their c-KIT/c-KIT1 G4 binding and inhibition of c-KIT expression (39–44) and it has been suggested that quadruplex targeting may be a novel approach to the therapy of GIST.

The topology and molecular structures of the c-KIT (28,31) and c-KIT1 (29,30) G4s have been characterized by nuclear magnetic resonance (NMR) studies and both show unique features compared to other known quadruplexes. In addition a crystal structure of the c-KIT G4 has been determined (32), with topology identical to that reported in the NMR study (on the identical sequence), albeit with differences in groove dimensions. The quadruplex has a parallel fold with two single-nucleotide propeller (double chain-reversal) loops and a long five-nucleotide lateral stem loop; unusually one non-G-tract guanine participates in the core of stacked G-quartets (Figure 1). The equivalence of solution and crystal structures suggests that the dynamics of the c-KIT quadruplex are limited and that the observed topology represents the global minimum structure. We report here on two further c-KIT crystal structures that have also been determined on the identical sequence, but crystallize in distinct space groups ($P3_121$ at 1.82-Å resolution and $H3$ at 2.73 Å, compared to space group $P2_12_12$ for the previously reported brominated form, which was determined at 1.62-Å resolution). These new crystal structures also show the unique c-KIT topology, providing added credence to this hypothesis. We further report on a series of molecular dynamics (MD) simulations, on a total of five distinct c-KIT quadruplex starting-points from the two high-resolution crystal structures and the NMR structure. This has enabled the flexibility of the c-KIT quadruplex to be examined in detail, and implications for the structure-based design of specific small molecules to be addressed.

MATERIALS AND METHODS

Crystallography

The c-KIT DNA quadruplex sequence d(AG₃AG₃CGCTG₃AGGAG₃) was synthesized and purified as described previously (32). A 2-mM DNA solution containing 20-mM potassium cacodylate buffer

at pH 6.5 and 50-mM KCl was heated to 358 K before annealing by slow cooling to room temperature. Crystals were grown by the hanging-drop vapor diffusion method. 1 μl of a premixed drop solution containing 10% MPD, 10-mM MgCl₂ 50-mM KCl and 50-mM sodium cacodylate at pH 6.5 was added to 1 μl of c-KIT DNA solution at a concentration of 2 mM. The drop solution was equilibrated against the well solution with 50% MPD, 20-mM MgCl₂ and 100-mM NaCl at 283 K. Hexagonal and rod-like crystals appeared after 2 weeks. The rod-like crystals were assigned the alternative hexagonal $H3$ setting for the rhombohedral space group $R3$ and the hexagonal crystals were assigned the trigonal space group $P3_121$. Data sets were collected at 105 K on a single flash-frozen crystal of each type at the Diamond synchrotron facility (UK).

Molecular replacement methods were used in solving these two crystal structures using the G-quartets extracted from the earlier ^{Br}U c-KIT DNA crystal structure (PDB code 3QXR). Other parts of the quadruplex structures were progressively built on the basis of 2F_o-F_c and F_o-F_c difference electron density maps. This was followed by the location of K⁺ ions in the central ion channels. A total of 162 water molecules were identified in the 1.82-Å hexagonal crystal structure, and three water molecules were located in the 2.73-Å trigonal crystal structure. Final R and R_{free} values are listed in Table 1. Coordinates and structure factors are available from the PDB as entries 4WO2 and 4WO3, respectively. Diffraction data were processed by the SCALE program in the CCP4 package (45). Programs COOT (46), PHASER (47) and REFMAC5 (48) were used in the processes of structure building, model fit and refinement, respectively. Programs CHIMERA (49) and PYMOL (50) were used for visualization and analysis. Table 1 details the crystallographic data.

MD simulations

Coordinates for the two higher-resolution crystal structures and the NMR structure of the c-KIT quadruplex DNA were obtained from the Protein Data Bank (PDB). Those were:

- (i) the crystal structure of ^{Br}U c-KIT G4 (PDB 3QXR) at 1.62-Å resolution (32);
- (ii) the crystal structure of the native c-KIT G4 at 1.82-Å resolution (PDB 4WO2);
- (iii) the NMR structure of monomeric c-KIT G4 (PDB 2O3M) (28).

All four crystallographically independent quadruplexes were used in the simulations since both of these X-ray structures comprising two intramolecular quadruplexes per asymmetric unit (chains A and B), together with model 1 from the NMR structure. These intramolecular 22-base quadruplexes all have the identical native sequence, 5'-d(AG₃AG₃CGCTG₃AG₂AG₃)-3', which has not been modified from that occurring in the c-KIT gene promoter sequence. Nucleotide numbering in this paper is sequential from the 5'-adenosine.

The sole exception to this is the sequence in the brominated X-ray structure (32), which differs at the 12th posi-

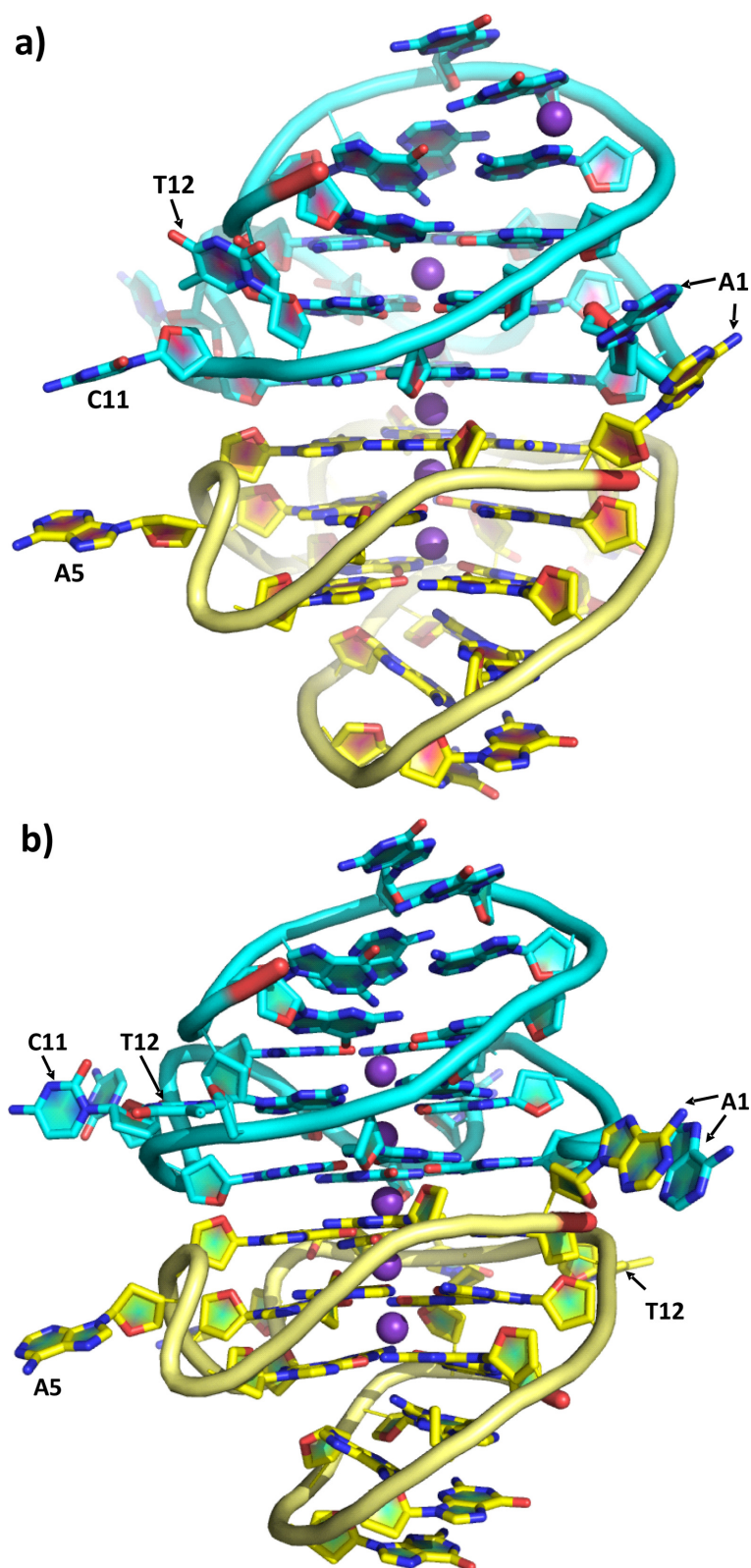


Figure 2. Cartoon representation of the crystallographic asymmetric unit of the two c-KIT quadruplex crystal structures. The two chains in each instance are colored cyan and yellow and the potassium ions are shown as purple spheres. (a) The 1.82-Å crystal structure in space group $P3_121$; (b) the 2.73-Å crystal structure in space group $H3$.

