UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of *Chemical Sciences*

Ph.D. COURSE IN MOLECULAR SCIENCES

CURRICULUM: Pharmaceutical Sciences – XXXI CYCLE

THE FINE ARCHITECTURE OF GUANINE-RICH REGIONS WITHIN ONCOGENE PROMOTERS

Coordinator: Ch.mo Prof. Leonard Prins

Supervisor: Ch.ma Prof.ssa Claudia Sissi

Ph.D. student: Riccardo Rigo

ABSTRACT

Suppression of oncogenes transcription represents an ideal tool to integrate the currently available therapeutics to treat several cancer types and to overcome the potential occurrence of resistance. An experimentally validated mechanism of intervention is represented by the induction/stabilization of G-quadruplex structures in genes promoter by small molecules. G-quadruplexes are DNA non-canonical secondary structures consisting of stacked G-quartets, cyclic arrangements of four guanine residues held together by Hoogsteen hydrogen bonds and stabilized by a central cation. At the moment, none of the identified G-quadruplex ligands reached the clinic. Several reasons can contribute to this poor outcome comprising both the plastic structural features of nucleic acids and the multiple metabolic pathways which might be affected when a small molecule interacts with G-quadruplex structures. Thus, one of the major issues lies on the proper structural analysis of targeted G-quadruplex.

To overcome this bias, in this Ph.D.'s thesis kinetic and thermodynamic behaviours of G-quadruplexes have been characterized to obtain an improved description of these structures as potential pharmaceutical targets. The study has been focused on G-rich sequences within c-KIT and EGFR oncogene promoters.

By applying a set of complementary structural and biophysical approaches, the folding pathways of these G-quadruplexes and influence of flanking regions in terms of structural stability and folding rearrangement have been described. The obtained information indicates that the promoter architecture might be not properly derived by analysis of minimal G-quadruplex forming sequences at the thermodynamic equilibrium, commonly used for screening assayes. Indeed, reported data suggest the existence of different unique mechanisms/pathways involved in the regulation of these oncogenes transcription which comprise kinetically favored folding intermediates or unprecedented structural arrangements.

The final outcome of Ph.D.'s project is a deeper understanding of nucleic acid tridimensional arrangement of EGFR and c-KIT promoters which might help in setting up new drug-design programs based on models of G-quadruplex target more closely related to the physiological ones.

TABLE OF CONTENTS

1. Preface

G-quadruplexes are non-canonical tetra-helices formed in guanine-rich regions of the genome. The building block of G-quadruplex is the G-quartet or G-tetrad, a planar array of four guanines interacting one to the other through H-bonds in Hoogsteen geometry (Figure 1).

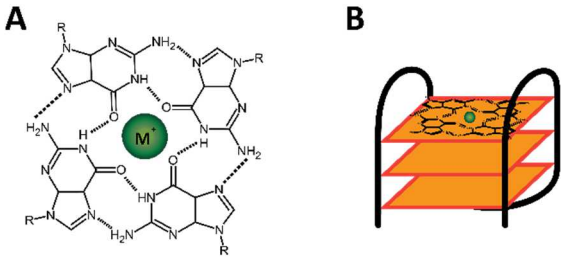


Figure 1. Building blocks of G-quadruplex structures. A) Four guanine residues bond through Hoogsteen base pairs forming a G-quartet. M^+ stands for monovalent cation, as K^+ or Na^+ ions. B) Schematic representation of a G-quadruplex structure.

Stacking interactions between two or more G-quartets constitute the core of G-quadruplex structures (named G-core) (Figure 2B). Carbonyl group of each guanine residue faces the internal cavity of the G-quartet making it strongly electronegative. Coordination of monovalent cations counterbalance this electron-density which in principle could lead to unfolding of the structure. Ionic and crowding conditions as well as the presence of small molecule ligands or G-quadruplex-binding proteins influence the stability of these DNA arrangements.

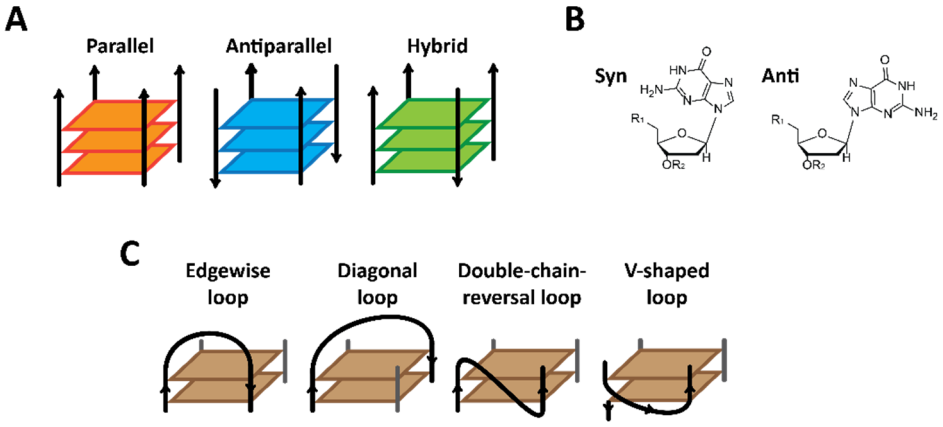


Figure 2. Schematic structural features of G-quadruplexes. A) Strands directionality. B) Syn/Anti base orientation of residues. C) Loop arrangements.

In a structural point of view, G-quadruplexes can be roughly divided into three main topologies depending on the relative orientation of the four strands (Figure 2A): parallel

structures which strands share the same orientation; antiparallel G-quadruplexes in which two strands show opposite directionality then the others; and hybrid (3+1) structures which present only one strand with opposite orientation. The torsion angle of glycosidic bond between nucleotide base and deoxyribose sugar pushes each residue into syn or anti conformation (Figure 2B). Syn/anti bases orientation and strands directionality contribute to define four grooves dimensions (narrow, medium or large size).

Another degree of structural complexity is provided by loops arrangement. They connect different strands in four principal manners (Figure 2C): edgewise loops, connecting adjacent strands with opposite directionality; diagonal loops, bridging two opposing anti-parallel strands; double-chain-reversal loops, connecting neighboring strands with same directionality; V-shaped loops, bridging two corners of the G-core in which one G-column is missing. Additionally, sequence composition of loops can increase structural diversity due to formation of base pairing within loops or interaction between loops and G-core.

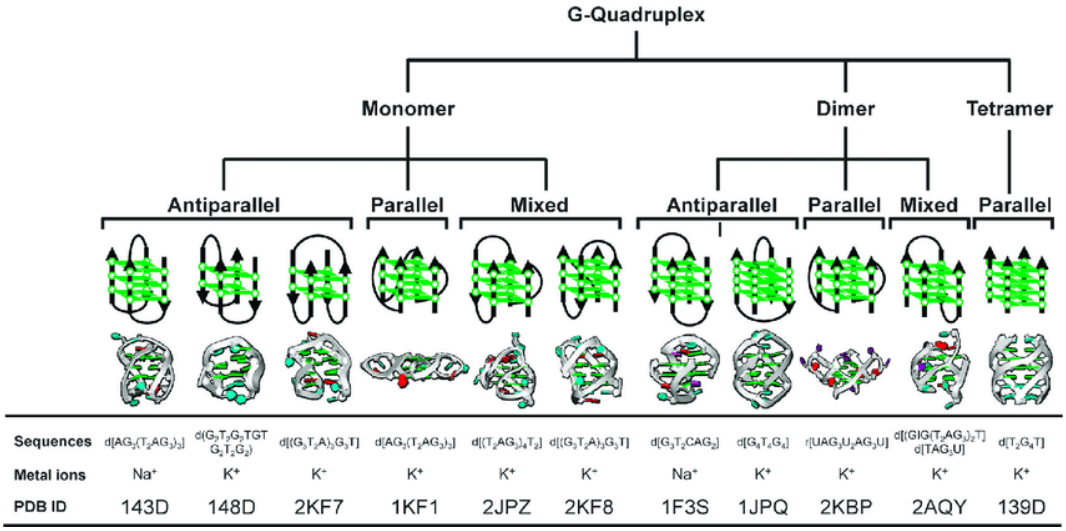


Figure 3. Examples of G-quadruplex polymorphisms (from Yaku H., *Chem. Commun.*, 2012, **48**, 6203-6216).

As a result, all these features determine a huge level of polymorphism among G-quadruplex structures (Figure 3) which can account for their different function roles and which can be selectively used to make them suitable therapeutic targets.

In the next chapters, I am going to present a structural perspective on different G-quadruplexes formed in promoter regions of two oncogenes, c-KIT and EGFR, focusing on their folding kinetics and thermodynamics.

2. Introduction

G-quadruplexes in human promoters: a challenge for therapeutic applications

Riccardo Rigo, Manlio Palumbo and Claudia Sissi*

Dept. of Pharmaceutical and Pharmacological Sciences, University of Padova, v. Marzolo, 5, 35131, Padova, Italy

* claudia.sissi@unipd.it

ABSTRACT

Background: G-rich sequences undergo unique structural equilibria to form G-quadruplexes (G4) both *in vitro* and *in cell* systems. Several pathologies emerged to be directly related to G4 occurrence at defined genomic portions. Additionally, G-rich sequences are significantly represented around transcription start sites (TSS) thus leading to the hypothesis of a gene regulatory function for G4. Thus, the tuning of G4 formation has been proposed as a new powerful tool to regulate gene expression to treat related pathologies. However, up-to date this approach did not provide any new really efficient treatment.

Scope of Review: Here, we summarize the most recent advances on the correlation between the structural features of G4 in human promoters and the role these systems physiologically exert. In particular we focus on the effect of G4 localization among cell compartments and along the promoters in correlation with protein interaction networks and epigenetic state. Finally, the intrinsic structural features of G4 at promoters are discussed to unveil the contribution of different G4 structural modules in this complex architecture.

Major Conclusions: It emerges that G4s play several roles in the intriguing and complex mechanism of gene expression, being able to produce opposite effects on the same target. This reflects the occurrence of a highly variegate network of several components working simultaneously.

General Significance: The resulting picture is still fuzzy, but some points of strength are definitely emerging, which prompts all of us to strengthen our efforts in view of a selective control of gene expression through G4 modulation.

Published – Rigo, R.; Palumbo, M.; Sissi, C. *Biochim. Biophys. Acta*. **2017**, 1861 (5 Pt B), 1399-1413

1. INTRODUCTION

G-quadruplexes (G4) are non-canonical tetrahelical nucleic acid structures in which the basic structural element is a G-tetrad formed by the association of four guanines through Hoogsteen hydrogen bonding. These nucleobases can be isolated or can derive from G-rich sequences (potential G4 forming sequences, PQS), thus opening the chance for G4 occurrence in cells.

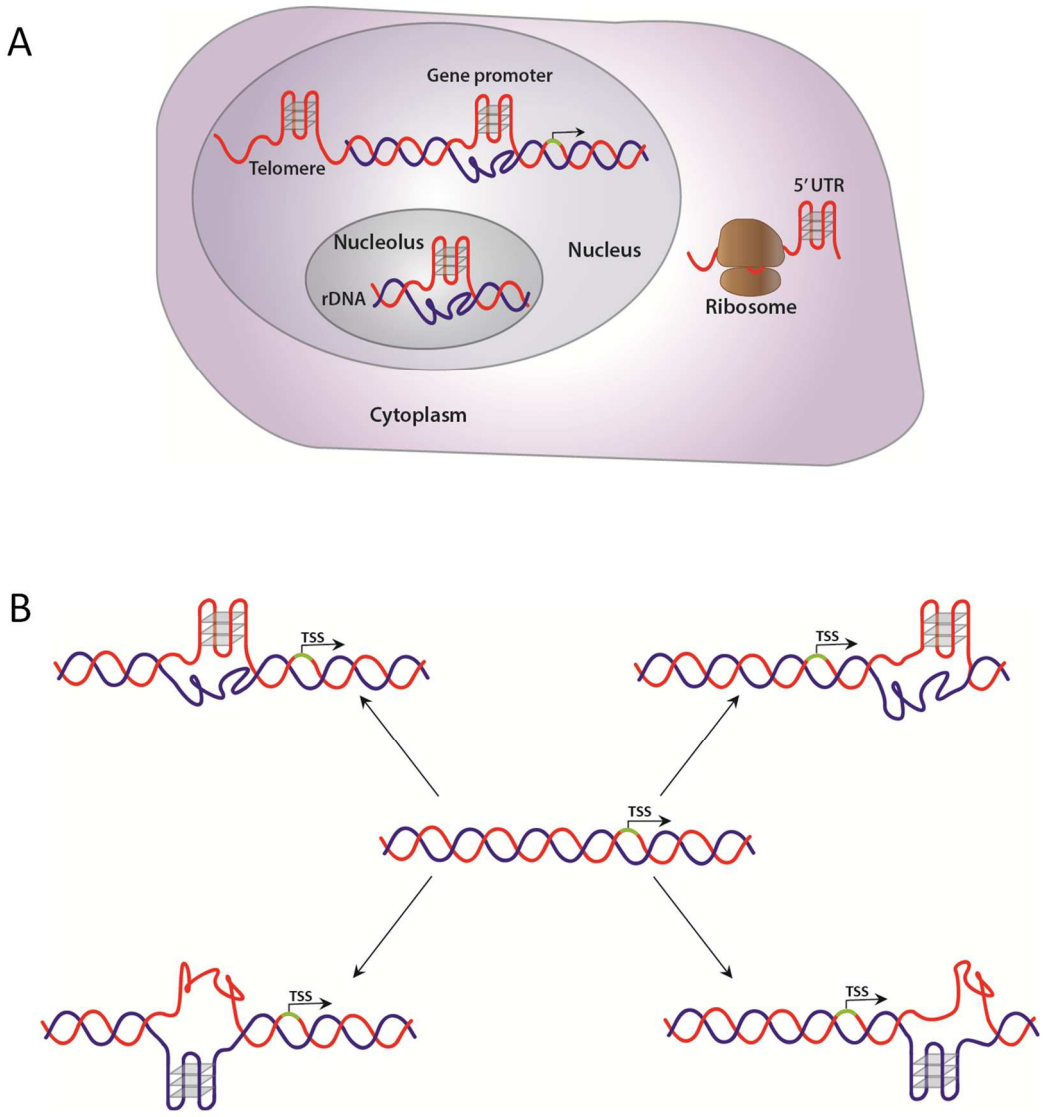


Figure 1. Distribution of potential G4 structures among different cellular compartments (PANEL A) and their possible location around TSS (PANEL B) [3].

The formation of G4 arrangements in physiologically relevant DNA sequences has been initially proposed to occur at telomeric level where G-rich repeats are present within an extended 3' single-strand protruding portion. This unique condition was considered suitable to allow the fold-back required for the formation of the tetrahelical intramolecular arrangement. Later on, G-rich sequences were identified in other portions of the genome (Figure 1A). Interestingly, bioinformatic analyses showed that they are not randomly distributed along the genome but cluster at defined regions like immunoglobulin switch regions, recombination sites and gene promoters [1,2]. In particular, the significant PQS enrichment around transcription start sites (TSS) lead to the hypothesis of a gene regulatory function for G4 [3]. Soon, studies with mutated promoters supported the importance of these sequences to control gene expression. However, the ultimate attribution of a transcription regulatory role to the conformational shift from the canonical double helix to the tetrahelical arrangement (and not merely to the presence of a G-rich sequence) derived only later, when formation of G4 at promoters was observed in living cells by multiple approaches [4,5,6,7].

Up-to-date, the G4-mediated regulation of transcription has been confirmed for genes involved in differentiation, in cancer progression and even in metabolic regulation. Interestingly, this mechanism appears to be a conserved feature not only in vertebrates but also in several microorganisms ranging from yeast, to bacteria down to viruses with a rapid evolutionary expansion starting in metazoan organisms [8,9,10,11,12]. Furthermore, recent studies highlight precise connections between G4 nucleic acid arrangements and several diseases, as a consequence of mis-regulation of gene expression as well as of genetic and epigenetic instability arising from difficult processing of non-canonical DNA structures [13]. The list of putative G4-related diseases is quite impressive. Moreover, the presence of G4 motifs in important pharmacogenes encoding metabolic enzymes, receptor proteins, folate transporter and proteins involved in potassium voltage-gated and sodium channels significantly impacts pharmacogenomics and drug development projects [14]. As a result, the emerging roles for G4 at gene promoters are highly variegate and with impressive therapeutic implications.

Although PQS distribution around TSS is clearly connected to G4 as transcription controllers, at the moment it is hard to predict the functional effects of the tetrahelical formation. G4s in proximal promoters mostly cause a down-regulation of gene expression but several examples in which they induce an up-regulation are reported [15]. A closer

look at PQS distribution showed that they can be located on the sense or the antisense strand, in front or behind TSS, in UTR as well as introns (Figure 1B). We expected that this distribution reflects distinct physiological effects, but it appears to be insufficient to safely predict them [16]. Consequently, a huge validation *in vitro* and *in cell* is required in order to consider each PQS as a good potential pharmacological target.

In addition, only if we are able to fully understand the processes ruling the structural shift of a double helix toward a G4, we can design proper tools to modulate it according to our needs. Focusing on the consequences on gene expression, this approach can be transformed into a novel therapeutic strategy. If we consider that for any hallmark of cancer one of more factors contain G-rich sequences in the promoter, the relevance for medicinal applications of G4-directed small molecules is striking. At the same time, this would represent a highly valuable molecular biology methodology to dissect the interconnection of several pathways in order to identify, ultimately, therapeutically most important sites of intervention.

The exponentially growing number of papers concerning G4 in promoters clearly reflects the willingness to turn on the light on these potentials. Thus, studies on G4 structure, stability, ligands recognition, protein recruitment and so on are rapidly accumulating. However, as it often happens in sciences, the more we discover the more unclear matter emerges. This review is not going to provide the reader a detailed report of all the main achievements on which we can finally refer on. Several recent outstanding reviews already cover these topics [17,18,19,20,21,22,23]. Conversely, we aim at using the most recent experiences to highlight the main issues concerning the correlations between the structural features of G4 formation in promoters and the role these systems physiologically exert.

1.1 An historical leading example: c-MYC promoter

In order to highlight the main issues concerning the comprehension of the molecular mechanisms related to a direct control of gene expression through a modulation of the promoter conformational state, a good overview is provided by *c-MYC*. It represents one of the first oncogene whose promoter was extensively investigated in connection to G4 formation but, actually, the interest for this genomic portion was anticipated by studies aimed at understanding the mechanisms of regulation of *c-MYC* expression in various

tumors. Chromatin mapping identified more than one DNAse I hypersensitive sites at the 5' of the human *c-MYC* gene [24]. Among them, the nuclease-hypersensitive element III_1 (NHE III_1) located upstream the P1 promoter received great attention since it is the preferred transcription origin and controls 80-90% of gene transcription [25]. Remarkably, only an unwound (or non-B) form of NHE III_1 was efficiently cleaved by nucleases, whereas the B-helix was not: this suggested a controversial connection between DNA structure and regulatory signal(s). The equilibrium between these two DNA conformational states was found to be slow and shifted towards the nucleases-sensitive conformation by helix supercoiling [26].

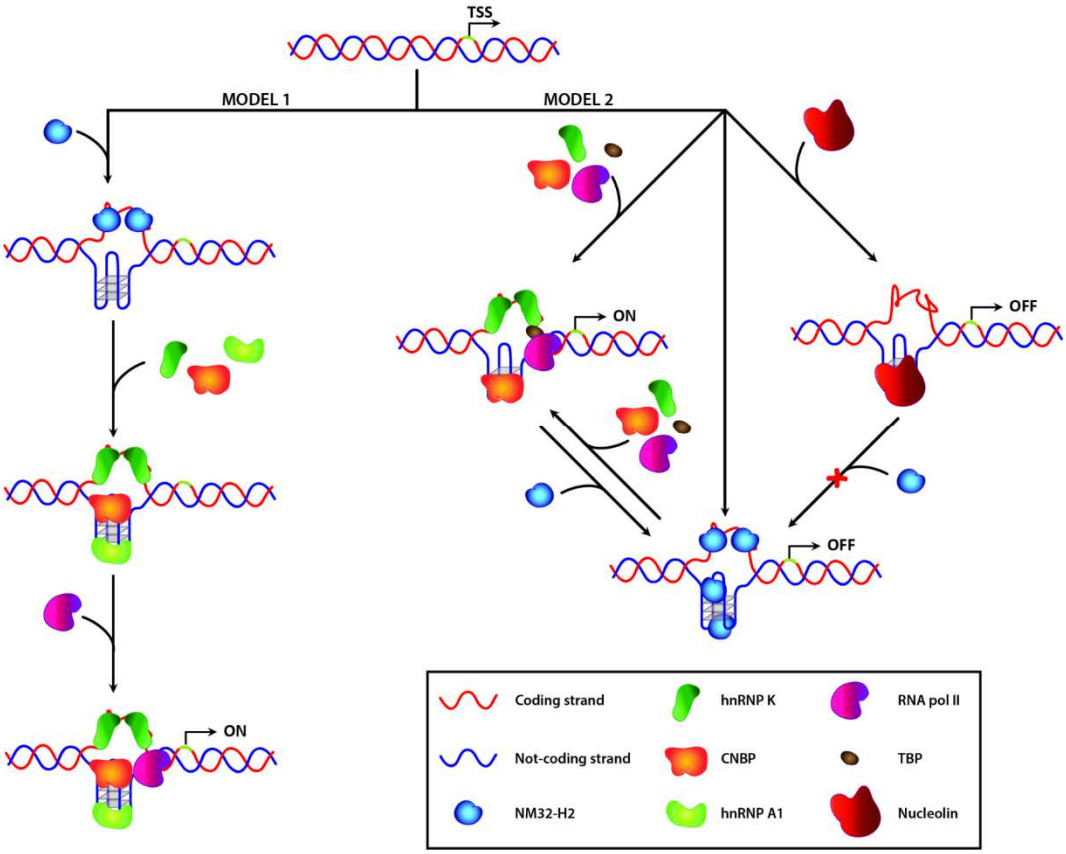


Figure 2. Schematic representation of the two models proposed for the regulation of c-MYC expression by G-quadruplex.

Notably, NHE III_1 comprises a pyrimidine-rich coding strand and a purine-rich noncoding strand and it was found to represent an arrest site for polymerase with an efficiency directly related to KCl concentration, thus suggesting a G4 arrangement for the nuclease-sensitive form. This was in line with chemical footprinting data of a 27-bp G-rich sequence

(Pu27) corresponding to the fragment located between position –142 and -115 upstream the P1 promoter in the presence of potassium: they fitted with an intra-strand fold-back DNA tetraplex formed by a core of 3 G-quartets further stabilized by two AT bp terminal capping domains [27]. Although all experimental evidences pointed to the formation of a G4 structure in the purine-rich strand, it was not clear how this structural element could regulate *c-MYC* transcription. The design of a model connecting the solved DNA structure and gene expression was preliminary prompted by the concomitant identification of a relevant number of proteins able to bind homopyrimidine or homopurine strands. Among them, those showing a significant sequence-selective affinity for the NHE III_1 in human *c-MYC* were sorted and used to fill the puzzle. This provided the model summarized in Figure 2 (MODEL 1). Accordingly, the first critical step is the recruitment of a helicase (Nucleotide diphosphate kinase B, NDPK-B or NM23-H2) to unwind the double helix and trap the homopyrimidine strand. This is important to separate the G-rich strand to let it free to fold into a G4. The resulting tetrahelix is further stabilized by other G4 selective proteins (i.e. cellular nucleic acid binding protein, CNBP, and heterogeneous nuclear ribonucleoprotein A1, hnRNP A1). At this point, the unpaired C-rich loop can recruit transcription factors with high affinity for the homopyrimidine strand (i.e. heterogeneous nuclear ribonucleoprotein K, hnRNP K) to ultimately attract the RNA polymerase machinery. The overall outcome is expected to be an increment of transcription. According to this picture, the use of competitor G4 (decoy, aptamers) was envisaged to trap the above-mentioned proteins and, consequently, to repress *c-MYC* expression.

The advantage of this first theoretical model rested in the conversion of the double-stranded G-rich segment in the promoter into a system more similar to the telomeric overhang where G4 formation was already confirmed. However, the requirement for specific proteins to regulate process apparently reduced the therapeutic targets. Indeed, proteins properly work only at a subset of sequences thus, in principle, excluding some G-rich sequences as potential sites of intervention irrespectively from their PQS content. This was soon overcome by the discovery that small molecules able to bind and stabilize *c-MYC* G4 were actually able to shift NHE III_1 towards the tetrahelical form even in the absence of proteins and with an efficiency sequence-dependent, thus opening the chance for targeted intervention [28,29]. Actually, these small molecules were also extremely valuable as molecular probes to clarify physiological effects of G4 on *c-MYC* expression and to highlight the structural behavior of NHE III_1.

In particular, the use of G4 binders confirmed G4 as a suppressor of gene expression This was in contrast to the first model but in agreement with *in cell* data acquired in the presence of properly designed constructs containing the *c-MYC* promoters in its wt or mutated form: thus, a second model was derived (Figure 2, MODEL 2). A critical analysis of these two models is useful for our purposes.

First of all, in both models, G4 regulates gene expression through a fine tuning of the recruitment of the transcriptional machinery. This differs from the original approach proposed for telomerase inhibition upon G4 formation where the tetra-helix was considered sufficient to block the enzymatic activity. Thus, the role of proteins is as important as the one of G4.

Interestingly, the two models derive from a different role attributed to NM23-H2, the helicase, with high affinity for the homopyrimidine strand. In the first model this protein was assumed to work as a transcription activator [30]. Conversely, upon ranking all the above mentioned proteins according to their affinity for specific DNA secondary structures, only an accessory role was attributed to NM23-H2, whereas CNBP and hnRNP K were indicated as the real regulatory effectors for transcriptional activation [31]. This means that modest variation in affinity or activity can dramatically alter the overall output in such complex systems.

Pu27 5’ – TGGGGAGGGTGGGGAGGGTGGGGAAGG – 3’

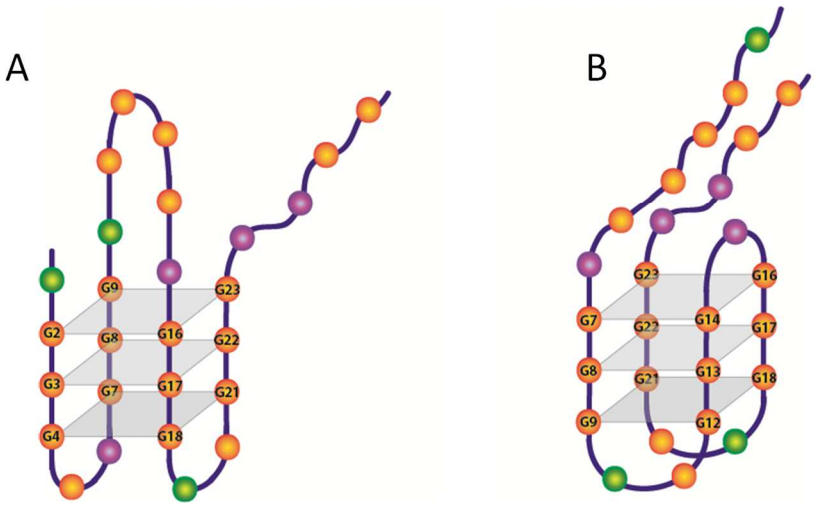


Figure 3. Models describing the basket (A) and the chair (B) G-quadruplex arrangements assumed by the c-MYC promoter sequence Pu27 in potassium containing solutions.

From a structural point of view, pioneering studies performed by Hurley with a perylene derivatives, PIPER, indicated the formations of two different secondary structures starting from the double helix form of NHE: a kinetically favored tetramolecular G4 and a thermodynamically favored intramolecular G4 [28]. Moreover, even in the absence of ligands, the presence of six G-tracts in Pu27 allows the formation of different intramolecular folded G4 structures among which two species appear to be the most relevant in solution (Figure 3) [29]:

- One "basket" form with tetrads formed by guanines in the 1st, 2nd, 4th and 5th G-rich tracts (from 5')
- One "chair" form, kinetically favored, derived from pairing of guanines of the 2nd, 3rd, 4th and 5th G-tracts. Mutational studies and luciferase reporter constructs indicated this form as the principal responsible for the control of *c-MYC* expression. As recently supported by single molecule techniques, this form is actually the most abundant in solution. Nevertheless, inclusion in the study of mutated and truncated sequences provided clear indication that the Gs in the first tract as well as the last two may contribute to finely tuning the kinetic and structural features of this folded species [32,33].

It is worth to note that several concerns are still present on the relevance of one single form in controlling the overall transcription process. Indeed, NHE III_1 contains more G-tracts than those considered in the Pu27 fragment: in principle, all of them can be involved in the formation of secondary structures with different effects on gene transcription [24,34,35]. In addition, it is not possible to rule out that the cellular environments, as well as the insertion of the G-rich tract into a long nucleic acid context, can actually promote the formation of G-tetrads deriving from the pairing of different guanines in comparison to those identified in a short single stranded sequence. As an example, this was experimentally observed by changing the DNA topology of the double helix containing the G-rich substrate [36].

To summarize what we learned from the *c-MYC* experience is that, in order to understand the multifaceted effects of G4, we have to keep a constant exchange of biophysical, biological, cellular and computational evidences focused at two main levels:

- the first one refers to an outlook of the complex cellular environment in which PQS are immersed in order to define the intricate network affecting G4 inside the cell;

- the second one is more strictly related to a fine description of the structural features of "physiological" G4, to delineate clearer structure-function correlations.

Their combination provides the fundamental hints to transform G-quadruplexes into successful targets for therapeutic intervention. These issues will be dissected in the following sections.

2. EFFECTS OF G4 FORMATION ON GENE TRANSCRIPTION: the contribution of the cellular environment

As above described, the rational explanation of cellular effects of G4 formation within *c-MYC* promoter was largely derived from a detailed exploration of proteins functions at this nuclear region. At G-rich sites, the two macromolecular components exert concomitant activities: proteins are important to modulate the nucleic acid structural equilibria as well as the G4 works as a "selector" for protein recruitment. This plastic interconnection might help to address the major limits of the G4 targeting approach. Indeed, data so far collected on several promoters containing G-rich domains confirmed that both up- or down-regulation can be promoted according to the system under investigation. The resulting outcome is even harder to be described since the modulation of one gene expression can cause a wide deregulation of several pathways apparently not directly related, through multifaceted loops.

A recent well described example is represented by the *RAS* family. The promoter region of KRAS contains multiple G-rich sequences. As confirmed by reporter gene assays each of them contributes to different extent to the suppression of gene expression according to its ability to fold into G4 [37]. However, in human pancreatic cancer cells, a more articulated mechanism involving multiple cellular pathways was dissected. In particular, it was observed that the oncogenic KRAS signaling is inserted in a complex loop in which Integrin-linked kinase (ILK) works both as downstream effector of KRAS and a regulator of *KRAS* expression through the recruitment of E2F1 (which supports KRAS-induced expression of ILK) and of hnRNP A1 (that destabilizes G4 in *KRAS* promoter) [38,39]. Thus, overall, also ILK inhibition blocks growth factor-stimulated *KRAS* expression and, as a result, the associated aggressive phenotype.

2.1 G4 and DNA topology

It is clear that the G4-protein interaction world is extremely wide since several families of proteins bind these DNA or RNA tetrahelices to exert a precise synergic regulation of almost all of the nucleic acid processing pathways [19,20]. Frequently this derives from a direct interaction between the two counterparts, although some peculiar exceptions are known.

Focusing on the theme related to G4 in promoters as mediators for transcription, one example is superhelicity. In the chromatin environment, negative supercoils cause a destabilization of the double helix that, consequently, melts more easily. At the same time, G4 formation counterbalances negative supercoiling. The sum of these two structural effects provides the needed energetic contribution for G4 formation in negatively supercoiled regions. Thus, the topological state of the promoter can control the distributions of the G4 population [40]. Interestingly, the progression of RNA polymerase machinery tends to accumulate positive and negative supercoils in front and behind, respectively. Consequently, as confirmed on *c-MYC* promoter, this stimulates G4 formation upstream the TSS, impairs the binding of transcription factor (TF) and suppresses transcription [41]. Changes in DNA topology are known to be produced by Topoisomerases, thereby affecting the probability of G4 formation in PQS. Additionally, a direct interaction between G4 and Topo I has been reported, suggesting mutual interference [42,43,44,45].

Helicases are effective G4 binder proteins which produce unwinding of DNA and RNA duplexes. These enzymes are currently divided into 6 superfamilies according to their sequence [46]. Additionally, the oligomeric state of the active form, the substrate preferences (DNA, RNA, DNA-RNA hybrids) as well as mechanistic differences can be used to further cluster them. Recently, an increasing number of helicases able to resolve non-canonical DNA structures emerged, mostly with specific activity at G4 level. Among them G4 helicases Pif1, FANCJ, BLM (Bloom syndrome) and WRN (Werner syndrome) are well characterized and help giving a general overview of their G4 processing abilities (for a recent review see [47]). Also Human helicases XPB and XPD, key members of the Transcription Factor IIH complex, are recruited atG4 forming sites since their binding sites overlap PQS [48]. However mechanistic information thus far available are not sufficient to assess their functional participation in G4-related events.

The high number of enzymes able to resolve G4 is considered necessary due to the large variability of G4 structures that can be formed along the genome. This implies that G4-targeted helicases are required during all the metabolic processes of nucleic acid, both in front and behind TSS. As far it concerns transcription, an interesting breakout in dissecting the function of G4-helicases interconnection has been recently obtained by using the fibroblasts of patients affected by the Bloom Syndrome (BS) and the Werner Syndrome (WS) as "natural knockdown" models [3]. The above severe pathologies are associated to loss-of-function mutations of BLM and WRN, respectively. These proteins are widely studied for their activity at telomeres where they bind G4s, solve them and recruit additional proteins. BLM and WRN are crucial to safely replicate these regions and to avoid telomere loss or chromosome aberrations. In patients affected by BS as well as WS unexpected up-regulation of genes with high PQS frequency at TSS and within first introns were observed [49,50]. This evidence clearly relates the loss of BLM and WRN to a reduced capacity to resolve G4 but it doesn't provide any explanation for the observed preferential gene up-regulation. A significant correlation of PQS presence and transcriptional changes emerged from a careful sequence analysis of the 4 kbp sequences centered at the TSS of all genes with differential expression in the two knockout models vs wt cell line [3]. In agreement with a generally accepted model in which G4 in promoter causes a reduction of gene expression, in BS and WS, PQS were found to be under-represented in up-regulated genes. However, this study highlighted that besides abundance, the position of PQS is crucial to drive the effect.

Indeed, whereas PQS in the antisense strand have a general downregulation effect, in the sense strand some slots located downstream the TSS (at positions 140-270; 1750-1770; 1900) significantly correlate to the up-regulated genes. To take account of these evidences a new model has been proposed in which G4 formation in the 140-270 bp region made the antisense strand more accessible to RNA-polymerase, thus promoting the initiation of the transcription (Figure 4).

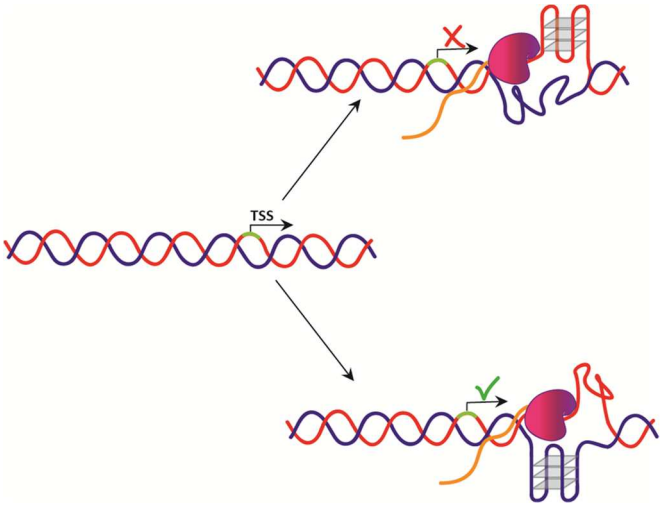


Figure 4. Differential consequences of BLM and WRN helicase suppression on the transcription of gene containing PQS downstream the TSS in the sense (blue) or antisense (red) strand.

2.2 G4 and protein trafficking

Among the large number of G4 interacting proteins a mention must be reserved to nucleolin (NCL), since it represents one of the most abundant phosphoproteins in eukaryotic cells. A peculiar feature of this protein is its involvement in the control of different processes: in addition to regulation of several genes' expression, ribosome biogenesis and reply to stress conditions are the most specialized. This panel of variegate activities has been attributed to its ability to recognize different nucleic acid targets thanks to its modular structure which comprises four highly conserved globular RNA binding domains and a basic C-terminal domain in addition to an acidic N-terminal portion [51]. Additionally, its phosphorylation state further implements the range of its abilities. It has been clearly demonstrated that nucleolin binds G4 and works as transcriptional regulator of G4-containing promoters. On c-MYC, experimental evidence localizes NCL at NHE III1 where it induces G4 formation and represses the transcription promoted by the transcription factor SP1 [52,53]. However, NCL effects are not conserved among genes containing G4 in the promoter. For example, it activates *VEGF* transcription upon binding to the G-rich domain of the promoter [54].

A more articulated NCL-mediated regulatory mechanism was observed on the expression of the non-selenocysteine containing phospholipid hydroperoxide glutathione peroxidase (NPGPx) [55,56]. In the gene coding for this protein, the promoter region ranging between

position -60 and +14 is highly G-rich and comprises up to 4 SP1 binding sites. NMR and footprinting analysis performed on the antisense strand indicate that the 5' portion folds into a G4 whereas at the 3' terminal base pairing supports a stem-loop hairpin. Both of these two structural domains are required to control gene expression. Indeed, in order to preserve the secondary structure content, the promoter activity and the ability to efficiently recruit nucleolin only the loops (in the hairpin or in the G4) can be mutated. Accordingly, it was proposed that in standard conditions, SP1 binds the promoter and allows transcription at a basal level. Under stress conditions, nucleolin (NCL) binds the proximal promoter through recognition of the hairpin-quadruplex element and displaces SP1. However, this new complex helps in keeping the dsDNA open thus finally resulting in transcription enhancement. The molecular factors concomitantly involved in this pathway are still object of debate. An interesting issue concerns the phosphorylation state of NCL that possibly controls shuffling of the phosphorylated form from the nucleolus to nucleoplasm and thus regulates the relative distribution of the protein between the two compartments. In this way, the kinase cascade behaves as a controller of NCL localization. Due to the high affinity of NCL for G4, the same function can be exerted by the tetrahelices, with G4 working as NCL recruiter and G4-ligands as competitors. This model was actually first proposed by studying the interplay between nucleolin and the G4 binder CX-3543 on rRNA biogenesis [57]. The activity profile of this drug rests on a high affinity for G4 and on its localization at nucleolar level. This is the compartment where the rDNA transcription leading to rRNA synthesis occurs. rDNA is highly GC rich and G4 formation in the non-template strand promotes the dense spacing of RNA Polymerase I (Pol I) required to skip the rate-limiting step of ribosome biogenesis [58]. In cells treated with CX-3543, the drug concentrates in the nucleoli where it displaces nucleolin from the rDNA. Consequently, the protein translocates into the nucleoplasm where it induces and stabilizes new G4s (i.e. at c-MYC promoter) [59]. At the same time, nucleolin depletion in the nucleoli reduces rRNA production.

A similar regulation of protein trafficking mediated by protein-G4 interactions has been proposed for another nucleolar phosphoprotein, nucleophosmin (NPM1). This protein is a multifunctional protein that interacts with several counterparts comprising both proteins and nucleic acid substrates. NPM1 continuously shuffles between the nucleus and the cytoplasm and it has an affinity for G4 in the low micromolar range [60]. Both the nucleolar localization and the high G4 affinity depend upon a three-helix bundle motif

located in the C-terminal domain of the protein. In analogy with nucleolin, the binding of NPM1 to PQS in the non-template strand of rDNA "traps" the enzyme in the nuclear organelles. Accordingly, mutation of the DNA binding domain of the protein causes its release from the rDNA and its migration into the cytosol. Extensive structural studies revealed that two of the three helices (H1 and H2) in the bundle domain are primarily responsible for G4 recognition and are in direct contact with it [61,62]. However, also a flanking unfolded region is crucial for the efficient formation of an encounter complex through long-range electrostatic interactions that contribute to enlarge the protein-DNA contact surface. They are not detectable in the NMR time scale but persist in molecular dynamics simulations: this area can represent the site of choice for a targeted therapy.

2.3 G4 location along the gene

Besides the requirement of further efforts to better define the detailed functional mechanisms of NCL, these studies open other interesting perspectives toward a rational description of down- vs up-regulation mediated by a single effector. For example, an assembled G4-hairpin motif is present in *NPGPx* and in *VEGF* promoters (both activated by nucleolin) but not in *c-MYC*. This envisages possible correlation between the geometry of the DNA-protein complexes and the transcriptional effects and suggests a novel rationale for an efficient drug design. Additionally, in *NPGPx* the G4 domain is closer to the TSS if compared to NHE III_1 in c-MYC. Since the transcription of the first gene is activated by NCL whereas the second one is suppressed, this points to differential effects based on their location with reference to TSS.

A closer look to PQS distribution around TSS showed that PQS frequency has one main peak located at about -100 and others in the 5'UTR and at the 5'-end of the first intron [16]. However, for a comprehensive analysis, their localization on the sense/antisense strands must be taken into account, too. As above anticipated, recent data suggest that G4s in the antisense strand substantially inhibit transcription, whereas the effect can be more variegate when they are on the sense strand since they can help in keeping the replication fork open [3,63]. In this context, a genome-wide analysis of the -500 to +500 region indicated a non-symmetric distribution of PQS between the coding and the template strand [64] with a predominance of G4 in the coding strand, downstream the TSS. Actually, this condition can generate a highly articulated output on the protein

expression. Indeed, if transcription is not suppressed, it will generate G-rich sites in the 5'UTR mRNA or in the pre-mRNA which potentially affect translation as well.

In other instances, conserved effects are observed irrespectively of the gene position where the protein-G4 complex is formed. This concept was nicely highlighted by dissecting the role of ATRX (alpha-thalassemia/mental retardation X-linked) on transcription. This protein is a chromatin remodeler helicase and its presence is enriched at heterochromatin regions. ATRX mutations or deletions are known to induce chromosome instability according to a modification of histone deposition. In the α-globin gene, chromatin immunoprecipitation sequencing studies revealed that ATRX efficiently recognizes G-rich variable number tandem repeats region (VNTR) located 1 kb upstream of the TSS. VNTR consists of $CGC(G_4CG_4)_n$ repeats: the length of tandem repeats is strictly related to the severity of the so-called α-thalassemia/mental retardation syndrome gene and to a downregulation of α-globin expression [65].

Conversely, in the ancestral pseudoautosomal region (aPAR) genes, ATRX binding sites are preferentially localized in the body gene [66]. In both cellular systems, ATRX depletion reduces gene expression likely due to stall of DNA replication fork as a consequence of G4 accumulation. This leads to the idea that ATRX helps transcription indirectly by binding and unwinding repressive G4 arrangements. Indeed, the protein contains a helicase domain and competes with 4 binders. However, *in vitro* data did not confirm the unwinding of the tetrahelices by ATRX. Conversely, ATRX promotes the incorporation of histone variant H3.3 at G-rich sites thus actually impairing G4 formation and allowing the transcriptional fork to proceed through G-rich sequences.

This proposal well fits the experimental evidences of G4 enrichment at well-defined chromosome domains i.e. nucleases hypersensitivity sites and heterochromatin where their presence has been related to nuclear organization and cellular differentiation [67].

2.4 G4 and Epigenetics

The mutual correlation of histone deposition and G4 formation clearly indicates that we cannot disclose other epigenetic modifications to better define and explain the variable gene expression patterns.

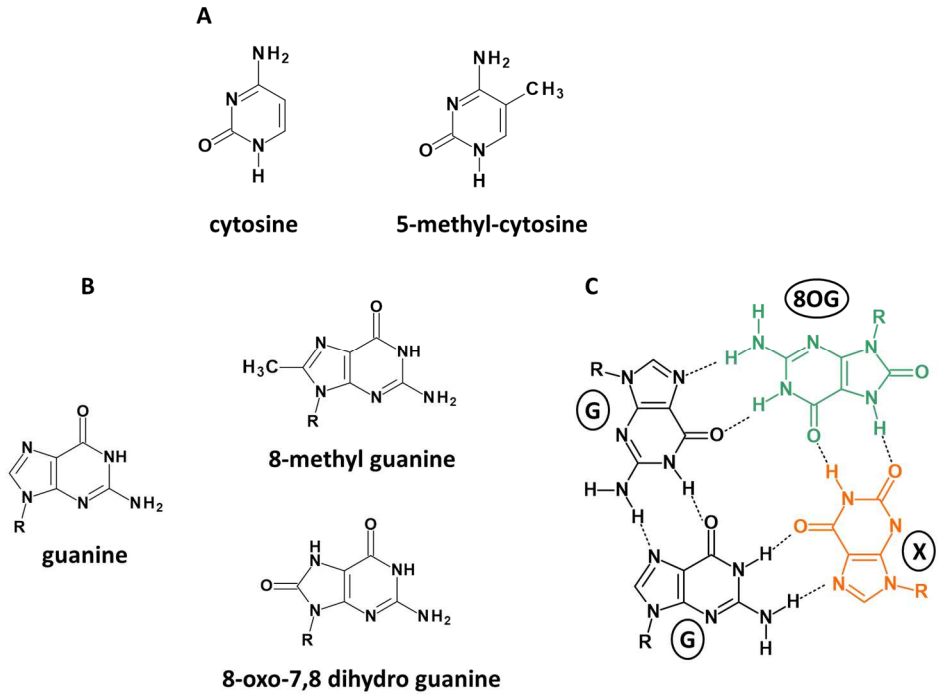


Figure 5. Principal modifications of cytosine (A) or guanine (B). In (C) the pairing of two G, one 8-oxo-7,8 dihydro guanine (green) and one xanthine (orange) is reported.

The principal molecular processes involved in epigenetics include DNA methylation at cytosines, histone modifications and chromatin organization. These natural, carefully coordinated events activate and deactivate parts of the genome at appropriate times and specific locations and become heavily influenced by age, environmental factors, lifestyle and state of disease. For a comprehensive review see [68]. A number of papers appeared anew, mutually linking the non-canonical G4 structure to epigenetic cellular processes. For example, in terms of distribution of replication origins, a high-resolution genome-wide investigation divided them into three main classes depending upon organization, chromatin environment and sequence features [69]. Class 3 origins are characterized by the presence of G-rich repeats potentially organized as G4. Interestingly, the latter are located at nucleosome-depleted regions upstream of the initiation sites. In this context, G4s are believed to operate at the pre-replication complex sites, rather than in origin opening, destabilizing proximal double stranded DNA structures. Not surprisingly, recent *in cell* analysis, showed that nucleosome-depleted chromatin and transcriptional activity shape the endogenous human G4 DNA landscape with clear evidences of enriched G4 formation in the promoters and 5' UTRs of highly transcribed genes [70].

This sequence distribution leads also to a partitioning of DNA polymerases. Indeed, enzymes specialized to work on specific DNA structures are known [71]. With reference to the *c-MYC* promoter, the η enzyme exhibits a 6.2-fold preference for binding to G4 over non-G4 DNA, while the ε polymerase binds both with nearly equal affinity and with drops to 4% in the activity on the G4 substrate. Quite remarkably, the relative replication fidelity for copying the c-MYC-related G4 is two orders of magnitude higher for the η polymerase. From these figures a model is proposed for the replication of G4 sequences, according to which kinetic partition of polymerase activity could facilitate fork progression across the non-canonical DNA structure.

It is worth to stress that as epigenetics affects G4 formation and stability, at the same time G4 affects epigenetic outcome. A significant example derives from recently disclosed notions on helicases that concertedly work with G4 to control dsDNA unwinding. Among them, Pif1 helicase shows superior efficiency in disrupting quadruplex arrangement and in suppressing G4 related epigenetic instability [72,73]. Interestingly, this stimulation derives from G4s directly. Indeed, they activate the dimerization of Pif1 that is necessary to perform efficient unwinding of the canonical double helix. The main consequence is that we can ultimately consider G4 as an epigenetic event, although transient. This explains why failure to resolve these DNA motifs generates genetic and epigenetic instability. The consequence of this connection is a modulation not only of gene expression but of all DNA processing events covering gene switching on/off, DNA replication origin and progression, recombination, damage repair responses as well as disease onset and progression with important physiological and pathological implications [74,75].

2.4.1 G4 and DNA methylation

Focusing on the latest significant advances in connection to transcription, we must mention cytosine methylation: it is one of the most significant epigenetic modifications and it profoundly affects transcriptome (Figure 5A). Its connection with G4 is easily predictable since CpG steps often overlap PQS. In general, in normal tissues, hypermethylation favors closed chromatin (thus reducing the chances of G4 conformations), while hypomethylation produces open chromatin. At the same time, G4s have the potential to directly affect methyl transfer to C [76,77]. Indeed, a genome-wide

analysis of CpG methylation in humans showed an even distribution of this modification according to the sequence composition. In particular, poorly methylated sites were enriched in G-quadruplex forming sequences. The picture however can still refer to yet unknown balance of multiple components. Indeed, C-5 methylation of cytosine in the *BCL-2* promoter causes unexpected stabilization of the G4 structure which makes it more difficult for DNA polymerases to proceed through the ordered secondary structure, producing an epigenetic regulation of gene transcription [78]. This connection has severe consequences in several pathologies. For example, the myotonic dystrophy type 1 is associated to an expansion of CTG repeats on the coding strand of *DMPK*, the gene encoding for the dystrophia myotonica protein kinase. A close by G-rich sequence ($G_4C_2G_4C_2G_4C_2G_3$) within the low methylation repeat, 2kb from the 3' end of *DMPK* is likely organized as a stable G4 thus favoring the repeat expansions [79].

A similar link between DNA methylation and G4 has been proposed for the D4Z4 repeat arrays. D4Z4 is an interesting genomic region as it is connected to fascioscapulohumeral muscular dystrophy, a very serious inherited disease. Within an overall hypermethylated genomic region, a reduced methylation is observed in a sub-region of D4Z4, characterized by a G-rich composition and predicted to form G4 structures. Thus, G4 might impair the DNA methylation process, while participating in the generation of high-order chromatin packaging [80].

A final remark concerns the effects of C methylation/C hydroxymehtylation of the C9orf72 repeat $(G_4C_2)_n$•$(G_2C_4)_n$ related to serious pathologies such as amyotrophic lateral sclerosis and frontotemporal dementia. Both C-rich and G-rich strands are characterized by heterogeneous structures with G4 features, quadruplexes being stabilized by G- and C-containing tetrads. Protein (hnRNP K) binding occurred at C-rich, but not at G-rich strands. As a general conclusion quadruplex formation in GpC islands should be considered irrespectively of the strand composition, C methylation playing a crucial regulatory role in structure stabilization and, consequently, in epigenetic processes modulation [81].

2.4.2 G4 and DNA damage/repair

Another significant issue refers to handling of endogenous or exogenous DNA damage occurring at the G4 level. Indeed, guanines represent hotspots of DNA damage produced by electrophilic or radical species, with potentially widespread consequences (Figure 5B).

These lesions caused by a variety of environmental toxicants like vincolozim, bis-phenols and phthalates actually result in epigenetic transgenerational inheritance of disease and phenotypic variation. From a molecular point of view the observed alterations include modifications in the programming of germline, which are transmitted to the next generation. The observed alterations affect methylation, zinc finger motifs and G4 structures, the latter being effective in increasing chromatin accessibility thus contributing to remodeling processes [82].

In *in vitro* systems the presence of 8-methyl-2'-deoxyguanine largely impacts the structure, stability and formation kinetics of both in intra- and inter-molecular G4 [83,84,85]. The outcome is highly dependent upon sequence context. Methylation at position 8 significantly stabilizes the parallel G-quadruplex structures, especially when the lesion is located at the 5'-end, consistent with findings referring to 8-amino or 8-bromo analogues. Interestingly, NMR studies indicate the formation of all-syn tetrads when the first or the fourth guanine of TG_3T and TG_4T are methylated at position 8, whereas modification of the second guanine leads to anti-glycosidic bonds. Moreover, replacement of the third guanine produces unstable G4.

The issue of the possible *in vivo* occurrence of 8-methyl dG adducts still awaits experimental confirmation. Nevertheless, the above data show subtle modulation of glycosidic bond conformation by local chemical changes in the G4 affecting both thermodynamics and kinetics of quadruplex formation, which anticipate significant effects *in vivo* as the result of G damage at position 8. This is extremely important since reactive oxygen species are highly effective in producing DNA base oxidation. Among modified Guanine derivative 8-oxo-7,8 dihydro guanosine (8OG) represents one of the major naturally occurring lesions hence, with expected effects on G4 assembly [86]. In terms of reactivity, oxidation of guanines paired in the quadruplex is twice as fast as oxidation of the same sequence in a duplex context. In terms of folding and stability, site-specific substitutions of 8OG for G are compatible with G4 but they impair the tetrahelical stability. Additionally, substitution in the middle of a GGG triplet produced multiple structures.

An interesting issue relates this guanine modification to the possible onset of mutations as a consequence of altered base-base recognition. Combination of 8OG with Xanthine (X), another important base lesion deriving from guanine deamination, was found to allow the formation of peculiar tetrads (i.e. G·G·X·8OG or ·X·8OG X·8OG, see Figure 5C) [87].

They can drive reversal of the hydrogen-bond polarity of the modified G-tetrad while preserving the original fold of the unmodified G4. The X-8OG pairing is hard to be generated biologically, as it would involve a cytosine to guanosine mutation on one strand, followed by different concomitant damage events (deamination or oxidation) at the mismatch. At the same time, the control exerted over G-tetrad polarity by joint X-8OG modifications will be suitable for the design of G4 with preset structures and properties and may be used for therapeutic or nanomaterials applications [88].

A key aspect concerning these damaged sites is related to their repair. Indeed, 8-oxo-7,8-dihydroguanine as well as other guanine oxidation derivatives (i.e. spiroiminodihydantoin) are not substrates for base-excision repair. Thus, the worry that accumulating base damage would soon threaten the genomic integrity of our DNA has clearly come into view [89,90]. A novel, unexpected finding consists in the observation that a large number of oncogenic G4s harbor an additional G-rich tract slightly apart from the four engaged in the quadruplex. The idea has emerged that this fifth track could play the role of a "spare tire" to favor repair of damaged G4s. Biophysical studies confirm that the fifth G-track is capable of replacing a damaged G-run containing an oxidized guanine, by placing the lesion in a large loop for effective removal of the damaged portion through a classical base excision repair mechanism, hence restoring full genetic functionality. This looping out mechanism likely occurs both at oncogene promoters and at the telomere level. In a sense, G-rich sequences can be considered as sensors of oxidative stress to which they counteract triggering epigenetic effects on gene expression to foster DNA repair.

3. EFFECTS OF G4 FORMATION ON GENE TRANSCRIPTION: the fine tuning by nucleic acid sequence and structure

This second section is well summarized by the concept that not all G4s are created to work comparably. This easily refers to the different base composition in PQS that ultimately produces distinct overall structures. From a medicinal chemistry point of view this feature is actually the key to realize selective drug recognition. However, it must pair with the fact that different G4 structures can play different roles according to their shape. Additionally, G4 stability and formation kinetics are crucial to select the physiologically relevant forms among a wide population of possible structures.

3.1 G4 polymorphism and stability

This variegate picture is strictly related to the well-known G4 polymorphism issue. Indeed, in analogy to the telomeric sequence for which numerous intramolecular conformations are possible, also short G-rich sequences derived from gene promoters frequently fold into multiple dissimilar structures. Each single member of the resulting population is different in terms of overall shape and location of chemical recognition elements, thus representing a unique biological entity eventually with specific biological functions. Up to date different strategies can be applied to prove G4 formation in physiological environment. However, the molecular features of the functional G4s remain vague. This is often linked to the experimental procedures used to resolve them: these can enrich the sample in forms that are not actually present *in vivo*, thus providing inappropriate models. Among several techniques designed to overcome this bias, Size Exclusion Chromatography (SEQ) provides a suitable balance between high and low resolution approaches [91,92]. SEC analysis performed on several G-rich promoter fragments actually succeeded in identifying several folded forms assumed by a single wt sequence. However, the structure/function correlation remains a tricky point: indeed, among the whole population of isoforms those endowed with too low formation kinetic or too fast unfolding rate can be assumed to be less relevant as regulatory elements for transcription. This aspect has been addressed in solution according to several experimental approaches. Among them, single molecule techniques appear to be the most innovative in dissecting these issues. As an example, when applied to the *c-MYC*-derived sequence pu27, magnetic tweezer experiments provided new insight into its thermodynamic profile. In agreement with in solution studies, they confirmed that the main “chair” form covers around 80% of the folded population. Additionally, it is characterized by an extremely low unfolding rate (10^{-6} s^{-1} at zero force) when compared to other PQS, thus sustaining its function as kinetic barrier to transcription [32].

An even more complex picture arises when multiple G-stretches are confined in a short sequence. In this condition, the formation of a heterogeneous G4 population in which each member derives from the pairing of a different combination of G-stretches can derive. As above mentioned, this is occurring in *c-MYC* but it represents a quite common event. Indeed, bioinformatics searches frequently identify sequences containing eight or more consecutive short G-tracts in the human genome [93,94,95]. In these conditions it

is even harder to identify the most relevant components for transcription regulation due to the potential overlap of possibly opposite effects promoted by each single folded species. This has been experimentally determined within the Human Tyrosine Hydroxylase Promoter. Here, a 45nts domain with seven G-stretches is present. Available data indicate that this full-length sequence is required to promote gene transcription [96]. However, selective mutation of the 5' or 3' G-run leads to G4 elements which behave as a repressor and as an activator of gene transcription, respectively. This behavior is even more intriguing if we consider that the introduced mutations do not affect any of the functionally confirmed transcription factor binding sites.

3.2 Structural elements deriving from the loops

Another level of complexity reflects the potential recruitment of apparently non-consecutive G-runs into a single G4. This idea was proposed in a work on the complementary strand of the 5-'UTR region of the *B-RAF* gene. In this instance, the monomeric G4 form was not the preferred one. Conversely, solution and crystallographic data agreed to indicate an intertwined dimer as the dominant species in potassium containing conditions [95]. The functional role of this dimeric arrangement might be limited *in cells* since only intramolecular structures are generally assumed to be involved in the regulation of gene transcription. Nevertheless, this model can represent a reliable functional structure for sequences contacting eight or more G-runs.

At the same time, these data turn on the light on the potential involvement into a single G4 of spatially separated G-repeats. This interesting novel perspective has been further exploited by extending the analysis of G-rich sequences distribution in terms of half–G4 (G2c) i.e. nucleic acid segments comprising only two runs of at least three consecutive guanines [97]. A correlation between distal cis regions and G2c in the sense strand was evidenced thus suggesting a potential functional role. Its explanation was based on the inclusion of a long loop into the G4 in which two columns are provided by G-repeats of the distal cis-regulatory region and two by the promoter itself: this "bimolecular" G4 should allow the promoter to be properly localized in proximity to the enhancer. The resulting structure is expected to be sufficiently stable to play a significant role but not to trap it, thus behaving exactly how it is required by a regulatory element.

VEGF 5′ – CGGGGCGGGCCGGGGCGGGGT – 3′

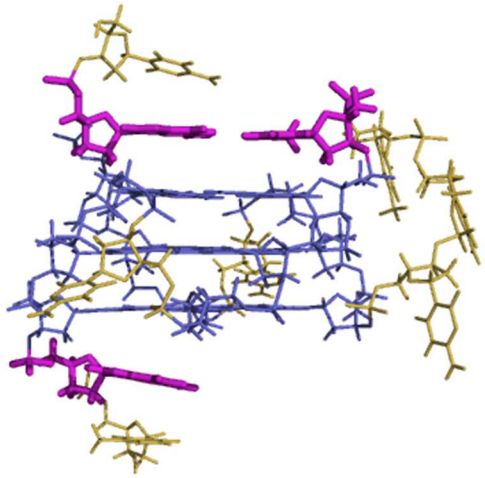


Figure 6. NMR structural model for VEGF promoter. Data were derived from the sequence containing G-T mutations at position 12 and 13 (PDB ID: 2M27). The residues highlighted in purple behave as capping elements, guanines involved in G-tetrads are in blue.

These novel models are sustained by several studies devoted to elucidate the role of the loops. G4s comprising short loops are generally highly stable: this allows to accommodate a very long central loop when the other two are sufficiently short [98]. In any events, on model sequences, a good negative correlation between the total number of bases in the loops and the G4 thermal stability was found in potassium containing solution. However, on natural sequences derived from gene promoters, this relationships is lost [99]. Lack of such correlation further points out that not only loop length but also loop base composition contributes to the thermodynamic features of the tested sequences. Indeed, several examples in which bases of the loops are structural elements directly involved in stabilizing the overall G4 arrangement are available. In these cases, the bases are principally forming base pairing or triplex arrangements which work as G4 capping elements. A well characterized model is provided by the 22mer human *VEGF* promoter sequence, Pu22 [100]. It prevalently folds into a G4 isomer comprising a 1:4:1 nts loops combination actually driven by an articulated capping architecture. This involves both the 3′ and 5′ flanking regions and the middle loop. In particular NMR studies identified a guanine (G21) stacked at the 3′ G-tetrad whereas the G13 of the middle loop stretches over the 5′ G-tetrads to cap it along with G2 (Figure 6). This overall architecture is unique, and thus it represents a useful site of intervention for therapeutic purposes.

As well as *VEGF*, also its receptor *VEGFR-2* contains a G4 -forming sequence located between positions -117 and -94 of the promoter. The resulting G4 element is characterized by an antiparallel arrangement, which can be stabilized by small interacting ligands, thereby impairing tumor angiogenesis. These findings support the idea of G4s representing valuable targets for therapeutic intervention at different steps of a common methabolic pathway [101].

In the case of G4s comprising long loops, their contribution to the overall structure can derive from more variegate interactions of different modules. A curious model is *BCL2*, where up to two G-rich segments in the P1 promoter were found to fold into G4. Both folded structures comprise long loops which actively contribute to defining the G4 core structural features.

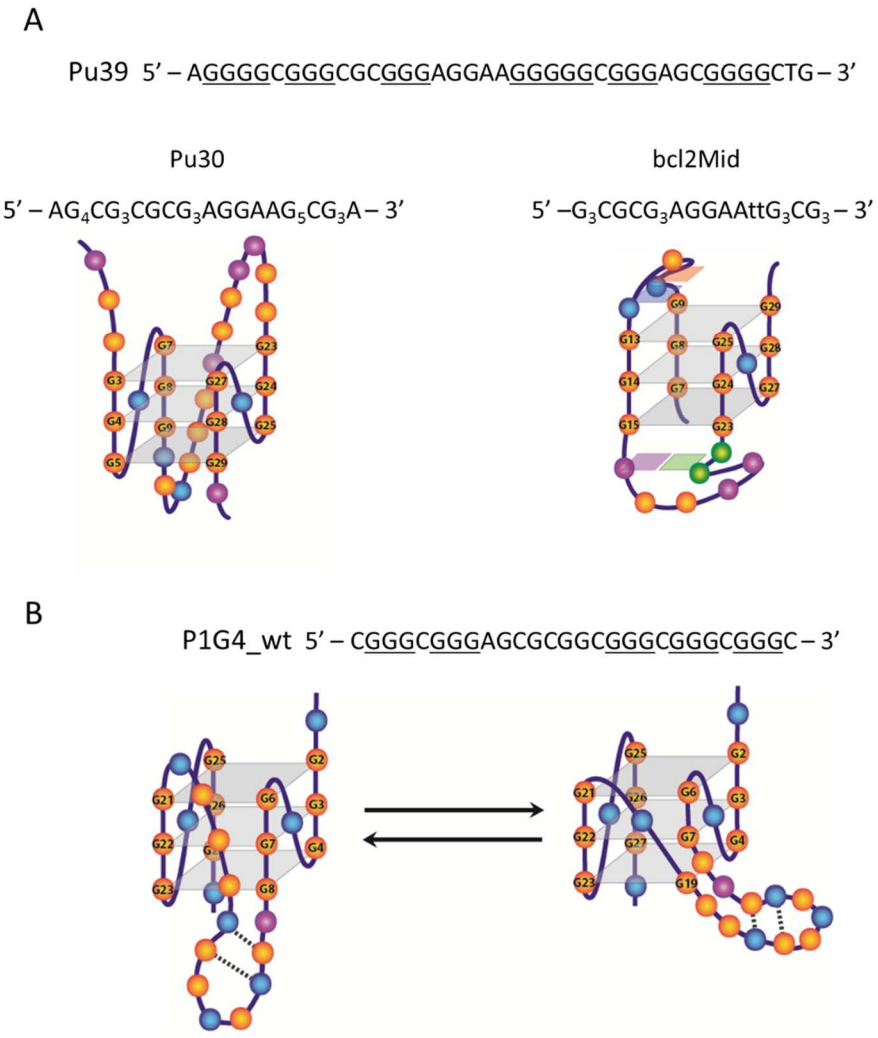


Figure 7. Schematic representation of G4 structures identified by NMR within the P1 promoter of BCL2. G-runs are underlined. PANEL A: folding of Pu39 identified by the

analysis of Pu30 and bcl2Mid (containing two G-T mutations) sequences; PANEL B: structural equilibria determined for the P1G4 sequence.

The so called Pu39 sequence located from position -1489 to -1451 upstream the TSS contains six G-runs (Figure 7A). The 5'-fragment comprising the first 30 nts (Pu30) was found to assume a main parallel folding which comprises a series of 1-13-1 nts-long loops and the pairing of guanines of the 1st, 2nd, 4th and 5th G-tracts [102]. This arrangement was unexpected since the formation of a stable three-tetrad mixed parallel/antiparallel G-quadruplex with three loops of 3, 7, and 1 nt was previously reported for a truncated sequence of 23 nts comprising the 2nd, 3rd, 4th and 5th G-runs (bcl2MidG4) [103]. The higher thermodynamic stability of 1245G4 counterbalanced by the faster kinetic formation of bcl2MidG4 suggested that the two different structures can interchange to allow a fine tuning of gene expression.

From a structural point of view, even more peculiar is the behavior of the P1G4 sequence located in the same promoter between position -1439 and -1412 [104]. In this case, the 28-nts wt sequence is in dynamic equilibrium between two forms which actually share two unique G4 stabilizing motifs (Figure 7B). The first one is a long middle loop (12 and 11 nts, respectively) which is arranged in a stem-loop duplex conformation supported by the presence of two GC Watson-Crick base pairs. The second one is a capping element provided by unpaired guanine likely hydrogen bonded on the 3' G-tetrad. Thus, the difference in the two structures is represented by the substitution of G8 (paired in the bottom tetrad in the structure with 1-12-1 nt loops) with G19 from the long loop: as a result, a G4 comprising a broken strand and with three 1-nt loops and one 11-nt is obtained.

The presence of hairpins in the loop is not odd since a limited number of base-pairs are sufficient to provide a relevant stability to the hairpin. Interestingly, this stabilizing effect is positively extended to the overall G4, as well described in a study on the promoter sequence of the gene coding for n-myc, a member of the myc family of transcription factors [105]. This observation was fully supported by a comprehensive analysis of the mechanical unfolding and refolding process for a G-rich fragment of *hTERT* (human telomerase reverse transcriptase) promoter. This 44 nts sequence (5'-$G_5CTG_3C_2G_4AC_3G_3AG_4TCG_3ACG_4CG_4$-3') is interesting since it folds into a G4 comprising a stem-loop duplex [106]. A laser-tweezers approach provided the analytical evidence for

the coexistence of sequential and cooperative refolding events. This means that the shortest hairpin motif works as seeding element to drive G4 formation, by bringing distant G-tracts in close proximity. In addition, it supports a tertiary interaction between the two DNA structural modules.

A completely different folding pattern is followed by a very peculiar sequence located in the *WNT1* promoter (WT22, 5'-$G_3C_2AC_2G_3CAG_5CG_3$-3') [107]. In the absence of K^+, the single stranded G rich sequence forms a stable hairpin; to allow G4 formation it must melt and this represents the rate limiting step for G4 formation since the hairpin-G4 transition is very slow (time scale of 4800 s). In contrast with most G-rich sequences at promoters, NMR structural characterization confirmed that this sequence folds according to a (3+1) topology, which closely resembles the arrangements found for the human telomeric sequence. This results from the combination of intermediate loops length (5, 4, 1) and of a GC pairing in the middle diagonal loop that has a critical role in distorting the Hoogsteen hydrogen bond network of the top G-tetrads. It is worth noting that this pairing is not a residual group deriving from the hairpin, but it represents a unique structural element of the G4 architecture.

These new duplex-quadruplex interfaces can represent an attractive novel platform for ligands recognition. Consequently, in addition to their features, also the genomic distribution of these "stem-loop/G4" ternary structures has been explored [108,109]. They appear to be clustered in genes associated with brain tissues with enrichment at transcriptional/mutagenesis hotspots and cancer-associated genes, thus reinforcing a potential physiological meaning for these peculiar structures.

3.3 G4-G4 neighbor modules

Another common condition at promoters is the presence of several G-rich stretches that are compatible with the formation of multiple close-by G4 modules. Relevant examples are found within *c-KIT*, *HRAS*, *KRAS*, *cMYB*, *BCL-2* [37,94,102,103,104,110,111,112,113,114,115,116,117,118]. In terms of regulation of gene expression, they can work independently one from each other. This indicates that a precise control of the transcription process might benefit by a modular arrangement of the promoter region. However, we cannot forget that these modular arrangements can

impact also on the structure and the thermodynamic stability of the overall system in comparison to the isolated units [119,120].

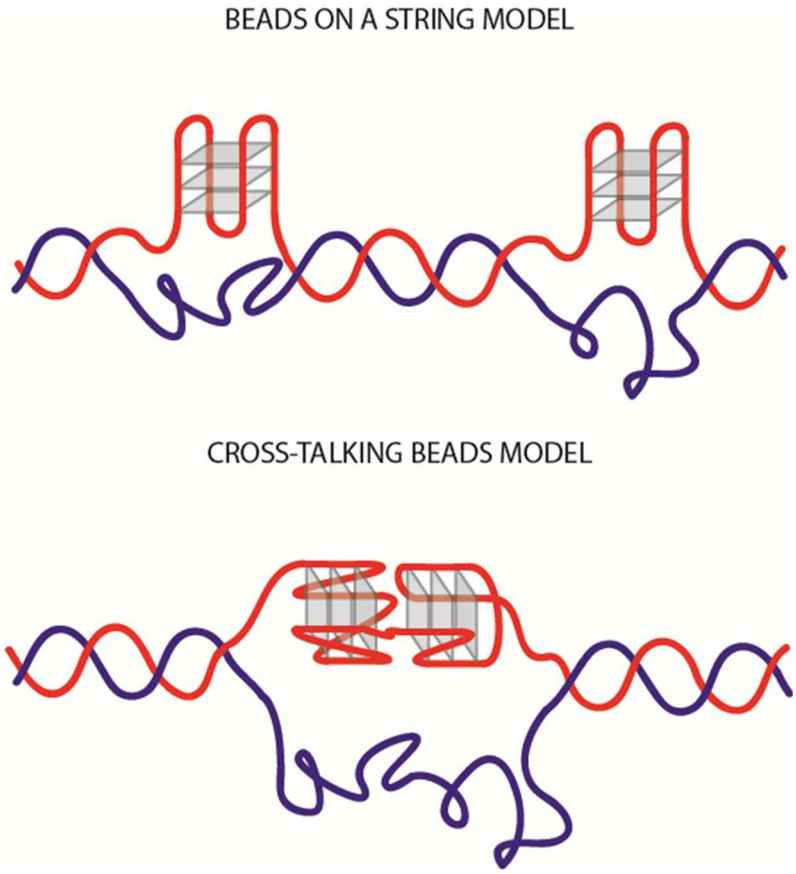


Figure 8. Models for proximal G4 modules.

Pioneering works about long sequences containing multiple PQS focused on the telomeric DNA and succeeded in identifying and characterizing the formation of G4s which behave as interacting units [106,121,122,123,124]. The output has been summarized into two main models that can be easily extended to multiple G4s in promoter sequences, too (Figure 8).

3.3.1 "Beads on a string" model

Depending upon the structures and the reciprocal distances, in one gene promoter we can find G4s that do not interact one to each other. They are expected to work independently on gene regulation: in this case, the overall effect consists of the sum of the different contributions of the single structures in silencing or enhancing the

transcription. This condition refers to the "beads on a string" model [94,102,103,104,114,116,118].

This model perfectly fits the organization of the *HRAS* promoter. This DNA portion contains several G-rich spots, four in the non-coding strand, and two in the coding strand [39]. The most investigated sequences are two PQS upstream TSS, named *hras-1* and *hras-2*, that, independently, fold into stable G-quadruplex structures [115].

In potassium containing solutions *hras-1*, assumes an antiparallel G4 conformation whereas *hras-2* that is located closer to the transcription start site, folds into a more stable parallel G-quadruplex. Chromatin immunoprecipitation and Electrophoretic Mobility Shift Assays showed that *hras-2* contains two binding sites for Sp1. In addition, *hras-1* and *hras-2* contain one and two MAZ (myc associated zinc finger protein) binding sites, respectively. MAZ binds both G4s with an affinity lower in comparison to the corresponding duplexes. This allows the protein to unfold them and to promote their pairing to the complementary strands [125]. Since both G4s suppress *HRAS* transcription, MAZ works as an activator. Interestingly, by luciferase assays, Xodo and coworkers observed that by inducing both quadruplexes it is possible to block completely the transcription of the gene whereas if only one G4 is stabilized, the effect on transcription is reduced to 50%. Thus, the G-quadruplex forms of *hras-1* and *hras-2* equally contribute to the downregulation of *HRAS*. This condition is not conserved for all sequences fitting the "beads on a string" model. Indeed, in some genes, each single G4 can contribute to a different extent to the regulation of gene expression. This is the case of Bcl-2 P1 promoter. Taking into account Pu39 and P1G4 it was shown that G4 induction leads in both instances to repression of gene transcription but with a dominant role exerted by P1G4 [104].

3.3.2 "Cross-talking beads" model

The effects of multiple G4s in a promoter region are not occurring only at functional level: if two or more G4s come close to each other, they can physically interact, forming higher order DNA structures and leading to variegate effects on gene regulation. This situation is described by the "cross-talking beads" model. Interestingly, it implies the formation of new interfaces mainly referring to π-π stacking of the external tetrads of the G4 modules and to the interaction of their loops [123,124]. In a medicinal chemistry perspective, they could be used as suitable targets to improve selective drug design [126,127]. Again, the

largest data-set refers to the telomeric sequence from which it was derived that cross-talks between different units deeply change the structural features of the participating units in terms of shape and thermodynamic behavior [120,128,129]. Besides the telomeric sequence, this model applies to a number of gene promoters too.

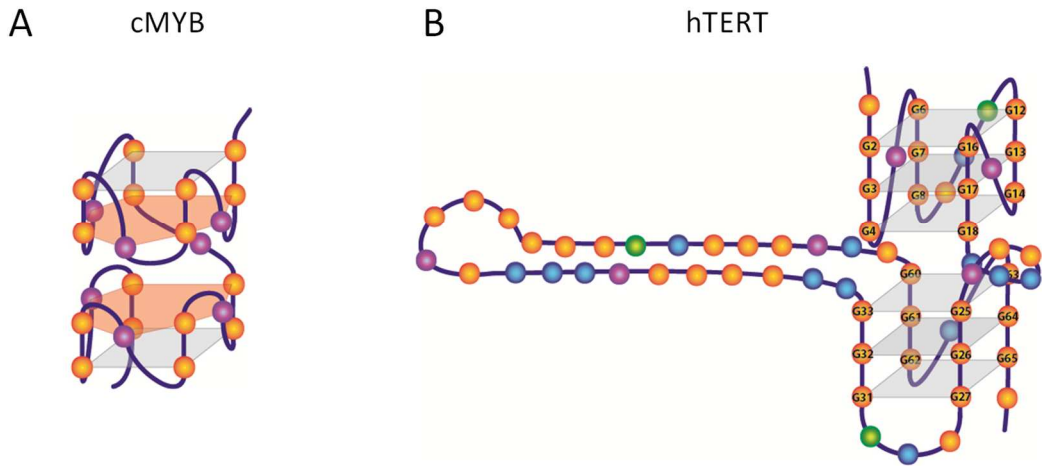


Figure 9. Models of peculiar G4-related interacting units.

An example is *c-MYB* promoter. It contains three $(GGA)_4$ repeat sequences, starting 17 nts downstream the TSS in the non-coding strand [117]. Each of these repeats can fold into a peculiar G4 structure that contains one G-tetrad and one planar heptad deriving from the insertion of three adenines among four guanines. These G4 units cross-talk forming a dimeric G4-G4 structure stabilized trough stacking interaction of the external heptads that are wider than the G-tetrads (Figure 9A). The final conformation corresponds to a peculiar architecture composed by tetrad:heptad:heptad:tetrad elements [130]. These GGA repeats act as cis-elements, regulating the DNA binding of transcription factors, i.e. MAZ. Indeed, if all the $(GGA)_4$ repeats are deleted gene expression no longer occurs, due to lack of TF binding sites. Conversely, the full-length sequence acts as a repressor of gene expression in the presence of KCl, likely blocking the progression of RNA-polymerase through G4 formation [131]. Interestingly, deletion of one single $(GGA)_4$ unit destabilizes the overall tetrahelical arrangement, leading to an increase in *c-MYB* transcription. This supports the physiological relevance of the dimeric tetrad:heptad:heptad:tetrad system as regulatory element.