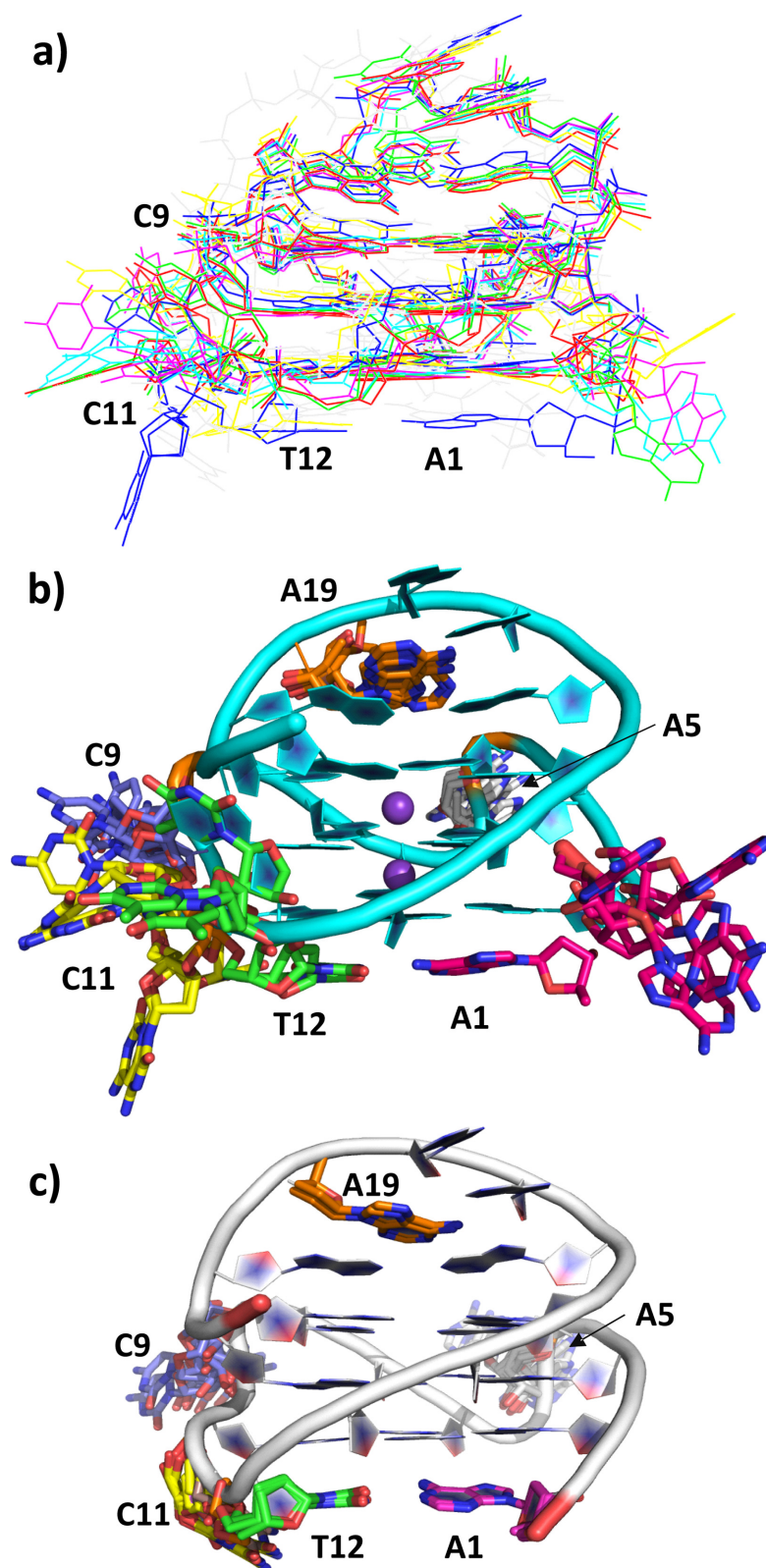


Figure 3. (a) Composite plot showing a superposition of the two c-KIT crystal structures, the previous ^{Br}U c-KIT crystal structure and the c-KIT NMR structure used in the MD analysis. (b) Composite plot showing a superposition of the three c-KIT crystal structures, highlighting the nucleotides with variable conformations. (c) composite plot showing a superposition of the ensemble of c-KIT NMR structures. In (b) and (c) all bases are shown but with only a single backbone.

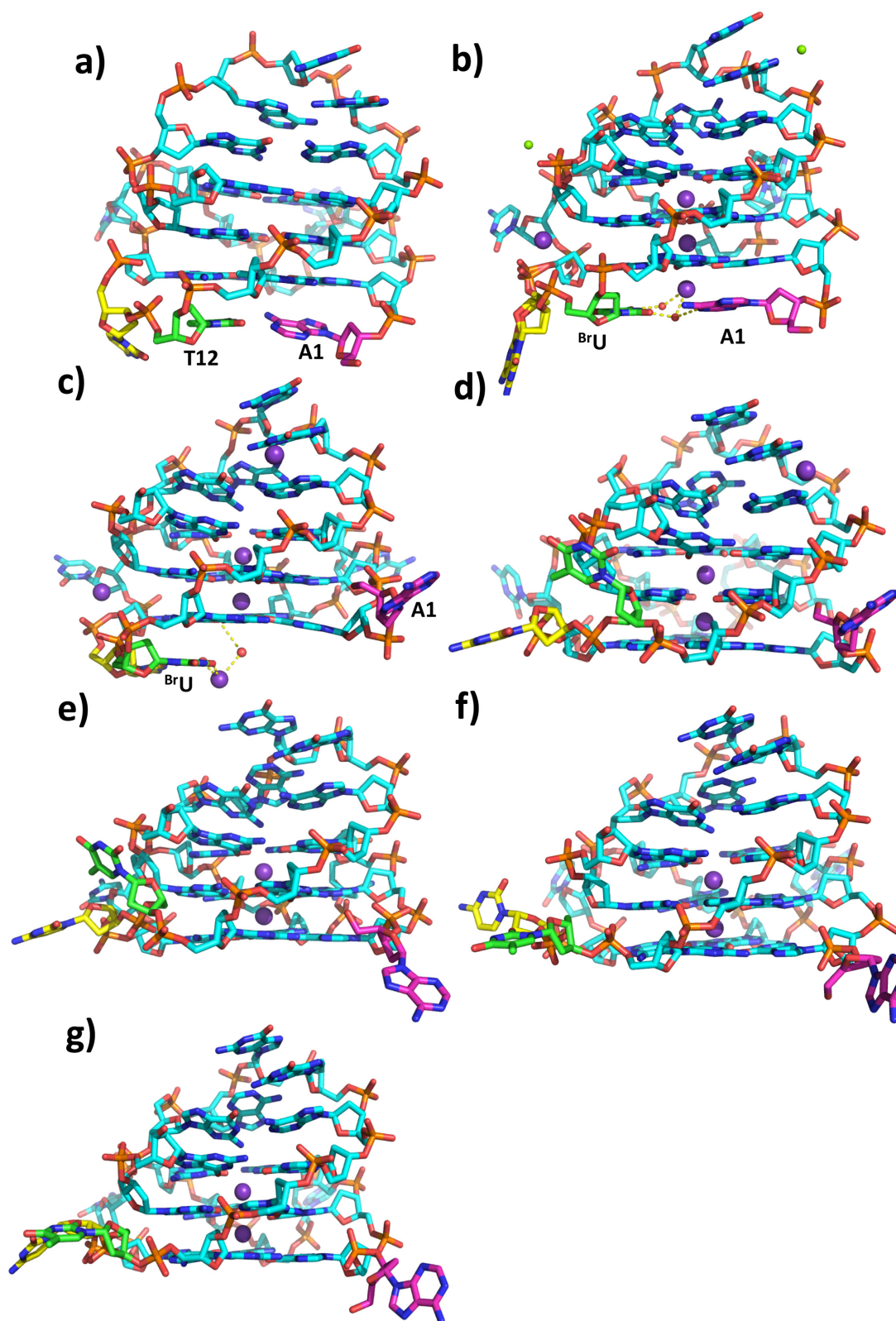


Figure 4. The flexibility of bases A1, C11 and T12 as observed in the experimental structures. A1 is colored purple, and C11 and T12 are in yellow and green, respectively. (a) In the NMR structure, (b) chain A in the 1.62-Å ^{Br}U c-KIT crystal structure, (c) chain B in the 1.62-Å ^{Br}U c-KIT crystal structure, (d) chain A in the 1.82-Å trigonal crystal structure, (e) chain B in the 1.82-Å trigonal crystal structure, (f) chain A in the 2.73-Å hexagonal crystal structure, (g) chain B in the 2.73-Å hexagonal crystal.

Table 1. Crystallographic data collection and refinement statistics

Sequence	d(AGGGAGGGCGCTGGGAGGAGGG)		
	BrU c-KIT	c-KIT	c-KIT
Data collection			
Space group	$P2_12_12$	$P3_121$	$H3$
Cell dimensions			
a, b, c (Å)	37.57, 51.35, 58.34	31.44, 31.44, 186.81	94.85, 94.85, 30.36
Wavelength (Å)	0.9763	0.9173	0.9686
Resolution (Å)	38.55–1.62	31.13–1.82	28.48–2.73
R_{merge}	0.057	0.083	0.115
I/σ	27.4	14.5	5.0
Completeness (%)	93.2	99.1	94.4
Redundancy	3.6	6.9	2.6
Total no. of reflections	50085	72204	6630
No. of unique reflections	14007	10436	2561
Refinement			
Resolution (Å)	26.9–1.62	10.28–1.82	28.48–2.73
No. of reflections	13279	9807	2439
$R_{\text{work}}/R_{\text{free}}$	0.176/0.221	0.210/0.223	0.183/0.246
No. of ions	11.5	6	5
No. of water molecules	210.5	162	3
Overall B factor (Å ²)	20.07	27.17	41.24
RMS deviations in			
Bond length (Å)	0.005	0.006	0.008
Bond angle (°)	1.0	0.694	1.3
PDB ID	3QXR	4WO2	4WO3

The data on the BrU c-KIT structure have been previously reported (32); these are shown here for comparison purposes.