Another sequence that fits the “cross-talking beads” model is the Insulin-like polymorphic region (ILPR). It is a G-rich sequence with variable number of tandem repeats (VNTR), located at -363bp upstream the insulin coding sequence. The number of these VNTR is

associated with the probability of developing type I diabetes mellitus. This G-rich sequence can accommodate multiple G4s. The minimal sequence required to form an intramolecular G4, in solution gives rise to a mixture of parallel and antiparallel conformers. This is very different from what observed on longer sequences comprising two G4 modules that fold mainly into a hybrid conformation [119]. Both spectroscopic and single molecule studies indicate that the melting profile of this dimeric system is appropriately described by a cooperative process involving two interacting cores [119,132]. This finding confirms that the interaction of the two G4 modules synergistically contributes to the overall stability of the system. Additionally, it highlights that it isn't possible to predict the structural features assumed by multiple tandem G4s starting from the characterization of the single participating G4 elements only.

A final mention should be deserved to *hTERT* promoter that comprises a core domain with five Sp1 binding sites located -180 to +1bp from the TSS. In the central region it contains twelve consecutive G-rich tracts able to fold into different G4 structures [133,134]. Two distinct models were suggested to describe the architecture of this sequence. In the first model, the system is characterized by two distinct G4s: the first one formed by G-tracts 1-4; the second one formed by G-tracts 5-6-11-12. The latter shows a 26 bases middle loop, forming a very stable hairpin structure (Figure 9B). In the second model, the twelve G-tracts are folded in three consecutive G4s [135]. Both models are based on G4-G4 contacts through stacking interactions that stabilize the structures. Interestingly, in the folded condition, the Sp1 consensus sites are hidden thus preventing the binding of the transcription factor to the DNA. This explains the observed repression of gene expression upon folding.

4. CONCLUSIONS

The data thus far presented show an intriguing and complex mechanism of gene expression, in which G4s play several roles, eventually being able to produce opposite effects on the same target. From one side, this is part of a fine-tuning regulatory machinery allowing timely and coordinated start and blockade of the transcription process, from the other it represents a major challenge for selective pharmacological intervention. At present, we know the ABC of G4 structure and recognition, but we are still unable to design a priori chemicals able to univocally interfere with the expression of

a single gene promoter, leaving the others unaffected. Two main issues to consider deal with the large planar arrangements of the external G tetrads, allowing efficient stacking with polycyclic condensed systems, and the high charge density of the tetrahelical system, both of which represent an effective, yet poorly specific, source of ligand binding stabilization. If we spoil the ligands thus far exploited from these two features, very little is left for efficient interaction. Even more, G4 are highly polymorphic, the energetic difference of stable conformations often not exceeding 1 kCal/mol. Hence, we are in the presence of multicomponent systems and of several structurally different complexes, each of which will interfere with target DNA to different extents and in different modes.

A further level of complexity derives from the location of the PQS in a global double helical context, where we are forced to consider equilibria involving Watson-Crick pairing with the complementary strand, which compete with G4 pairing [136,137]. Although it was not considered in this review the concomitant occurrence of other non-canonical arrangements on the purine rich-strand (i. e. I-motif) must be taken into account. Indeed, they can implement or disfavor the G4 formation [138,139]. Furthermore, kinetics of G4 folding can effectively interfere with thermodynamics at PQS locations producing unexpected G-quadruplex distributions *in cell*. A dynamic picture must be considered by the medicinal chemist, according to which competition by drug molecules involves a number of yet ill-defined players [140]. And, until now, we haven't mentioned DNA-binding factors, which are the terminal addressees of the physiological plasticity of G-rich genomic sequences. G-quadruplexes are the target of G4-stabilizing protein chaperones as well as of G4-unwinding proteins. If we statistically consider this issue, a prevalence of G4 disrupting biomolecules have been described until now, whereas known G4 stabilizing proteins are less frequent. Of course, we know that G4s must be unfolded to allow transcription (or replication) of the G-rich strand and correctly refold the DNA double helix, which eventually justifies the ubiquitous incidence of G4 unwinding species such as the family of helicases. Surely, much more needs to be learned in this connection.

A further structurally (and pharmacologically) relevant motif is the frequent occurrence of clusters of interacting G4 sequences, which generate a superior level of complexity, hence producing novel, more complicated and more extended targets for drug design and development. In fact, in the presence of two or more close-by PQS, we may think of more sophisticated ligands logically devised to exploit simultaneous interference with grouped G4s, which might remarkably improve our ability to produce selective drugs.

In terms of G4-mediated biological response, the outcome of drug interference could be agonistic or antagonistic, i.e. G4 stabilization or destabilization might increase or decrease the original level of gene expression. Unfortunately, we are at present unable to predict the result given the promoter sequence. Moreover, as foreshadowed in this review, indirect mechanisms can account for additional regulatory roles played by PQS along the genome, which cannot thus far be foreseen with accuracy.

Evidently, the job for a medicinal chemist in order to produce selective interference at G4 of a given promoter thereby developing novel effective drug molecules able to activate/deactivate transcription of a specific gene is a very hard one. We badly need deeper knowledge of the molecular mechanisms tuning gene expression and G4 recognition, along with the influence of cellular environment on these processes, including epigenetic events, like those mentioned earlier in this review.

A final, possibly optimistic, remark comes from the observed changes in location and density of PQS at the onset of several diseases (see above). In this case we can envisage the possibility to distinguish healthy cells from diseased cells in terms of G4 targeting and to devise novel drugs specifically aimed at a desired site producing more effective treatments devoid of the severe side-effects caused by poorly specific recognition. Hopefully, we will succeed in a not too distant future.

5. ACKNOWLEDGEMENTS

This work was supported by University of Padova grant (CPDA147272/14) and PhD studentship (RR).

6. REFERENCES

[1] J.L. Huppert, S. Balasubramanian, G-quadruplexes in promoters throughout the human genome, Nucleic Acids Res 35 (2007) 406-413.

[2] A. Verma, V.K. Yadav, R. Basundra, A. Kumar, S. Chowdhury, Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells, Nucleic Acids Res 37 (2009) 4194-4204.

[3] J.A. Smestad, L.J. Maher, 3rd, Relationships between putative G-quadruplex-forming sequences, RecQ helicases, and transcription, BMC Med Genet 16 (2015) 91.

[4] D. Sun, K. Guo, Y.J. Shin, Evidence of the formation of G-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene, Nucleic Acids Res 39 (2011) 1256-1265.

[5] G. Biffi, D. Tannahill, J. McCafferty, S. Balasubramanian, Quantitative visualization of DNA G-quadruplex structures in human cells, Nat Chem 5 (2013) 182-186.

[6] A. Henderson, Y. Wu, Y.C. Huang, E.A. Chavez, J. Platt, F.B. Johnson, R.M. Brosh, Jr., D. Sen, P.M. Lansdorp, Detection of G-quadruplex DNA in mammalian cells, Nucleic Acids Res 42 (2014) 860-869.
[7] A. Laguerre, K. Hukezalie, P. Winckler, F. Katranji, G. Chanteloup, M. Pirrotta, J.M. Perrier-Cornet, J.M. Wong, D. Monchaud, Visualization of RNA-Quadruplexes in Live Cells, J Am Chem Soc 137 (2015) 8521-8525.
[8] H. Abe, N.J. Gemmell, Abundance, arrangement, and function of sequence motifs in the chicken promoters, BMC Genomics 15 (2014) 900.
[9] S. Amrane, A. Kerkour, A. Bedrat, B. Vialet, M.L. Andreola, J.L. Mergny, Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development, J Am Chem Soc 136 (2014) 5249-5252.
[10] S. Da Ros, E. Zorzan, M. Giantin, L. Zorro Shahidian, M. Palumbo, M. Dacasto, C. Sissi, Sequencing and G-quadruplex folding of the canine proto-oncogene KIT promoter region: might dog be used as a model for human disease?, PLoS One 9 (2014) e103876.
[11] I.T. Holder, J.S. Hartig, A matter of location: influence of G-quadruplexes on Escherichia coli gene expression, Chem Biol 21 (2014) 1511-1521.
[12] S.S. Smith, Evolutionary expansion of structurally complex DNA sequences, Cancer Genomics Proteomics 7 (2010) 207-215.
[13] N. Maizels, G4-associated human diseases, EMBO Rep 16 (2015) 910-922.
[14] S.L. Cree, M.A. Kennedy, Relevance of G-quadruplex structures to pharmacogenetics, Front Pharmacol 5 (2014) 160.
[15] H.P. Gu, S. Lin, M. Xu, H.Y. Yu, X.J. Du, Y.Y. Zhang, G. Yuan, W. Gao, Up-regulating relaxin expression by G-quadruplex interactive ligand to achieve antifibrotic action, Endocrinology 153 (2012) 3692-3700.
[16] N. Maizels, L.T. Gray, The G4 genome, PLoS Genet 9 (2013) e1003468.
[17] S. Neidle, A Personal History of Quadruplex-Small Molecule Targeting, Chem Rec 15 (2015) 691-710.
[18] S. Neidle, Quadruplex Nucleic Acids as Novel Therapeutic Targets, J Med Chem 59 (2016) 5987-6011.
[19] V. Brazda, L. Haronikova, J.C. Liao, M. Fojta, DNA and RNA quadruplex-binding proteins, Int J Mol Sci 15 (2014) 17493-17517.
[20] C. Sissi, B. Gatto, M. Palumbo, The evolving world of protein-G-quadruplex recognition: a medicinal chemist's perspective, Biochimie 93 (2011) 1219-1230.
[21] S. Balasubramanian, L.H. Hurley, S. Neidle, Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?, Nat Rev Drug Discov 10 (2011) 261-275.
[22] 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.
[23] D. Rhodes, H.J. Lipps, G-quadruplexes and their regulatory roles in biology, Nucleic Acids Res 43 (2015) 8627-8637.
[24] U. Siebenlist, L. Hennighausen, J. Battey, P. Leder, Chromatin structure and protein binding in the putative regulatory region of the c-myc gene in Burkitt lymphoma, Cell 37 (1984) 381-391.
[25] T.L. Davis, A.B. Firulli, A.J. Kinniburgh, Ribonucleoprotein and protein factors bind to an H-DNA-forming c-myc DNA element: possible regulators of the c-myc gene, Proc Natl Acad Sci U S A 86 (1989) 9682-9686.
[26] T.C. Boles, M.E. Hogan, DNA structure equilibria in the human c-myc gene, Biochemistry 26 (1987) 367-376.
[27] T. Simonsson, P. Pecinka, M. Kubista, DNA tetraplex formation in the control region of c-myc, Nucleic Acids Res 26 (1998) 1167-1172.

[28] A. Rangan, O.Y. Fedoroff, L.H. Hurley, Induction of duplex to G-quadruplex transition in the c-myc promoter region by a small molecule, J Biol Chem 276 (2001) 4640-4646.
[29] A. Siddiqui-Jain, C.L. Grand, D.J. Bearss, L.H. Hurley, 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 (2002) 11593-11598.
[30] E.F. Michelotti, T. Tomonaga, H. Krutzsch, D. Levens, Cellular nucleic acid binding protein regulates the CT element of the human c-myc protooncogene, J Biol Chem 270 (1995) 9494-9499.
[31] D. Ma, Z. Xing, B. Liu, N.G. Pedigo, S.G. Zimmer, Z. Bai, E.H. Postel, D.M. Kaetzel, NM23-H1 and NM23-H2 repress transcriptional activities of nuclease-hypersensitive elements in the platelet-derived growth factor-A promoter, J Biol Chem 277 (2002) 1560-1567.
[32] H. You, J. Wu, F. Shao, J. Yan, Stability and kinetics of c-MYC promoter G-quadruplexes studied by single-molecule manipulation, J Am Chem Soc 137 (2015) 2424-2427.
[33] R.I. Mathad, E. Hatzakis, J. Dai, D. Yang, c-MYC promoter G-quadruplex formed at the 5'-end of NHE III1 element: insights into biological relevance and parallel-stranded G-quadruplex stability, Nucleic Acids Res 39 (2011) 9023-9033.
[34] J. Seenisamy, E.M. Rezler, T.J. Powell, D. Tye, V. Gokhale, C.S. Joshi, A. Siddiqui-Jain, L.H. Hurley, The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4, J Am Chem Soc 126 (2004) 8702-8709.
[35] A.T. Phan, V. Kuryavyi, H.Y. Gaw, D.J. Patel, Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter, Nat Chem Biol 1 (2005) 167-173.
[36] D. Sun, L.H. Hurley, The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression, J Med Chem 52 (2009) 2863-2874.
[37] R.K. Morgan, H. Batra, V.C. Gaerig, J. Hockings, T.A. Brooks, Identification and characterization of a new G-quadruplex forming region within the kRAS promoter as a transcriptional regulator, Biochim Biophys Acta 1859 (2016) 235-245.
[38] P.C. Chu, M.C. Yang, S.K. Kulp, S.B. Salunke, L.E. Himmel, C.S. Fang, A.M. Jadhav, Y.S. Shan, C.T. Lee, M.D. Lai, L.A. Shirley, T. Bekaii-Saab, C.S. Chen, Regulation of oncogenic KRAS signaling via a novel KRAS-integrin-linked kinase-hnRNPA1 regulatory loop in human pancreatic cancer cells, Oncogene (2015).
[39] S. Cogoi, L.E. Xodo, G4 DNA in ras genes and its potential in cancer therapy, Biochim Biophys Acta 1859 (2016) 663-674.
[40] L. Baranello, D. Levens, A. Gupta, F. Kouzine, The importance of being supercoiled: how DNA mechanics regulate dynamic processes, Biochim Biophys Acta 1819 (2012) 632-638.
[41] T.A. Brooks, L.H. Hurley, The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics, Nat Rev Cancer 9 (2009) 849-861.
[42] P.B. Arimondo, J.F. Riou, J.L. Mergny, J. Tazi, J.S. Sun, T. Garestier, C. Helene, Interaction of human DNA topoisomerase I with G-quartet structures, Nucleic Acids Res 28 (2000) 4832-4838.
[43] P. Yadav, N. Owiti, N. Kim, The role of topoisomerase I in suppressing genome instability associated with a highly transcribed guanine-rich sequence is not restricted to preventing RNA:DNA hybrid accumulation, Nucleic Acids Res 44 (2016) 718-729.

[44] C. Marchand, P. Pourquier, G.S. Laco, N. Jing, Y. Pommier, Interaction of human nuclear topoisomerase I with guanosine quartet-forming and guanosine-rich single-stranded DNA and RNA oligonucleotides, J Biol Chem 277 (2002) 8906-8911.
[45] A.M. Ogloblina, V.A. Bannikova, A.N. Khristich, T.S. Oretskaya, M.G. Yakubovskaya, N.G. Dolinnaya, Parallel G-Quadruplexes Formed by Guanine-Rich Microsatellite Repeats Inhibit Human Topoisomerase I, Biochemistry (Mosc) 80 (2015) 1026-1038.
[46] M.R. Singleton, M.S. Dillingham, D.B. Wigley, Structure and mechanism of helicases and nucleic acid translocases, Annu Rev Biochem 76 (2007) 23-50.
[47] O. Mendoza, A. Bourdoncle, J.B. Boule, R.M. Brosh, Jr., J.L. Mergny, G-quadruplexes and helicases, Nucleic Acids Res 44 (2016) 1989-2006.
[48] L.T. Gray, A.C. Vallur, J. Eddy, N. Maizels, G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD, Nat Chem Biol 10 (2014) 313-318.
[49] J.E. Johnson, K. Cao, P. Ryvkin, L.S. Wang, F.B. Johnson, Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential, Nucleic Acids Res 38 (2010) 1114-1122.
[50] G.H. Nguyen, W. Tang, A.I. Robles, R.P. Beyer, L.T. Gray, J.A. Welsh, A.J. Schetter, K. Kumamoto, X.W. Wang, I.D. Hickson, N. Maizels, R.J. Monnat, Jr., C.C. Harris, Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs, Proc Natl Acad Sci U S A 111 (2014) 9905-9910.
[51] H. Ginisty, H. Sicard, B. Roger, P. Bouvet, Structure and functions of nucleolin, J Cell Sci 112 (Pt 6) (1999) 761-772.
[52] V. Gonzalez, K. Guo, L. Hurley, D. Sun, Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein, J Biol Chem 284 (2009) 23622-23635.
[53] V. Gonzalez, L.H. Hurley, The C-terminus of nucleolin promotes the formation of the c-MYC G-quadruplex and inhibits c-MYC promoter activity, Biochemistry 49 (2010) 9706-9714.
[54] D.J. Uribe, K. Guo, Y.J. Shin, D. Sun, Heterogeneous nuclear ribonucleoprotein K and nucleolin as transcriptional activators of the vascular endothelial growth factor promoter through interaction with secondary DNA structures, Biochemistry 50 (2011) 3796-3806.
[55] Y.I. Chen, P.C. Wei, J.L. Hsu, F.Y. Su, W.H. Lee, NPGPx (GPx7): a novel oxidative stress sensor/transmitter with multiple roles in redox homeostasis, Am J Transl Res 8 (2016) 1626-1640.
[56] P.C. Wei, Z.F. Wang, W.T. Lo, M.I. Su, J.Y. Shew, T.C. Chang, W.H. Lee, A cis-element with mixed G-quadruplex structure of NPGPx promoter is essential for nucleolin-mediated transactivation on non-targeting siRNA stress, Nucleic Acids Res 41 (2013) 1533-1543.
[57] D. Drygin, A. Siddiqui-Jain, S. O'Brien, M. Schwaebe, A. Lin, J. Bliesath, C.B. Ho, C. Proffitt, K. Trent, J.P. Whitten, J.K. Lim, D. Von Hoff, K. Anderes, W.G. Rice, Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis, Cancer Res 69 (2009) 7653-7661.
[58] I. Grummt, G. Langst, Epigenetic control of RNA polymerase I transcription in mammalian cells, Biochim Biophys Acta 1829 (2013) 393-404.
[59] V. Gonzalez, L.H. Hurley, The c-MYC NHE III(1): function and regulation, Annu Rev Pharmacol Toxicol 50 (2010) 111-129.
[60] S. Chiarella, A. De Cola, G.L. Scaglione, E. Carletti, V. Graziano, D. Barcaroli, C. Lo Sterzo, A. Di Matteo, C. Di Ilio, B. Falini, A. Arcovito, V. De Laurenzi, L. Federici,

Nucleophosmin mutations alter its nucleolar localization by impairing G-quadruplex binding at ribosomal DNA, Nucleic Acids Res 41 (2013) 3228-3239.
[61] A. Arcovito, S. Chiarella, S. Della Longa, A. Di Matteo, C. Lo Sterzo, G.L. Scaglione, L. Federici, Synergic role of nucleophosmin three-helix bundle and a flanking unstructured tail in the interaction with G-quadruplex DNA, J Biol Chem 289 (2014) 21230-21241.
[62] P.L. Scognamiglio, C. Di Natale, M. Leone, M. Poletto, L. Vitagliano, G. Tell, D. Marasco, G-quadruplex DNA recognition by nucleophosmin: new insights from protein dissection, Biochim Biophys Acta 1840 (2014) 2050-2059.
[63] T. Agarwal, S. Roy, S. Kumar, T.K. Chakraborty, S. Maiti, In the sense of transcription regulation by G-quadruplexes: asymmetric effects in sense and antisense strands, Biochemistry 53 (2014) 3711-3718.
[64] Z. Du, Y. Zhao, N. Li, Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription, Genome Res 18 (2008) 233-241.
[65] Y. Li, J. Syed, Y. Suzuki, S. Asamitsu, N. Shioda, T. Wada, H. Sugiyama, Effect of ATRX and G-Quadruplex Formation by the VNTR Sequence on alpha-Globin Gene Expression, Chembiochem 17 (2016) 928-935.
[66] M.A. Levy, K.D. Kernohan, Y. Jiang, N.G. Berube, ATRX promotes gene expression by facilitating transcriptional elongation through guanine-rich coding regions, Hum Mol Genet 24 (2015) 1824-1835.
[67] R.F. Hoffmann, Y.M. Moshkin, S. Mouton, N.A. Grzeschik, R.D. Kalicharan, J. Kuipers, A.H. Wolters, K. Nishida, A.V. Romashchenko, J. Postberg, H. Lipps, E. Berezikov, O.C. Sibon, B.N. Giepmans, P.M. Lansdorp, Guanine quadruplex structures localize to heterochromatin, Nucleic Acids Res 44 (2016) 152-163.
[68] C.D. Allis, T. Jenuwein, The molecular hallmarks of epigenetic control, Nat Rev Genet 17 (2016) 487-500.
[69] C. Cayrou, B. Ballester, I. Peiffer, R. Fenouil, P. Coulombe, J.C. Andrau, J. van Helden, M. Mechali, The chromatin environment shapes DNA replication origin organization and defines origin classes, Genome Res 25 (2015) 1873-1885.
[70] R. Hansel-Hertsch, D. Beraldi, S.V. Lensing, G. Marsico, K. Zyner, A. Parry, M. Di Antonio, J. Pike, H. Kimura, M. Narita, D. Tannahill, S. Balasubramanian, G-quadruplex structures mark human regulatory chromatin, Nat Genet 48 (2016) 1267-1272.
[71] S. Eddy, L. Maddukuri, A. Ketkar, M.K. Zafar, E.E. Henninger, Z.F. Pursell, R.L. Eoff, Evidence for the kinetic partitioning of polymerase activity on G-quadruplex DNA, Biochemistry 54 (2015) 3218-3230.
[72] X.L. Duan, N.N. Liu, Y.T. Yang, H.H. Li, M. Li, S.X. Dou, X.G. Xi, G-quadruplexes significantly stimulate Pif1 helicase-catalyzed duplex DNA unwinding, J Biol Chem 290 (2015) 7722-7735.
[73] K. Paeschke, M.L. Bochman, P.D. Garcia, P. Cejka, K.L. Friedman, S.C. Kowalczykowski, V.A. Zakian, Pif1 family helicases suppress genome instability at G-quadruplex motifs, Nature 497 (2013) 458-462.
[74] V. Cea, L. Cipolla, S. Sabbioneda, Replication of Structured DNA and its implication in epigenetic stability, Front Genet 6 (2015) 209.
[75] C.M. Wickramasinghe, H. Arzouk, A. Frey, A. Maiter, J.E. Sale, Contributions of the specialised DNA polymerases to replication of structured DNA, DNA Repair (Amst) 29 (2015) 83-90.
[76] S. De, F. Michor, DNA secondary structures and epigenetic determinants of cancer genome evolution, Nat Struct Mol Biol 18 (2011) 950-955.

[77] R. Halder, K. Halder, P. Sharma, G. Garg, S. Sengupta, S. Chowdhury, Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide, Mol Biosyst 6 (2010) 2439-2447.
[78] J. Lin, J.Q. Hou, H.D. Xiang, Y.Y. Yan, Y.C. Gu, J.H. Tan, D. Li, L.Q. Gu, T.M. Ou, Z.S. Huang, Stabilization of G-quadruplex DNA by C-5-methyl-cytosine in bcl-2 promoter: implications for epigenetic regulation, Biochem Biophys Res Commun 433 (2013) 368-373.
[79] L. Buckley, M. Lacey, M. Ehrlich, Epigenetics of the myotonic dystrophy-associated DMPK gene neighborhood, Epigenomics 8 (2016) 13-31.
[80] K. Tsumagari, L. Qi, K. Jackson, C. Shao, M. Lacey, J. Sowden, R. Tawil, V. Vedanarayanan, M. Ehrlich, Epigenetics of a tandem DNA repeat: chromatin DNaseI sensitivity and opposite methylation changes in cancers, Nucleic Acids Res 36 (2008) 2196-2207.
[81] B. Zamiri, M. Mirceta, K. Bomsztyk, R.B. Macgregor, Jr., C.E. Pearson, Quadruplex formation by both G-rich and C-rich DNA strands of the C9orf72 (GGGGCC)8*(GGCCCC)8 repeat: effect of CpG methylation, Nucleic Acids Res 43 (2015) 10055-10064.
[82] C. Guerrero-Bosagna, S. Weeks, M.K. Skinner, Identification of genomic features in environmentally induced epigenetic transgenerational inherited sperm epimutations, PLoS One 9 (2014) e100194.
[83] C.J. Lech, J.K. Cheow Lim, J.M. Wen Lim, S. Amrane, B. Heddi, A.T. Phan, Effects of site-specific guanine C8-modifications on an intramolecular DNA G-quadruplex, Biophys J 101 (2011) 1987-1998.
[84] C.S. Mekmaysy, L. Petraccone, N.C. Garbett, P.A. Ragazzon, R. Gray, J.O. Trent, J.B. Chaires, Effect of O6-methylguanine on the stability of G-quadruplex DNA, J Am Chem Soc 130 (2008) 6710-6711.
[85] P.L. Tran, A. Virgilio, V. Esposito, G. Citarella, J.L. Mergny, A. Galeone, Effects of 8-methylguanine on structure, stability and kinetics of formation of tetramolecular quadruplexes, Biochimie 93 (2011) 399-408.
[86] V.A. Szalai, M.J. Singer, H.H. Thorp, Site-specific probing of oxidative reactivity and telomerase function using 7,8-dihydro-8-oxoguanine in telomeric DNA, J Am Chem Soc 124 (2002) 1625-1631.
[87] V.V. Cheong, B. Heddi, C.J. Lech, A.T. Phan, Xanthine and 8-oxoguanine in G-quadruplexes: formation of a G.G.X.O tetrad, Nucleic Acids Res 43 (2015) 10506-10514.
[88] V.V. Cheong, C.J. Lech, B. Heddi, A.T. Phan, Inverting the G-Tetrad Polarity of a G-Quadruplex by Using Xanthine and 8-Oxoguanine, Angew Chem Int Ed Engl 55 (2016) 160-163.
[89] A.M. Fleming, J. Zhou, S.S. Wallace, C.J. Burrows, A Role for the Fifth G-Track in G-Quadruplex Forming Oncogene Promoter Sequences during Oxidative Stress: Do These "Spare Tires" Have an Evolved Function?, ACS Cent Sci 1 (2015) 226-233.
[90] N. An, A.M. Fleming, C.J. Burrows, Human Telomere G-Quadruplexes with Five Repeats Accommodate 8-Oxo-7,8-dihydroguanine by Looping out the DNA Damage, ACS Chem Biol 11 (2016) 500-507.
[91] H.T. Le, M.C. Miller, R. Buscaglia, W.L. Dean, P.A. Holt, J.B. Chaires, J.O. Trent, Not all G-quadruplexes are created equally: an investigation of the structural polymorphism of the c-Myc G-quadruplex-forming sequence and its interaction with the porphyrin TMPyP4, Org Biomol Chem 10 (2012) 9393-9404.

[92] M.C. Miller, H.T. Le, W.L. Dean, P.A. Holt, J.B. Chaires, J.O. Trent, Polymorphism and resolution of oncogene promoter quadruplex-forming sequences, Org Biomol Chem 9 (2011) 7633-7637.
[93] J.M. Dettler, R. Buscaglia, V.H. Le, E.A. Lewis, DSC deconvolution of the structural complexity of c-MYC P1 promoter G-quadruplexes, Biophys J 100 (2011) 1517-1525.
[94] T.S. Dexheimer, D. Sun, L.H. Hurley, Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter, J Am Chem Soc 128 (2006) 5404-5415.
[95] D. Wei, A.K. Todd, M. Zloh, M. Gunaratnam, G.N. Parkinson, S. Neidle, Crystal structure of a promoter sequence in the B-raf gene reveals an intertwined dimer quadruplex, J Am Chem Soc 135 (2013) 19319-19329.
[96] M.M. Farhath, M. Thompson, S. Ray, A. Sewell, H. Balci, S. Basu, G-Quadruplex-Enabling Sequence within the Human Tyrosine Hydroxylase Promoter Differentially Regulates Transcription, Biochemistry 54 (2015) 5533-5545.
[97] H. Hegyi, Enhancer-promoter interaction facilitated by transiently forming G-quadruplexes, Sci Rep 5 (2015) 9165.
[98] 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.
[99] N. Kumar, S. Maiti, A thermodynamic overview of naturally occurring intramolecular DNA quadruplexes, Nucleic Acids Res 36 (2008) 5610-5622.
[100] P. Agrawal, E. Hatzakis, K. Guo, M. Carver, D. Yang, Solution structure of the major G-quadruplex formed in the human VEGF promoter in K+: insights into loop interactions of the parallel G-quadruplexes, Nucleic Acids Res 41 (2013) 10584-10592.
[101] E. Salvati, P. Zizza, A. Rizzo, S. Iachettini, C. Cingolani, C. D'Angelo, M. Porru, A. Randazzo, B. Pagano, E. Novellino, M.E. Pisanu, A. Stoppacciaro, F. Spinella, A. Bagnato, E. Gilson, C. Leonetti, A. Biroccio, Evidence for G-quadruplex in the promoter of vegfr-2 and its targeting to inhibit tumor angiogenesis, Nucleic Acids Res 42 (2014) 2945-2957.
[102] P. Agrawal, C. Lin, R.I. Mathad, M. Carver, D. Yang, 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 (2014) 1750-1753.
[103] J. Dai, T.S. Dexheimer, D. Chen, M. Carver, A. Ambrus, R.A. Jones, D. Yang, An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution, J Am Chem Soc 128 (2006) 1096-1098.
[104] B. Onel, M. Carver, G. Wu, D. Timonina, S. Kalarn, M. Larriva, D. Yang, A New G-Quadruplex with Hairpin Loop Immediately Upstream of the Human BCL2 P1 Promoter Modulates Transcription, J Am Chem Soc 138 (2016) 2563-2570.
[105] S. Benabou, R. Ferreira, A. Avino, C. Gonzalez, S. Lyonnais, M. Sola, R. Eritja, J. Jaumot, R. Gargallo, Solution equilibria of cytosine- and guanine-rich sequences near the promoter region of the n-myc gene that contain stable hairpins within lateral loops, Biochim Biophys Acta 1840 (2014) 41-52.
[106] Z. Yu, V. Gaerig, Y. Cui, H. Kang, V. Gokhale, Y. Zhao, L.H. Hurley, H. Mao, Tertiary DNA structure in the single-stranded hTERT promoter fragment unfolds and refolds by parallel pathways via cooperative or sequential events, J Am Chem Soc 134 (2012) 5157-5164.

[107] M.H. Kuo, Z.F. Wang, T.Y. Tseng, M.H. Li, S.T. Hsu, J.J. Lin, T.C. Chang, Conformational transition of a hairpin structure to G-quadruplex within the WNT1 gene promoter, J Am Chem Soc 137 (2015) 210-218.
[108] K.W. Lim, P. Jenjaroenpun, Z.J. Low, Z.J. Khong, Y.S. Ng, V.A. Kuznetsov, A.T. Phan, Duplex stem-loop-containing quadruplex motifs in the human genome: a combined genomic and structural study, Nucleic Acids Res 43 (2015) 5630-5646.
[109] K.W. Lim, Z.J. Khong, A.T. Phan, Thermal stability of DNA quadruplex-duplex hybrids, Biochemistry 53 (2014) 247-257.
[110] 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.
[111] S. Rankin, A.P. Reszka, J. Huppert, M. Zloh, G.N. Parkinson, A.K. Todd, S. Ladame, S. Balasubramanian, S. Neidle, Putative DNA quadruplex formation within the human c-kit oncogene, J Am Chem Soc 127 (2005) 10584-10589.
[112] S. Cogoi, M. Paramasivam, V. Filichev, I. Geci, E.B. Pedersen, L.E. Xodo, Identification of a new G-quadruplex motif in the KRAS promoter and design of pyrene-modified G4-decoys with antiproliferative activity in pancreatic cancer cells, J Med Chem 52 (2009) 564-568.
[113] S. Cogoi, L.E. Xodo, G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, Nucleic Acids Res 34 (2006) 2536-2549.
[114] J. Dai, D. Chen, R.A. Jones, L.H. Hurley, D. Yang, NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region, Nucleic Acids Res 34 (2006) 5133-5144.
[115] A. Membrino, S. Cogoi, E.B. Pedersen, L.E. Xodo, G4-DNA formation in the HRAS promoter and rational design of decoy oligonucleotides for cancer therapy, PLoS One 6 (2011) e24421.
[116] M.I. Onyshchenko, T.I. Gaynutdinov, E.A. Englund, D.H. Appella, R.D. Neumann, I.G. Panyutin, Quadruplex formation is necessary for stable PNA invasion into duplex DNA of BCL2 promoter region, Nucleic Acids Res 39 (2011) 7114-7123.
[117] S.L. Palumbo, R.M. Memmott, D.J. Uribe, Y. Krotova-Khan, L.H. Hurley, S.W. Ebbinghaus, A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity, Nucleic Acids Res 36 (2008) 1755-1769.
[118] H. Sun, J. Xiang, Y. Shi, Q. Yang, A. Guan, Q. Li, L. Yu, Q. Shang, H. Zhang, Y. Tang, G. Xu, A newly identified G-quadruplex as a potential target regulating Bcl-2 expression, Biochim Biophys Acta 1840 (2014) 3052-3057.
[119] J.D. Schonhoft, R. Bajracharya, S. Dhakal, Z. Yu, H. Mao, S. Basu, Direct experimental evidence for quadruplex-quadruplex interaction within the human ILPR, Nucleic Acids Res 37 (2009) 3310-3320.
[120] B. Pagano, A. Randazzo, I. Fotticchia, E. Novellino, L. Petraccone, C. Giancola, Differential scanning calorimetry to investigate G-quadruplexes structural stability, Methods 64 (2013) 43-51.
[121] L. Payet, J.L. Huppert, Stability and structure of long intramolecular G-quadruplexes, Biochemistry 51 (2012) 3154-3161.
[122] L. Petraccone, C. Spink, J.O. Trent, N.C. Garbett, C.S. Mekmaysy, C. Giancola, J.B. Chaires, Structure and stability of higher-order human telomeric quadruplexes, J Am Chem Soc 133 (2011) 20951-20961.
[123] S. Haider, G.N. Parkinson, S. Neidle, Molecular dynamics and principal components analysis of human telomeric quadruplex multimers, Biophys J 95 (2008) 296-311.

[124] L. Petraccone, J.O. Trent, J.B. Chaires, The tail of the telomere, J Am Chem Soc 130 (2008) 16530-16532.
[125] S. Cogoi, A.E. Shchekotikhin, L.E. Xodo, HRAS is silenced by two neighboring G-quadruplexes and activated by MAZ, a zinc-finger transcription factor with DNA unfolding property, Nucleic Acids Res 42 (2014) 8379-8388.
[126] A.R. Cousins, D. Ritson, P. Sharma, M.F. Stevens, J.E. Moses, M.S. Searle, Ligand selectivity in stabilising tandem parallel folded G-quadruplex motifs in human telomeric DNA sequences, Chem Commun (Camb) 50 (2014) 15202-15205.
[127] A. Cummaro, I. Fotticchia, M. Franceschin, C. Giancola, L. Petraccone, Binding properties of human telomeric quadruplex multimers: a new route for drug design, Biochimie 93 (2011) 1392-1400.
[128] I. Fotticchia, C. Giancola, L. Petraccone, G-quadruplex unfolding in higher-order DNA structures, Chem Commun (Camb) 49 (2013) 9488-9490.
[129] L. Petraccone, Higher-order quadruplex structures, Top Curr Chem 330 (2013) 23-46.
[130] A. Matsugami, T. Okuizumi, S. Uesugi, M. Katahira, Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA, J Biol Chem 278 (2003) 28147-28153.
[131] C. Broxson, J. Beckett, S. Tornaletti, Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene, Biochemistry 50 (2011) 4162-4172.
[132] L. Bauer, K. Tluckova, P. Tohova, V. Viglasky, G-quadruplex motifs arranged in tandem occurring in telomeric repeats and the insulin-linked polymorphic region, Biochemistry 50 (2011) 7484-7492.
[133] S.L. Palumbo, S.W. Ebbinghaus, L.H. Hurley, Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands, J Am Chem Soc 131 (2009) 10878-10891.
[134] S. Selvam, Z. Yu, H. Mao, Exploded view of higher order G-quadruplex structures through click-chemistry assisted single-molecule mechanical unfolding, Nucleic Acids Res 44 (2016) 45-55.
[135] J.B. Chaires, J.O. Trent, R.D. Gray, W.L. Dean, R. Buscaglia, S.D. Thomas, D.M. Miller, An improved model for the hTERT promoter quadruplex, PLoS One 9 (2014) e115580.
[136] 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.
[137] O. Mendoza, J. Elezgaray, J.L. Mergny, Kinetics of quadruplex to duplex conversion, Biochimie 118 (2015) 225-233.
[138] Y. Cui, D. Kong, C. Ghimire, C. Xu, H. Mao, Mutually Exclusive Formation of G-Quadruplex and i-Motif Is a General Phenomenon Governed by Steric Hindrance in Duplex DNA, Biochemistry 55 (2016) 2291-2299.
[139] M.L. Greco, M. Folini, C. Sissi, Double stranded promoter region of BRAF undergoes to structural rearrangement in nearly physiological conditions, FEBS Lett 589 (2015) 2117-2123.
[140] A.N. Lane, J.B. Chaires, R.D. Gray, J.O. Trent, Stability and kinetics of G-quadruplex structures, Nucleic Acids Res 36 (2008) 5482-5515.

3. Aim of the project

G-quadruplexes are promising therapeutic targets in anticancer therapy due to their ability in regulating oncogenes expression. Despite a great scientific effort in identification of selective G-quadruplex binders, nowadays none of them reached the clinic. Several reasons can contribute to this poor outcome. In particular, one of the major issues in G-quadruplex ligands discovery and development remains the fine description of the physiological targets. Indeed, little information is reported concerning kinetic and thermodynamic aspects of G-quadruplex folding. Deeper studies in these research areas are expected to provide with a more comprehensive understanding of the structural arrangement of oncogene promoters which, ultimately, could lead to a more precise definition of the potential targets.

The aim of this Ph.D.'s project lies within this premise and the herein reported works have been designed to provide a detailed characterization of G-quadruplexes focusing on their kinetic and thermodynamic behaviors in order to obtain a clear picture of their structures in physiological conditions. The attention has been centered on promoter regions of two oncogenes: c-KIT and EGFR. Both encode for tyrosine kinase receptors which are physiologically involved, once activated by endogenous ligands, in cell growth and proliferation. Overexpression and mutations of these oncogenes have been associated to a number of human cancers, such as GIST (gastro-intestinal stromal tumor), melanoma and leukemia (in the case of c-KIT) and glioblastoma, breast cancer and lung cancer (in the case of EGFR). Tyrosine kinases inhibitors are the main available therapeutics. However, intrinsic or acquired drug resistance are often developed making the treatment ineffective and new therapeutic strategies are needed. In this context, G-quadruplexes within their promoters have been considered as interesting targets, so that small molecules could exert anticancer activity as a result of DNA structure recognition.

At the beginning of this Ph.D.'s project, three G-rich regions within c-KIT promoter, named kit2, kit* and kit1, were found to be able to fold into G-quadruplexes and high-resolution structures of kit1 and kit2 were available. Moreover, all of them were reported to be involved in modulating c-KIT transcription. However, some pieces of the puzzle were still missing. From a medicinal chemistry point of view, this lack of knowledge led to poor correlation between G-quadruplex recognition by small molecules and regulation of c-KIT transcription, thus making the prediction of effect of G-quadruplex binders, in terms of

anticancer activity, a real challenge. The effort in completing the characterization of c-KIT promoter has been summarized in three scientific publications which have been reported in the following chapter. Three main topics have been addressed:

- The existence and importance of kit* G-quadruplex have been neglected for long time. Nevertheless, kit* sequence overlaps the consensus site of SP1 transcription factor which is an essential protein in c-KIT transcription and thus kit* G-quadruplex might be severely involved in modulating oncogene expression. In this Ph.D.'s project, high-resolution structure of kit* G-quadruplex has been determined by using NMR spectroscopy.
- High-resolution structures of kit2 G-quadruplex have been obtained on mutated sequences and previous studies highlighted kinetic issues in the folding process of this sequence. This Ph.D.'s work provided a detailed analysis of the folding pathway of kit2 G-quadruplex in physiologically relevant experimental conditions.
- kit2 and kit* G-quadruplexes are separated by only two nucleotides. This proximity could enable positive or negative influences of each G-quadruplex unit on the folding features of the neighboring structure. Part of the Ph.D.'s project has been devoted to evaluate the folding behavior of these two G-quadruplex arrangements when they are inserted within the same sequence.

On the contrary, EGFR promoter had not been studied yet. An *in silico* prediction tool identified two G-rich regions, EGFR-272 and EGFR-37, with the potential to fold into G-quadruplex structures. Thus, in the next section the results of Ph.D. work on EGFR promoter are reported. Therein, EGFR-272 has been considered as suitable starting point to explore the architecture of this oncogene promoter. In this case, a set of biophysical approaches, which considered kinetic and thermodynamic aspects in G-quadruplex folding, have proved essential in obtaining structural data when high-resolution techniques failed in determining EGFR-272 structure because of its complex equilibria in solution.

4. Scientific publications

Two-quartet kit* G-quadruplex is formed via double-stranded pre-folded structure

Anita Kotar[1], Riccardo Rigo[2], Claudia Sissi[2,*] and Janez Plavec[1,3,4,*]

[1] Slovenian NMR Center, National Institute of Chemistry, SI-1000, Ljubljana, Slovenia,

[2] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

[3] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia and

[4] EN-FIST Center of Excellence, Ljubljana, Slovenia

* To whom correspondence should be addressed. Tel: +386 1 4760353; Fax: +386 1 4760300; E-mail: janez.plavec@ki.si. Correspondence may also be addressed to Claudia Sissi. Tel: +39 049 827 5711; Fax: +39 049 827 5366; E-mail: claudia.sissi@unipd.it.

ABSTRACT

In the promoter of *c-KIT* proto-oncogene, whose deregulation has been implicated in many cancers, three G-rich regions (kit1, kit* and kit2) are able to fold into G-quadruplexes. While kit1 and kit2 have been studied in depth, little information is available on kit* folding behavior despite its key role in regulation of *c-KIT* transcription. Notably kit* contains consensus sites for SP1 and AP2 transcription factors. Herein, a set of complementary spectroscopic and biophysical methods reveals that kit*, d[**GG**CGA**GG**AGG**GG**CGT**GG**CCGGC], adopts a chair type antiparallel G-quadruplex with two G-quartets at physiological relevant concentrations of KCl. Heterogeneous ensemble of structures is observed in the presence of Na^+ and NH_4^+ ions, which however stabilize pre-folded structure. In the presence of K^+ ions extensive stacking interactions of adenine and thymine residues on the top G-quartet contribute to structural stability together with a G10•C18 base pair and a fold-back motif of the five residues at the 3'-terminal under the bottom G-quartet. The 3'-tail enables formation of a bimolecular pre-folded structure that drives folding of kit* into a single G-quadruplex. Intriguingly, kinetics of kit* G-quadruplex formation matches timescale of transcriptional processes and might

Submitted – Kotar, A.; Rigo, R.; Sissi, C.; Plavec, J. *Nucleic Acids Res.* **2018**

demonstrate interplay of kinetic and thermodynamic factors for understanding regulation of c-KIT proto-oncogene expression.

INTRODUCTION

The *c-KIT* proto-oncogene encodes a transmembrane tyrosine kinase receptor (c-kit) (1,2), which participates, once activated by endogenous ligands, in a broad range of physiological processes, including cell proliferation, migration, maturation and survival (3). The overexpression and/or mutations of *c-KIT* gene have been implicated in a number of human cancers, like gastrointestinal stromal tumors, pancreatic cancer, leukemia and melanoma (4,5). In the proximal promoter of *c-KIT* three guanine (G)-rich regions have been identified that are able to fold into G-quadruplexes (3,6,7). In general, G-quadruplexes consist of stacked G-quartets, cyclic arrangements of four guanine residues held together by Hoogsteen hydrogen bonds and stabilized by a central cation (8). The G-quadruplex structures are highly polymorphic and are influenced by factors such as oligonucleotide sequence and concentration, type and concentration of cations in solution, crowding conditions and presence of other biomolecules (9-18). In general, under *in vitro* conditions G-quadruplexes adopt parallel, antiparallel and hybrid (3+1) topologies characterized by different orientations of the four G-rich tracts. Three types of loops that connect guanine residues involved in G-quartets are typical for G-quadruplexes and can adopt propeller, diagonal and edgewise orientations (19). G-quadruplexes have been clearly shown to exist inside human cells (20-23). Moreover, bioinformatic analysis revealed that G-rich repeats are not randomly distributed in the human genome but are mostly located in regions associated with a number of essential biological processes such as transcription, replication and telomere maintenance (24,25). In addition to *c-KIT* proto-oncogene, more than 40% of genes that encode human proteins contain one or more putative G-quadruplex forming motifs in the promoter regions (26). Stabilization of G-quadruplex structures is a novel promising strategy to regulate gene expression at transcriptional and translational levels (27-31). For example, stabilization of G-quadruplex structures by small molecule ligands in the promoter region of *c-KIT* gene has been linked with inhibition of its transcriptional activity and reduction of the expression of c-kit tyrosine kinase receptor which possibly lead to favorable anticancer effects (32-34).

The G-quadruplex-forming regions within the *c-KIT* promoter have been named kit1, kit* and kit2 (3,6,7). They are closely clustered and separated from each other by only few nucleotides (six between kit1 and kit*, and only one between kit* and kit2, Figure 1A). kit1 and kit2 oligonucleotides contain four GGG-repeats each that fold into parallel G-quadruplexes with three G-quartets under *in vitro* conditions. Their structural characteristics as well as thermodynamic behavior have been extensively studied (6,35-42). Worth of note, kit1 G-quadruplex contains unique structural features: an isolated guanine that is involved in the formation of the G-quartet core and a stem-loop consisting of five nucleotides (36). In the case of kit2, complex folding pathway was observed involving stable folding intermediates that convert into thermodynamically stable monomeric and dimeric parallel G-quadruplex structures (37,38). High-resolution structure of kit* is not available. It is noteworthy that kit* contains consensus sites for SP1 and AP2 transcription factors (3,7,43). Therefore, its duplex–quadruplex equilibrium might modulate *c-KIT* transcription. Additionally, we have demonstrated recently that cross-talking interactions stabilize kit* and kit2 G-quadruplexes and impair their refolding into duplex form in the presence of the complementary strands (44). We deemed it essential to get deeper insights into structural details of kit* G-quadruplex.

Herein, we describe the results of our studies on structural features of kit*, d[**GG** CGA **GG** AGG **GG** CGT **GG** CCGGC], performed by using a set of complementary spectroscopic and biophysical methods. A kit * sequence containing a GGGG-tract as well as four GG-repeats could be expected to fold into multiple G-quadruplex structures. In contrast, kit* adopts a single antiparallel two G-quartet G-quadruplex at physiologically relevant temperature (37°C) and concentration of K^+ ions (100 mM) as revealed by NMR. Loop regions consisting of three regions of three residues each are prone to interact with each other as well as with the 3'-tail and thus contribute to stabilization of the structure. Two adenine residues in the loop regions provide excellent opportunity to evaluate their stacking interactions or exposure to solvent with fluorescence spectroscopy. Searching for a rationale for the observed conformational selection of kit* uncovered importance of the last five residues at the 3'-terminal. Shortening of the 3'-tail disturbed its ability to promote formation of a dimeric assembly in the absence of K^+ ions and was expected to reveal a novel pre-folded structure that drives folding into a single G-quadruplex. We deemed it essential to explore folding behavior of kit*. Upon addition of K^+ ions, kit* folds into a G-quadruplex within seconds, but if the 3'-tail is absent folding pathway changes

and many parallel G-quadruplexes are formed. Interesting interplay of kinetic and thermodynamic factors orchestrated by the 3'-tail uncovered its crucial role in folding process, thermal stabilization and structural integrity of kit* G-quadruplex thus alike contributing to fine regulation of c-KIT proto-oncogene expression in physiological conditions.

MATERIALS AND METHODS

NMR sample preparation

The isotopically unlabeled and residue-specific partially ^{15}N, ^{13}C-labeled (10% guanine and 4% cytosine residues) kit* and modified oligonucleotides (Table 1) were synthesized and prepared as described before (45,46). Oligonucleotides were purified and desalted with the use of Amicon Ultra-15 Centrfugal Filter Units to give NMR samples with concentration between 0.1 and 2.0 mM per strand. NMR samples of kit* and modified oligonucleotides were prepared by dissolving desalted oligonucleotides in 300 µL of 90%/10% mixture of $H_2O/^2H_2O$ in 20 mM potassium phosphate buffer (pH 7.4) or in 20 mM lithium cacodylate buffer (pH 7.2) and in the presence of KCl with varying concentrations from 10 to 250 mM as well as in 100 mM concentrations of LiCl, NaCl, NH_4Cl and KCl. For samples dissolved in 100% ^2H_2O we lyophilized solutions of oligonucleotides, KCl and buffer separately and subsequently dissolved and mixed them together in 300 µL 100% ^2H_2O.

NMR experiments

NMR data were collected on Agilent (Varian) NMR System 600 and 800 MHz spectrometers equipped with triple-resonance HCN cryogenic and PFG One NMR probes in the temperature range from 5 to 65°C. The majority of NMR spectra were recorded at 37°C with samples dissolved in 90%/10% $H_2O/^2H_2O$. DPFGSE_NOESY spectra were acquired at mixing times of 40, 80, 150, 200, 300 and 450 ms. ^{13}C- and ^{15}N-edited HSQC experiments were recorded on residue-specifically ^{15}N, ^{13}C-labelled oligonucleotides. DQF-COSY, TOCSY (τ_m 40 and 80 ms), ^1H–^{31}P COSY and 1D ^{31}P NMR spectra NMR spectra were acquired on NMR samples prepared in 100% ^2H_2O solution. NMR spectra were

processed and analyzed by using VNMRJ (Varian Inc.) and the Sparky (UCSF) software. Cross-peak assignment and integration was achieved using Sparky (UCSF) software.

Circular dichroism (CD) spectroscopy

Circular dichroism spectra were acquired on a JASCO J-810 spectropolarimeter equipped with a Peltier temperature controller. CD spectra were recorded from 230 to 320 nm with the following parameters: scanning speed of 100 nm/min, band width of 2 nm, data interval of 0.5 nm and response of 2 s. Measurements were performed using a 1 cm path length quartz cuvette at oligonucleotide concentration of ~4 µM in 10 mM TRIS (pH 7.5). CD titrations with KCl or LiCl were performed at 25°C. After each titration step the system was left to equilibrate before spectra acquisition. CD signal variations as a function of the concentration of cations at the maximum were fitted by applying one-site saturation model (Supplementary Data, Equation 1). For CD kinetic experiments, KCl was added manually to the cuvette from a stock solution and mixing was provided by an in-cuvette magnetic stirring bar. After a mixing time of 10 s spectra acquisition was initiated using an interval scan of 30 min. The temperature was maintained at 37°C (Supplementary Data, Equation 2). CD melting studies were performed on kit* and kit*17. The experiments were carried out between 25 and 95°C in 150 mM KCl. The heating rate was 50°C/hour, each 2°C the temperature was held for 5 minutes, the corresponding CD signals were recorded at 294 nm and 264 nm, in the case of kit* and kit*17, respectively, and fitted according to van't Hoff formalism (Supplementary Data, Equations 3 and 4).

Fluorescence spectroscopy

Fluorescence measurements were performed on kit*-5-2AP and kit*-8-2AP by using a JASCO-FP-6500 spectrofluorometer equipped with a Peltier Temperature controller. Emission spectra were recorded in the 320-460 nm range with the excitation wavelength fixed at 305 nm. The scanning speed was 200 nm/min, the response was 1 s, and both the emission and the excitation band widths were 3 nm. The fluorescently labeled oligonucleotides were diluted to a concentration of 0.4 µM in 10 mM TRIS (pH 7.5). The effect of reaching 150 mM KCl on the fluorescence spectra was monitored at 25°C. Acrylamide quenching experiments were performed on previously folded

oligonucleotides in 150 mM KCl. The quenching efficiency was monitored at 370 nm at 25°C and plotted as a function of the acrylamide concentration. Data points were fitted according to Stern-Volmer formalism to obtain quenching parameters (Supplementary Data, Equations 5 and 6).

Restraints, structure calculations and molecular dynamics simulations

For structural calculations we used NOE-derived distance (force constant 20 kcal mol^{-1} $Å^{-2}$), hydrogen bonds (force constant 20 kcal mol^{-1} $Å^{-2}$) and torsion angles χ (force constant 200 kcal mol^{-1} rad^{-2}) restraints. NOE (Nuclear Overhauser Effect) distance restraints for protons were obtained from a 2D NOESY spectrum recorded with a mixing time of 450 ms. The average volume of H7-H6 cross-peaks of T15 was used as the distance reference of 3.0 Å. Cross-peaks were classified as strong (1.8 - 3.6 Å), medium (2.5 - 5.0 Å), weak (3.5 - 6.5 Å) and very weak (4.5 - 7.5 Å). The torsion angle χ restraints were applied for all of residues based on the intensity of respective H8-H1' 2D NOESY cross-peaks, (for region 200° - 280° for adenines and guanines with *anti*-orientation along glycosidic bonds, 25° - 95° for guanines with *syn*-orientation along glycosidic bonds, 170° - 310° for thymine and cytosine residues). The structures of kit* G-quadruplex were calculated by the simulated annealing (SA) simulations. The force field parameters were adopted from the Generalized Amber force field (47) (Supplementary Data). For molecular dynamics simulations, the kit* G-quadruplex was placed in a truncated octahedral box of TIP3P water molecules with the box border at least 10 Å away from any atoms of the G-quadruplex. Extra K^+ ions were added to neutralize the negative charges of the G-quadruplex. Prior to MD simulations, the system was subjected to a series of minimizations and equilibrations (Supplementary Data).

RESULTS

K^+ ions induce folding of kit* into a well-defined G-quadruplex, while Li^+, Na^+ and NH_4^+ ions stabilize pre-folded structure through GC base pairs

^{1}H NMR spectrum of kit* in the presence of 100 mM LiCl at 37°C exhibits two broad signals at δ 12.94 and 13.14 ppm that are typical for imino protons of guanine residues involved

in GC base pair formation (Figure S1). No imino signals are detected in the region between δ 10.5 and 12.5 ppm which would be characteristic for imino protons of guanine residues involved in G-quartets. Thus, the two signals arise from formation of a pre-folded structure of kit* (pre-kit*) that does not contain stacked G-quartets (*vide infra*). In the presence of 100 mM NH_4Cl at 37°C similar NMR fingerprint of imino signals has been detected for kit* as in the presence of LiCl suggesting formation of pre-kit*, and absence of G-quadruplex structure (Figure S1). Pre-kit* is predominant species also in the presence of 100 mM NaCl at 37°C. However, we also observe eight weak and broad imino signals between δ 11.1 and 12.2 ppm in ^{1}H NMR spectrum at 100 mM NaCl suggesting a minor formation of G-quadruplex (Figure S1). In contrast, ^{1}H NMR spectrum of kit* in the presence of 100 mM KCl at 37°C exhibits eight narrow and well-resolved signals between δ 11.20 and 12.13 ppm, which are consistent with formation of a single G-quadruplex with two G-quartets (Figures 1B and S1). The additional signal at δ 13.50 ppm suggests the presence of a GC base pair in Watson-Crick geometry. Interestingly, in the presence of K^+ ions we also detect two broad signals at δ 12.94 and 13.14 ppm of lower intensity (~10%) corresponding to pre-kit* that are observed in LiCl, NH_4Cl and NaCl containing solutions (Figures S1-S3). The coexistence of the two distinct species in the presence of the K^+ ions has been confirmed by DOSY NMR experiment at 5°C displaying two different translational diffusion coefficients of 0.87×10^{-10} (kit* G-quadruplex) and 0.45×10^{-10} $m^2 s^{-1}$ (pre-kit*) (Figure S4). In full agreement, analysis of kit* electrophoretic mobility showed co-presence of two species, monomeric G-quadruplex and dimeric pre-kit* (Figure S5). Despite the observation of minor species in solution, kit* G-quadruplex is predominantly favored structure in the presence of K^+ ions.

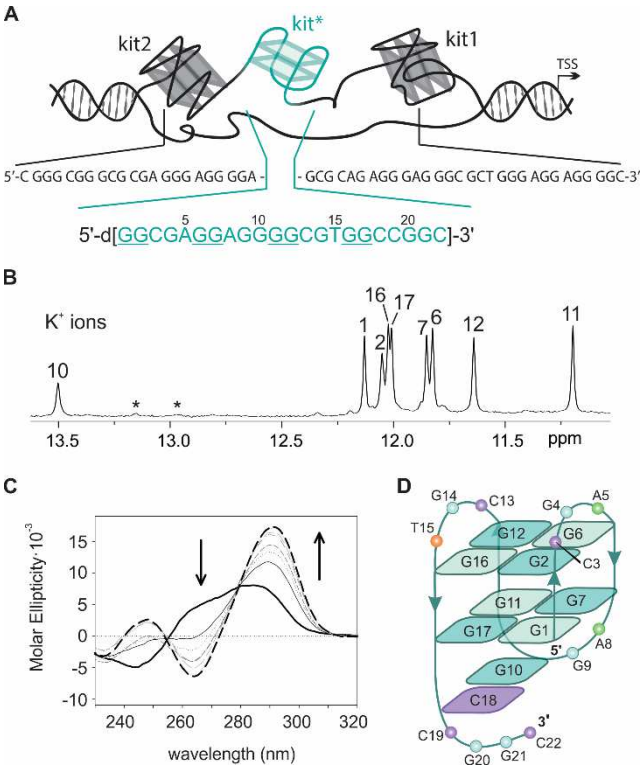


Figure 1. Characterization of a G-quadruplex adopted by kit*. A) Schematic representation of G-rich region in the promoter of *c-KIT* gene and its sequence. The arrow indicates the transcription start site (TSS). B) Imino region of ^{1}H NMR spectrum of kit* in the presence of 100 mM KCl, 0.5 mM kit* per strand, pH 7.4 and 37°C on an 800 MHz NMR spectrometer. Assignments are shown above the individual signals. Signals marked with * correspond to pre-kit*. C) CD spectra of kit* at 4 µM concentration per strand titrated with KCl in 10 mM TRIS, pH 7.5 at 25°C. Arrows indicate changes in CD spectra from 0 mM (black solid line) to 350 mM KCl (black dashed line). D) Topology of G-quadruplex adopted by kit*. *Anti* and *syn* guanines in G-quartets are marked with darker and lighter shades of cyan, respectively.

Titration of kit* with KCl (0-250 mM) revealed nine (8+1) resolved imino ^{1}H NMR signals (Figures 1B and S6). By increasing K^{+} ion concentration the intensity of imino signals increased and reached a plateau at 50 mM KCl, while no new imino signals appeared in ^{1}H NMR spectra. The CD spectrum of kit* in the presence of KCl suggests formation of an antiparallel topology with the minimum and maximum at 264 and 294 nm, respectively (Figures 1C and S7). Isodichroic points at 255 and 279 nm are conserved between 0 and 350 mM KCl indicating that the initial folded state(s) interconvert into the final antiparallel form without participation of stable kinetic intermediates. Analysis of CD signal variation

at 294 nm as a function of KCl concentration provided a dissociation constant (K_D) of 14.7 ± 1.0 mM for K^+ ions at 25°C (Equation 1). Furthermore, in ^{1}H NMR spectra no significant variation of the signals associated with changes in kit* G-quadruplex structure was observed at concentrations of kit* between 0.1 and 1.7 mM (Figure S8). The above results indicate that kit* forms a monomeric G-quadruplex. At concentrations of kit* above 1.7 mM per strand we detected a broad and unresolved hump in the imino region of ^{1}H NMR spectra that might be related to formation of higher order structures (Figure S9). The intermolecular association is observed also after overnight annealing (95→25°C) at 1.0 mM and higher concentrations of kit* per strand. Noteworthy, after keeping NMR samples (0.4 mM kit* per strand) for few months at room temperature we observed decrease of signals of kit* G-quadruplex and significant increase of signals at δ 12.94 and 13.14 ppm associated with GC base pairs involving G20 and G21.

An antiparallel G-quadruplex with a two G-quartet core, three edgewise loops and a 3'-tail

Residue-specific, partial ^{13}C,^{15}N-labeling of guanine (10%) and cytosine (4%) residues enabled unambiguous assignment of H1 and H8 as well as H6 proton resonances of guanines and cytosines, respectively (Figures S10-S14). H8 and H2 protons of A8 were assigned by its substitution with thymine residue (i.e. in kit*A8T, Table 1) which resulted in minimal changes in ^{1}H NMR spectrum with respect to kit* with exception of signals for A8 (Figure S15). After assignment of A8 we identified the aromatic protons of A5 with the help of characteristic heteronuclear correlations involving C8 and C2 atoms in ^{1}H-^{13}C HSQC spectra of kit* (Figure S16). The spectral assignment of resonances of kit* were completed with the use of through-bond (DQF-COSY and TOCSY experiments) and through-space (NOESY) correlations (Figures 2 and S17-S21, Table S1).

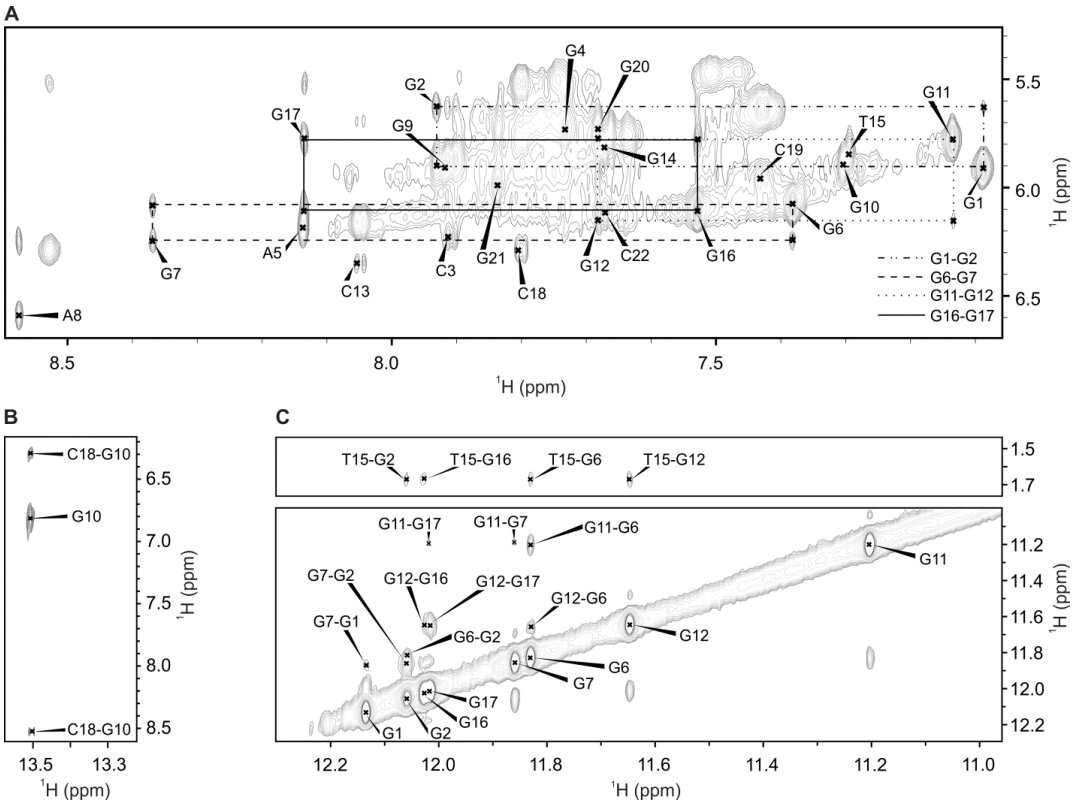


Figure 2. Expansions of NOESY spectrum (τ_m 450 ms) of kit*. A) The aromatic-anomeric region with intra-residue NOE cross-peaks marked with residue numbers. For clarity, only sequential NOE cross-peaks of G1-G2, G6-G7, G11-G12 and G16-G17, characteristic for 5'-*syn-anti*-3' steps, are marked with different line styles. B) NOE cross-peaks between G10 imino and C18 amino protons. C) Imino-methyl and imino-imino regions. The NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4 and 37°C on an 800 MHz NMR spectrometer.

The characteristic H1-H8 connectivities in NOESY spectra allowed us to establish topology of kit* G-quadruplex involving two G-quartets with the following hydrogen-bond directionalities: G1→G17→G11→G7 and G2→G6→G12→G16 (Figures 1D and S18). The guanine residues of G1-G17-G11-G7 quartet exhibit a clockwise donor-acceptor hydrogen-bonding directionality, while those of G2-G6-G12-G16 quartet display an anti-clockwise directionality. Orientation of G-quartets that are stacked one above the other is additionally supported by inter-quartet H1-H1 NOE cross-peaks of G2-G7, G6-G11 and G12-G17 residues (Figure 2C). Four distinct and strong cross-peaks observed in the H8-H1' region of NOESY spectrum indicate that G1, G6, G11 and G16 residues predominantly adopt a s*yn* conformation along glycosidic torsion angles (Figure S20). The "rectangular"

pattern of NOE cross-peaks of H1'$_{(n)}$-H8$_{(n+1)}$-H1'$_{(n+1)}$-H8$_{(n)}$ for G1-G2, G6-G7, G11-G12 and G16-G17 demonstrates that these sequential connectivities are characteristic of 5'-*syn-anti*-3' steps in antiparallel topology (Figure 2A).

Three edgewise loops are arranged in an anticlockwise manner forming a '-(lll)' topology (17). Each loop connects the two neighboring G-tracts and consists of three nucleotides (1: C3-G4-A5, 2: A8-G9-G10 and 3: C13-G14-T15). A5 and T15 from the first and the third loop are stacked on G2-G6-G12-G16 quartet which is supported by the observation of NOE cross-peaks between A5 H8 and G6 H1 as well as between T15 Me protons and H1 of G2, G6, G12 and G16 (Figure 2C), T15 H6 and G16 H1. The NOE cross-peak between G4 H8 and G6 H8 indicates that G4 residue from the first loop is oriented into the groove. A terminal 3'-tail comprises C18-C19-G20-G21-C22 segment. The NOE cross-peaks between G10 H1 and C18 NH_2 suggest formation of G10•C18 base pair in Watson-Crick geometry (Figure 2B). Stacking of G10•C18 base pair below G1-G17-G11-G7 quartet is supported by NOE cross-peaks between G11 H1 and G10 H8, H2'/2", H5'/5" as well as between G7 H8 and G10 H8.

Table 1. Sequences of kit* and its analogues

Oligonucleotide[#]	1	5	10	15	20
kit*[##]	GG	CGA GG	AGG GG	CGT GG	CCGGC
kit*17[##]	GG	CGA GG	AGG GG	CGT GG	
kit*18[##]	GG	CGA GG	AGG GG	CGT GG	C
kit*19[##]	GG	CGA GG	AGG GG	CGT GG	CC
kit*19TTT	GG	CGA GG	AGG GG	CGT GG	CCTTT
kit*20	GG	CGA GG	AGG GG	CGT GG	CCG
kit*18T	GG	CGA GG	AGG GG	CGT GG	CT
Tkit*18	TGG	CGA GG	AGG GG	CGT GG	C
Akit*18	AGG	CGA GG	AGG GG	CGT GG	C
kit*T15A	GG	CGA GG	AGG GG	CGA GG	CCGGC
kit*A8T	GG	CGA GG	TGG GG	CGT GG	CCGGC
kit*C18,19T	GG	CGA GG	AGG GG	CGT GG	TTGGC
kit*C19T	GG	CGA GG	AGG GG	CGT GG	CTGGC
kit*C18T	GG	CGA GG	AGG GG	CGT GG	TCGGC
kit*C13T	GG	CGA GG	AGG GG	TGT GG	CCGGC
kit*C3T	GG	TGA GG	AGG GG	CGT GG	CCGGC
Tkit*	TGG	CGA GG	AGG GG	CGT GG	CCGGC
Akit*	AGG	CGA GG	AGG GG	CGT GG	CCGGC
c	GC	CGG			
		5	10	15	20
kit*5-2AP[###]	GG	CG-2AP-GG	AGG GG	CGT GG	CCGGC
		5	10	15	20
kit*8-2AP[###]	GG	CGA GG-2AP-GG	GG	CGT GG	CCGGC

[#] Sequences are reported in the 5'-3' direction. Guanine residues involved in G-quartets of kit* are underlined. [##] Melting temperatures are 60.5±0.1°C (TRIS buffer) and 55.6±0.1°C (K-phosphate buffer) for kit*, 50.6±0.6°C, 75.9±0.1°C (TRIS buffer) and

48.8±0.6°C, 65.1±0.6°C (K-phosphate buffer) for kit*17, 45.7±0.4°C, 58.1±0.5°C (K-phosphate buffer) for kit*18, 52.7±0.1°C (K-phosphate buffer) for kit*19. ### 2AP stands for 2-aminopurine, a fluorescent analogue of adenine.

A5 is stacked on G2-G6-G12-G16 quartet, while A8 is exposed to solvent

To map the relative positions of adenine residues with respect to G2-G6-G12-G16 and G1-G17-G11-G7 quartets, we alternatively substituted A5 and A8 with a 2-aminopurine (2-AP) in kit*5-2AP and kit*8-2AP without changing the structure (Figure S22), respectively. The fluorescence of 2-AP is very sensitive to the environment and, in particular, it is strongly quenched by stacking interactions with guanine residues (48). Fluorescence signals of 2-AP in kit*5-2AP and in kit*8-2AP increase upon addition of 150 mM KCl suggesting that fluorophores are more exposed to solvent in the kit* G-quadruplex in comparison to the pre-folded structure (Figure 3A). Moreover, kit*8-2AP exhibits a significantly higher increase of fluorescence signal compared to kit*5-2AP indicating the tendency of 2-AP in the former oligonucleotide to protrude into solution. To fully investigate this difference, the accessibility of the adenines to solvent was further assessed by monitoring the fluorescence quenching of kit*8-2AP and kit*5-2AP upon titration with acrylamide (Figure 3B). In the case of kit*8-2AP we observe a linear fluorescence quenching of 2-AP (Figure 3C) and only one Stern-Volmer constant (K_{sv}) is derived by applying Stern-Volmer formalism (Equation 5). This K_{sv} value (15.04 ±0.15 M^{-1}) indicates that A8 experiences only one microenvironment in which it is exposed to solvent and thus, it does not participate in stacking interactions with the G-quartet. On the other hand, kit*5-2AP exhibits a downward curvature of fluorescence quenching plot (Figure 3C) and can be properly fitted using two different K_{sv} values (Equation 6). The derived value of K_{sv1} (1.14 ± 0.51 M^{-1}) corresponds to a microenvironment in which A5 is engaged in stacking interactions with a G-quartet. Conversely, the K_{sv2} value (36.06 ± 4.05 M^{-1}) is associated with a position where A5 is more exposed to solvent. Observations of two K_{sv} values for adenines stacked on a G-quartet, such as A5, have been reported earlier (48). Even though the NMR data indicate a single fold of G-quadruplex, the structural heterogeneity revealed by fluorescence measurements probably results from highly localized, nanosecond fluctuations in positioning of A5 that are confined to the respective loop regions.

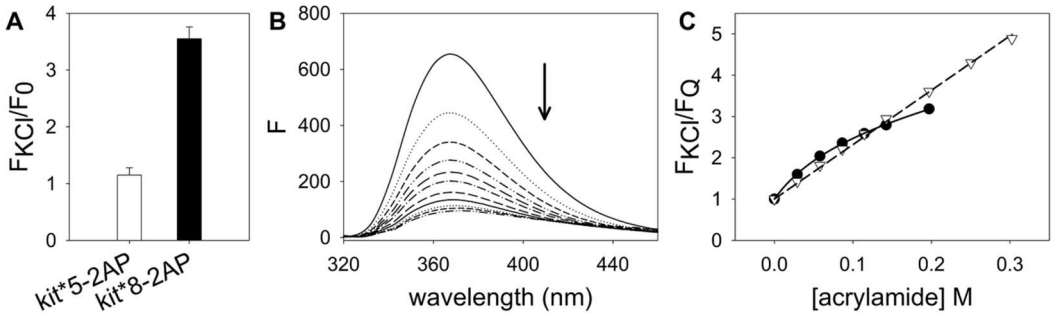


Figure 3. Fluorescence analysis of exposure of 2-AP in kit* to solvent. A) Fold-change emission of 2-AP in kit*5-2AP and kit*8-2AP at 370 nm upon addition of 150 mM KCl in 10 mM TRIS, pH 7.5 at 25°C. B) Fluorescence emission variation of 2-AP in kit*8-2AP upon titration with acrylamide in 10 mM TRIS, pH 7.5, 150 mM KCl at 25°C. C) Fluorescence emission of kit*5-2AP (black dots) and kit*8-2AP (white triangles) at 370 nm plotted as a function of the concentration of acrylamide.

High-resolution structure of kit* G-quadruplex exhibits a well-defined fold-back motif of the 3'-tail

An ensemble of ten structures of kit* G-quadruplex (Figures 4A and S23) was calculated with a simulated-annealing method based on 235 NOE-derived distance, 18 hydrogen bond and 22 torsion-angle restraints (Table S2).

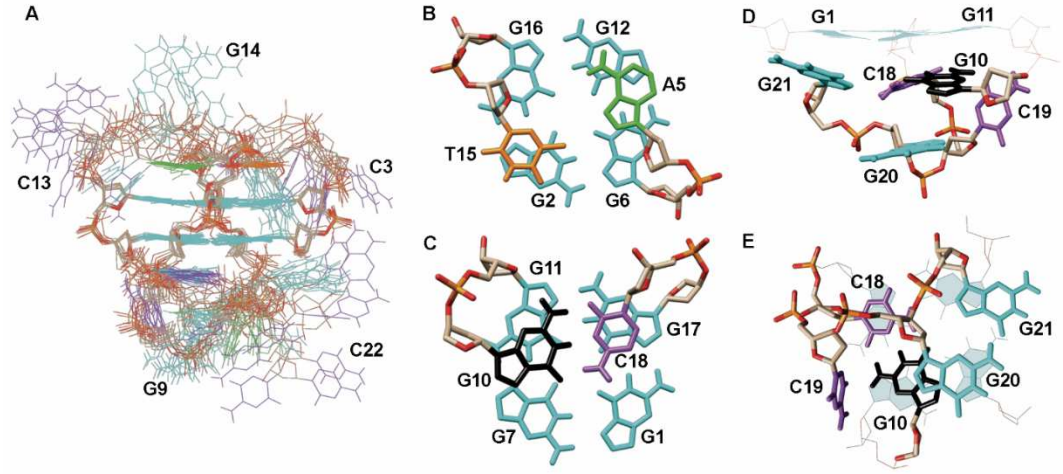


Figure 4. Structure of kit* G-quadruplex (PDB ID: 6GH0). A) An ensemble of 10 structures. B) Stacking of A5 and T15 on G2-G6-G12-G16 quartet. C) G10•C18 base pair stacked on G1-G17-G11-G7 quartet. D) Side and E) bottom view of the fold-back motif of the 3'-tail and stacking of G10•C18 base pair on G1-G17-G11-G7 quartet. Guanines are shown in cyan and black, adenines in green, cytosines in purple and thymines in orange.