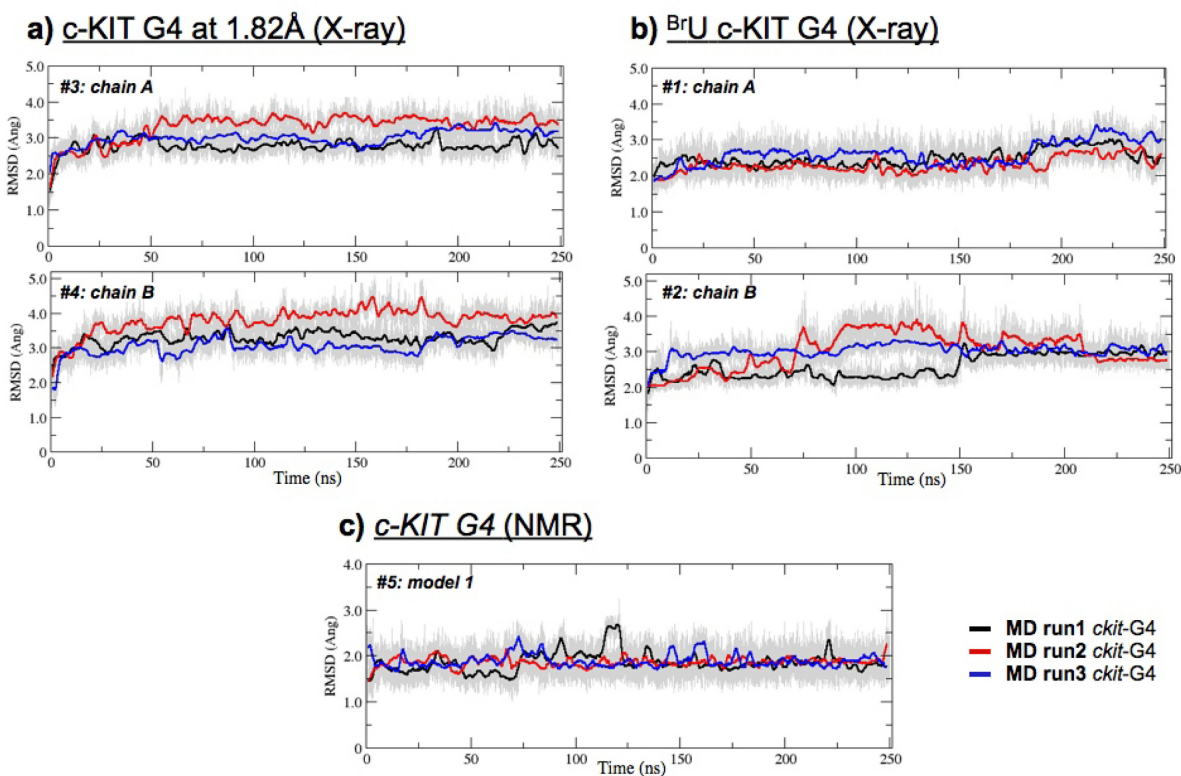


Figure 5. RMSD plots showing the structural convergence of the simulated systems during multiple MD runs of the c-KIT 22-mer G-quadruplexes. Each of the 250-ns MD runs was performed three times, for all five simulated systems: chains A and B of the native c-KIT (X-ray) (a) chains A and B of the BrU c-KIT X-ray structure, pdb 3QXR (b) model 1 of the c-KIT NMR structure, pdb 2O3M (c) Running averages over 500 frames of the RMSD plots (gray) for MD run 1 (black), MD run 2 (red) and MD run 3 (blue).

tion, where 5-brominated uridine (^{Br}U) replaced a thymine (T). The ^{Br}U residue was modified back to T for the simulations using the Accelrys Discovery Studio package (www.accelrys.com; v.3.5), to match the original native sequence. Overall, five individual c-KIT 22-mer quadruplex units were used as starting points for the MD studies, i.e.

- #1 chain A and #2 chain B, both derived from the ^{Br}U c-KIT crystal structure,
- #3 chain A and #4 chain B, derived from the high-resolution native c-KIT crystal structure reported here,
- #5 model 1 from the native c-KIT NMR structure.

Since the individual conformers available in the NMR ensemble have an average deviation in their structural alignment of <0.5 Å, it was considered that any one of them would be equally suitable to be used as a starting point for the MD study. The NMR structure was used for comparison purposes—to explore the dynamic behavior of c-KIT structures of identical sequence, but determined by two different techniques (X-ray and NMR).

Consecutive K⁺ ions vertically aligned within the central core of the quadruplexes and mid-way between each G-quartet were retained at their respective crystallographic positions, as well as the structural K⁺ ions present in the loops of individual 1, 2 and 3 quadruplexes. Two additional Mg²⁺ ions observed in the loops of quadruplex 1 were also retained for the simulations. In the NMR system, the missing structural K⁺ ions were manually added into the central core of the G4, within their respective positions as reported in the X-ray structures. The explicit water molecules, present within the crystallographic structures, were retained for the simulation setups, even though they only represent a small fraction of the total numbers of water molecules in these structures.

All of the MD simulations were all-atom ones, performed with the GROMACS v.4.6 program (51), employing the AMBER *parmbsc0* force field (52) ported into GROMACS. The TIP3P water model was used and for the ions, AMBER *parm99sb* parameters were applied, both for the structural ions and the counter-ions. The simulation protocols, reported elsewhere (53,54), were consistent for all five c-KIT quadruplex systems. The production step of each of the five MD simulations was carried out in triplicate for 250 ns in order to improve sampling and statistical reliability, resulting in 750 ns of simulation data per quadruplex for subsequent analysis. Altogether, 3750 ns of MD trajectories were generated and analyzed by a robust clustering algorithm that automatically provides the desired number of clusters, in order to identify the most prevalent conformations sampled throughout the simulation time at 300 K. All the simulations were performed on an in-house AMD CPU-based Linux cluster (IBM Blade Center H; 16 CPUs per simulation) at the Italian Institute of Technology (Genoa, Italy). A robust clustering algorithm (55), adapted to cope with MD data (56), was employed here (see the Supplementary data for further information). The advantage of using this particular clustering algorithm is that multiple non-consecutive trajectories of an individual system can be clustered together, while it ensures that only those transitions between individual clusters are counted/reported if the frames be-

long to the same/corresponding trajectory, and also ensures that no ‘artificial’ (or unreal) transitions between clusters are formed.

Clustering of the MD trajectories (3 × 250 ns) was carried out for each of the five systems, employing a total of 150 000 frames, each having a 5-ps time step. The maximum number of clusters was set up to 10 (this was found to be an optimal number, based on preliminary clustering analysis where various numbers of resulting clusters were identified—ranging from 5 to more than 20).

RESULTS

Crystal structures

The c-KIT quadruplex in both the 1.82-Å and the 2.73-Å structures (Table 1) forms head-to-head dimers in the crystals, closely analogous to that previously observed in the ^{Br}U form. Each monomeric intramolecular quadruplex in a dimer has been designated as chain A or B. Thus the two new crystal structures comprise four crystallographically independent c-KIT quadruplexes. The quadruplex topology in each instance is identical to that previously reported for c-KIT in the crystal and in NMR solution. In each instance the 22-nucleotide DNA sequence folds into a parallel four-stranded G-quadruplex, and the isolated non-G-tract guanine (G10) is embedded in the G-quartet core. Two single-residue linkers (A5 and C9) form two propeller (double-chain reversal) loops that bridge three G-quartet layers. C11 and T12 form the third loop, which connects two G-quartet corners (G10 and G13). The stem-loop formed by the five-residue sequence A16–G17–G18–A19–G20 allows the terminal G21–G22 to be inserted back to the G-quartet core. The fourth loop, the long lateral stem loop retains its secondary particular features of two A:G hydrogen bonds. Potassium ions were located in the central channel of both crystal structures (Figure 2). No additional potassium or magnesium ions were observed in the grooves or loops of these quadruplexes. This contrasts with their location in the higher-resolution ^{Br}U form (strong electron density was observed in the 1.82-Å crystal structure at the boundary between two stacked asymmetric units, which is consistent with a potassium ion rather than a water molecule).

The bases at the dimer stacking interfaces (A1, C11 and T12) show the most striking difference between the four individual c-KIT G-quadruplexes, the ones reported in the earlier ^{Br}U crystal structure, and the NMR structure (Figures 3a and b and 4). In the NMR c-KIT structure, the two bottom bases A1 and T12 form a Watson–Crick base pair and stack onto the bottom quartet surface (Figure 4a). In the earlier ^{Br}U (orthorhombic) G4 crystal structure, there is hydrogen bonding in chain A between N6 of A1 and O2 of ^{Br}U12, and a water molecule bridges between N1 of A1 and N3 of ^{Br}U12 (Figure 4b). A1 of the ^{Br}U G4 in chain B is oriented away of the structure, and bases C11 and ^{Br}U12 are stacking on the quartet surface, beneath the contact between a potassium ion and a water molecule (Figure 4c).