The two G-quartet core in the kit* G-quadruplex consisting of G2-G6-G12-G16 (top) and G1-G17-G11-G7 (bottom) quartets is well-defined (RMSD: 0.60 ± 0.13 Å, Figures 4A and S19). G-quartets are not perfectly planar as G7, G16 and G17 residues are oriented out of the plane for ~7° (Figure S23C). A5 and T15 residues are coplanar and stacked on the top G-quartet (Figure 4B). A5 is positioned above G12 and G6 residues, while T15 is stacked over G2. The formation of A5•T15 Watson-Crick base pair is prevented because the amino group of A5 is oriented away from T15 and at the same time T15 is positioned with Watson-Crick edge away from A5. A5 is part of the first edgewise loop, C3-G4-A5, where G4 is oriented in the groove defined by G2-G6 and G1-G7 stems. C3 protrudes into solution and exhibits much more conformational freedom than G4 and A5. T15 is part of the third edgewise loop, C13-G14-T15, in which C13 and G14 are exposed to solution and highly flexible. On the other side of the G-quartet core, the middle edgewise loop (A8-G9-G10) and the 3'-tail (C18-C19-G20-G21-C22) are oriented below the bottom G-quartet. In particular, G10 and C18 are base paired in Watson-Crick geometry and stacked on the bottom G-quartet (Figure 4C). Other residues in the 3'-tail, C19-G20-G21-C22, are arranged in a fold-back motif (Figure 4D,E), in which C19 is perpendicular to the G10•C18 base pair, G20 stacked on it, while G21 is in co-planar position to G10•C18 base pair in nine out of ten structures. Only the 3'-terminal residue, C22, is highly flexible and protrudes into solution. A8 and G9 are oriented towards the fold-back motif of the 3'-tail. Three independent 200 ns unrestrained molecular dynamic (MD) simulations in the presence of K^+ ions and explicit water molecules displayed no transitions of the kit* G-quartet core as well as preservation of G10•C18 base pair and the fold-back motif of the 3'-tail (Figure S24). During MD simulations G20 was stacked on the stable G10•C18 base pair, while C19 was oriented below G20. Interestingly, during one of the simulations, formation of a G10•C18-G21 base triple below the bottom G-quartet was observed, where G21 was hydrogen bonded to C18 (G21 N7 to C18 H42). MD trajectories are consistent with highly dynamic nature of C22 as indicated by NMR data. A8 and G9 of the second loop are oriented towards residues of the fold-back motif adopted by the 3'-tail. Moreover, during MD simulations we observed that A8 can occupy additional conformation, in which it protrudes into solution. However, in both conformations A8 is exposed to solvent. These results are consistent with results of 2-AP fluorescence spectroscopy for A8. On the other side of kit* G-quadruplex, A5 and T15 are preferentially stacked on the top G-quartet, although some buckling from the G-quartet plane is

observed during MD simulations. Local motions of A5 in the structure of kit* G-quadruplex revealed by MD simulations are in good agreement with observation of two K_{sv} constants in 2-AP fluorescence experiments. C3, G4 and C13 from the first and the third loop are oriented towards grooves, while G14 is more flexible and during MD simulations adopts two major conformations in which it protrudes into solution or is oriented over A5.

Modification and shortening of the 3'-tail crucially alter structure of kit* G-quadruplex, while changes in the loop regions are less significant

High-resolution structure of kit* G-quadruplex provides detailed information about residues located in the loops and the 3'-tail. To investigate how they influence integrity of kit* G-quadruplex structure we prepared several analogs of kit* with C→T, A→T and T→A modifications in the loops and the 3'-tail (Figures 5A and S25). The replacement of C3, A8 and C13 with thymines (i.e. kit*C3T, kit*A8T and kit*C13T) results in eight imino ^{1}H NMR signals characteristic for G-quartet formation and a signal at around δ 13.4 ppm indicating presence of G10•C18 base pair. Similarities between ^{1}H NMR spectra of kit*C3T, kit*A8T and kit*C13T with the parent kit* suggests that these modifications have minimal effect on G-quadruplex structure. For all these modified oligonucleotides we observe also imino signals assigned to pre-kit*. Its imino signals in ^{1}H NMR spectra are more intense in kit*C3T, kit*A8T and kit*C13T compared to kit* indicating the presence of pre-kit* in higher percentage with regards to kit*. Modification T15→A in kit*T15A results in better dispersion of imino signals in comparison to kit*, although G-quadruplex formation is reduced by 70% (Figure S26). Modifications of the residues in the 3'-tail in kit*C18T, kit*C19T and kit*C18,19T result in substantial changes in their ^{1}H NMR spectra in comparison to kit*. A hump together with more than eight, broad and overlapped signals in the imino region of ^{1}H NMR spectra of kit*C18T, kit*C19T and kit*C18,19T indicates formation of multiple species, while no imino signal characteristic for G10•C18 base pair is detected. The latter is expected for kit*C18T, but not for kit*C19T, because only C18 is found to be involved in the G10•C18 base pair, stabilizing element of the kit* G-quadruplex.

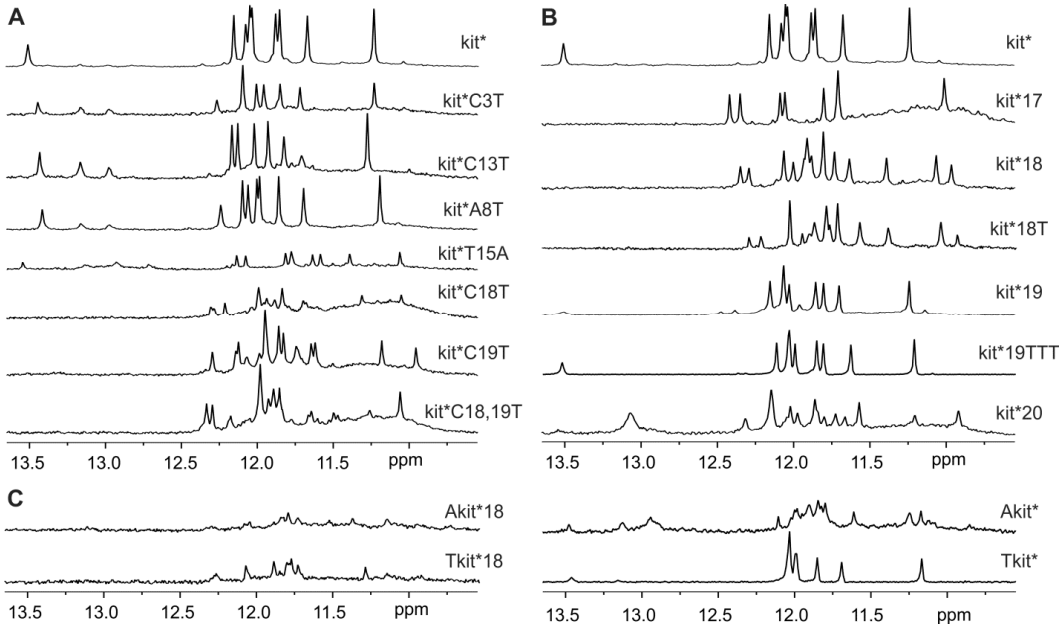


Figure 5. Imino region of ^{1}H NMR spectra of A) modified, B) 3'-shortened and C) at 5'-extended analogs of kit*. NMR spectra were recorded at ~0.3 mM concentration of oligonucleotides per strand, 100 mM KCl, pH 7.4 on 800 and 600 MHz NMR spectrometers. Sequences of oligonucleotides are reported in Table 1.

Since sequence modifications in the 3'-tail crucially affect structure of kit* G-quadruplex we prepared seven oligonucleotides of different lengths whereupon the 3'-tail has been shortened (Figures 5B and S25, Table 1). ^{1}H NMR spectrum of kit*17 without the entire 3'-tail reveals eight signals between δ 11.00 and 12.49 ppm indicating formation of a G-quadruplex with two G-quartets. No imino signals characteristic of GC base pairs have been observed. At the same time, a hump in the imino region increased within hours indicates slow formation of species with poorly defined structures (*vide infra*). In the case of kit*18 that contains C18 residue, which is involved in G10•C18 base pair in kit* G-quadruplex, sixteen overlapped signals are detected in the imino region of ^{1}H NMR spectrum. This indicates formation of two distinct G-quadruplex structures. It is noteworthy that no signal characteristic for the GC base pair is detected in ^{1}H NMR spectrum of kit*18 at 37°C. Thus, no GC base pair is formed, or it is very exposed to solvent because it is not protected by the 3'-tail. In kit*18T an additional thymine residue at the 3'-end does not induce formation of expected GC base pair or lead to a single G-quadruplex structure. In the case of kit*19, eight imino signals and a broad signal at δ 13.51 ppm were observed in ^{1}H NMR spectrum. Moreover, we detected at least four

additional imino signals with lower intensity indicating the presence of a minor G-quadruplex structure in the sample of kit*19. Formal extension with three additional thymine residues at the 3'-end was favorable for formation of one G-quadruplex structure and formation of GC base pair as revealed by nine (8+1) imino signals in ^{1}H NMR spectrum of kit*19TTT. TTT-end helps to stabilize kit* G-quadruplex structure trough additional hydrogen bonds, electrostatic and hydrophobic interactions with residues from the second loop similarly as G20-G21-C22 fragment in kit*. Formation of multiple G-quadruplex species was observed for kit*20.

Extension of oligonucleotide kit*18 at the 5'-end with thymine (i.e. Tkit*18) or adenine (i.e. Akit*18) residues reduced formation of G-quadruplexes by 80% as revealed by weak signals in the imino region of ^{1}H NMR spectrum (Figures 5C and S25, Table 1). Interestingly, for Tkit*18 eight imino signals detected in ^{1}H NMR spectrum suggest presence of one major G-quadruplex. Contrary, extension at the 5'-end of kit* with thymine (i.e. Tkit*) was not critical for G-quadruplex structure. Elongation with adenine residue (i.e. Akit*) of kit* is more relevant in the natural context of *c-KIT* gene, but weak imino signals in ^{1}H NMR spectrum indicating that is less favorable for formation of G-quadruplex. We detected eight resolved and many broad imino signals between δ 10.9 and 12.1 ppm in ^{1}H NMR spectrum of Akit* suggesting formation of two, major and minor, G-quadruplex structures. Signal at δ 13.48 ppm assigned to G10•C18 base pair was also detected in addition to signals of pre-folded structure at around δ 13.0 ppm.

The folding pathway is affected by the length of kit*

To better unveil the role of the 3'-tail in the G-quadruplex folding pathway we performed CD kinetic experiments on kit*, kit*17, kit*18 and kit*19. Upon the addition of 150 mM KCl, kit*, kit*18 and kit*19 exhibit fast folding kinetic into antiparallel G-quadruplexes that is completed within the sample mixing time of 10 s (Figure S27A-C). Although NMR data show that kit*18 and kit*19 fold into two distinct G-quadruplex structures each, the CD spectra of equilibrated kit*, kit*18 and kit*19 overlap. This suggests that all their G-quadruplexes share a common antiparallel topology.

The behavior of kit*17 is different. This oligonucleotide initially folds into an antiparallel G-quadruplex (Figure S27D, grey solid line), which then slowly evolves to other folded species, as revealed by CD signal variations over time (Figure S27D, black dashed line). The

folding pathway of kit*17 was studied by applying Singular Value Decomposition (SVD) analysis (49,50) to the CD dataset of its folding process (Figures S27D and S28). The initial folding stage (F0→F1) is a fast event, thus its accurate analysis was not possible with our equipment (Figure S28D). For this reason, we analyzed only the slower folding step (F1→F2) (Figure S28D). Singular values in the S matrix and the U and V autocorrelation coefficients (meaningful above 0.75) indicate that, in the F1→F2 folding process, two main species contribute to the dichroic signal variation (Figure S28A). Significant V eigenvectors were fitted by applying a mono-exponential kinetic model (Equation 2, Figure S28B). The fitting parameters allowed us to obtain the basic spectra of the species in solution (Figure S28C). F1 displays the dichroic spectrum typical of an antiparallel G-quadruplex (Figure S28C, solid line). It is formed immediately upon addition of KCl and it slowly converts into F2, whose dichroic spectrum presents a positive peak at 264 nm and a negative one at 248 nm consistent with a parallel G-quadruplex arrangement (Figure S28C, dashed line). The time needed to achieve the equilibrium (τ) of the F1→F2 process is 4.03 ± 0.24 h.

CD melting experiments reveal that kit* exhibits a clear single-transition melting profile (Figure 6) with a melting temperature (T_m) of 60.5 ± 0.1°C in 150 mM KCl (Equation 3). These observations suggest that kit* folds mainly into one conformation at these experimental conditions. In comparison, kit*17 melting profile displays two different melting transitions (Figure 6, Equation 4) at 50.6 ± 0.6°C (T_{m1}) and at 75.9 ± 0.4°C (T_{m2}) confirming that at least two structures are present in the solution and participate in the melting process. T_{m1} is associated with the kinetically-favored antiparallel G-quadruplex, while T_{m2} is related to thermodynamically-favored parallel forms. Interestingly, melting of antiparallel kit*17 G-quadruplex can be followed through the increase of CD melting profile, because its negative optical contribution in CD spectra at 264 nm is reduced (Figure 6C). At 56°C more stable parallel structure starts to melt, which is reflected in decreasing CD melting profile at 264 nm. We propose that T_{m1} of kit*17 is lower compared to the T_m of the kit* structure because of the lack of the 3'-tail which contributes to the stability of the entire G-quadruplex. UV melting analysis acquired on kit*18 and kit*19 at experimental conditions used for NMR experiments revealed comparable results as CD melting data. Two thermal transitions are detected also for kit*18, while kit*19 shows a reduction of the thermal stability by 3°C compared to kit* (Figure S29). These data fully confirm the destabilizing influence of shortening of the 3'-tail in kit*.

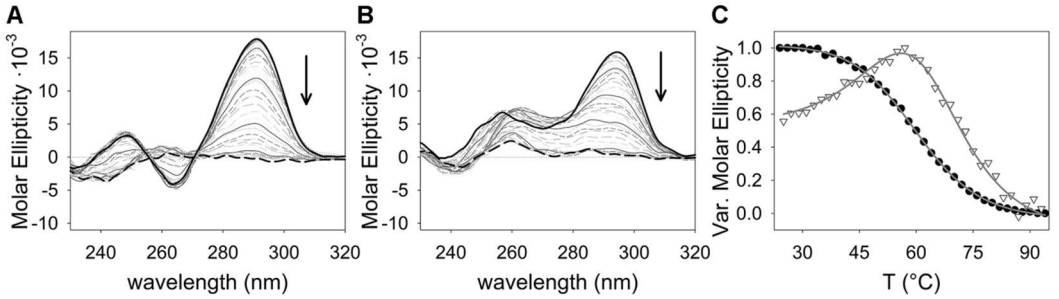


Figure 6. Temperature-dependent variation between 25 (black solid line) and 95°C (black dashed line) of the CD spectra of A) kit* and B) kit*17 in 150 mM KCl. C) Melting profile of kit* (black dots) and kit*17 (white triangles) monitored at 294 nm and 264 nm, respectively.

Two molecules of kit* are partially paired in a homo-dimeric structure of pre-folded structure

Two signals at δ 12.94 and 13.14 ppm in ^{1}H NMR spectrum at 37°C of kit* in the presence of Li^{+} ions indicate formation of a pre-folded structure, named pre-kit*, that does not contain stacked G-quartets (Figure 7A). Unambiguous assignment of imino signals, using residue-specific, partially ^{13}C,^{15}N-labeled kit*, reveals that G20 and G21 residues of the 3'-tail are involved in GC Watson-Crick base pairs in pre-kit* (Figure S30). The base pairing in pre-kit* occurs between G20 and G21 of one molecule and C18 and C19 of another kit* molecule (Figure S31). Thus, the proposed partially paired segment of kit* molecules consists of four GC base pairs and is highly symmetric, which is in agreement with the observation of only two imino signals in ^{1}H NMR spectrum. CD spectrum of pre-kit* in the presence of Li^{+} ions at 25°C displays a maximum at 284 nm and a minimum at 245 nm confirming an antiparallel double-stranded secondary structure (Figures 1B and S32). Existence of an isodichroic point at 255 nm suggests that there is only one conformational step from unfolded oligonucleotide towards pre-kit* form. Noteworthy, decrease of temperature of the Li^{+} ion containing solution from 37 to 5°C enables detection of pre-folded structures, besides pre-kit*, that are involved in conformational exchange. Indeed, at 37°C only signals in ^{1}H NMR spectrum corresponding to pre-kit* are observed, while upon lowering the sample temperature to 5°C new signals appear between δ 12.2-13.8 and 10.2-11.8 ppm in ^{1}H NMR spectrum (Figure 7B). In ^{15}N-edited HSQC NMR spectra in the presence of LiCl at 5°C two imino signals were observed for every residue of kit* that

was ^{13}C,^{15}N-labeled (Figure S33). These results suggest coexistence of two different secondary structures in the presence of Li^+ ions at 5°C. Pre-kit* is mostly stabilized by GC base pairs, while the other structure contains more base pairs in non-Watson-Crick geometry.

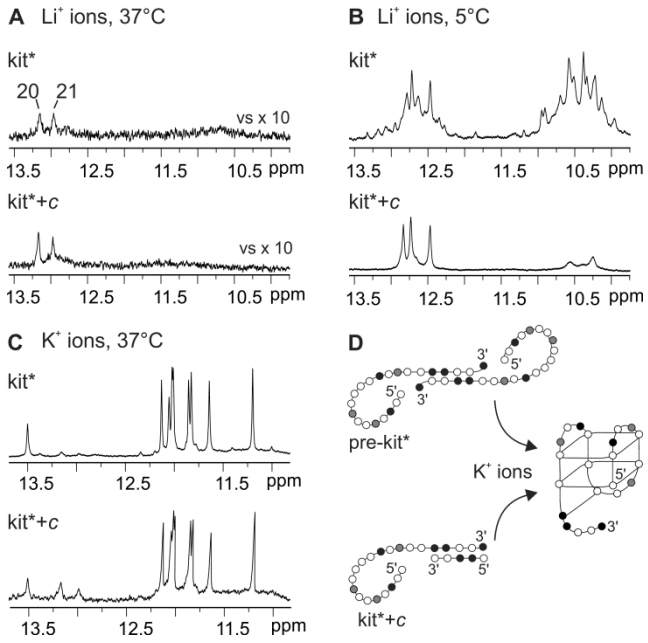


Figure 7. Imino region of ^1H NMR spectra of kit* and equimolar mixture of kit* and construct *c* (kit*+*c*) recorded in the presence of Li^+ ions at A) 37 and B) 5°C as well as C) in the presence of K^+ ions at 37°C. D) Schematic presentation of forms involved in formation of kit* G-quadruplex. Guanines are marked in white, cytosines in black, adenines and thymines in grey.

In order to explore interactions of residues in the 3'-tail of pre-kit* structure, we designed a five nucleotides long construct *c,* d[GCCGG], that is complementary to C18-C19-G20-G21-C22 segment of kit*. Our results show that base pairing of the 3'-tail of kit* with the construct *c* blocks its other possible interactions (Figure 7B). ^1H NMR spectrum of kit* hybridized with the construct *c* (marked as kit*+*c*) in equimolar amounts displays similar pattern of imino signals compared to kit* at 37°C (Figure 7A). Upon lowering the temperature to 5°C, three well-resolved imino signals between δ 12.4 and 13.0 ppm have been detected for kit*+*c* instead of many overlapped peaks in the presence of Li^+ ions for kit* (Figure 7B). Formation of self-complementary structures adopted by construct *c* was not observed (Figure S34). Interestingly, in the case of kit* as well as kit*+*c* formation of the same G-quadruplex is detected after addition of K^+ ions (Figure 7C). These results

indicate that kit* in the pre-folded structure pre-kit* and kit*+*c* share a comparable inter-molecular arrangement of the 3'-tail in the presence of LiCl, where the 3'-tails of two distinct kit* molecules are paired through GC Watson-Crick interactions forming a homo-dimer (Figures 7D and S5). Signals assigned to pre-folded structures were observed in NMR spectra of kit* as well as kit*+*c* also in the presence of K^+ ions (Figures 7C and S35).

DISCUSSION

c-KIT proto-oncogene exhibits three G-rich regions in its promoter, kit1, kit* and kit2, that are able to fold into G-quadruplexes. kit* region plays a key role in regulation of *c-KIT* transcription since it contains consensus sites for SP1 and AP2 transcription factors. The NMR derived kit* structure presented herein is, to the best of our knowledge, the first high-resolution structure of a two G-quartet G-quadruplex originating from a promoter region. At physiologically relevant temperature (37°C) and concentration of K^+ ions (100 mM) kit* adopts a chair type antiparallel G-quadruplex comprising three edgewise loops and a 3'-tail. A5 and T15 residues from the first and the third loops stack on the top G-quartet, while other residues from these two loops protrude into solution or are oriented into the grooves. On the other side of the G-quartet core, G10•C18 base pair is stacked on the bottom G-quartet. C19, G20, G21 and C22 residues from the 3'-tail are arranged into a fold-back motif below the G10•C18 base pair and the bottom G-quartet. To compare, in the case of two G-quartet G-quadruplexes originating from telomeric regions base triples are stacked on both G-quartets (51-54). The fold-back motif consisting of three terminal residues, in which two of them are stacked on G-quartet, has been described for G-quadruplex originating from human *c-MYC* promoter (55). Four residue overhang has been shown to adopt an extra-G-quartet fold-back element that contributes to stabilization of VEGF aptamer (56).

kit* in its sequence contains one GGGG-tract and four GG-repeats, of which G20-G21 repeat, located in the 3'-terminal, is not involved in the formation of G-quartet core. Contrary to expectations, we observed that 3'-tail plays a crucial role in the folding process, structural integrity and thermal stability of kit* G-quadruplex. Firstly, the 3'-tail is involved in the formation of a stable pre-folded structure in the absence of K^+ ions. In this form, residues in the 3'-tail of two kit* molecules are base paired through GC Watson-Crick interactions forming an inter-molecular partially paired homo-dimeric structure.

Upon addition of K^+ ions, organization of the 3'-tail in pre-kit* prevents Hoogsteen base pairing of the terminal GG-repeat (G20-G21). It is noteworthy that modifications in the 3'-tail disturb GC base pairs in pre-kit* and as a result lead to formation of more than two G-quadruplexes. Once G1-G17 residues of kit* get involved in the G-quadruplex, the preservation of GC base pairing of two 3'-tails is most likely sterically unfavored and the 3'-tails dissociate from the complementary strand leading to formation of a monomeric G-quadruplex. In full agreement, the replacement of G20-G21 repeat with thymine residues results in one G-quadruplex structure. Secondly, G10•C18 base pair and the fold-back motif of the 3'-tail stabilize kit* G-quadruplex by covering the bottom G-quartet. Perusal of the high-resolution structure of kit* G-quadruplex show that the fold-back structural element enables additional hydrogen bonds, electrostatic and hydrophobic interactions between residues of the 3'-tail and the middle loop as well as the bottom G-quartet. Examination of the folding behavior has revealed that the 3'-tail plays an important role in folding of kit*. For kit*17 without the 3'-tail alteration of folding pathway was observed together with formation of antiparallel structure and poorly defined G-quadruplexes with parallel topology that were formed within few hours after addition of K^+ ions. It is noteworthy that even preservation of the G10•C18 stabilizing element alone is not sufficient for formation of a single kit* G-quadruplex if the 3'-tail is not long enough. Thus, the conformational selection of only one kit* G-quadruplex is controlled by the presence of all residues of the 3'-tail in kit*. These observations are unique as terminal regions often change folding kinetics of oligonucleotide and destabilize formation of the final G-quadruplex structure (57-59). On the other hand, elongation of kit* at the 5'-end with thymine residue did not disturb kit* G-quadruplex structure, while adenine residue is less favorable and lead to formation of additional minor structure.

Considering the whole G-rich region in promoter of *c-KIT* gene, kit1 and kit2 G-rich oligonucleotides in contrast to kit* fold into three G-quartet parallel G-quadruplexes. It is worth mentioning that differences in structures of kit1, kit2 and kit* G-quadruplexes are associated with differences of kinetics of their formation. kit* folds in seconds, which is on a timescale of transcriptional processes (57,60). This is faster compared to folding kinetics of kit2, where the initial fast folding kinetic intermediate converts into the thermodynamically stable structures only within hours (38). Kinetic data for kit1 are not available, but since it forms a parallel G-quadruplex slower folding kinetic can be predicted compared to antiparallel kit* G-quadruplex (57). Therefore, kit* could modulate the

conformational organization of the neighboring kit1 and kit2 regions and thus play a functional role in regulation of transcription. The dynamic interplay of G-rich promoter region of c-KIT gene suggests that kit* could serve as an optimal site of intervention to finely tune the level of c-kit expression. Interestingly, this potential regulatory role merges with the complexity of the herein solved structure where the stability of the kit* G-quadruplex depends upon a network of interactions involving the loops and the 3'-tail. The structural features allow for foreseeing a possibility of selective recognition where ligands with even modest affinity may be able to induce cellular effects. Notably, selective stabilization of c-KIT compared to human telomeric G-quadruplex has been already reported for isoalloxazines ligands (62). Discrimination of ligands between kit1 and kit2 G-quadruplexes have been reported showing specific recognition and structural stabilization (61-64). Structural data of kit* presented herein provide new knowledge for design of ligands with more comprehensively and specifically targeting G-rich region of c-KIT gene, which could result in reduction of off-targeting of multiple G-quadruplexes frequently occurred.

In conclusion, structural characterization of the kit* G-quadruplex, the identification of a pre-folded structure and insights into its folding behavior expand on our knowledge about the processes involved in the regulation of the *c-KIT* proto-oncogene transcription. In particular, faster folding kinetic of kit* in comparison to kit1 and kit2 G-quadruplexes raises new questions about the role of kit* in affecting the folding processes of neighboring G-rich regions. Finally, the high-resolution structure of kit* G-quadruplex with two G-quartet antiparallel topology, the GC base pair and the fold-back motif of the 3'-tail, can be a suitable starting point for rational design of tailor-made ligands that will modulate c-kit expression.

ACCESSION NUMBERS

The coordinates for structure of kit* G-quadruplex have been deposited in the Protein Data Bank (accession number: 6GH0) and Biological Magnetic Resonance Data Bank (accession number: 34269).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors A.K. and J.P. acknowledge the financial support from the Slovenian Research Agency [research core funding No. P1-0242]. R.R. and C.S. were founded by University of Padova. [Ph.D. fellowship and CPDA147272/14]. The authors acknowledge the CERIC-ERIC Consortium for the access to experimental facilities and financial support.

FUNDING

This work was supported by Slovenian Research Agency [P1-0242 to A.K. and J.P.]; and University of Padova [Ph.D. fellowship to R.R., CPDA147272/14 to C.S.]. Funding for open access charge: Slovenian Research Agency [P1-0242].

Conflict of interest statement. None declared.

REFERENCES

1. d'Auriol, L., Mattei, M.G., Andre, C. and Galibert, F. (1988) Localization of the human c-kit protooncogene on the q11-q12 region of chromosome 4. *Hum. Genet.*, **78**, 374-376.
2. 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.
3. Yamamoto, K., Tojo, A., Aoki, N. and Shibuya, M. (1993) Characterization of the promoter region of the human c-kit proto-oncogene. *Jpn. J. Cancer Res.*, **84**, 1136-1144.
4. Metcalfe, D.D. (2008) Mast cells and mastocytosis. *Blood*, **112**, 946-956.
5. Gregory-Bryson, E., Bartlett, E., Kiupel, M., Hayes, S. and Yuzbasiyan-Gurkan, V. (2010) Canine and human gastrointestinal stromal tumors display similar mutations in c-KIT exon 11. *BMC Cancer*, **10**, 559-568.
6. 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.
7. Raiber, E.A., Kranaster, R., Lam, E., Nikan, M. and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res.*, **40**, 1499-1508.
8. Hud, N.V. and Plavec, J. (2006) *The Role of Cations in Determining Quadruplex Structure and Stability in Quadruplex Nucleic Acids*. The Royal Society of Chemistry, Cambridge.
9. Fujii, T., Podbevšek, P., Plavec, J. and Sugimoto, N. (2017) Effects of metal ions and cosolutes on G-quadruplex topology. *J. Inorg. Biochem.*, **166**, 190-198.
10. 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.
11. Karsisiotis, A.I., O'Kane, C. and Webba da Silva, M. (2013) DNA quadruplex folding formalism--a tutorial on quadruplex topologies. *Methods*, **64**, 28-35.

12. Čeru, S., Šket, P., Prislan, I., Lah, J. and Plavec, J. (2014) A New Pathway of DNA G-Quadruplex Formation. *Angew. Chem. Int. Ed.*, **53**, 4881-4884.
13. Marušič, M., Hošnjak, L., Krafčikova, P., Poljak, M., Viglasky, V. and Plavec, J. (2017) The effect of single nucleotide polymorphisms in G-rich regions of high-risk human papillomaviruses on structural diversity of DNA. *Biochim. Biophys. Acta*, **1861**, 1229-1236.
14. Marušič, M. and Plavec, J. (2015) The Effect of DNA Sequence Directionality on G-Quadruplex Folding. *Angew. Chem. Int. Ed.*, **54**, 11716-11719.
15. Šket, P. and Plavec, J. (2010) Tetramolecular DNA Quadruplexes in Solution: Insights into Structural Diversity and Cation Movement. *J. Am. Chem. Soc.*, **132**, 12724-12732.
16. Trajkovski, M., Webba da Silva, M. and Plavec, J. (2012) Unique Structural Features of Interconverting Monomeric and Dimeric G-Quadruplexes Adopted by a Sequence from the Intron of the N-myc Gene. *J. Am. Chem. Soc.*, **134**, 4132-4141.
17. Webba da Silva, M. (2007) Geometric Formalism for DNA Quadruplex Folding. *Chem. Eur. J.*, **13**, 9738-9745.
18. Greco, M.L., Kotar, A., Rigo, R., Cristofari, C., Plavec, J. and Sissi, C. (2017) Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter. *Nucleic Acids Res.*, **45**, 10132-10142.
19. Marušič, M., Šket, P., Bauer, L., Viglasky, V. and Plavec, J. (2012) Solution-state structure of an intramolecular G-quadruplex with propeller, diagonal and edgewise loops. *Nucleic Acids Res.*, **40**, 6946-6956.
20. 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.
21. Shivalingam, A., Izquierdo, M.A., Marois, A.L., Vysniauskas, A., Suhling, K., Kuimova, M.K. and Vilar, R. (2015) The interactions between a small molecule and G-quadruplexes are visualized by fluorescence lifetime imaging microscopy. *Nat. Commun.*, **6**: 8178.
22. Henderson, A., Wu, Y., Huang, Y.C., Chavez, E.A., Platt, J., Johnson, F.B., Brosh, R.M., Jr., Sen, D. and Lansdorp, P.M. (2014) Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.*, **42**, 860-869.
23. Hansel-Hertsch, R., Di Antonio, M. and Balasubramanian, S. (2017) DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Biol.*, **18**, 279-284.
24. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotech.*, **33**, 877-881.
25. Maizels, N. and Gray, L.T. (2013) The G4 genome. *PLoS Genet.*, **9**: e1003468.
26. Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406-413.
27. Neidle, S. (2017) Quadruplex nucleic acids as targets for anticancer therapeutics. *Nat. Rev. Chem.*, **1**: 41.
28. Rigo, R., Palumbo, M. and Sissi, C. (2017) G-quadruplexes in human promoters: A challenge for therapeutic applications. *Biochim. Biophys. Acta*, **1861**, 1399-1413.
29. 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.

30. Collie, G.W. and Parkinson, G.N. (2011) The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chem. Soc. Rev.*, **40**, 5867-5892.
31. Xu, H., Di Antonio, M., McKinney, S., Mathew, V., Ho, B., O'Neil, N.J., Santos, N.D., Silvester, J., Wei, V., Garcia, J. *et al.* (2017) CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat. Commun.*, **8**: 14432.
32. 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.
33. 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.
34. Zorzan, E., Da Ros, S., Musetti, C., Shahidian, L.Z., Coelho, N.F., Bonsembiante, F., Letard, S., Gelain, M.E., Palumbo, M., Dubreuil, P. *et al.* (2016) Screening of candidate G-quadruplex ligands for the human c-KIT promotorial region and their effects in multiple in-vitro models. *Oncotarget*, **7**, 21658-21675.
35. 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.
36. 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.
37. Kuryavyi, 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.
38. Rigo, R., Dean, W.L., Gray, R.D., Chaires, J.B. and Sissi, C. (2017) Conformational profiling of a G-rich sequence within the c-KIT promoter. *Nucleic Acids Res.*, **45**, 13056-13067.
39. Miller, M.C., Le, H.T., Dean, W.L., Holt, P.A., Chaires, J.B. and Trent, J.O. (2011) Polymorphism and resolution of oncogene promoter quadruplex-forming sequences. *Org. Biomol. Chem.*, **9**, 7633-7637.
40. Da Ros, S., Zorzan, E., Giantin, M., Zorro Shahidian, L., Palumbo, M., Dacasto, M. and Sissi, C. (2014) Sequencing and G-quadruplex folding of the canine proto-oncogene KIT promoter region: might dog be used as a model for human disease? *PLoS One*, **9**: e103876.
41. Wei, D., Husby, J. and Neidle, S. (2015) Flexibility and structural conservation in a c-KIT G-quadruplex. *Nucleic Acids Res.*, **43**, 629-644.
42. 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.
43. Park, G.H., Plummer, H.K. and Krystal, G.W. (1998) Selective Sp1 Binding Is Critical for Maximal Activity of the Human c-*kit* Promoter. *Blood*, **92**, 4138-4149.
44. Rigo, R. and Sissi, C. (2017) Characterization of G4-G4 Crosstalk in the c-KIT Promoter Region. *Biochemistry*, **56**, 4309-4312.
45. Kocman, V. and Plavec, J. (2017) Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. *Nat. Commun.*, **8**: 15355.

46. Kotar, A., Wang, B., Shivalingam, A., Gonzalez-Garcia, J., Vilar, R. and Plavec, J. (2016) NMR Structure of a Triangulenium-Based Long-Lived Fluorescence Probe Bound to a G-Quadruplex. *Angew. Chem. Int. Ed. Engl.*, **55**, 12508-12511.
47. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A. (2004) Development and testing of a general amber force field. *J. Comput. Chem.*, **25**, 1157-1174.
48. Gray, R.D., Petraccone, L., Buscaglia, R. and Chaires, J.B. (2010) 2-aminopurine as a probe for quadruplex loop structures. *Methods Mol. Biol.*, **608**, 121-136.
49. Hendler, R.W. and Shrager, R.I. (1994) Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. *J. Biochem. Biophys. Methods*, **28**, 1-33.
50. DeSa, R.J. and Matheson, I.B. (2004) A practical approach to interpretation of singular value decomposition results. *Methods Enzymol.*, **384**, 1-8.
51. Galer, P., Wang, B., Sket, P. and Plavec, J. (2016) Reversible pH Switch of Two-Quartet G-Quadruplexes Formed by Human Telomere. *Angew. Chem. Int. Ed. Engl.*, **55**, 1993-1997.
52. 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.
53. Zhang, Z., Dai, J., Veliath, E., Jones, R.A. and Yang, D. (2010) Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K+ solution: insights into the interconversion of human telomeric G-quadruplex structures. *Nucleic Acids Res.*, **38**, 1009-1021.
54. Amrane, S., Ang, R.W., Tan, Z.M., Li, C., Lim, J.K., Lim, J.M., Lim, K.W. and Phan, A.T. (2009) A novel chair-type G-quadruplex formed by a Bombyx mori telomeric sequence. *Nucleic Acids Res.*, **37**, 931-938.
55. Ambrus, A., Chen, D., Dai, J., Jones, R.A. and Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, **44**, 2048-2058.
56. Marušič, M., Veedu, R.N., Wengel, J. and Plavec, J. (2013) G-rich VEGF aptamer with locked and unlocked nucleic acid modifications exhibits a unique G-quadruplex fold. *Nucleic Acids Res.*, **41**, 9524-9536.
57. Marchand, A. and Gabelica, V. (2016) Folding and misfolding pathways of G-quadruplex DNA. *Nucleic Acids Res.*, **44**, 10999-11012.
58. Gray, R.D. and Chaires, J.B. (2008) Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures. *Nucleic Acids Res.*, **36**, 4191-4203.
59. Arora, A., Nair, D.R. and Maiti, S. (2009) Effect of flanking bases on quadruplex stability and Watson-Crick duplex competition. *Febs J.*, **276**, 3628-3640.
60. Morisaki, T., Muller, W.G., Golob, N., Mazza, D. and McNally, J.G. (2014) Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat. Commun.*, **5**: 4456.
61. Duarte, A.R., Cadoni, E., Ressurreicao, A.S., Moreira, R. and Paulo, A. (2018) Design of Modular G-quadruplex Ligands. *ChemMedChem*, **13**, 869-893.
62. 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.
63. Diveshkumar, K.V., Sakrikar, S., Rosu, F., Harikrishna, S., Gabelica, V. and Pradeepkumar, P.I. (2016) Specific Stabilization of c-MYC and c-KIT G-Quadruplex

DNA Structures by Indolylmethyleneindanone Scaffolds. *Biochemistry*, **55**, 3571-3585.

64. Bejugam, M., Gunaratnam, M., Muller, 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.

SUPPLEMENTARY INFORMATION

Two-quartet kit* G-quadruplex is formed via double-stranded pre-folded structure

Anita Kotar[1], Riccardo Rigo[2], Claudia Sissi[2,*] and Janez Plavec[1,3,4,*]

[1] Slovenian NMR Center, National Institute of Chemistry, SI-1000, Ljubljana, Slovenia,

[2] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

[3] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia and

[4] EN-FIST Center of Excellence, Ljubljana, Slovenia

* To whom correspondence should be addressed. Tel: +386 1 4760353; Fax: +386 1 4760300; Email: janez.plavec@ki.si. Correspondence may also be addressed to Claudia Sissi. Tel: +39 049 827 5711; Fax: +39 049 827 5366; Email: claudia.sissi@unipd.it.

Submitted – Kotar, A.; Rigo, R.; Sissi, C.; Plavec, J. *Nucleic Acids Res.* **2018**

MATERIALS AND METHODS

Circular dichroism (CD) spectroscopy

CD titrations were performed by increasing concentrations of KCl or LiCl in solution of oligonucleotide at 25 °C. After each titration step the system was left to equilibrate before spectra acquisition. CD signal variations at the maximum (for titration with K^+ ions monitored at 294 nm, while in Li^+ ions at 266 nm) were plotted as a function of the concentration of cations. Data points were fitted by applying one-site saturation model.

$$\theta_i = \frac{\theta_\infty \cdot [M^+]}{K_D + [M^+]} \qquad \text{Equation 1}$$

where $[M^+]$ is the cation concentration, θ_i is the signal at $[M^+]$, θ_∞ is the signal at the saturation and K_D is the dissociation constant.

For CD kinetic experiment, KCl was added manually to the cuvette from a stock solution and mixing was provided by an in-cuvette magnetic stirring bar. After a mixing time of 10 s, spectra acquisition was initiated using an interval scan of 30 min. The temperature was maintained at 37 °C. Data related to the folding process of kit*17 were analyzed by means of SVD analysis (1,2). The D matrix, defined by the optical signals in the whole wavelength range (rows) acquired at each single temperature (columns), was divided into three submatrices U, S, V, so that D = U x S x V. The singular values in S matrix and the autocorrelation coefficients of the U and V matrices give information about the number of species in solution that contribute to the signal variation. Significant U and V autocorrelation coefficients values above 0.75 were considered. Significant V eingenvectors in V matrix were globally fitted by applying different kinetic model and the best fitting was obtained by applying mono-exponential kinetic model,

$$\theta t = a + b \cdot e^{(t/\tau)} \qquad \text{Equation 2}$$

where θ_t is CD signal at time t, a and b are fitting parameters and τ is the relaxation time of the process.

The fitting parameters formed the H matrix. By multiplying the H matrix for the U x S matrix, the actual shapes of the dichroic signals of species in solution were obtained.