In the two new c-KIT crystal structures, two quadruplexes stack each other through the bottom G-quartet, and the interface bases A1, C11 and T12 are oriented away from the common surface (Figure 4d and e). The two A1 bases in each chain asymmetric unit form a stacking pair in both two

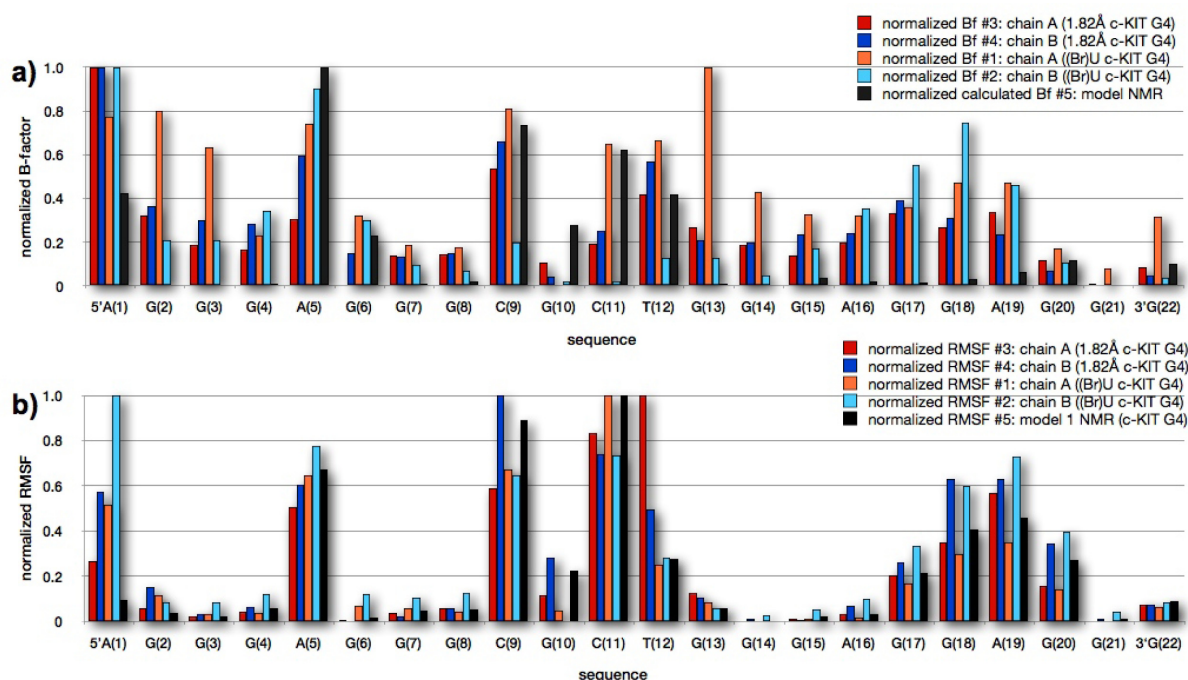


Figure 6. (a) Plot showing the normalized B-factor values from the experimental crystal structures, on a per-residue basis. (b) RMS fluctuations from the MD simulations, calculated on a per-residue basis for the five c-KIT G4 systems studied. #1 G4 and #2 G4 from the native c-KIT G4 crystal structure are shown in red and blue, respectively, #3 G4 and #4 G4 from the ^{Br}U c-KIT G4 are shown in orange and cyan, respectively, and #5, the c-KIT G4 NMR structure is shown in black.

crystal structures, while C11 and T12 have different roles. In the trigonal crystal form, C11 and A5 from different asymmetric units form a long stacking unit. Two C11 bases in the 2.7-Å hexagonal crystal structure form a hydrogen bond with a G7 phosphate oxygen atom in the other asymmetric unit, and the T12 base stacks with the G18 base from another asymmetric unit.

These differences between the total of six independent c-KIT quadruplexes in the three crystal structures are also seen in the diverse conformations of several others of the external, non-hydrogen-bonded bases (Figure 3a and b). Backbone torsion angles for the extra-helical nucleotides adopt a wide range of values (see Table 2 and Supplementary Table S1), with very few torsion angles being conserved across the six independent quadruplex chains (values for the nucleotides comprising the stacked G-quartet core are in general highly conserved—see the Supplementary data for a full listing of torsion angles). Loop angles δ and to a lesser extent ζ have conserved values, the most notable exception being in the quadruplex A chain of the 2.73-Å crystal structure. In the ^{Br}U form, bases A1, A5 and C9 are directed out from the quadruplex core, base C11 from chain B is tucked into the core, under the G10 deoxyribose, and the ^{Br}U at position 12 is also tucked in, partially stacking onto G13. In the high-resolution 1.82-Å structure bases A1, A5 and C9 are also oriented outward, as is C11. Base T12 however adopts a distinct, partially swung-out conformation, with the edge of the thymine coming closer to the backbone at G22. By contrast, these bases in the ensemble of NMR structures adopt a significantly narrower range of conformations (Figure 3c).

The most pronounced difference between the crystal structures and the NMR models involves the large cleft between the stem-loop (G18, A19 and G20) and the adjacent G-quartet. The net effect of these differences is that the large cleft is consistently narrower in all of the crystal structures.

MD simulations

Multiple-trajectory MD simulations of the intramolecular 22-mer c-KIT G4 in five distinct systems were carried out for 250 ns, each repeated in triplicate, for a total of 3750-ns simulation time. The systems were:

- #1: chain A (with 4 K⁺ and 2 Mg²⁺ ions) of the ^{Br}U 1.62 Å c-KIT G4
- #2: chain B (with 4 K⁺ ions) of the ^{Br}U 1.62 Å c-KIT G4
- #3: chain A (with 3 K⁺ ions) of the 1.82 Å c-KIT G4
- #4: chain B (with 2 K⁺ ions) of the 1.82 Å c-KIT G4
- #5: model 1 (with 2 K⁺ ions) from the c-KIT G4 NMR structure

The multiple-trajectory approach enabled improved sampling of the conformational space to be obtained for a given set of conditions/local minima, resulting in more statistically robust conclusions than analyses based on a single trajectory approach. The structural convergence of the five systems over the course of the multiple 250 ns simulations is graphically represented as RMSD plots in Figure 5 and summarized in Supplementary Table S2. Table 3 summarizes the ions and water molecules in the simulations. In the absence of the structural central channel ions that were manually placed into the NMR model prior to starting the simulations (#5), the system collapsed. (Data not shown)

Table 2. The backbone torsion angles for the quadruplex loop residues of the NMR structure (model 1 only) and the three crystal structures

		NMR	BrU c-KIT 1.62 Å		c-KIT 1.82 Å		c-KIT 2.73 Å	
			Quad-A	Quad-B	Quad-A	Quad-B	Quad-A	Quad-B
A1	γ	−63	−63	−74	−125	−118	−142	−57
	δ	118	157	131	91	149	101	91
	ε	180	−76	−155	74	−101	104	55
	ζ	−81	−158	63	−155	−64	−145	142
	χ	39	−100	−80	−108	−136	−107	−72
A5	α	52	59	−89	71	81	119	80
	β	128	154	−169	152	168	179	150
	γ	55	45	168	48	26	−29	31
	δ	149	149	110	93	128	125	109
	ε	179	−119	−96	−150	−151	−162	−145
	ζ	156	83	160	−169	173	−163	176
	χ	−51	−153	−159	−133	−81	−87	−97
C9	α	75	57	66	72	69	64	−90
	β	176	155	140	169	156	135	−101
	γ	59	49	57	50	52	56	158
	δ	131	107	106	145	98	104	76
	ε	−108	−153	−143	−125	−155	−129	−115
	ζ	−149	−172	169	112	−172	158	152
	χ	−128	−155	−100	−160	−113	−161	−140
C11	α	10	45	151	50	53	−101	−54
	β	128	98	−143	−160	−176	−120	−128
	γ	−48	175	50	67	70	161	127
	δ	135	152	151	150	142	106	82
	ε	−144	−99	−140	−163	−151	−118	−166
	ζ	−127	−79	−170	149	161	143	−134
	χ	−175	29	−170	−128	−146	−114	−176
T12	α	−106	−74	−80	−138	−166	74	−114
	β	−103	−163	−157	−146	−125	102	−131
	γ	63	52	56	68	68	−7.1	77
	δ	145	111	151	141	144	85.2	78
	ε	173	−150	−173	−163	−134	36.6	62
	ζ	−104	−123	−76	133	82	158.4	−105
	χ	−121	−156	−133	−126	−112	25.2	13

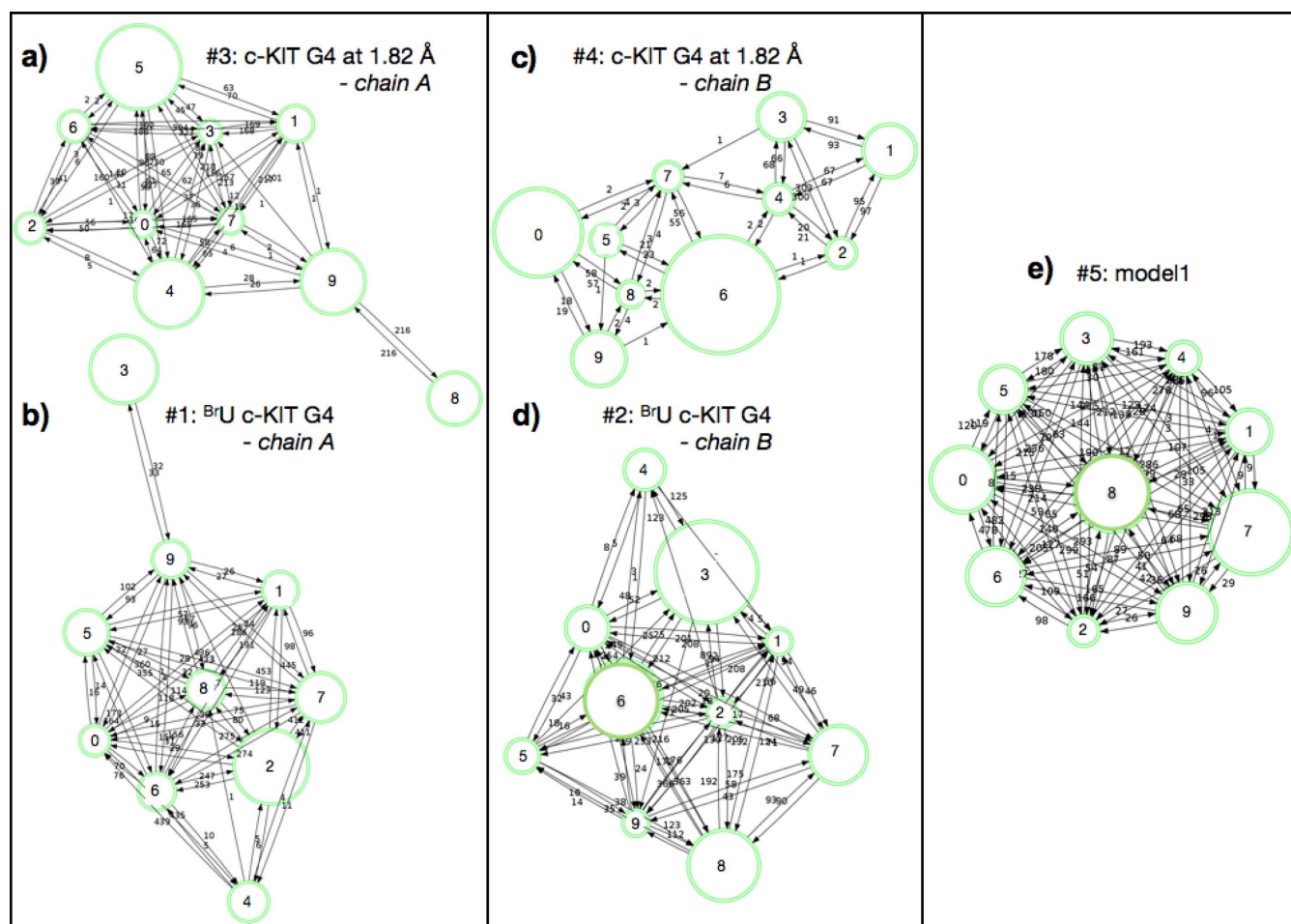


Figure 7. Clusters of c-KIT G4 conformations sampled through the multiple MD runs. The triplicate MD trajectories (i.e. 750 ns) of the two c-KIT G4 chain A were clustered by the robust clustering algorithm to yield (a) 10 clusters of #3 G4, (b) 10 clusters of #1 G4, (c) 10 clusters of #4 G4, (d) 10 clusters of #2 G4 and (e) 10 clusters of #5 G4, in order to identify the predominant conformations sampled throughout the simulations. In the graphical representation of the clusters (a–e) transitions between individual clusters were counted and indicated by arrows. The cluster sizes (i.e. the number of structures within a cluster) are represented by circles.

here, but during preliminary 250-ns simulations the NMR structure became unwound and unstable in the absence of K^+ ions).

The flexibility of the bases within these systems was explored prior to starting the MD simulations, by plotting the averaged normalized B factor values per residue that were obtained from the X-ray structures (Figure 6a). Although experimental B factors are purely crystallographic quantities, theoretical B-factors (per residue) were also calculated, and normalized, for the c-KIT G4 NMR structure, employing all the conformations within the NMR bundle. These theoretical B factors were calculated by means of the GROMACS *g_rmsf* tool. This shows a consistent pattern of increased B values for those nucleotides with extrahelical bases, especially at and around A1, A5, C9, C11 and T12. Nucleotide A19 is also flexible, even though the adenine base is stacked onto G18. The RMSD values averaged per residue over all of the models within the NMR structure were calculated and normalized by means of the GROMACS *g_rmsf* tool (Figure 6a). Comparison of the crystallographic and NMR-derived fluctuations shows that the

flexibility of residues A5, C9, C11, T12 and A19 is apparent both in the experimental and simulated structures. The RMS fluctuations of individual bases were calculated on completion of the MD simulations from all three trajectories available for each of the five systems, and were normalized (Figure 6b). There is consistently good agreement between the experimental and simulation data, with the same regions and nucleotides in all of the c-KIT quadruplexes showing equivalent flexibility and stability.

The three 250-ns MD runs per system were clustered together, resulting in 10 clusters per system. Each of the identified clusters is represented by a medoid structure (i.e. a representative structure/object of the cluster whose average dissimilarity to all other structures/trajectory frames in the cluster is minimal). Each of the medoids corresponds to a physical frame from the trajectory. The probability of formation (i.e. the cluster population) of each cluster was also calculated and is reflected in the graphical representation of the clusters (Figures 7 and 8).

Although clustering was performed solely for the nucleic acid atoms, the complete systems—including the structural

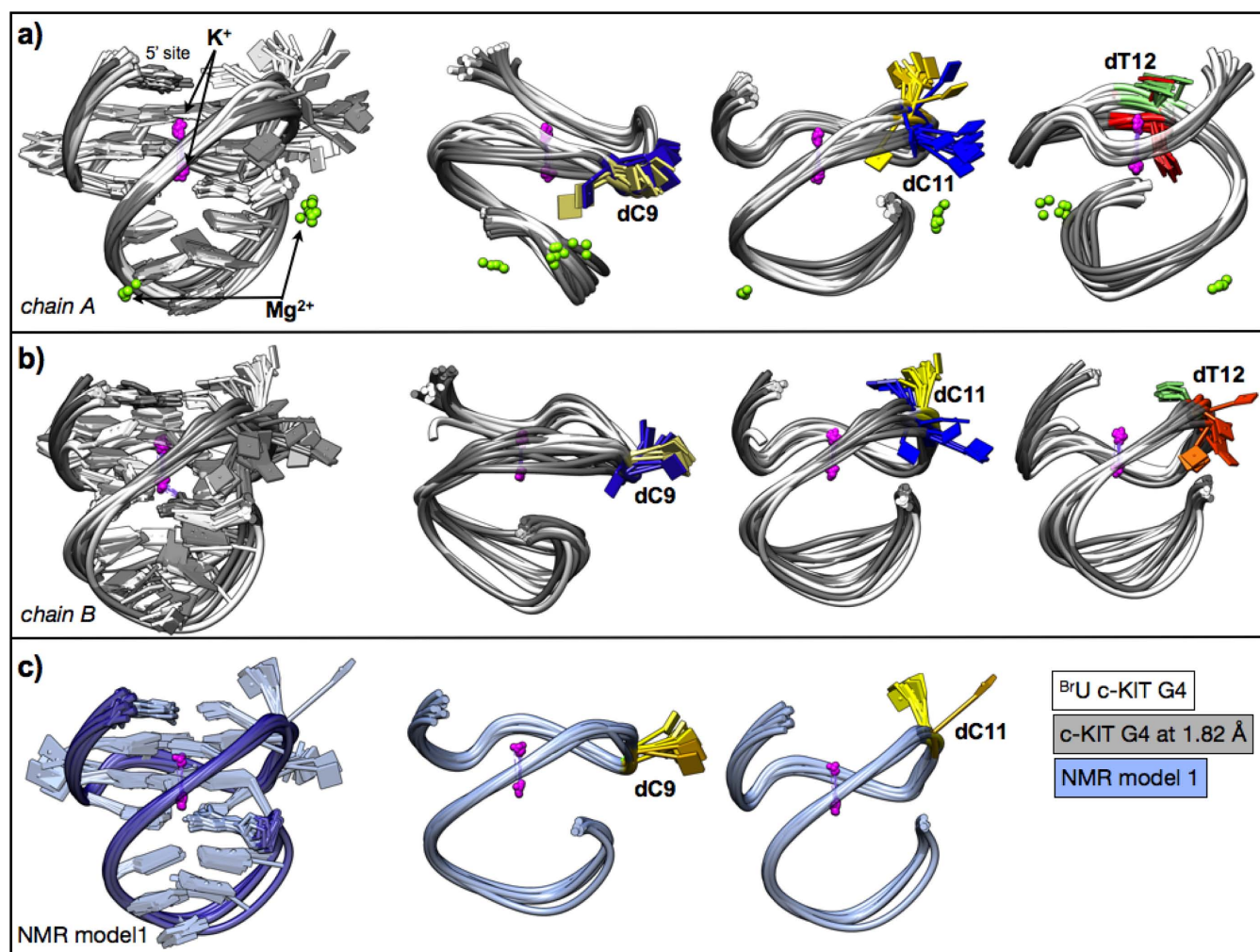


Figure 8. Conformations of the c-KIT G4 conformations sampled through the multiple MD runs. (a) Upon superimposition of all 20 c-KIT G4 chain A cluster-representative medoid structures (10 from #3 G4 and 10 from #1 G4), the highest level of conformational flexibility was found for bases C9, C11 (#3 G4 in blue, #1 G4 in yellow) and T12 (#3 G4 in orange, #1 G4 in green). (b) All c-KIT G4 chain B cluster-representative medoid structures (i.e. 20 conformations) were superimposed and compared; bases C9, C11 (#4 G4 in blue, #2 G4 in yellow) and T12 (#4 G4 in orange, #2 G4 in green) displayed the highest level of conformational variability (base flipping). (c) The superposition of the NMR-based c-KIT conformations (#5 G4, 10 conformations) shows base flipping at bases C9 and C11 (yellow).