CD melting studies were performed on kit* and kit*17. The experiments were carried out between 25 and 95°C in 150 mM KCl. The heating rate was 50°C/hour where at each 2°C the temperature was held for 5 minutes and the corresponding CD signal was recorded at 294 and 264 nm for kit* and kit*17, respectively. The melting profile of kit* that shows only one fully reversible melting transition was analyzed by fitting the signal variation according to a single-transition model based on van't Hoff formalism (3,4),

$$\theta_T = \frac{u * e^{-\left(\frac{\Delta H}{RT} * \left(\left(\frac{T}{Tm}\right) - 1\right)\right)} + l}{e^{-\left(\frac{\Delta H}{RT} * \left(\left(\frac{T}{Tm}\right) - 1\right)\right)} + 1} \qquad \text{Equation 3}$$

where T is temperature, θ_T is the CD signal at temperature T, u and l are fitting parameters, ΔH is the enthalpy of the unfolding process, R is the ideal gas constant and Tm is the melting temperature of the folded oligonucleotide.

The melting profile of kit*17, presenting two distinct melting transitions, was analyzed by using a two-transitions model based on van't Hoff formalism (3,4),

$$\theta_T = \frac{u * e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right)\right)} + a * e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right)\right)} + l}{e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right)\right)} + e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right)\right)} + 1} \qquad \text{Equation 4}$$

where T is temperature, θ_T is the CD signal at temperature T, u, a and l are fitting parameters, ΔH1 and ΔH2 are the enthalpies associated to each melting step, R is the ideal gas constant, T_{m1} and T_{m2} are the melting temperatures.

UV spectroscopy

UV melting experiments were recorded on a Varian CARY-100 BIO UV-VIS spectrophotometer (Varian Inc.) equipped with a thermoelectric temperature controller at 20 μM concentration of oligonucleotides per strand in 20 mM potassium phosphate buffer (pH 7.4) and 100 mM KCl. Folding/unfolding processes were followed between 10 and 95 °C by measuring absorbance at 260 and 295 nm using scanning rate of 0.5 °C min^{-1}. Data points were fitted by applying Equations 3 and 4.

Fluorescence spectroscopy

Acrylamide quenching experiments were performed on previously folded oligonucleotides in 150 mM KCl. The quenching efficiency was monitored at 370 nm at 25°C and plotted as a function of the acrylamide concentration. Data points were fitted according to Stern-Volmer formalism to obtain quenching parameters. Linear quenching correlation was analyzed using a one-component system model (5),

$$\frac{F_0}{F_Q} = 1 + K_{sv} \cdot [Q] \qquad \text{Equation 5}$$

where [Q] is the acrylamide concentration, F_0 is the initial value of fluorescence, F_Q is the fluorescence signal at [Q] and K_{sv} is the Stern-Volmer constant.
Non-linear quenching correlation was analyzed according to a two-component system model (5),

$$\frac{F_Q}{F_0} = \frac{f}{1+K_{sv,1}\cdot[Q]} + \frac{1-f}{1+K_{sv,2}\cdot[Q]} \qquad \text{Equation 6}$$

where [Q] is the acrylamide concentration, F_0 is the initial value of fluorescence, F_Q is the fluorescence signal at [Q], f is the accessible fluorophore fraction, $K_{sv,1}$ and $K_{sv,2}$ are the Stern-Volmer constants.

Polyacrylamide gel electrophoresis (PAGE)

A solution containing 3 µM kit* labelled with 6-FAM at 3'-end and 500 µM kit* (not labelled) in 10 mM TRIS, 100 mM KCl, pH 7.5 was incubated at room temperature for 1 h. Then the sample was loaded on a native 15% polyacrylamide gel (19:1 acrylamide: bisacrylamide) in 1× TBE and the gel was run at 10 °C for 3 h at 150 V. 22 and 44 bases long Scrambled oligonucleotides (22b, 5'-GGATGTGAGTGTGAGTGTGAGG-6-FAM-3'; 44b 5'-GGATGTGAGTGTGAGTGTGAGG GGATGTGAGTGTGAGTGTGAGG-6-FAM-3) were used as electrophoretic mobility markers. The resolved bands were visualized on an image acquisition system (Geliance 600 Imaging system, Perkin-Elmer).

Structure calculations and molecular dynamics simulations

The structures of kit* G-quadruplex were calculated by the simulated annealing (SA) simulations. The force field parameters were adopted from the Generalized Amber force field (6). SA simulations were performed using the CUDA version of pmemd module of AMBER 14 program suites (7,8) and Cornell et al. force field basic version parm99 (9) with the bsc0 (10), χOL4 (11), ε/ζOL1 (12) and βOL1 (13) refinements. The initial extended single-stranded DNA structure was obtained using the leap module of AMBER 14. A total of 100 structures were calculated in 80 ps of NMR restrained SA simulations using the generalized Born implicit model (14,15). The cut-off for non-bonded interactions was 999 Å and the SHAKE algorithm (16) for hydrogen atoms was used with the 0.4 fs time steps. For each SA simulation, a random velocity was used. The SA simulation was performed as described (17): in 0-2 ps, the temperature was raised from 300 K to 1000 K and held constant at 1000 K for 38 ps. Temperature was scaled down to 500 K in the next 24 ps and reduced to 100 K in the next 8 ps and was further reduced to 0 K in the last 8 ps. A family of 10 structures was selected based on the smallest restraints violations and lowest energy. Figures were visualized and prepared with UCSF Chimera (18).

Molecular dynamics simulations

For molecular dynamics simulations the kit* G-quadruplex was placed in a truncated octahedral box of TIP3P water molecules with the box border at least 10 Å away from any atoms of the G-quadruplex. Extra K^+ ions were added to neutralize the negative charges of the G-quadruplex. The force field for MD simulation was the same as for the SA simulation. The simulations were performed with the CUDA version of pmemd module of AMBER 14 (7,19-22). Prior to MD simulation, the system were subjected to a series of minimizations and equilibrations. The equilibration protocol started with 1000 steps of steepest descent minimization followed by 4000 steps of conjugate gradient minimization with 10 kcal mol^{-1} $Å^{-2}$ position restraints on DNA atoms. Then the whole system was minimized by 1000 steps of steepest descent minimization and 4000 steps of conjugate gradient minimization without restraint. The system was heated from 0 to 300 K during 100 ps with position restraints of 10 kcal mol^{-1} $Å^{-2}$ on G-quadruplex. Afterwards, the system was equilibrated during 50 ps at constant temperature of 300 K and pressure of 1

atm with 2 kcal mol^{-1} $Å^{-2}$ restraints on DNA atoms. Finally, the system was equilibrated using starting velocities from the previous equilibration without position restraints. Pressure coupling used during equilibration was set to 0.2, coupling during the last molecular dynamics phase was set to 5. The production simulation were carried out at constant pressure of 1 atm and constant temperature of 300 K maintained using Langevin dynamics with a collision frequency of 2.0. Periodic boundary conditions were used, and electrostatic interactions were calculated by the particle mesh Ewald method (23,24) with the non-bonded cutoff set to 8 Å. The SHAKE algorithm (16) was applied to bonds involving hydrogens, and a 2 fs integration step was used. The production run was carried out for continuous 200 ns and the snapshots were written at every 1 ps. Three MD simulations were performed with different initial velocity distributions. Trajectories were analyzed using the CPPTRAJ module of AMBER.

RESULTS

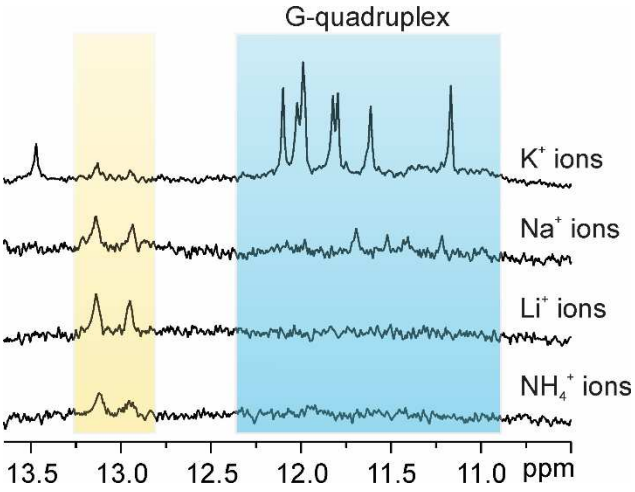


Figure S1. Imino region of ^{1}H NMR spectra of kit* in the presence of 100 mM concentration of K^+, Na^+, Li^+ and NH_4^+ ions as indicated on the right. ^{1}H NMR spectra were recorded in lithium cacodylate buffer (pH 7.2) at 0.3 - 0.5 mM kit* concentration per strand, 37°C on a 600 MHz NMR spectrometer. The region characteristic for imino signals of guanine residues involved in G-quartets is marked with blue. Signals of pre-folded state(s) of kit* are marked with yellow.

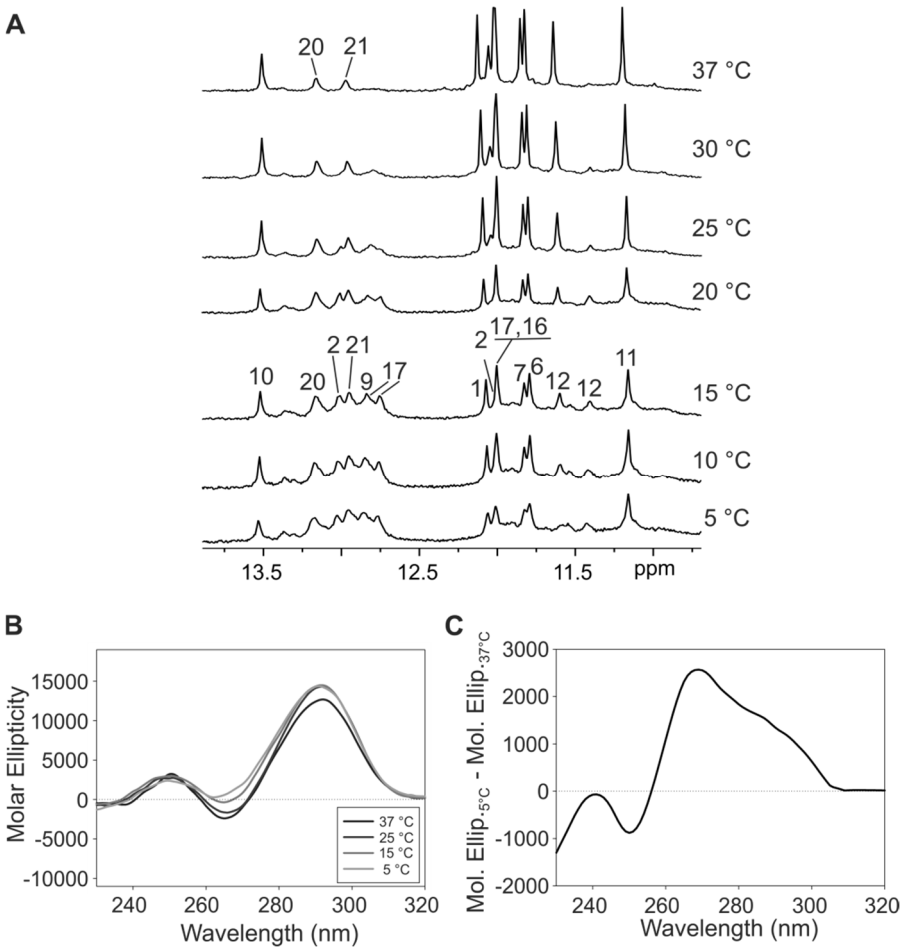


Figure S2. A) Imino region of ^{1}H NMR spectra of kit* between 5 and 37°C. Assignments are shown above individual signals. The NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz NMR spectrometer. B) CD spectra of kit* in 100 mM KCl at different temperatures. C) CD signal of kit* at 5°C subtracted of the CD contribution at 37°C.

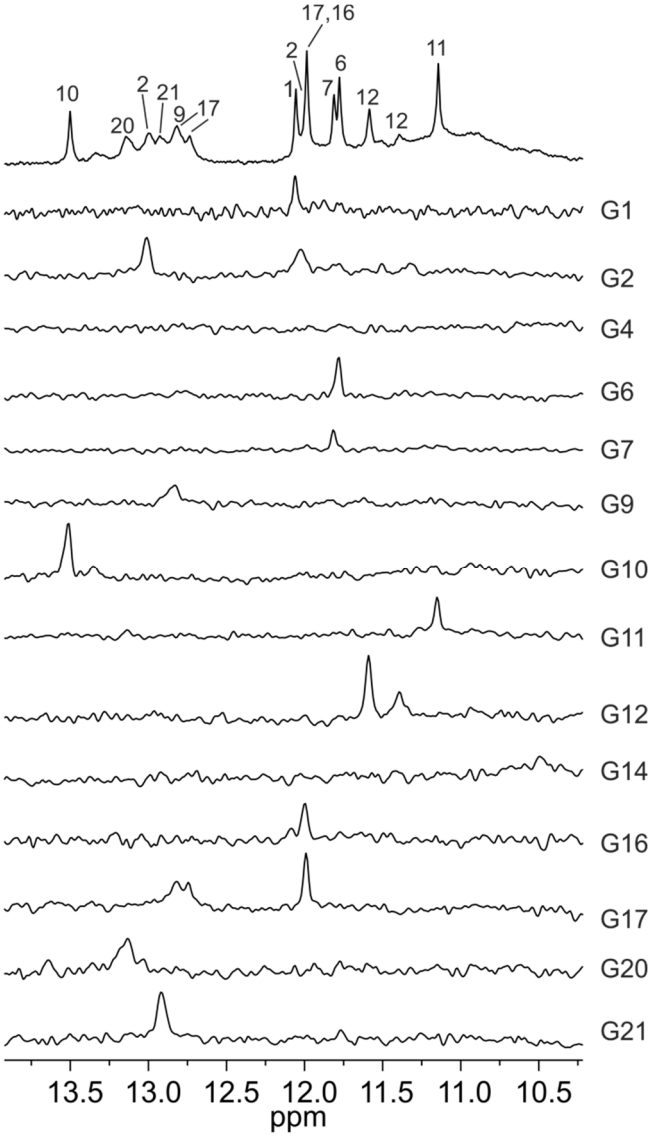


Figure S3. ^{1}H and 1D ^{15}N-edited HSQC NMR spectra of kit* at 15°C. The HSQC spectra were acquired on partially (10%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotides. Assignment of H1 proton resonances is indicated on the right side of each 1D HSQC spectrum. NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4 on a 600 MHz NMR spectrometer.

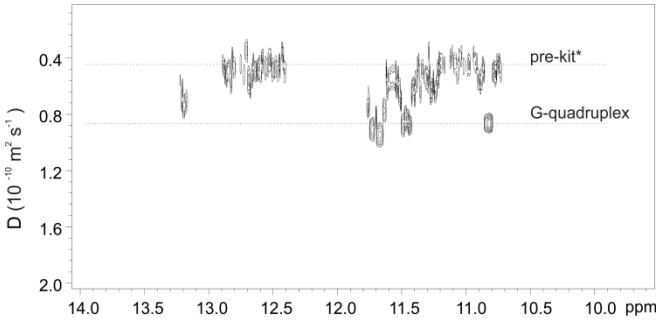


Figure S4. DOESY NMR spectrum of kit*. The determined translational diffusion coefficients were 0.87 x 10^{-10} for kit* G-quadruplex and 0.45 x 10^{-10} m^2 s^{-1} for pre-kit*. The NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 5°C on a 600 MHz NMR spectrometer.

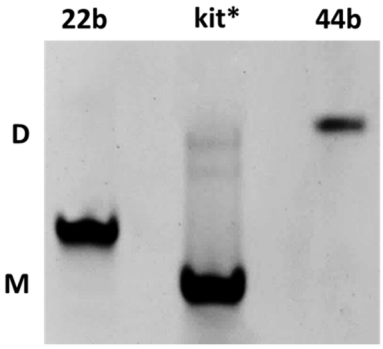


Figure S5. Distribution of kit* between monomeric and dimeric (pre-kit*) species. PAGE resolution of 500 µM kit* in 100 mM KCl, 10 mM TRIS, pH 7.5 performed on a 15% polyacrylamide gel in TBE 1x. 22b and 44b were electrophoretic mobility markers of 22 and 44 bases, respectively.

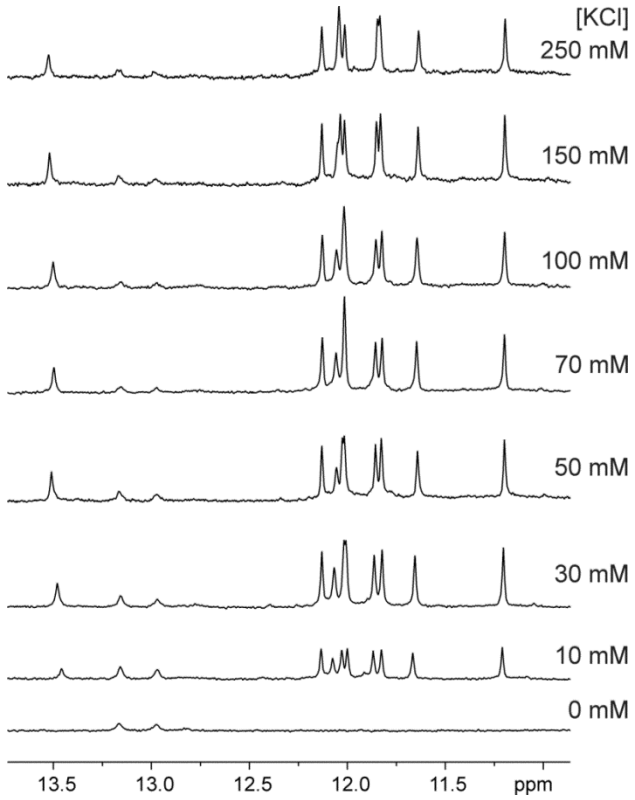


Figure S6. Imino region of ^{1}H NMR spectra of kit* at different concentrations of KCl indicated on the right. ^{1}H NMR spectra were recorded in potassium phosphate buffer (pH 7.4), at 0.4 mM kit* concentration per strand, 37°C on a 600 MHz NMR spectrometer.

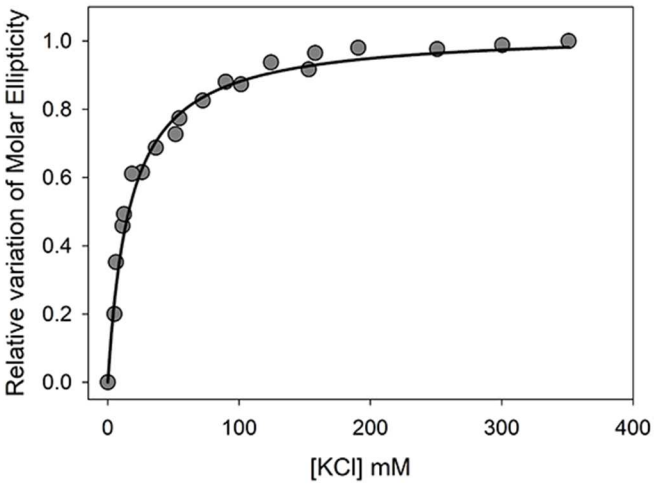


Figure S7. Relative variation of molar ellipticity obtained by monitoring the spectral changes induced by the addition of KCl at 294 nm. Data were fitted according to one-site saturation model (Equation 1).

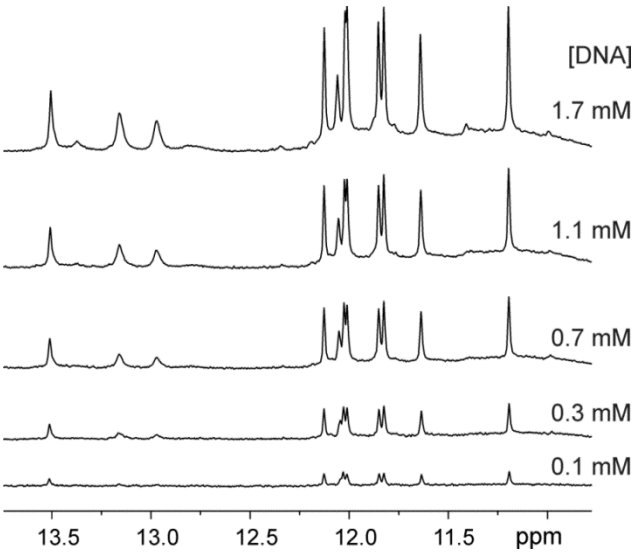


Figure S8. Imino region of ^{1}H NMR spectra of kit* in the presence of different concentrations per strand of kit* (DNA) as indicated on the right. ^{1}H NMR spectra were recorded in potassium phosphate buffer (pH 7.4), at 100 mM KCl, 37°C on a 600 MHz NMR spectrometer.

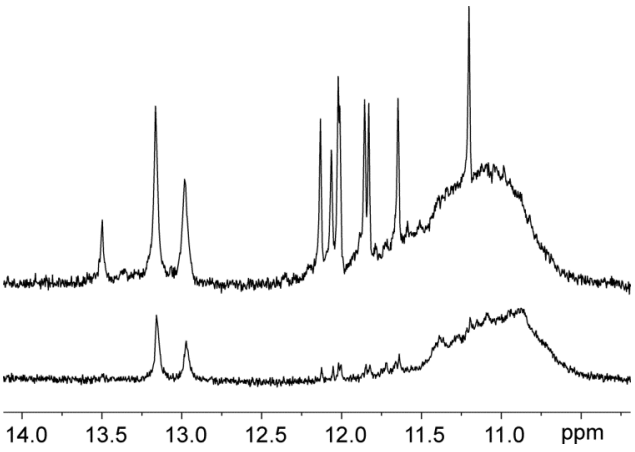


Figure S9. Imino region of ^{1}H NMR spectra of kit* before (top) and after (bottom) overnight annealing. ^{1}H NMR spectra were recorded at 2 mM kit* concentration per strand in potassium phosphate buffer (pH 7.4), at 100 mM KCl, 37°C on a 600 MHz spectrometer.

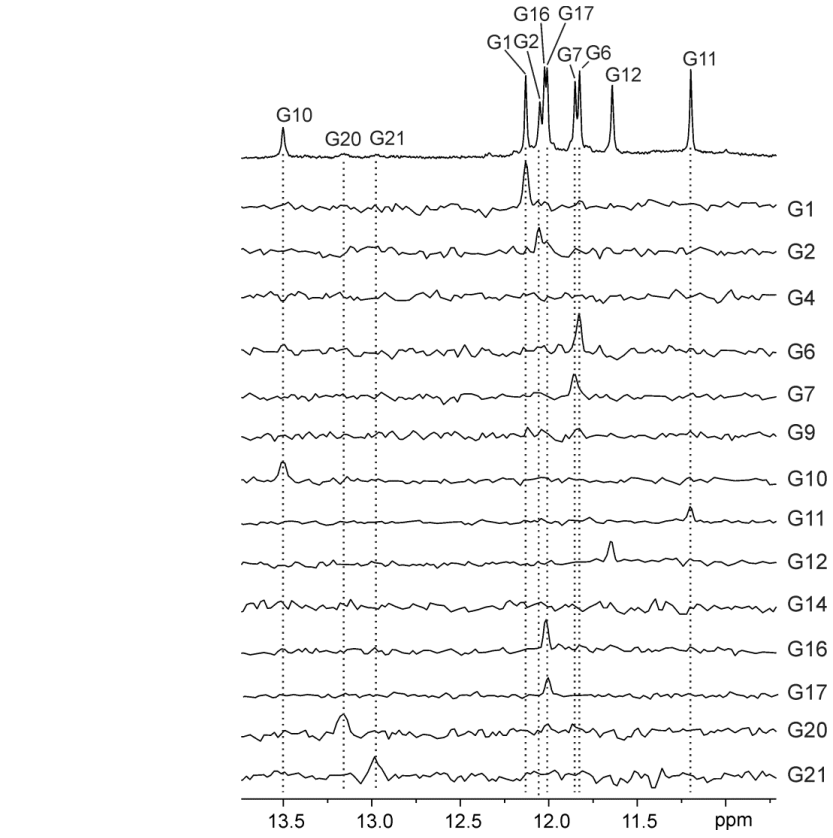


Figure S10. Imino region of ^{1}H and 1D ^{15}N-edited HSQC NMR spectra of kit* at 37°C. The HSQC spectra were acquired on partially (10%) residue-specifically ^{15}N- and ^{13}C-labelled oligonucleotides. Assignment of H1 proton resonances is indicated on the right of each 1D HSQC spectrum. NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4 on a 600 MHz spectrometer.

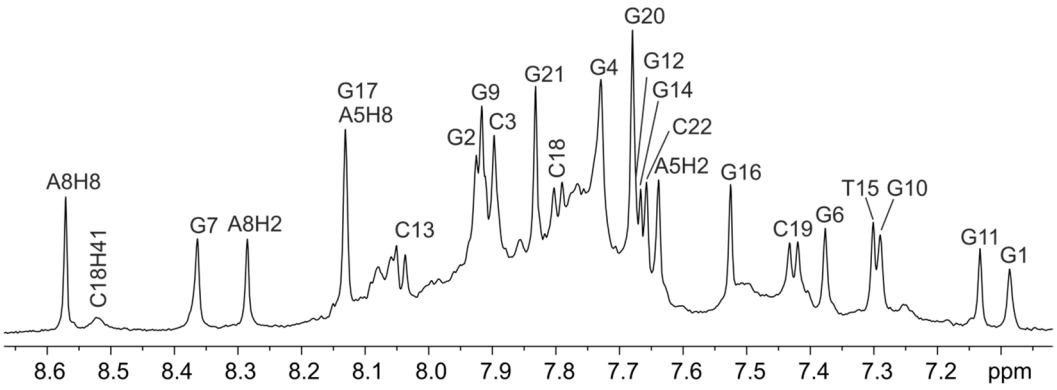


Figure S11. Aromatic region of ^{1}H NMR spectra of kit*. Assignments are shown above individual signals. NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on a 600 MHz spectrometer.

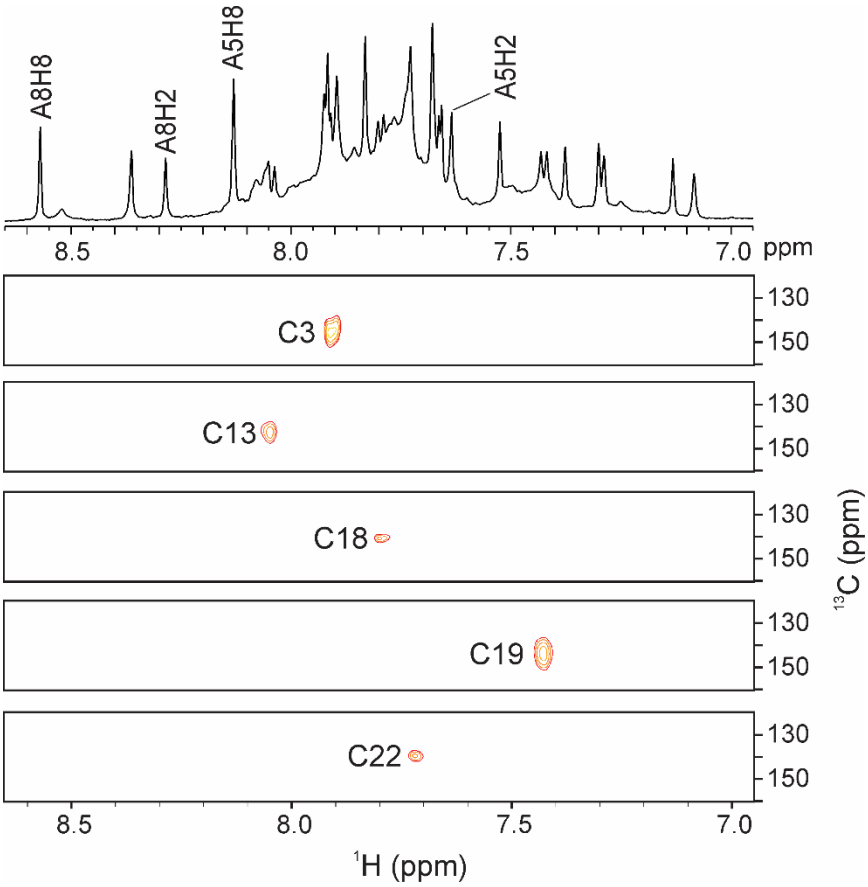


Figure S12. Aromatic region of 1D ^{1}H and 2D ^{13}C-edited HSQC NMR spectra of kit*. The HSQC spectra were acquired on partially (4%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotides. Assignment of H6 proton resonances is indicated next to the 2D cross-peaks. NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on a 600 MHz spectrometer.

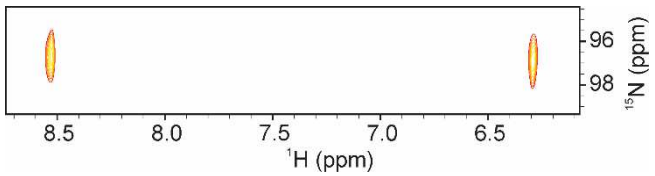


Figure S13. 2D ^{15}N-edited HSQC NMR spectrum of amino group of C18 residue of kit*. The HSQC spectrum was acquired on partially (4%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotide. NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on a 600 MHz spectrometer.

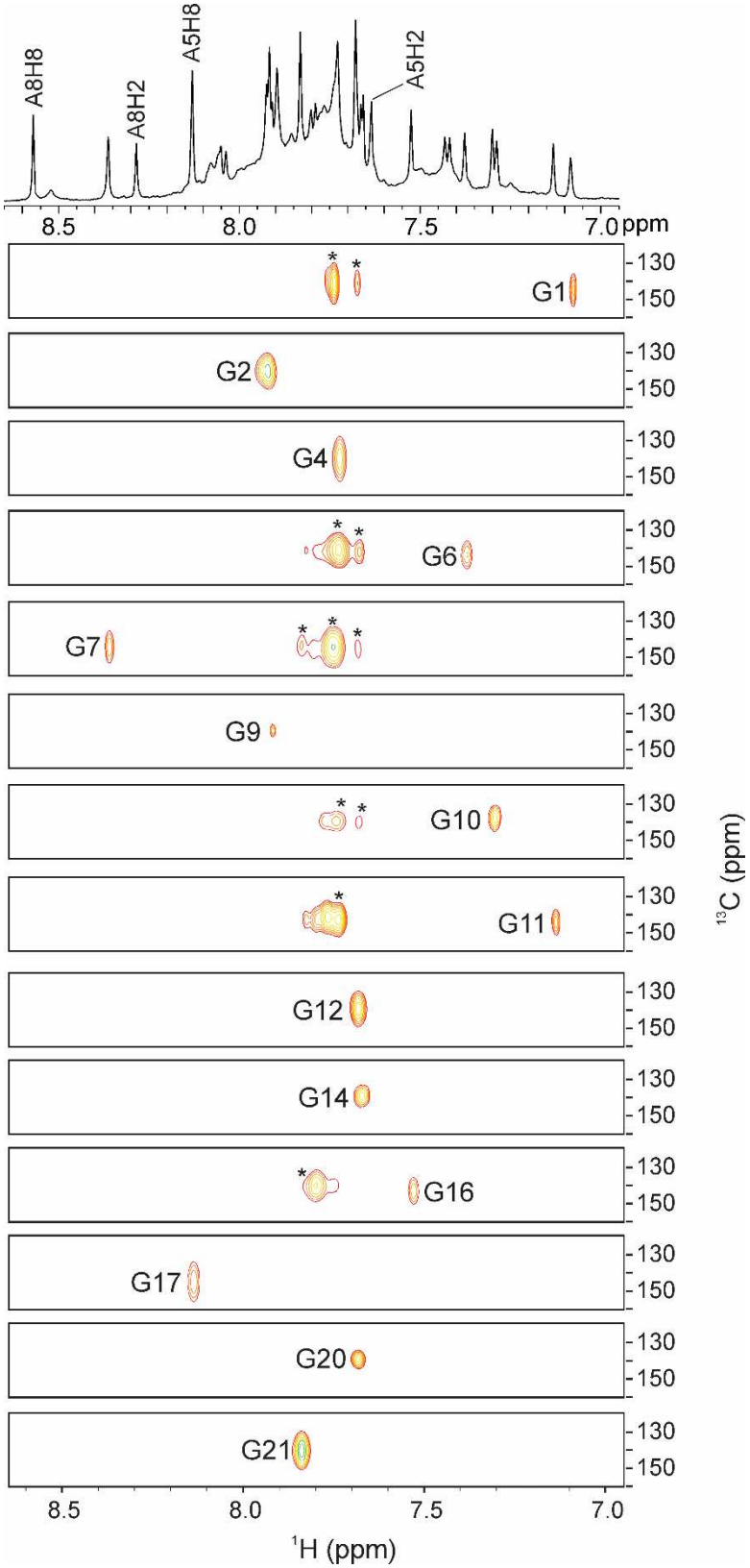


Figure S14. Aromatic region of 1D ^{1}H and 2D ^{13}C-edited HSQC NMR spectra of kit*. The HSQC spectra were acquired on partially (10%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotides. Assignment of H8 proton resonances is indicated next to the 2D cross-

peaks. With the star are marked signals that arise from unfolded or pre-folded oligonucleotide. NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on a 600 MHz spectrometer.

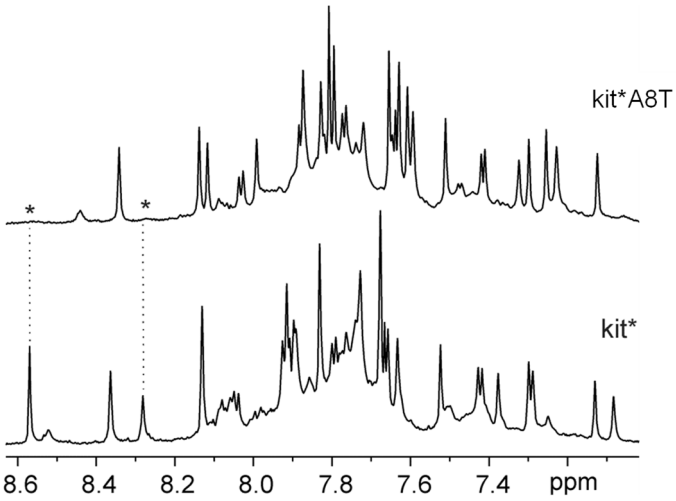


Figure S15. Aromatic region of ^{1}H NMR spectra of kit*A8T and kit*. The missing aromatic signals of A8 in kit*A8T are marked with stars. NMR spectra were recorded at 0.4 mM oligonucleotide concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer.

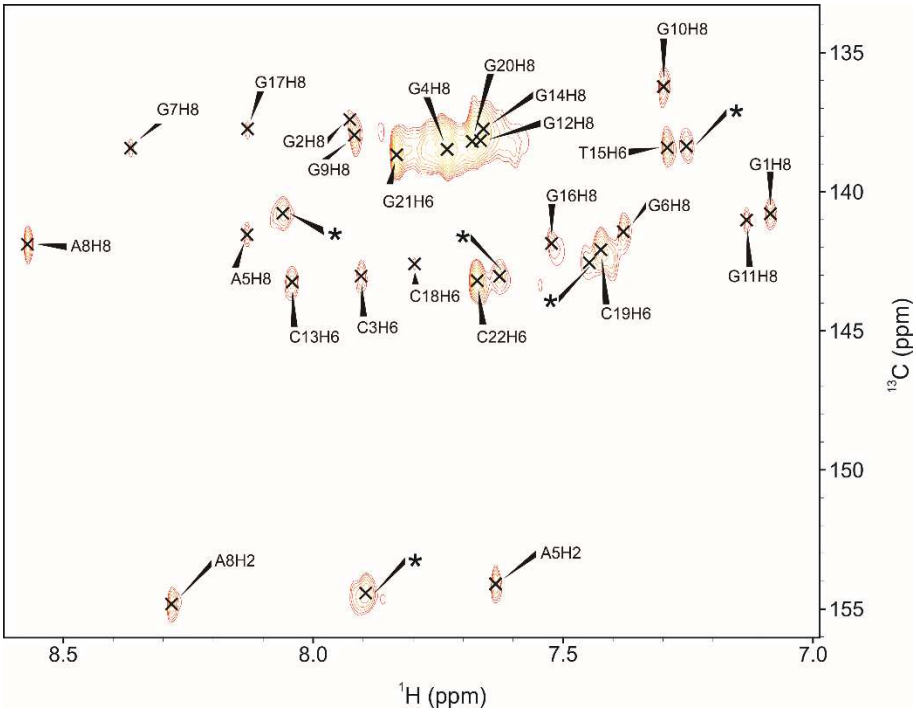


Figure S16. Aromatic region of 2D ^{1}H-^{13}C HSQC NMR spectrum of kit*. Assignment of individual cross-peaks is indicated. NMR spectrum was recorded at 0.4 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer. The signals marked with stars arise from the pre-folded structure.

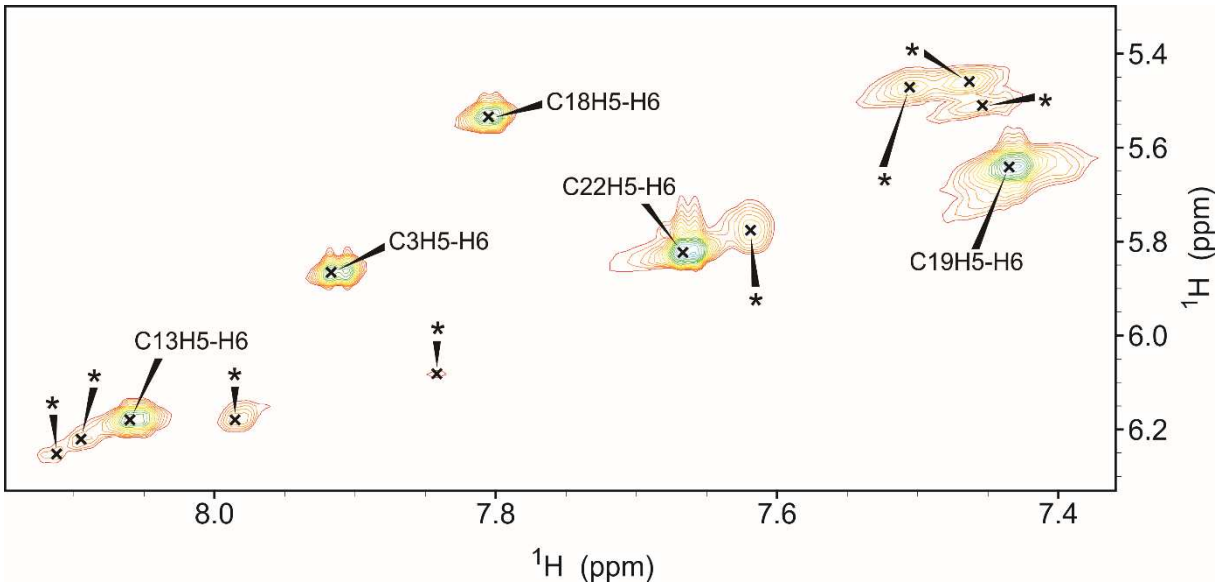


Figure S17. H6-H5 cross-peaks of cytosine residues in TOCSY spectrum (τ_m 80 ms) of kit*. Assignment of individual cross-peaks of kit* G-quadruplex is indicated. The cross-peaks belonging to pre-kit* are marked with *. NMR spectrum was recorded at 0.4 mM kit* concentration per strand, 100 mM KCl, 100% $^{2}H_2O$, 37°C on an 800 MHz spectrometer.