Table 3. Listing of the ions and water molecules in the simulations

System	Timescale	Residues	Ions		TIP3P waters	
			Structural	Counter-ions	Structural	Total
#1 BrU c-KIT G4 chain A	3 × 250 ns	22 bases	4 K ⁺ , 2 Mg ²⁺	13 K ⁺	123	4900
#2 BrU c-KIT G4 chain B	3 × 250 ns	22 bases	4 K ⁺	17 K ⁺	88	6176
#3 1.82-Å c-KIT G4 chain A	3 × 250 ns	22 bases	3 K ⁺	18 K ⁺	164	6092
#4 1.82-Å c-KIT G4 chain B	3 × 250 ns	22 bases	2 K ⁺	19 K ⁺	164	6096
#5 model 1 c-KIT G4 NMR	3 × 250 ns	22 bases	2 K ⁺	19 K ⁺	/	4775

ions—were retrieved for visualization. The structural K⁺ ions within the central core in between the quartets always remained at their respective positions (Figures 7 and 8), as well as the Mg²⁺ ions, which remained within the loops of the #1 G4 structure (Figure 7). However the K⁺ ions initially present in the loops of systems #1, #2 and #3 did not remain at their sites throughout the course of the MD runs.

Clusters of chains #1 and #3 (Figure 7) from both X-ray structures displayed a common pattern. Nine of the 10 clusters are highly interconnected, i.e. undergoing multiple transitions between structural conformations, but one cluster is isolated (cluster no. 8 for #3, and cluster no. 3 for #1; the population in both cases ranges between 12 and 15%, respectively), and is only connected to one of the other clusters. Superimposition of all the 20 representative structures

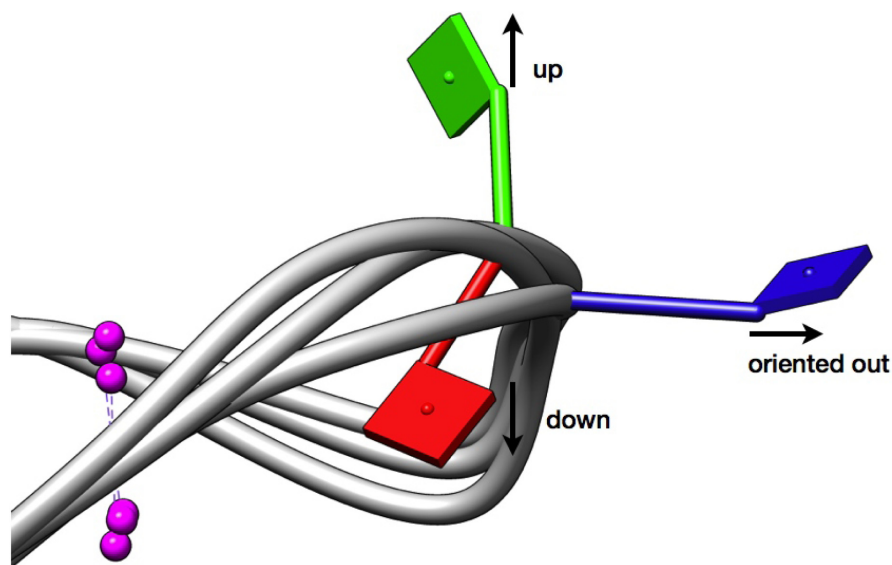


Figure 9. Different positions/orientations of bases identified in cluster representatives sampled throughout the MD runs. Bases C9, C11 and T12 were found to be oriented up (green), down (red) and away from the backbone (blue). The bases were also found to be stacking on the bases of other G-quartets (not shown here).

of #1 and #3 shows that their conformations are overall in good agreement, with the exception of bases C9, C11 and T12 where significant base flipping occurs (Figure 7).

Superimposition of the cluster representatives (medoid structures) of #2 and #4 G4s reveals more conformational flexibility/variability compared to #1, in particular in the loop region formed by bases A16-G20. Clear base-flipping occurs at the corresponding bases C9, C11 and T12 (Figure 7).

In terms of medoid structures obtained from the NMR model MD simulations (Figure 8), very good agreement was observed for the conformations, in accord with high stability of the simulation data (trajectories). The clusters are also very well interconnected with high frequencies of transitions between them (Figure 8a). Bases C9 and C11 were the only two where some conformational variability of the bases was found (Figure 8b). However that was to a much smaller extent compared to the X-ray structures.

The flipping (and change of orientation) of bases C9, C11 and T12, determined from the clustering data of all five systems, was further examined and quantified, and is summarized in Table 4. Visualizing/keeping the G-quartets in plane, orthogonal to the K^+ ions in between the G-quartets, has enabled four orientations of these bases to be identified with respect to the DNA backbone: (i) up, (ii) down, (iii) oriented out, (iv) stacking, as shown in Figure 9. These positions correspond to those described in Table 4.

Base C9 remained predominantly in the same orientation in all five simulated systems, as in all three experimental structures: it is oriented away from the DNA backbone. This base did sometimes adopt other orientations, although not in structure #2 or the NMR model. Base C11 has the highest degree of conformational variability, in both #3 and #4 G4s, as well as #1. T12, on the other hand, was found to be stacking on base G13 in the NMR model as well as

in both #1 and #2 structures. In #3 and #4, T12 was also found to be oriented away from the DNA backbone.

DISCUSSION

The short c-KIT loops are very flexible

The consistent experimental observations of only very limited conformational flexibility for the core of the c-KIT quadruplex in all the crystal and NMR structures are in overall excellent accord with the MD data, as shown in particular in Figure 6, comparing experimental B factors with calculated RSMD values. Similarly the conclusions from the simulations that the loop nucleotides on the exterior of the structures are conformationally flexible and show extensive base flipping are in good accord with the experimental observations from the three separate crystal structures and the NMR structure.

The trinucleotide loops in human telomeric quadruplexes have been extensively explored by MD simulation methods (57–63) and these have also been found to be conformationally mobile (60,63). Although a range of conformations have been observed in these simulations, dependent on loop type, base–base stacking in the loops does stabilize some of them for at least part of the simulation time. Loop flexibility has also been reported in a previous c-KIT quadruplex simulation (64), and the present study extends this finding by showing that single or dinucleotide loops are less likely to form discrete transiently stabilized base-stacked loops than are trinucleotide ones, as in human telomeric quadruplexes. By contrast the pentanucleotide AGGAG loop in the c-KIT quadruplexes is rather stable, with again experiment and simulation in broad agreement.

The potential ligand binding sites of the c-KIT quadruplex

The flexibility of the external unpaired bases, in particular those forming the two single-nucleotide loops A5 and C9

Table 4. Overview of the conformational changes of bases C9, C11 and T12 sampled throughout the three repetitive MD runs of each of the five studied systems

Base	Chain A		Chain B		#5 model 1 c-KIT G4 NMR
	#1 ^{Br} U c-KIT G4	#3 1.82-Å c-KIT G4	#2 ^{Br} U c-KIT G4	#4 1.82-Å c-KIT G4	
C9 →	85%	93%	100%	74%	100%
C9 ↑	-	-	-	15% (cl. 2, 3)	-
C9 ↓	15% (cl. 0, 4)	7% (cl. 2)	-	11% (cl. 9)	-
C11 →	8% (cl. 9)	73%	-	62% (cl. 0, 6–9)	16% (cl. 7)
C11 ↑	77%	15% (cl. 9)	100%	32% stack on G10	84%
C11 ↓	15% (cl. 3)	12% (cl. 8)	-	6% (cl. 5)	-
T12 →	-	16% (cl.4)	-	24% (cl. 6)	-
T12 ↑	-	-	-	-	-
T12 ↓	-	43%	-	76%	-
T12 stack (G13)	100%	27% (cl. 8, 9)	100%	-	100%

and the dinucleotide loop C11, T12, is also of potential significance for small-molecule binding to the termini of the c-KIT quadruplex. For example, one site may involve the stacking of T12 onto the bottom G-quartet, as observed in the ^{Br}U c-KIT crystal structure and the NMR structure. This feature also has 27% occurrence in the simulated chain A of the 1.82 Å c-KIT G4 trigonal form (Table 4), even though the stacking is not found in the initial starting point for this particular structure. Some ligands may stack on top of the *in situ* T12-A1 stack rather than onto this terminal G-quartet, which would involve displacing the T12-A1 base pair.

The large cleft in the c-KIT quadruplex is conserved in all the structures, experimental and theoretical, although its dimensions do have some variability (see, in particular, Figure 8). The cleft diameter can be defined by P...P inter-strand separation distances across the cleft. Distance P(G8)...P(G20) is 14.8 Å in the NMR structure, 8.1 Å in the ^{Br}U crystal structure and 8.7 Å in the 1.82 Å crystal structure reported here. The cleft is of sufficient size and nature to be a suitable small-molecule binding site. Even though *in silico* docking methods need to take the flexibility of cleft dimensions into account, the nature of the site is well-suited to experimental fragment-based approaches (53,65) to discover novel c-KIT quadruplex-selective small molecules capable of down-regulating c-KIT expression.

The c-KIT quadruplex core topology is conserved

Quadruplex DNA and RNA nucleic acids can in principle adopt a wide range of topologies, and X-ray and NMR methods have to date almost certainly only accessed a small cohort of them. The potential for a large, but as yet undefined number of folds is due to several factors, including (i) the inequality of G-tract lengths in many quadruplexes coupled with (ii) the diversity of loop lengths and sequence. It is apparent, for example, that when a guanosine nucleotide occurs in a loop region, it can become directly involved in quartet formation and thus induces more complex arrangements than is the case with simple quadruplex-forming sequences containing non-guanosine loops. The

c-KIT quadruplexes is an exemplar of this complexity of structure.

The present study addresses the question of whether the c-KIT quadruplex shows conformational diversity in differing environments. The earlier observations were that its fold as found by NMR in solution was also observed in the crystalline state. These findings have been reinforced by the determination of the two further c-KIT crystal structures reported here, each of which occurs in a distinct space group. These thus in total provide visualizations of the c-KIT quadruplex in three distinct crystal lattice environments. The differing lattices have not imposed restrictions on the conformations of those external bases of the c-KIT quadruplex, which are unencumbered by hydrogen bonding or internal stacking. Instead these bases are found in a diversity of conformations, comparable to those in the NMR ensemble (compare Figure 3b and c). The CD spectrum of the c-KIT DNA quadruplex in K⁺ solution shows the features of a typical parallel folded quadruplex (data not shown), with a prominent positive signal at around 260 nm and a negative signal at around 240 nm, fully consistent with the fold observed in the crystallographic and NMR c-KIT quadruplex structures.

It can thus be concluded that the core fold of the c-KIT quadruplex is highly conserved, stable and is not environmentally sensitive, at least in K⁺ conditions. This is in striking contrast to the high topological variability shown by human telomeric quadruplex sequences (see, for example (66–76)). This suggests that topological variability is not inherent to quadruplexes as a general class of nucleic acid structures. Rather, only those quadruplexes that have loops containing at least two nucleotides can readily show such variability since single-nucleotides are generally constrained to form propeller loops and are unable to refold to form other types of loop, for stereochemical reasons (a single nucleotide is of insufficient dimensions to form a diagonal or lateral loop). The findings here are in accord with single-molecule studies of the c-KIT quadruplex (77) embedded in a duplex, which show that its dynamics are much reduced compared to the human telomeric quadruplex, and reinforce the view that the stability of the c-KIT

topology makes it a suitable druggable quadruplex target for small molecules. The majority of promoter and other quadruplexes for which there is experimental folding data, from CD studies and in some instances from NMR studies, have at least one single-nucleotide loop and parallel folds have mostly been assigned to these structures (see, for example (24,30,78–81)). We speculate from this that intramolecular quadruplexes can be grouped into two broad classes (i) those with at least one single-nucleotide loop, which often show singular topologies even though loops are highly flexible, and (ii) with all loops comprising at least two nucleotides, which is likely to lead to folding dynamism, and with loops having more stable and less dynamic base-stacked secondary structures.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We are grateful to the Diamond Light Source for access to Synchrotron facilities, to Ambrose Cole (Birkbeck College) and Gavin Collie (CRUK Group, School of Pharmacy) for expert assistance with data collection, to Tony Reszka for help with oligonucleotide synthesis and spectroscopy and to Gary Parkinson for general discussions.

FUNDING

Cancer Research UK [C129/A4489 to S.N.]; China Fellowship [to D.W.]. Funding for open access charge: University College London.

Conflict of interest statement. None declared.

REFERENCES

- Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
- Bochman,M.L., Paeschke,K. and Zakian,V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770–780.
- Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
- Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
- Cao,K., Ryzkin,P. and Johnson,F.B. (2012) Computational detection and analysis of sequences with duplex-derived interstrand G-quadruplex forming potential. *Methods*, **57**, 3–10.
- Williamson,J.R. (1994) G-tetrad structures in telomeric DNA. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 703–730.
- Postberg,J., Tsytlonok,M., Sparvoli,D., Rhodes,D. and Lipps,H.J. (2012) A telomerase-associated RecQ protein-like helicase resolves telomeric G-quadruplex structures during replication. *Gene*, **497**, 147–154.
- Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
- Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress *c-MYC* transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11593–11598.
- Balasubramanian,S., Hurley,L.H. and Neidle,S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261–275.
- Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
- Bugaut,A. and Balasubramanian,S. (2012) 5'-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Res.*, **40**, 4727–4741.
- Katapadi,V.K., Nambiar,M. and Raghavan,S.C. (2012) G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, **100**, 72–80.
- Dong,D.W., Pereira,F., Barrett,S.P., Kolesar,J.E., Cao,K., Damas,J., Yatsunyk,L.A., Johnson,F.B. and Kaufman,B.A. (2014) Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics*, **15**, 677.
- Bharti,S.K., Sommers,J.A., Zhou,J., Kaplan,D.L., Spelbrink,J.N., Mergny,J.-L. and Brosh,R.M. Jr (2014) DNA sequences proximal to human mitochondrial DNA deletion breakpoints prevalent in human disease form G-quadruplexes, a class of DNA structures inefficiently unwound by the mitochondrial replicative twinkle helicase. *J. Biol. Chem.*, **289**, 29975–29993.
- Bose,P., Hermetz,K.E., Conneely,K.N. and Rudd,M.K. (2014) Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. *PLoS ONE*, **9**, e101607.
- Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
- Henderson,A., Wu,Y., Huang,Y.C., Chavez,E.A., Platt,J., Johnson,F.B., Brosh,R.M. Jr, Sen,D. and Lansdorf,P.M. (2013) Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.*, **42**, 860–869.
- Biffi,G., Tannahill,D., Miller,J., Howat,W.J. and Balasubramanian,S. (2014) Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PLoS One*, **9**, e102711.
- Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of *c-myc*. *Nucleic Acids Res.*, **26**, 1167–1172.
- Cogoi,S., Paramasivam,M., Membrino,A., Yokoyama,K.K. and Xodo,L.E. (2010) The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element. *J. Biol. Chem.*, **285**, 22003–22016.
- Wei,D., Todd,A.K., Zloh,M., Gunaratnam,M., Parkinson,G.N. and Neidle,S. (2013) Crystal structure of a promoter sequence in the B-raf gene reveals an intertwined dimer quadruplex. *J. Am. Chem. Soc.*, **135**, 19319–19329.
- Rodriguez,R., Miller,K.M., Forment,J.V., Bradshaw,C.R., Nikan,M., Britton,S., Oelschlaegel,T., Xhemalce,B., Balasubramanian,S. and Jackson,S.P. (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.*, **8**, 301–310.
- Dai,J., Chen,D., Jones,R.A., Hurley,L.H. and Yang,D. (2006) NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res.*, **34**, 5133–5144.
- Agrawal,P., Lin,C., Mathad,R.I., Carver,M. and Yang,D. (2014) The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K⁺-solution. *J. Am. Chem. Soc.*, **136**, 1750–1753.
- Rankin,S., Reszka,A.P., Huppert,J., Zloh,M., Parkinson,G.N., Todd,A.K., Ladame,S., Balasubramanian,S. and Neidle,S. (2005) Putative DNA quadruplex formation within the human *c-kit* oncogene. *J. Am. Chem. Soc.*, **127**, 10584–10589.
- Fernando,H., Reszka,A.P., Huppert,J., Ladame,S., Rankin,S., Venkitaraman,A.R., Neidle,S. and Balasubramanian,S. (2006) A conserved quadruplex motif located in a transcription activation site of the human *c-kit* oncogene. *Biochemistry*, **45**, 7854–7860.
- Phan,A.T., Kuryavyi,V., Burge,S., Neidle,S. and Patel,D.J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human *c-kit* promoter. *J. Am. Chem. Soc.*, **129**, 4386–4392.
- Hsu,S.T., Varnai,P., Bugaut,A., Reszka,A.P., Neidle,S. and Balasubramanian,S. (2009) 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**, 13399–13409.