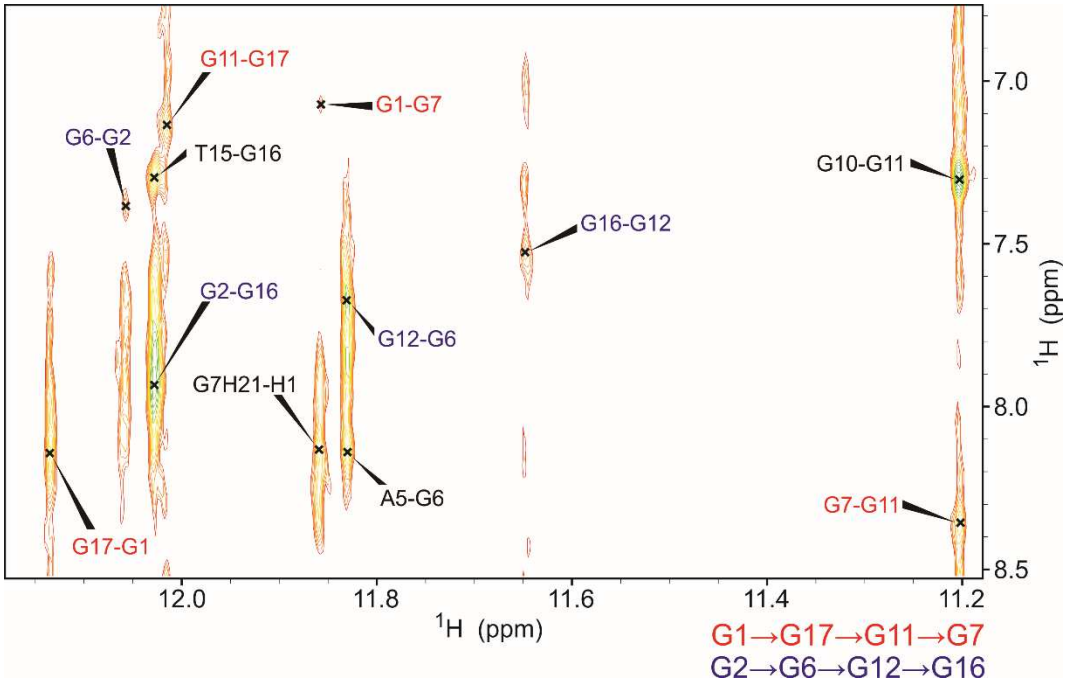


Figure S18. Imino-aromatic region of NOESY spectrum (τ_m 450 ms) of kit*. Cross-peaks identifying connectivities within G1-G17-G11-G7 and G2-G6-G12-G16 quartets are colored red and blue, respectively. With arrows are marked H1-H8 and G7H1-H21 connectivities. NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer.

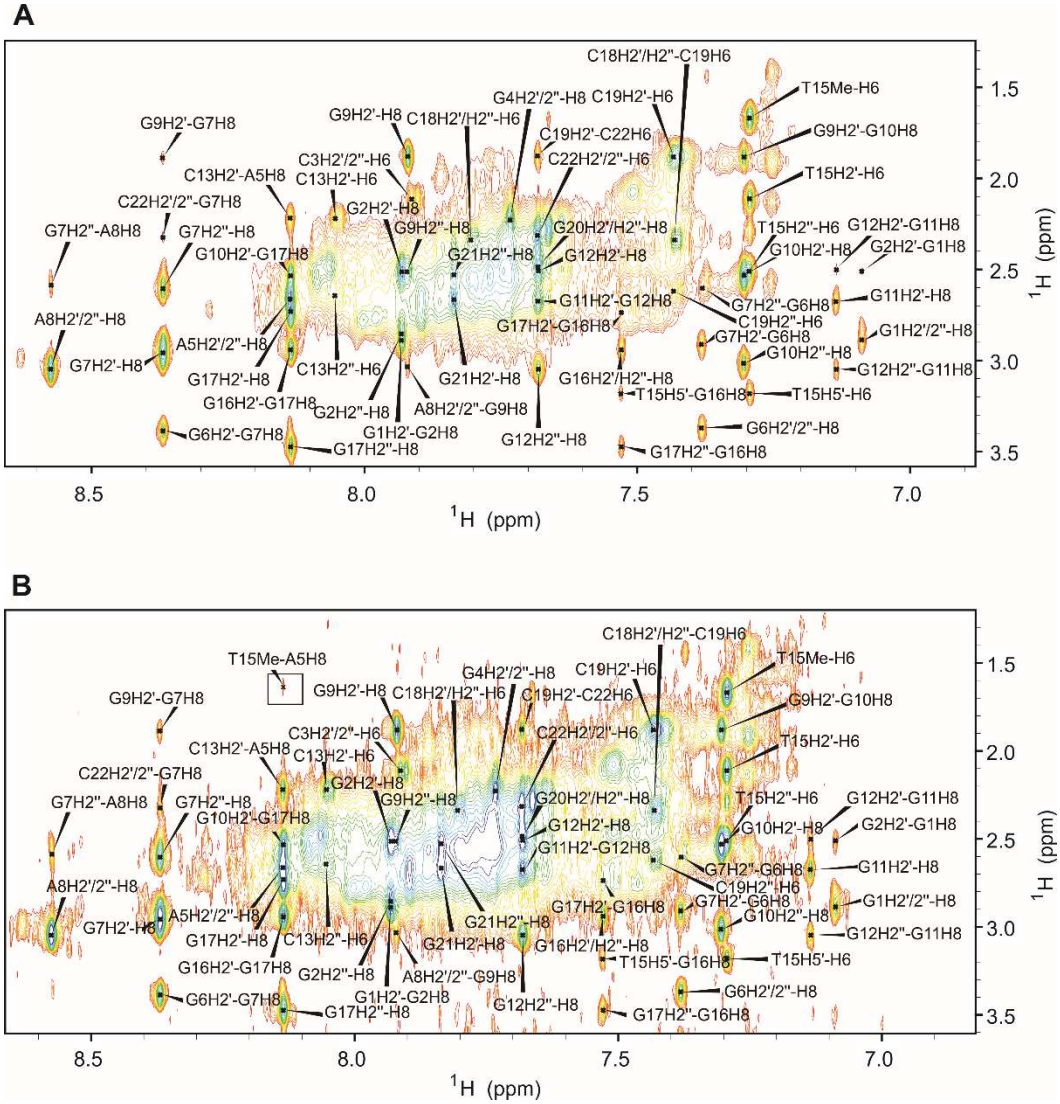


Figure S19. A) Aromatic-H2'/H2" region of NOESY spectrum (τ_m 450 ms) of kit*. B) Aromatic-H2'/H2" region of NOESY spectrum with 2..5 lover levels of contours compared to A) to present cross-peak between A8 H8 and T15 me (marked with square).NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer.

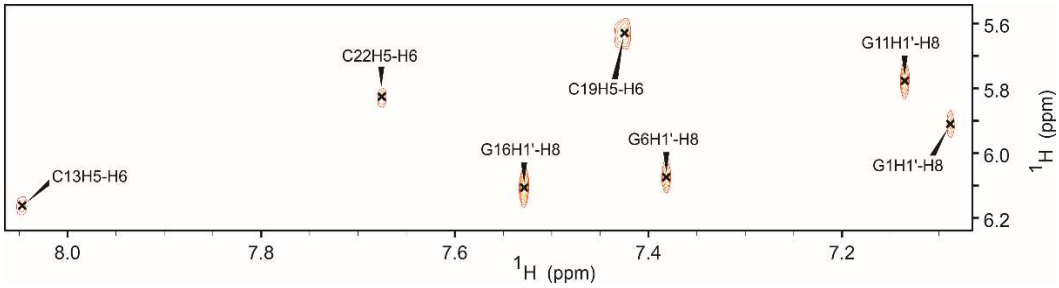


Figure S20. Aromatic-anomeric region of NOESY spectrum (τ_m 40 ms) of kit*. NMR spectrum was recorded at 0.5 mM kit* kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer.

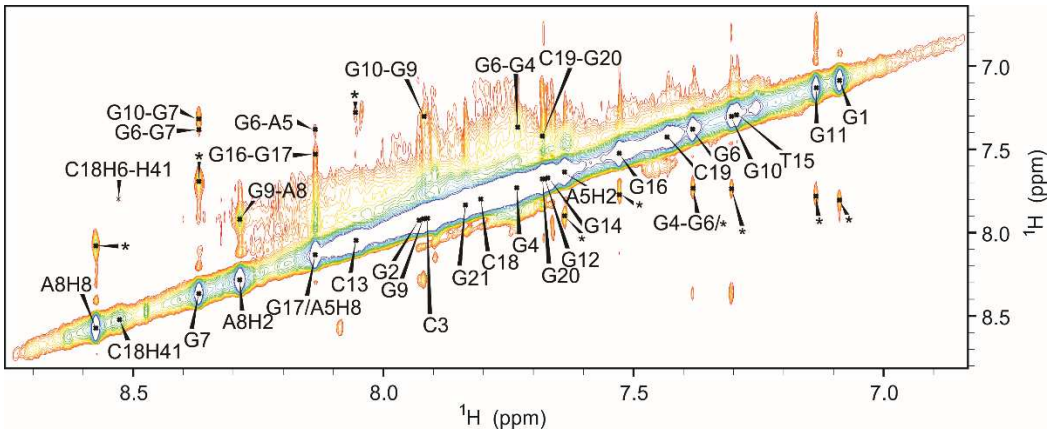


Figure S21. Aromatic-aromatic region of NOESY spectrum (τ_m 450 ms) of kit*. NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on an 800 MHz spectrometer. Unambiguously assigned cross-peaks are marked above the diagonal of NOESY spectrum. Signals that could not be assigned due to overlapping or are assigned to pre-folded structure are marked with stars.

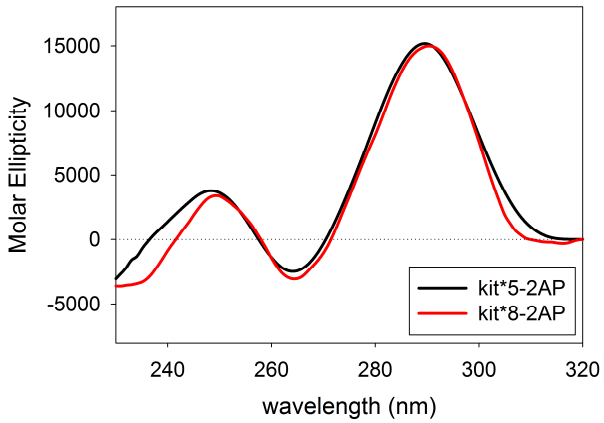


Figure S22. CD spectra of kit*5-2AP and kit*8-2AP in 10 mM TRIS, 100 mM KCl, pH 7.5 at 25 °C.

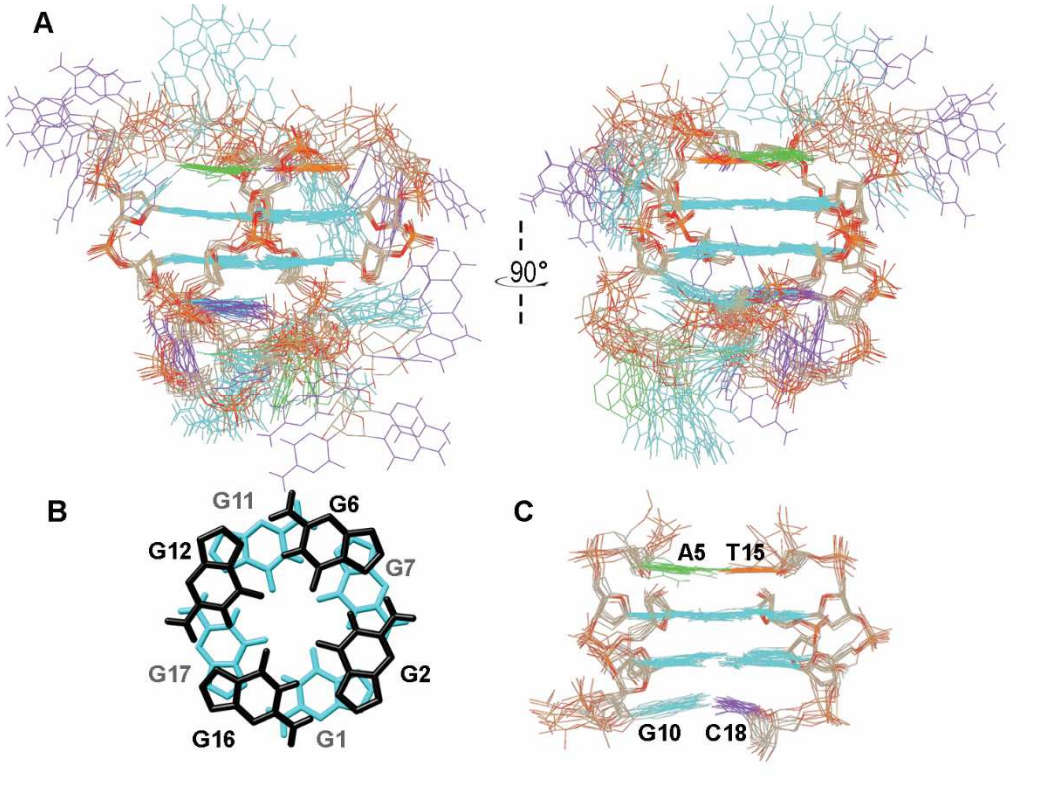


Figure S23. A) Stereoscopic view of the 10 refined superimposed structures of kit* G-quadruplex (PDB ID: 6GH0). B) Stacking of the two G-quartets. C) Side view of the G-quartet core, A5 and T15 stacked on the top G-quartet and G10•C18 base pair stacked on the bottom G-quartet.

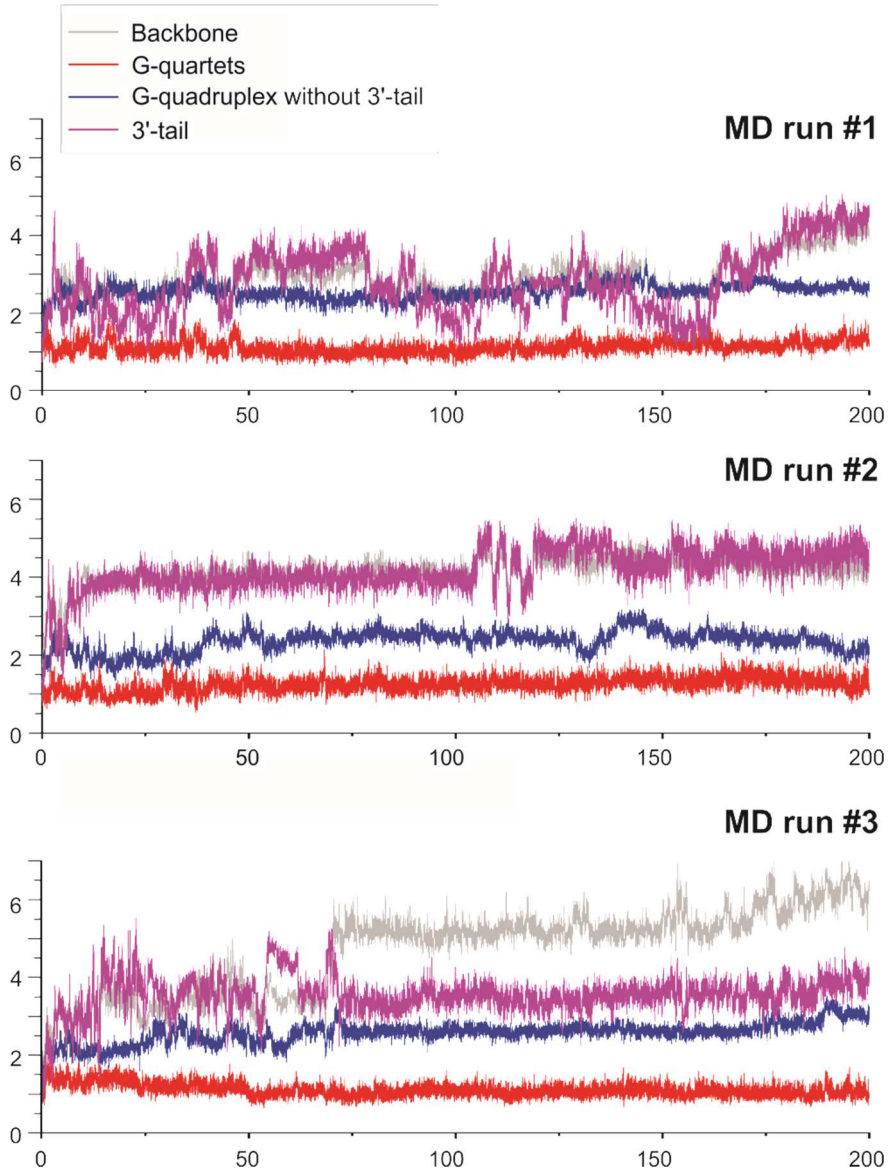


Figure S24. RMSD values of G-quadruplex backbone, G-quartet core, G-quadruplex without the 3'-tail and the 3'-tail alone as a function of simulation time during three independent 200 ns MD simulations.

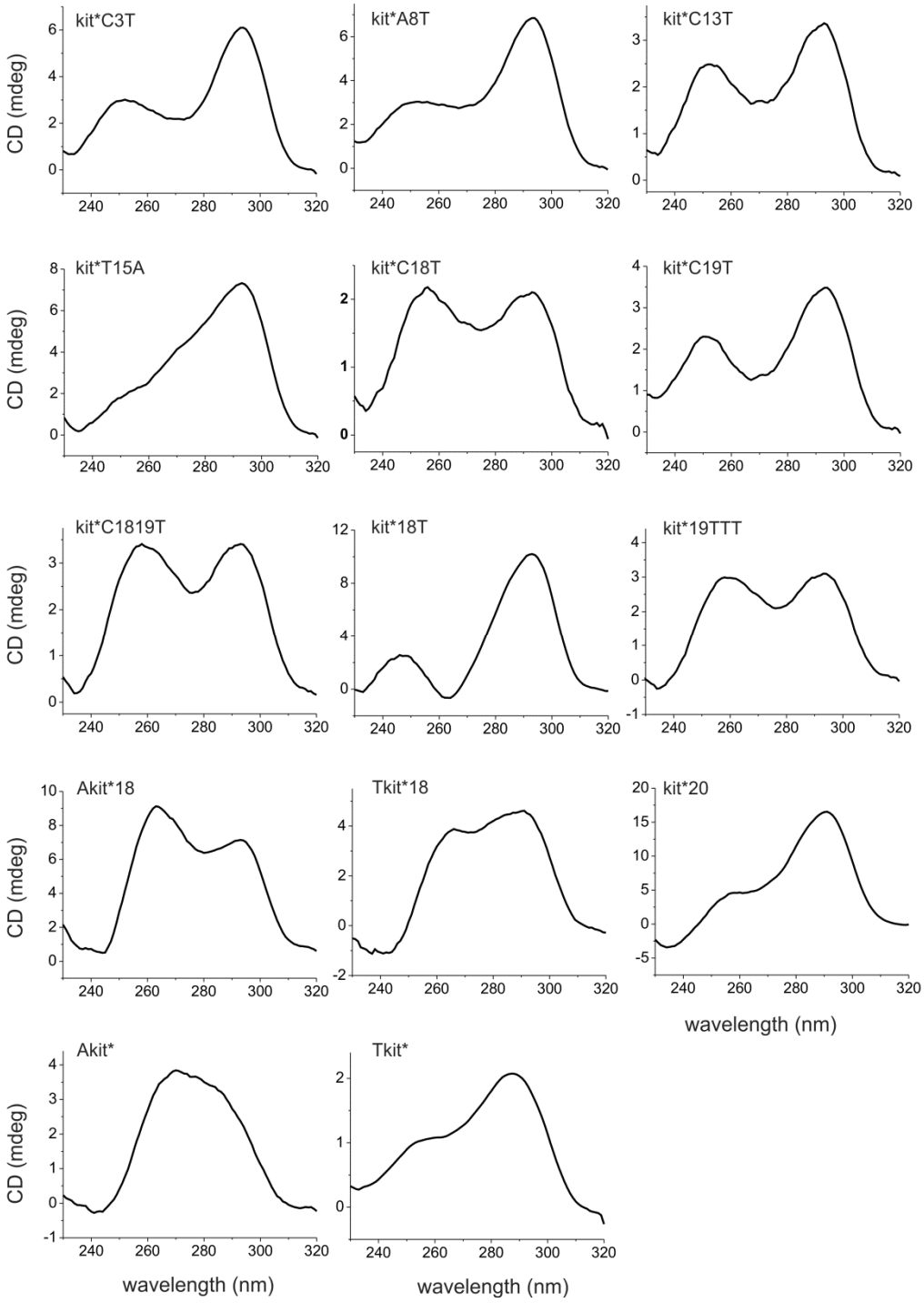


Figure S25. CD spectra of analogues of kit* oligonucleotide. CD spectra were recorded in 20 mM phosphate buffer (pH 7.4), 100 mM KCl at 37°C. The concentrations of oligonucleotides per strand were between 10 and 40 µM.

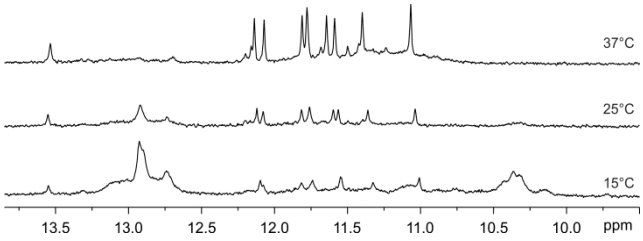


Figure S26. Imino region of ^{1}H NMR spectra of kit*T15A. NMR spectra were recorded at 20 mM phosphate buffer (pH 7.4), 100 mM KCl on a 600 MHz spectrometer.

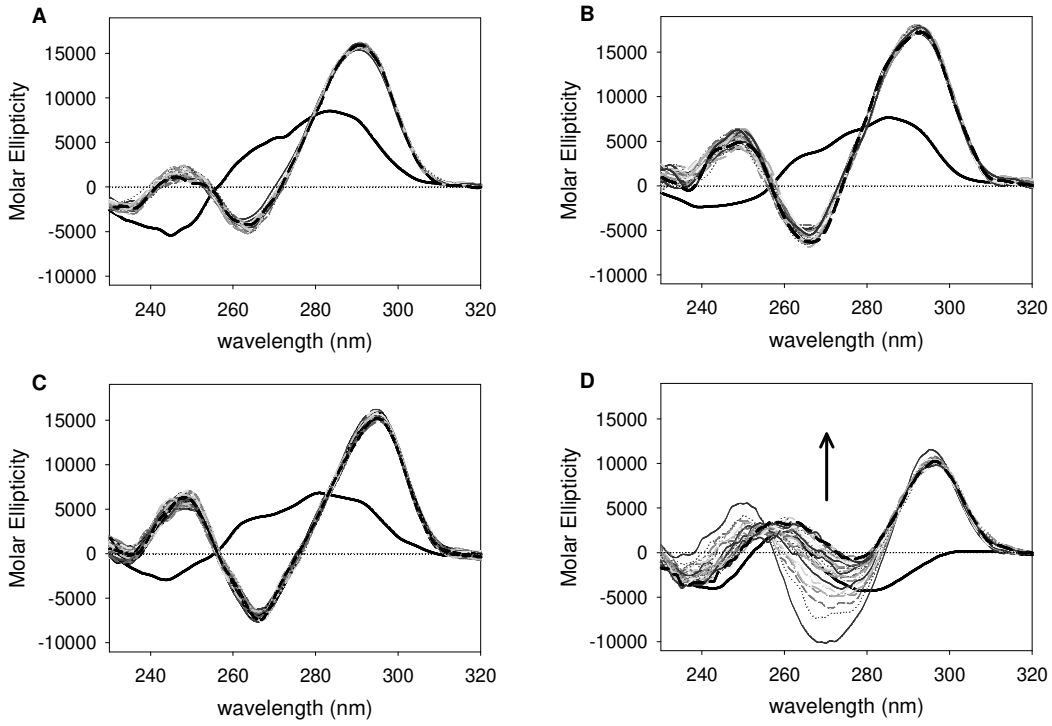


Figure S27. Time-dependent variation of CD spectra of A) kit*, B) kit*19, C) kit*18 and D) kit*17 recorded in 10 mM TRIS, pH 7.5, induced by the addition of 150 mM KCl at 37°C. The solid black lines correspond to CD spectra in the absence of K^{+} ions; the black dashed lines are related to dichroic signals after 15 hours upon addition of KCl.

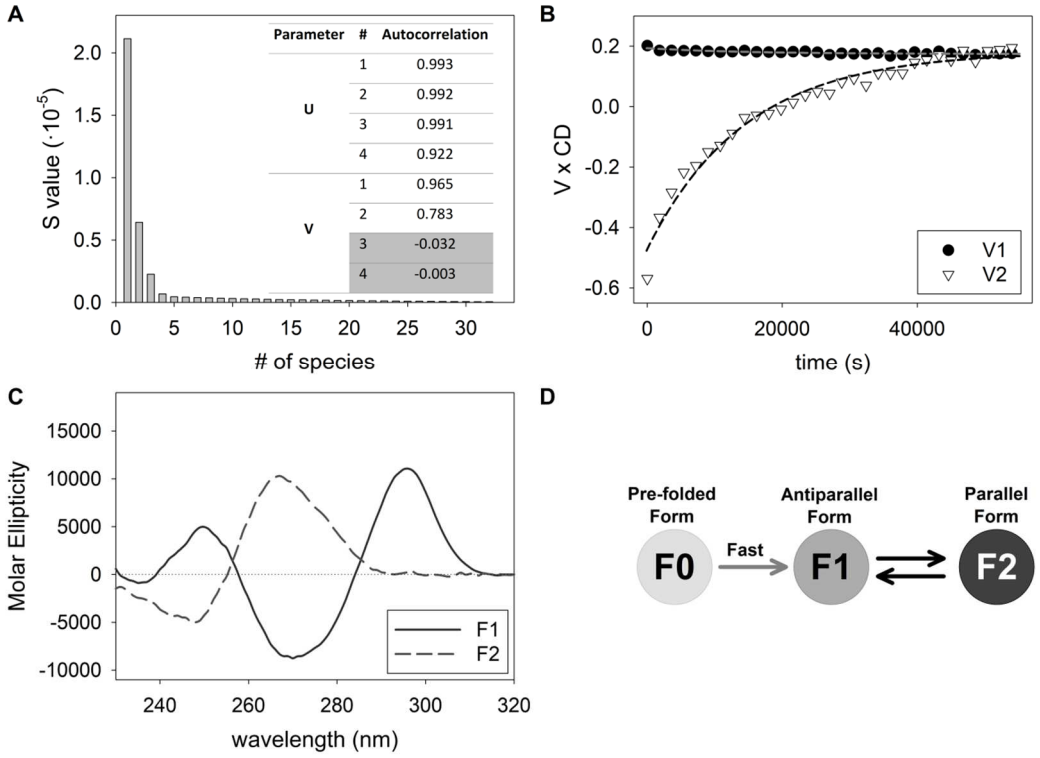


Figure S28. SVD analysis of CD folding kinetics of kit*17 promoted by 150 mM KCl. A) S matrix values and autocorrelation coefficients of U and V matrices indicating the relevance of species in solution participating to the overall dichroic signal variations along time. B) Plot of the significant V eigenvectors as a function of time and their global fitting by using mono-exponential kinetic model. C) CD spectra of species in solution that contribute to the overall changes of the dichroic signal as derived from SVD analysis. D) Proposed scheme of folding of kit*17.

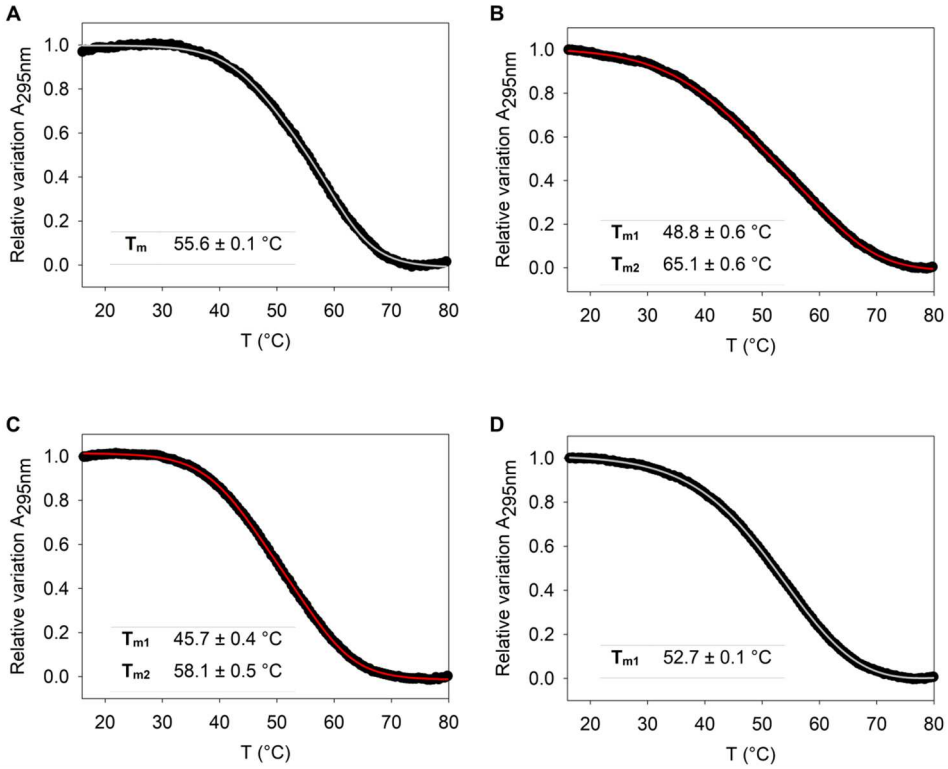


Figure S29. Thermal profiles of A) kit*, B) kit*17, C) kit*18 and D) kit*19. The UV spectra were recorded at 20 µM concentration of an oligonucleotide per strand, in 20 mM potassium phosphate buffer (pH 7.4) and 100 mM KCl. Data points were fitted by applying Equations 3 and 4 to obtain the best fit. Melting temperatures are reported in the graphs.

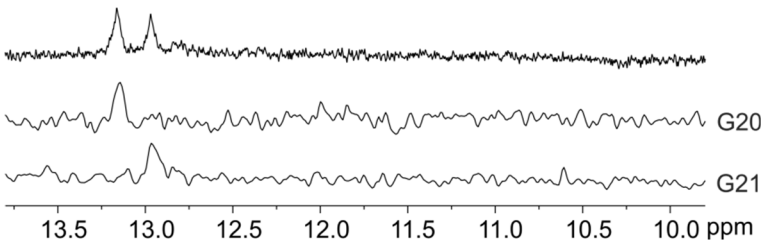


Figure S30. Imino region of ^{1}H and 1D ^{15}N-edited HSQC NMR spectra of pre-folded state of kit* (pre-kit*). The HSQC spectra were acquired on partially (10%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotides. Assignment of H1 proton resonances is indicated on the right of each 1D HSQC spectrum. NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 37°C on a 600 MHz spectrometer.

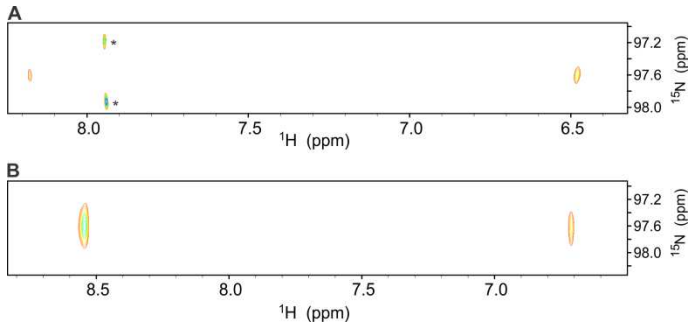


Figure S31. 2D ^{15}N-edited HSQC NMR spectrum of amino group of A) C18 and B) C19 residues of pre-kit*. The HSQC spectrum was acquired on partially (4%) residue-specifically ^{15}N- and ^{13}C-labeled oligonucleotide. NMR spectrum was recorded at 0.5 mM kit* concentration per strand, 37°C on a 600 MHz spectrometer. The signals in A) marked with stars are the F1 artifacts.

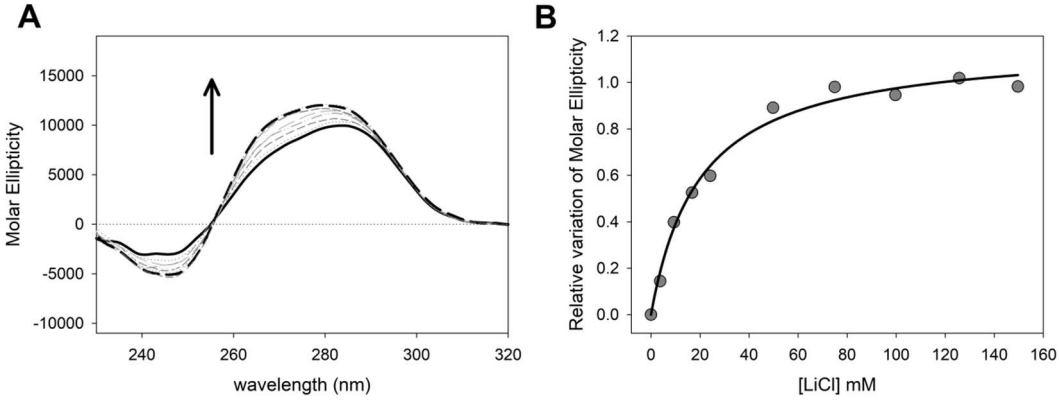


Figure S32. A) CD spectra of kit* titrated with increasing concentration of LiCl in 10 mM TRIS, pH 7.5 at 25°C. B) Relative variation of molar ellipticity obtained monitoring the spectral changes induced by the addition of LiCl at 266 nm. Data were fitted according to one-site saturation model (Equation 1). The obtained K_D value is 19.9 ± 2.4 mM.

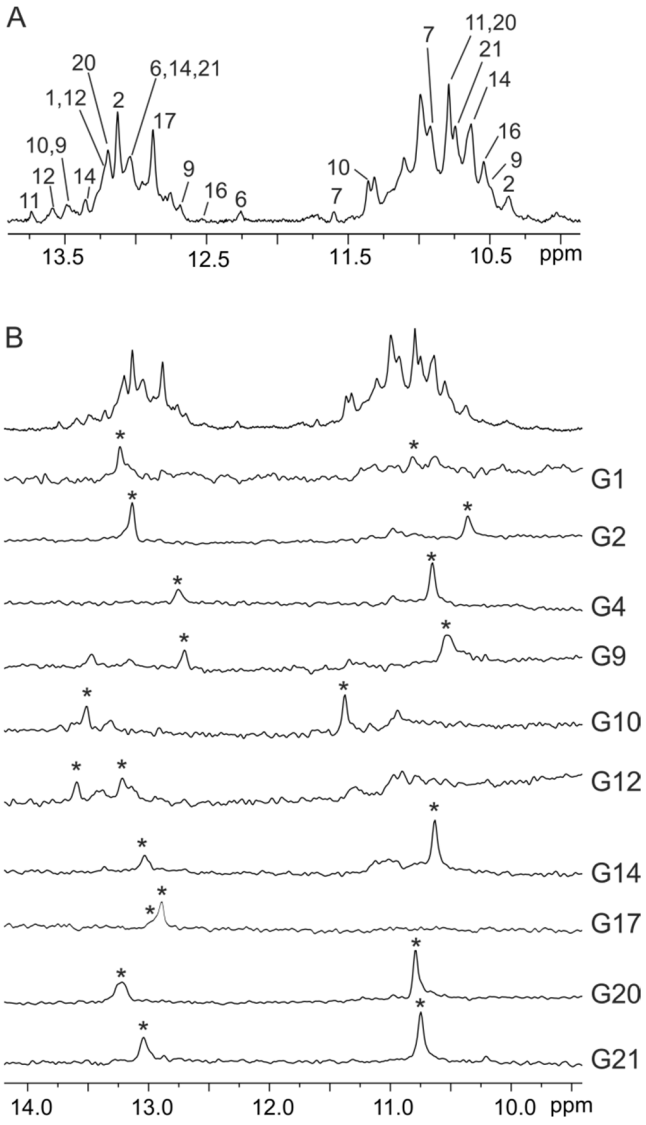


Figure S33. A) Imino region of ^{1}H NMR spectra of pre-folded states of kit*. Assignments are shown above individual signals. B) Imino region of ^{1}H and 1D ^{15}N-edited HSQC NMR spectra of pre-folded states of kit*. The HSQC spectra were acquired on partially (10%) residue-specifically ^{15}N- and ^{13}C-labelled oligonucleotides. Assignment of H1 proton resonances is indicated on the right of each 1D HSQC spectrum. For G6, G7, G11 and G16 residues the 2D ^{15}N-edited HSQC NMR spectra were recorded to obtain better resolution (not shown). NMR spectra were recorded at 0.5 mM kit* concentration per strand, 100 mM KCl, pH 7.4, 5°C on a 600 MHz spectrometer.

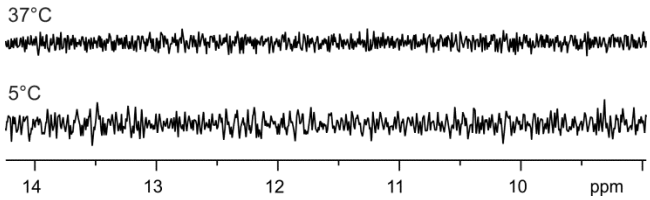


Figure S34. Imino region of ^{1}H NMR spectra of *c* construct at 37 and 5°C. NMR spectra were recorded at 0.3 mM concentration of DNA on a 600 MHz spectrometer.