30. Kuryavii, V., Phan, A.T. and Patel, D.J. (2010) Solution structures of all parallel-stranded monomeric and dimeric G-quadruplex scaffolds of the human c-kit2 promoter. *Nucleic Acids Res.*, **38**, 6757–6773.
31. Todd, A.K., Haider, S.M., Parkinson, G.N. and Neidle, S. (2007) Sequence occurrence and structural uniqueness of a G-quadruplex in the human c-kit promoter. *Nucleic Acids Res.*, **35**, 5799–5808.
32. Wei, D., Parkinson, G.N., Reszka, A.P. and Neidle, S. (2012) Crystal structure of a c-kit promoter quadruplex reveals the structural role of metal ions and water molecules in maintaining loop conformation. *Nucleic Acids Res.*, **40**, 4691–4700.
33. Neidle, S. (2012) *Therapeutic Applications of Quadruplex Nucleic Acids*. Academic Press, London.
34. Bidzinska, J., Cimino-Reale, G., Zaffaroni, N. and Folini, M. (2013) G-quadruplex structures in the human genome as novel therapeutic targets. *Molecules*, **18**, 12368–12395.
35. Boddupally, P.V., Hahn, S., Beman, C., De, B., Brooks, T.A., Gokhale, V. and Hurley, L.H. (2012) Anticancer activity and cellular repression of c-MYC by the G-quadruplex-stabilizing 11-piperazinyloquinoline is not dependent on direct targeting of the G-quadruplex in the c-MYC promoter. *J. Med. Chem.*, **55**, 6076–6086.
36. Yarden, Y., Kuang, W.J., Yang-Feng, T., Coussens, L., Munemitsu, S., Dull, T.J., Chen, E., Schlessinger, J., Francke, U. and Ullrich, A. (1987) Human proto-oncogene C-Kit—a new cell-surface receptor tyrosine kinase for an unidentified ligand. *EMBO J.*, **6**, 3341–3351.
37. Fletcher, J.A. and Rubin, B.P. (2007) KIT mutations in GIST. *Curr. Opin. Genet. Dev.*, **17**, 3–7.
38. Nannini, M., Biasco, G., Astolfi, A. and Pantaleo, M.A. (2013) An overview on molecular biology of KIT/PDGFR α wild type (WT) gastrointestinal stromal tumours (GIST). *J. Med. Genet.*, **50**, 653–661.
39. Bejugam, M., Sewitz, S., Shirude, P.S., Rodriguez, R., Shahid, R. and Balasubramanian, S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression. *J. Am. Chem. Soc.*, **129**, 12926–12927.
40. Gunaratnam, M., Swank, S., Haider, S.M., Galesa, K., Reszka, A.P., Beltran, M., Cuenca, F., Fletcher, J.A. and Neidle, S. (2009) Targeting human gastrointestinal stromal tumor cells with a quadruplex-binding small molecule. *J. Med. Chem.*, **52**, 3774–3783.
41. Bejugam, M., Gunaratnam, M., Müller, S., Sanders, D.A., Sewitz, S., Fletcher, J.A., Neidle, S. and Balasubramanian, S. (2010) Targeting the c-Kit Promoter G-quadruplexes with 6-Substituted Indenoisoquinolines. *ACS Med. Chem. Lett.*, **1**, 306–310.
42. McLuckie, K.I., Waller, Z.A., Sanders, D.A., Alves, D., Rodriguez, R., Dash, J., McKenzie, G.J., Venkitaraman, A.R. and Balasubramanian, S. (2011) G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells. *J. Am. Chem. Soc.*, **133**, 2658–2663.
43. Wang, X., Zhou, C.-X., Yan, J.-W., Hou, J.-Q., Chen, S.-B., Ou, T.-M., Gu, L.-Q., Huang, Z.-S. and Tan, J.-H. (2013) Synthesis and evaluation of quinazolone derivatives as a new class of c-KIT G-quadruplex binding ligands. *ACS Med. Chem. Lett.*, **4**, 909–914.
44. Castor, K.J., Liu, Z., Fakhoury, J., Hancock, M.A., Mittermaier, A., Moitessier, N. and Sleiman, H.F. (2013) A platinum(II) phenylphenanthroimidazole with an extended side-chain exhibits slow dissociation from a c-Kit G-quadruplex motif. *Chemistry*, **19**, 17836–17845.
45. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A. et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr.*, **D67**, 235–242.
46. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr.*, **D66**, 486–501.
47. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
48. Vagin, A.A., Steiner, R.A., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F. and Murshudov, G.N. (2004) REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr.*, **D60**, 2184–2195.
49. Petersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comp. Chem.*, **25**, 1605–1612.
50. DeLano, W.L. (2008) The PyMOL Molecular Graphics System.
51. Hess, B., Kutzner, C., van der Spoel, D. and Lindahl, E. (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.
52. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T.E., Laughton, C.A. and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
53. Nasiri, H.R., Bell, N.M., McLuckie, K.I., Husby, J., Abell, C., Neidle, S. and Balasubramanian, S. (2014) Targeting a c-MYC G-quadruplex DNA with a fragment library. *Chem. Commun. (Camb.)*, **50**, 1704–1707.
54. Husby, J., Todd, A.K., Platts, J.A. and Neidle, S. (2013) Small-molecule G-quadruplex interactions: systematic exploration of conformational space using multiple molecular dynamics. *Biopolymers*, **99**, 989–1005.
55. Park, H.S. and Jun, C.H. (2008) A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.*, **36**, 3336–3341.
56. Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W. and Cavalli, A. (2014) The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun.*, in press.
57. Cang, X., Šponer, J. and Cheatham, T.E. III (2011) Insight into G-DNA structural polymorphism and folding from sequence and loop connectivity through free energy analysis. *J. Am. Chem. Soc.*, **133**, 14270–14279.
58. Cang, X., Šponer, J. and Cheatham, T.E. III (2011) Explaining the varied glycosidic conformational, G-tract length and sequence preferences for anti-parallel G-quadruplexes. *Nucleic Acids Res.*, **39**, 4499–4512.
59. Šponer, J., Cang, X. and Cheatham, T.E. III (2012) Molecular dynamics simulations of G-DNA and perspectives on the simulation of nucleic acid structures. *Methods*, **57**, 25–39.
60. Islam, B., Sgobba, M., Laughton, C., Orozco, M., Šponer, J., Neidle, S. and Haider, S. (2013) Conformational dynamics of the human propeller telomeric DNA quadruplex on a microsecond time scale. *Nucleic Acids Res.*, **41**, 2723–2735.
61. Šponer, J., Mládek, A., Špačková, N., Cang, X., Cheatham, T.E. III and Grimme, S. (2013) Relative stability of different DNA guanine quadruplex stem topologies derived using large-scale quantum-chemical computations. *J. Am. Chem. Soc.*, **135**, 9785–9796.
62. Zgarbová, M., Luque, F.J., Šponer, J., Cheatham, T.E. III, Otyepka, M. and Jurečka, P. (2013) Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput.*, **9**, 2339–2354.
63. Zhu, H., Xiao, S. and Liang, H. (2013) Structural dynamics of human telomeric G-quadruplex loops studied by molecular dynamics simulations. *PLoS ONE*, **8**, e71380.
64. Stebbins, J.D., Lunec, J. and Larcombe, L.D. (2012) An in silico study of the differential effect of oxidation on two biologically relevant G-quadruplexes: possible implications in oncogene expression. *PLoS One*, **7**, e43735.
65. Warner, K.D., Homan, P., Weeks, K.M., Smith, A.G., Abell, C. and Ferré-D'Amaré, A.R. (2014) Validating fragment-based drug discovery for biological RNAs: lead fragments bind and remodel the TPP riboswitch specifically. *Chem. Biol.*, **21**, 591–595.
66. Wang, Y. and Patel, D.J. (1993) Solution structure of the human telomeric repeat d[AG₃(T₂AG₃)₃] G-tetraplex. *Structure*, **1**, 263–282.
67. Phan, A.T. and Patel, D.J. (2003) Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *J. Am. Chem. Soc.*, **125**, 15021–15027.
68. Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R.A. and Yang, D. (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.*, **34**, 2723–2735.
69. Dai, J., Carver, M., Punchihewa, C., Jones, R.A. and Yang, D. (2007) Structure of the hybrid-2 type intramolecular human telomeric G-quadruplex in K⁺ solution: insights into structure polymorphism of the human telomeric sequence. *Nucleic Acids Res.*, **35**, 4927–4740.
70. Lim, K.W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D.J., Luu, K.N. and Phan, A.T. (2009) Structure of the human telomere in K⁺ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. *J. Am. Chem. Soc.*, **131**, 4301–4309.

71. Luu, K.N., Phan, A.T., Kuryavyi, V., Lacroix, L. and Patel, D.J. (2006) Structure of the human telomere in K^+ solution: an intramolecular (3+1) G-quadruplex scaffold. *J. Am. Chem. Soc.*, **128**, 9963–9970.
72. Parkinson, G.N., Lee, M.P.H. and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876–880.
73. Phan, A.T., Kuryavyi, V., Luu, K.N. and Patel, D.J. (2007) Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K^+ solution. *Nucleic Acids Res.*, **35**, 6517–6525.
74. Phan, A.T., Luu, K.N. and Patel, D.J. (2006) Different loop arrangements of intramolecular human telomeric (3+1) G-quadruplexes in K^+ solution. *Nucleic Acids Res.*, **34**, 5715–5719.
75. Renčičuk, D., Kejnovská, I., Školáková, P., Bednářová, K., Motlová, J. and Vorlíčková, M. (2009) Arrangements of human telomere DNA quadruplex in physiologically relevant K^+ solutions. *Nucleic Acids Res.*, **37**, 6625–6634.
76. Heddi, B. and Phan, A.T. (2013) Structure of human telomeric DNA in crowded solution. *J. Am. Chem. Soc.*, **133**, 9824–9833.
77. Shirude, P.S., Okumus, B., Ying, L., Ha, T. and Balasubramanian, S. (2007) Single-molecule conformational analysis of G-quadruplex formation in the promoter DNA duplex of the proto-oncogene c-kit. *J. Am. Chem. Soc.*, **129**, 7484–7485.
78. Kuryavyi, V., Cahoon, L.A., Seifert, H.S. and Patel, D.J. (2012) RecA-binding pilE G4 sequence essential for pilin antigenic variation forms monomeric and 5' end-stacked dimeric parallel G-quadruplexes. *Structure*, **20**, 2090–2102.
79. Adrian, M., Ang, D.J., Lech, C.J., Heddi, B., Nicolas, A. and Phan, A.T. (2014) Structure and conformational dynamics of a stacked dimeric G-quadruplex formed by the human CEB1 minisatellite. *J. Am. Chem. Soc.*, **136**, 6297–6305.
80. Waller, Z.A., Howell, L.A., Macdonald, C.J., O'Connell, M.A. and Searcey, M. (2014) Identification and characterisation of a G-quadruplex forming sequence in the promoter region of nuclear factor (erythroid-derived 2)-like 2 (Nrf2). *Biochem. Biophys. Res. Commun.*, **447**, 128–132.
81. Martadinata, H. and Phan, A.T. (2014) Formation of a stacked dimeric G-quadruplex containing bulges by the 5'-terminal region of human telomerase RNA (hTERC). *Biochemistry*, **53**, 1595–1600.