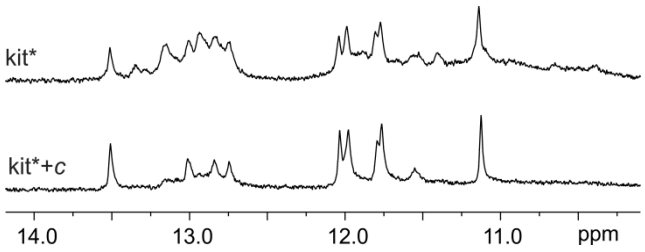


Figure S35. Imino region of ^{1}H NMR spectra of kit* and kit*+*c* at 5°C. NMR spectra were recorded at 100 mM KCl, pH 7.4 on an 800 MHz spectrometer.

Table S1. ^{1}H NMR chemical shifts of kit* G-quadruplex.[a]

Proton / Residue	H6/H8	H1/H2 H5/Me	H1'	H2'/H2"	H3'	H4'	H5'/H5"
G1	7.09	12.13	5.91	2.88/2.51	5.01	4.36	4.15
G2	7.93	12.06	5.63	2.51/2.86	5.01	4.36	4.15
C3	7.91	5.87	6.24	2.12/2.73	n.a.	4.36	4.11
G4	7.73	/	5.74	2.22	4.86	4.29	4.13
A5	8.14	7.64	6.19	2.66	4.84	n.a.	4.17
G6	7.38	11.83	6.08	3.38/2.91	4.83	4.33	3.78/3.64
G7	8.37	11.86	6.25	2.98/2.60	5.21	4.96	4.34
A8	8.57	8.29	6.59	3.04	4.84	4.49	4.18
G9	7.92	/	5.92	1.88	4.81	4.36	4.05
G10	7.30	13.50	5.89	2.54/3.02	5.08	4.79	4.22/4.23
G11	7.14	11.20	5.78	2.67/3.05	4.96	4.37	4.20
G12	7.68	11.65	6.13	2.50	4.83	4.51	4.11
C13	8.05	6.16	6.35	2.22/2.65	4.89	4.41	4.25
G14	7.67	/	5.76	2.28	4.89	4.26	4.08
T15	7.29	1.67	5.85	2.11/2.51	4.68	4.05	3.18/3.74
G16	7.53	12.03	6.11	2.94/3.47	4.91	4.51	4.30/4.07
G17	8.14	12.02	5.78	2.73	5.04	4.51	4.30
C18	7.80	5.53	6.30	2.34/2.62	4.82	4.35	4.28
C19	7.43	5.63	5.96	1.88/2.34	4.81	4.26	4.10
G20	7.68	/	5.73	2.45	4.84	n.a.	4.05
G21	7.84	/	6.00	2.67/2.53	4.89	4.22	4.06
C22	7.68	5.82	6.11	2.32	5.01	n.a.	n.a.

[a] ^{1}H NMR chemical shifts given in ppm were measured in 90% H_2O / 10% $^{2}H_2O$, 37 °C, 0.5 mM concentration of kit* oligonucleotide per strand, 100 mM KCl and 20 mM potassium phosphate buffer (pH 7.4) and referenced to TMS. "n.a." stands for not assigned chemical shifts.

Table S2. Structural statistics for G-quadruplex structure adopted by kit*

NMR distance and torsion angle restraints	
NOE-derived distance restraints	
Total	235
Intra-residue	173
Inter-residue	62
Sequential	50
Long-range	12
Hydrogen bond restraints	18
Torsion angle restraints	22
G-quartet planarity restraints	24
Structure statistics	
*Violation*s	
Mean NOE restraint violation (Å)	0.10 ± 0.02
Max NOE restraint violation (Å)	0.20
Max torsion angle NOE restraint violation (°)	2.36
Deviation from idealized geometry	
Bonds (Å)	0.01 ± 0.00
Angles (°)	2.48 ± 0.05
Pairwise heavy atom RMSD (Å)	
Overall	2.83 ± 0.33
G-quartets	0.60 ± 0.13
G-quartets and G10•C18	0.70 ± 0.12
G-quartets, A5 and T15	0.87 ± 0.19

REFERENCES

1. DeSa, R.J. and Matheson, I.B. (2004) A practical approach to interpretation of singular value decomposition results. *Methods Enzymol.*, **384**, 1-8.
2. Hendler, R.W. and Shrager, R.I. (1994) Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. *J. Biochem. Biophys. Methods*, **28**, 1-33.
3. Gray, R.D., Buscaglia, R. and Chaires, J.B. (2012) Populated intermediates in the thermal unfolding of the human telomeric quadruplex. *J. Am. Chem. Soc.*, **134**, 16834-16844.
4. Greenfield, N.J. (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protoc.*, **1**, 2527-2535.
5. Gray, R.D., Petraccone, L., Buscaglia, R. and Chaires, J.B. (2010) 2-aminopurine as a probe for quadruplex loop structures. *Methods Mol. Biol.*, **608**, 121-136.
6. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A. (2004) Development and testing of a general amber force field. *J. Comput. Chem.*, **25**, 1157-1174.
7. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. and Walker, R.C. (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878-3888.
8. Case, D.A., Berryman, J.T., Betz, R.M., Cerutti, D.S., T.E. Cheatham, T.E., III,, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W. *et al.* (2015) AMBER 2015, University of California, San Francisco.
9. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **118**, 2309-2309.
10. Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T.E., 3rd, 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.
11. Krepl, M., Zgarbova, M., Stadlbauer, P., Otyepka, M., Banas, P., Koca, J., Cheatham, T.E., 3rd, Jurecka, P. and Sponer, J. (2012) Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA. *J. Chem. Theory Comput.*, **8**, 2506-2520.
12. Zgarbova, M., Luque, F.J., Sponer, J., Cheatham, T.E., 3rd, Otyepka, M. and Jurecka, P. (2013) Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.*, **9**, 2339-2354.
13. Zgarbova, M., Sponer, J., Otyepka, M., Cheatham, T.E., 3rd, Galindo-Murillo, R. and Jurecka, P. (2015) Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.*, **11**, 5723-5736.
14. Onufriev, A., Bashford, D. and Case, D.A. (2000) Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B*, **104**, 3712-3720.
15. Onufriev, A., Bashford, D. and Case, D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, **55**, 383-394.
16. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. (1977) Numerical integration of Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.*, **23**, 327-341.

17. Neidle, S. (2010) *Principles of Nucleic Acid Structure*. Academic Press:, London.
18. Pettersen, 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. Comput. Chem.*, **25**, 1605-1612.
19. Ding, Y., Fleming, A.M., He, L. and Burrows, C.J. (2015) Unfolding Kinetics of the Human Telomere i-Motif Under a 10 pN Force Imposed by the α-Hemolysin Nanopore Identify Transient Folded-State Lifetimes at Physiological pH. *J. Am. Chem. Soc.*, **137**, 9053-9060.
20. Phan, A.T., Kuryavyi, V., Gaw, H.Y. and Patel, D.J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nat. Chem. Biol.*, **1**, 167-173.
21. Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., Emeson, R.B. *et al.* (2010) The Solution Structure of the ADAR2 dsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove. *Cell*, **143**, 225-237.
22. Kotar, A., Wang, B., Shivalingam, A., Gonzalez-Garcia, J., Vilar, R. and Plavec, J. (2016) NMR Structure of a Triangulenium-Based Long-Lived Fluorescence Probe Bound to a G-Quadruplex. *Angew. Chem. Int. Ed. Engl.*, **55**, 12508-12511.
23. Darden, T., York, D. and Pedersen, L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.*, **98**, 10089-10092.
24. Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H. and Pedersen, L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577-8593.

Conformational profiling of a G-rich sequence within the *c-KIT* promoter

Riccardo Rigo[1], William L. Dean[2], Robert D. Gray[2], Jonathan B. Chaires[2] and Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

[2] James Graham Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA.

* To whom correspondence should be addressed. Tel: +39-049-827-5711; Fax: +39-049-827-5366; Email: claudia.sissi@unipd.it

Present Address: Claudia Sissi, Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

ABSTRACT

G-quadruplexes (G4) within oncogene promoters are considered to be promising anticancer targets. However, often they undergo complex structural rearrangements that preclude a precise description of the optimal target. Moreover, even when solved structures are available, they refer to the thermodynamically stable forms but little or no information is supplied about their complex multistep folding pathway. To shed light on this issue, we systematically followed the kinetic behavior of a G-rich sequence located within the *c-KIT* proximal promoter (kit2) in the presence of monovalent cations K^+ and Na^+. A very short-lived intermediate was observed to start the G4 folding process in both salt conditions. Subsequently, the two pathways diverge to produce distinct thermodynamically stable species (parallel and antiparallel G-quadruplex in K^+ and Na^+, respectively). Remarkably, in K^+-containing solution a branched pathway is required to drive the wild type sequence to distribute between a monomeric and dimeric G-quadruplex. Our approach has allowed us to identify transient forms whose relative abundance is regulated by the environment; some of them were characterized by a half-life within the timescale of physiological DNA processing events and thus may represent possible unexpected targets for ligands recognition.

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

INTRODUCTION

Nucleic acids can fold into a variety of secondary structures (1). Among them, G-quadruplexes (G4), formed within G-rich DNA or RNA sequences, are of current interest. The basic unit of G4 is the G-tetrad, a square planar array of four guanines interacting through Hoogsteen hydrogen bonds. Two or more G-tetrads assemble through π stacking interactions to build the final G4 arrangement (2). G4 DNA may be comprised of mono-, bi- or tetra-molecular arrangements, although within genomic DNA unimolecular forms are expected to predominate. Overall G4 may also be clustered into parallel, antiparallel or mixed hybrid forms, as defined by participating strand orientation and loop topology.

Recent studies highlighted frequent G-quadruplex formation at oncogene promoters and suggested their relevant role in transcription regulation (3,4). Based on these observations, much effort has been directed towards the identification of small molecules able to modulate the conformational equilibria of these G-rich sequences in a specific and controlled way. The general starting point of modern drug design is the availability of a high-resolution structure of the target obtained by NMR and/or crystallographic techniques. These data provide essential information concerning the final thermodynamically stable structures assumed by the target sequences, but little or no information is supplied about their folding pathway. Nevertheless, oligonucleotide folding into a G4 tends to be a very complex, multistep pathway, and this can be critical since the final structure is achieved passing through different folding intermediates (5).

Another common feature of many G4s is that the thermodynamic equilibrium is reached very slowly. In contrast, recent studies showed that transcriptional processes are rapid events, considering that RNA polymerase II elongation rates in mammalian cells range between 1.3 to 4.3 kb/min (6,7). Additionally, transcription factors bind their specific sites on a time scale of seconds and their residence time varies from seconds to minutes (7-13). It can be concluded that in order to affect the binding of transcription factors to their consensus sequences and to stop the transcription, a G4 in the gene promoter should fold in a timescale close to the activity of the transcriptional machinery. This makes it clear that investigation of short-lived intermediates and folding rates are crucially important. Such short-lived structures might represent a suitable target for G4 ligands leading to unique downstream effects.

A good model for these studies is represented by kit2, a well-known G-rich sequence within the *c-KIT* gene promoter. This gene codes for a tyrosine-kinase receptor which, once activated, participates in a broad range of physiological processes, including cell proliferation, migration, maturation and survival (14,15). Recently, induction of G4 structures within this promoter has been associated with reduction of the expression of the receptor, possibly leading to important anti-cancer effects (16-18). Previous work highlighted the strong polymorphic behavior of kit2 (19,20). In fact, using proper mutants, two different G-quadruplex forms, a monomer and an intertwined dimer, have been resolved, both arranged according to a parallel conformation (21). In solution, the wild type sequence can assume both of these structures resulting in slow folding processes that progressively affect the population distribution (20). Once formed, the monomer and dimer seemingly do not interconvert on a fast time scale, suggesting that they might be derived from different folding pathways (22). All of this evidence suggests that a discrepancy between the solved structures and the pharmacological target(s) could actually exist.

The main purpose of this study was to dissect the G4 folding pathway of kit2 under different ionic conditions by following its kinetic behavior through the application of analytical ultracentrifugation, spectroscopic and electrophoretic techniques in order to derive information about the topology of intermediates that might be relevant during the physiological remodeling of the promoter occurring during the cell cycle. The resulting picture is expected to enable the design of selective ligands able to stop the transcription of *c-KIT* in a more targeted and efficient way.

MATERIALS AND METHODS

Materials

kit2 oligonucleotide 5'-CGGGCGGGCGCGAGGGAGGGG-3' was purchased from Eurogentec (Liège, Belgium) that performed an RP-HPLC purification of the products. We checked the quality of the oligonucleotides by means of PAGE and HPLC without further purification. The oligonucleotide was dissolved in 10 mM TRIS, pH 7.5, to prepare a 1 mM stock solution (strand concentration). Before use, each sample was heated at 95°C for 10

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

minutes in the required buffer and then slowly cooled down at room temperature to equilibrate the system.

Circular Dichroism

Circular dichroism experiments were performed on a JASCO J-810 spectropolarimeter equipped with a Peltier temperature controller. Measurements were obtained using a 1 cm path-length quartz cuvette. CD spectra were recorded from 230 nm to 320 nm, with the following parameters: scanning speed 100 nm/min; band width of 2 nm; data interval of 0.5 nm; response of 2 s. The contribution of the buffer was subtracted from the sample spectra after each acquisition. The nucleic acid concentration, the salt concentration, the buffer composition and the temperature used for the experiments varied according to the purpose of the assays.

For CD kinetic experiments, the selected cation was added manually to the cuvette from a stock solution and mixing was provided by an in-cuvette magnetic stirring bar. After a mixing time of 5 seconds, spectra acquisition was initiated, using an interval scan of 60 seconds. Observed ellipticities were converted to molar ellipticity [Θ] which is equal to $deg \cdot cm^2 \cdot dmol^{-1}$ (Mol. Ellip.) calculated using the DNA residue concentration in solution.

Polyacrylamide gel electrophoresis (PAGE)

For each sample, 200 ng of oligonucleotide were heated at 95°C for 10 minutes in 10 mM TRIS pH7.5 and allowed to cool to room temperature overnight. Afterwards, increasing concentrations of KCl (0-50 mM) were added to the samples and they were left to equilibrate for 24 h at room temperature. Then the samples were loaded on a native 15 % polyacrylamide (19:1 acrylamide: bisacrylamide) PAGE in 1x TBE. A scrambled oligonucleotide (M, which sequence is 5'-GGATGTGAGTGTGAGTGTGAGG-3') of the same molecular weight of the studied sequence was used as an electrophoretic mobility marker. The gel was stained using SYBR green II, and the resolved bands were visualized on an image acquisition system (Geliance 600 Imaging system, Perkin-Elmer).

Analytical Ultracentifugation

AUC was carried out in a Backman Coulter ProteomeLab XL-A analytical ultracentrifuge (Beckman Coulter Inc., Brea, CA) at 25°C overnight at 50000 rpm in standard 2 sector cells. Each sample contained 2 μM kit2 in 10 mM TRIS pH 7.5 equilibrated with/without 50 mM KCl or NaCl. Data were analysed using the program Sedfit (www.sedfit.com). The concentration-dependent distributions of sedimenting species were calculated using the c(S) continuous distribution model considering measured values for buffer density and viscosity. Buffer density was measured on a Mettler/Paar Calculating Density Meter DMA SSA at 20°C, and viscosity was measured using an Anton Paar AMVn Automated Microviscometer. For the calculation of frictional ratio, 0.55 ml/g was used for partial specific volume and 0.3 g/g was assumed for the amount of water bound.

Stopped-flow UV kinetic experiments

Fast folding steps were studied by using a UV stopped flow apparatus equipped with a rapid scanning monochromator (On-Line Instrument System, Borgat, GA, USA). Each sample was prepared in 10 mM TRIS pH 7.5. DNA concentrations and cations were changed according to the purpose of the experiments. The UV-absorbance spectra were recorded from 270 nm to 330 nm with an acquisition rate of 1 ms/spectrum. A water bath maintained the system at a constant temperature of 25 °C. Three mixing experiments were averaged for each data analysis. In order to avoid artifacts, buffer-buffer and buffer-oligonucleotide mixing experiments were performed as controls.

Data analysis

Single-wavelength kinetic experiments were analyzed by nonlinear least square using single or multiple exponential fitting functions.

The best fits for UV stopped flow experiments were obtained using a bi-exponential fitting function (Equation 1),

$$\theta_{t,\lambda} = \theta_{\infty,\lambda} + \theta_{1,\lambda} \cdot e^{(-k_1 \cdot t)} + \theta_{2,\lambda} \cdot e^{(-k_2 \cdot t)}$$

The best fitting for single wavelength CD kinetic experiments were obtained by applying a model that considers concurrent, independent first-order and second order-kinetic

processes. The terms describing these two processes (reported in Table 1 in Ref (23)) were linearly combined to derive Equation 2:

$$\theta_{t,\lambda} = \theta_{\infty,\lambda} + \theta_{4,\lambda} \cdot e^{(-k_4 \cdot t)} + \frac{{\theta_{\infty,\lambda}}^2 \cdot k_3 \cdot t}{1 + \theta_{\infty,\lambda} \cdot k_3 \cdot t}$$

where $\theta_{t,\lambda}$ is the value of signal at time t and $\theta_{\infty,\lambda}$ is the final value of the signal. $\theta_{1\text{-}2\text{-}4,\lambda}$ are amplitude factors for each exponential, while $k_{1\text{-}2\text{-}3\text{-}4}$ are kinetic constants.

Singular Value Decomposition (SVD)

Multiple wavelength CD kinetic experiments were analysed using SVD. This dataset consists of a matrix $\theta_{i,j}$, where $\theta_{i,j}$ is the ellipticity at wavelength i and time j. The data matrix containing the time-dependent CD spectra, named D matrix, is broken up into three different sub-matrices: S matrix, that keeps information about how much every single species contributes to the dichroic signal; the U matrix, that maintains data concerning the spectral shapes of all the conformations in solution; and the V matrix, that indicates how the spectral changes occur over time. The combination of S, U and V matrices provides the initial D matrix, so that D = U x S x V. The S values and U and V autocorrelation coefficients allow determination of the number of species that significantly contribute to the dichroic changes. The best fitting of the significant V eingenvectors was obtained by applying a model considering first order and second order kinetic processes (Equation 2) and determining the kinetic constants and other fitting parameter that contribute to the H matrix. By the linear combination of UxS matrix and H matrix, we obtained the actual spectra of the species in solution significantly contributing to the dichroic signal (24-26). Matlab, Olis Globalworks and Sigmaplot packages were used to perform SVD analysis.

RESULTS

Folded species distribution in potassium-containing solution

As highlighted in the introduction, kit2 can fold into different G-quadruplex structures with different kinetic parameters (20). Our first aim was to assess whether the major

forms present in solution at equilibrium are consistent with the high-resolution structures obtained by NMR on mutated kit2 sequences.

We started the evaluation of the folding process of kit2 by considering the system at thermodynamic equilibrium. In order to monitor the relative distribution of the species in solution in this condition, PAGE experiments were performed following the folding of kit2 in the presence of increasing concentrations of KCl.

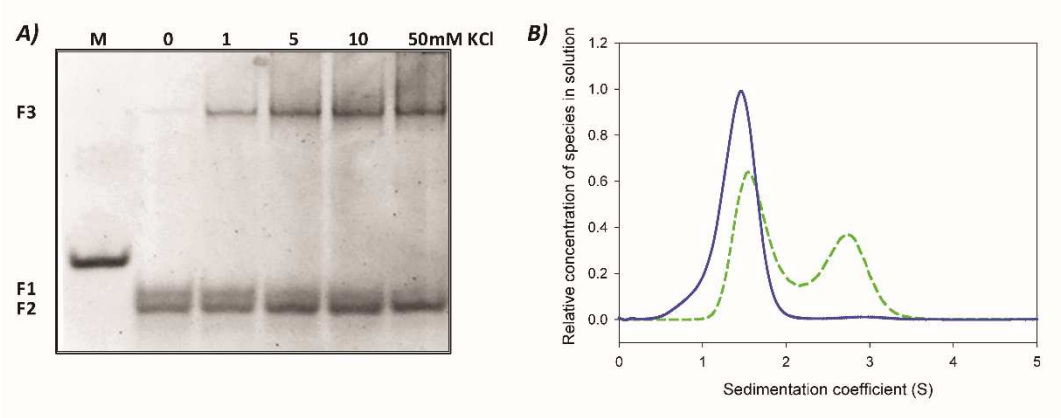


Figure 1. A) Electrophoretic resolution of kit2 in presence of increasing KCl concentration (0-50 mM) performed on 15% acrylamide native PAGE in 1x TBE. M refers to a scrambled oligonucleotide of the same molecular weight, B) relative distribution of kit2 species in solution as function of their sedimentation coefficient with/without 50 mM KCl (respectively, dashed and solid lines) determined in 10 mM TRIS pH 7.5.

As shown in Figure 1A, in the absence of metal ion (lane 0), kit2 shows two different electrophoretic bands (F1 and F2), both with higher mobility when compared to a scrambled oligonucleotide of the same molecular weight (first lane, M). After addition of potassium, the upper band (F1) disappears and converts into F2 and into a new band (F3) with remarkably lower electrophoretic mobility. At 50 mM KCl, the conversion of F1 is complete. Previous data confirmed that these bands can be extracted from the gel with negligible interconversion, thus allowing one to acquire the corresponding CD spectra, both of which are compatible with a parallel G-quadruplex (22). Consistently with the involvement of the monovalent cation, we can assign F2 and F3 to monomeric and dimeric G-quadruplexes, respectively. These conclusions were confirmed by analytical ultracentrifugation (AUC) that allows one to determine the sedimentation coefficient of each species in solution (Figure 1B).

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

AUC data showed that kit2 is mostly monomeric in the absence of monovalent cations (Figure 1B, solid line). In the presence of 50 mM KCl (Figure 1B, dashed line), the population of oligonucleotide splits into two groups, probably monomeric and dimeric forms. The experimentally determined sedimentation coefficients match those calculated using HydroPro software based on the reported NMR solved structures of the monomer and the dimer (PDB: 2KYP and PDB: 2KYO; respectively; Table 1). The integration of the obtained peaks showed that at the equilibrium 60% of kit2 in solution assumes the monomeric form whereas the remaining corresponds to the dimer.

Table 1. Distribution of the kit2 species in solution as determined by AUC analysis in 10 mM TRIS pH 7.5 with/without 50 mM KCl and comparison of their experimentally determined sedimentation coefficients with those calculated by HydroPro software on solved structures.

Species	**Distribution**	**Sedimentation coefficient (S)**	
in solution	**in solution**	**HydroPro**	**AUC**
Monomer	60%	1.50[a]	1.52
Dimer	40%	2.53[b]	2.65
Monomer (no KCl)	100%	/	1.47

[a] data calculated on PDB 2KYP

[b] data calculated on PDB 2KYO

It is important to remember that the dimeric form does not simply derived from the interaction of two single monomeric G-quadruplexes but is a bimolecular inter-strand G4. This unique structural feature, along with the poor monomer-dimer interconversion, reasonably suggests that they might be derived from different folding pathways likely involving other unrevealed DNA intermediates. To test this hypothesis, we have observed the kinetics of the conformational changes of kit2 induced by the addition of a physiological concentration (150 mM) of KCl by CD spectroscopy (Figure 2).

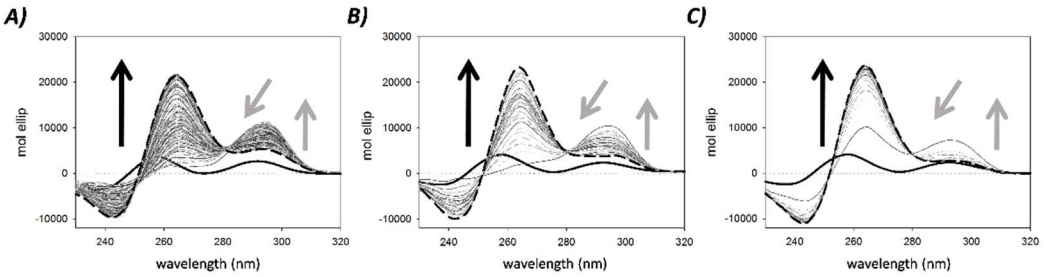


Figure 2. Time dependent changes of CD spectra of 4 µM kit2 induced by the addition of 150 mM KCl in 10 mM TRIS pH 7.5, at A) 10°C, B) 25°C and C) 37°C. Data were collected for 27 h. Arrows indicate the variation of the main peaks as a function of time, solid and dashed lines correspond to the oligonucleotide in the absence of metal ion and at the end of the process, respectively.

In the absence of potassium ions, the CD spectrum of kit2 is not that expected for an unfolded sequence, since it exhibits two positive peaks at 258 nm and 298 nm (Figure 2, solid black lines). Immediately after the addition of the metal ion, a positive peak centered at 294 nm increased whereas the positive peak at 258 nm decreased. Subsequently, the contribution at 294 nm lowered and an intense positive band at 264 nm appeared, thus leading to the final dichroic profile of kit2. According to gel electrophoresis and AUC results, this final spectrum is comprised of contributions from both the monomeric and the dimeric forms (Figure 2, dashed line). Indeed, as previously reported, both of them exhibit a positive peak at 264 nm and a negative one at 245nm (22). As expected, by increasing the temperature, the processes became faster. Nevertheless, the overall spectral datasets (Figure 2) showed the same behavior, thus suggesting that the oligonucleotide undergoes the same structural variations at all the tested temperatures.
This first data set showed that the folding process of kit2 in KCl is not a one-step process, but that it is comprised of at least two processes: an initial fast folding into a kinetic intermediate that then converts into the thermodynamically stable structures. These two steps were then subsequently analysed separately.

The fast folding step analysis

As described above, the addition of KCl caused a rapid structural rearrangement of kit2 that occurred within the mixing time and it was not possible to characterize it by our CD

equipment. In order to properly characterize this intermediate, stopped-flow experiments with UV detection were done. The dead time mixing of potassium was 10-20 ms, allowing for the resolution of faster steps. The absorbance variation in the 270-330 nm range was thus recorded over time.

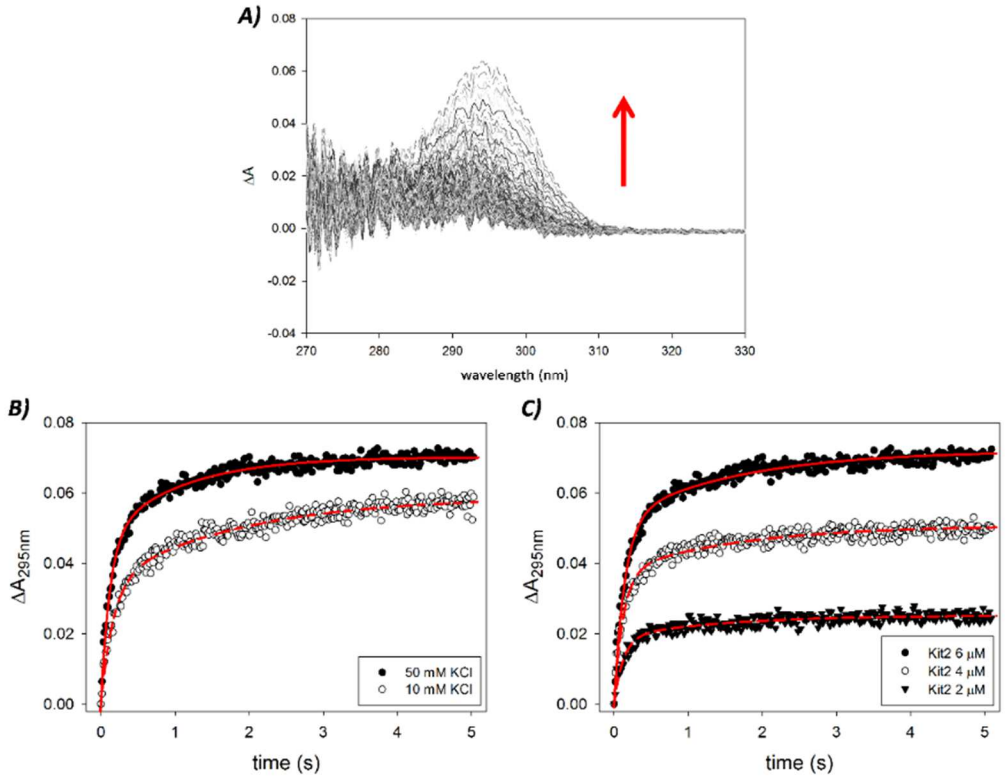


Figure 3. Time dependence variation of UV absorbance of kit2 upon addition of KCl in 10 mM TRIS pH 7.5, 25°C. A) Changes of UV spectra of 6 µM kit2 recorded after addition of 50 mM KCl (data were recorded for 5 s), B) variation of the absorbance at 295nm induced by different concentrations of potassium ion (10 mM and 50 mM) and C) using different concentrations of kit2 (2 µM, 4 µM, 6 µM) in 50 mM KCl.

This approach allowed us to identify a major hyperchromic effect at 295 nm (Figure 3A) that is consistent with the formation of a G-quadruplex structure (27). Thus, in order to derive the kinetic parameters relative to the tetrahelix formation and to minimize the contribution of other forms to the data processing, we analysed the UV variation at 295 nm as a function of the incubation time. Additionally, the stopped-flow experiments were performed at 10 mM and 50 mM KCl to determine if this initial step required potassium (Figure 3B). All experimental data were analysed according to various kinetic models and, as supported by the Shapiro-Wilk normality test, the best fit was always obtained using a

two-step exponential model (Figure 3B, red lines) (28). The derived kinetic constants are reported in Table 2. They indicate that within few seconds after the addition of K^+, kit2 undergoes to at least two conformational changes (leading to forms thereafter named I1 and I2). Interestingly, with respect to G-quadruplex formation, the folding process is faster when the potassium concentration is increased, as reported for telomeric sequences (29,30).

Table 2. Kinetic constants obtained by fitting the absorbance variation recorded at 295 nm after the addition of KCl using a "two kinetic processes" model. Data were acquired in 10 mM TRIS at 25°C.

[KCl]	[kit2]	k_1 (s^{-1})	k_2 (s^{-1})
10 mM	**6 µM**	6.00 ± 0.33	0.54 ± 0.04
50 mM	**6 µM**	6.96 ± 0.20	0.62 ± 0.03
50 mM	**4 µM**	8.13 ± 0.34	0.62 ± 0.04
50 mM	**2 µM**	8.12 ± 0.62	0.70± 0.09

To check if these initial folding steps were DNA concentration-dependent, stopped-flow experiments were repeated at 2 µM, 4 µM and 6 µM of oligonucleotide. The results are reported in Figure 3C. As expected, by increasing the amount of DNA, the recorded UV signals increased whereas the derived kinetic constants did not change proportionally with the oligonucleotide concentration (Table 2). This is evidence that the species involved along these initial steps are monomeric.

The slow folding step analysis

Since our evidence indicated that the very first part of the folding process is complete within 5 seconds which represent the lag-time for our CD experimental protocol, we were only able to use CD for a fine analysis of the slower step. Data were acquired in 50 mM KCl at different oligonucleotide concentrations (20 and 6 µM, Figure S1). Immediately

after the addition of potassium, the concentration of DNA in solution did not affect the molar ellipticity at 294 nm that reached a constant intensity of about 11000 deg·cm^2·dmol^{-1}. This data perfectly matches with our stopped-flow kinetic experiments performed at different oligonucleotide concentrations and further supports the hypothesis that in the first fast-steps no dimeric species are formed.

Interestingly, after complete equilibration (about 2.5 h), the molar ellipticity at 264 nm was much more intense in 20 µM kit2. This is not unexpected since the dimer should provide a higher dichroic signal than the monomer (31). In our system, this phenomenon was associated with a decrease of the signal at 294 nm that was directly related to the oligonucleotide concentration. Nevertheless, it remains much lower than the change at 264 nm. Thus, to analyse the slow kinetic rearrangements of our sequence, we took into account the signal at the lower wavelength. Considering the monomer-dimer formation, the experimental data points were well fitted by using an equation that contains first order (k_4) and second order (k_3) kinetic processes (Eq. 2, see Materials and Methods). The derived constants are summarized in Table 3.

Table 3. Kinetic constants determined applying a "first order + second order kinetic processes" fitting model to the dichroic signal variation at 264 nm of kit2 at different concentrations (6 µM and 20 µM) induced by the addition of 50 mM KCl in 10 mM TRIS pH 7.5, at 25°C.

[cKit2]	**k_3 ($s^{-1}M^{-1}$) ($\cdot 10^2$)**	**k_4 (s^{-1}) ($\cdot 10^2$)**
6 µM	0.33 ± 0.07	0.040 ± 0.008
20 µM	7.30 ± 1.86	0.076 ± 0.002

In both conditions, the first process (k_3) is much faster than the second one. Interestingly this constant is strongly influenced by the amount of kit2 in solution (Table 3). This suggests that the process associated with k_3 is the one responsible for the dimer formation.

Profiling the slower process using SVD analysis

Due to the simultaneous formation of monomeric and dimeric G4s in the slower steps, it was worthwhile to derive not only the kinetic parameters but also a reliable prediction of the number of participating species in solution as well as their basic dichroic spectra. Thus, we decided to apply Singular Value Decomposition (SVD) analysis to the data matrix formed by the time-dependent CD spectra of kit2 in the whole wavelength range. Indeed, SVD is an analytical tool that is very useful for characterizing complex kinetics events (24-26). In order to refer our analysis to kit2 behavior in a more physiological environment, we chose to perform SVD analysis on a dataset obtained in 150 mM KCl. Since to perform an accurate SVD analysis, we needed to acquire significant data points for all the forms involved in the folding process, we decided to work at 10 °C. Indeed, as above reported (Figure 2), by reducing the temperature the folding process becomes slower, but the conformational changes are conserved.

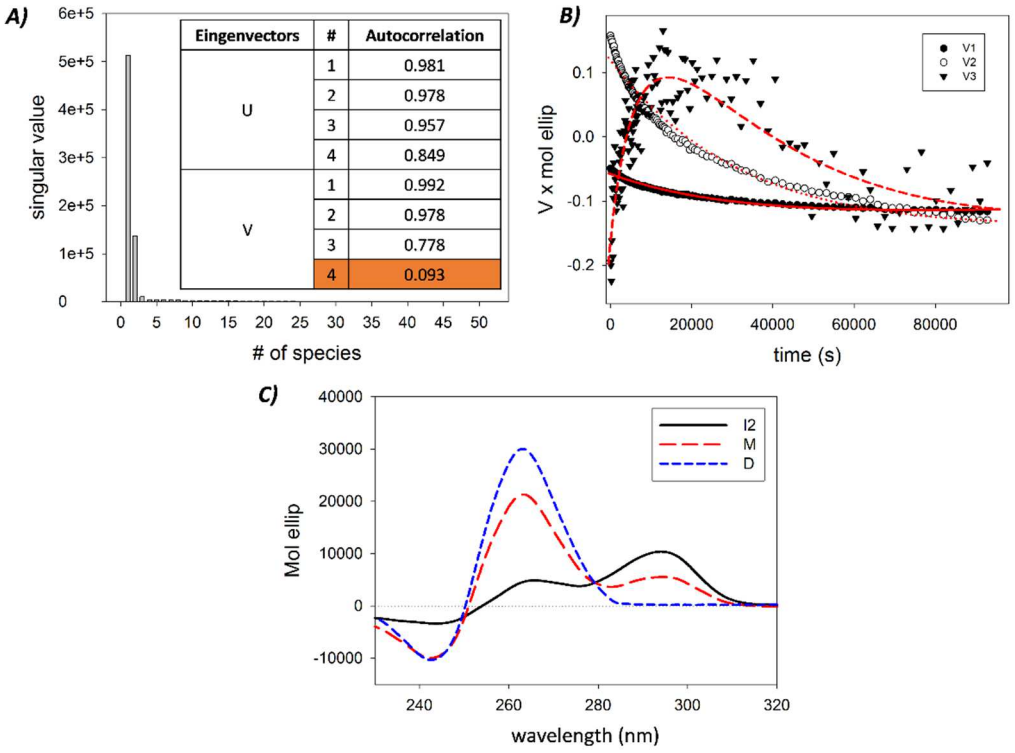


Figure 4. A) S matrix values and autocorrelation coefficients of U and V matrices indicating the relevance of species in solution participating to the overall dichroic signal variations, B) plot of the significant V eingenvectors as a function of time and their global fitting (red line) (Eq. 2, see Materials and Methods), C) CD basic spectra of the species in solution that contribute to the overall changes of the dichroic signal as derived from SVD analysis,

where I2 refers to the second intermediate, M to the final monomeric species and D to the dimer.

In agreement with the result of data analysis performed at a single wavelength, the singular values contained in the S matrix as well as the U and V autocorrelation coefficients (meaningful above 0.75) indicated that there are at least three species in solution that contribute to the variation of the CD signal (Figure 4A). These species should account for the thermodynamically stable structures (the monomeric and the dimeric G-quadruplexes observed in AUC and PAGE) as well as for the initial intermediate, derived from the fast-initial steps. The significant V eingenvectors were well fitted by Equation 2 (Figure 4B). This model provided the kinetic constants (Table 4) and the parameters forming the H matrix. The linear combination of H matrix and UxS matrix provides the actual spectra of the species recurring along the process.

Table 4. Fitting parameters derived by applying the "first order + second order independent processes" kinetic model to the significant V-eingenvectors obtained using SVD analysis. Spectral Dataset acquired upon the addition of 150 mM KCl in 10 mM TRIS at 10°C.

Fitting model	**R^2**	**Kinetic constants ($\cdot 10^5$)**
First order + Second order independent processes	0.91	$k_3 = 9.02 \pm 3.52\ s^{-1}M^{-1}$
		$k_4 = 2.90 \pm 0.20\ s^{-1}$

The so obtained spectral shapes are reported in Figure 4C. One species presents a positive peak at 294 nm and a shoulder at 270 nm. It corresponds to the starting point of these kinetic experiments and we can relate it to the second folding intermediate formed at the end of the fast folding phase (I2). Its dichroic features indicate it as a hybrid G-quadruplex. The other two derived spectra are attributable to the thermodynamically stable species, the monomer and the dimer. Both present a main positive chiroptical contribution at 264 nm and a minor negative one at 245 nm, in line with their parallel conformation. As reported in the literature, due to the higher number of bases participating in the structure

formation, the dimeric species is expected to exhibit a more intense dichroic signal (31). This observation confirms that the highest dichroic signal is related to the dimeric form.

Characterization of the species in sodium-containing solution

It is well documented that different ionic species can contribute to G-quadruplex formation while often leading to different specific structures. Even though sodium is less representative of the physiological environment in the nucleolus, this cation can interfere in the final folding of the studied sequence. Thus, we evaluated the ability of sodium ions to induce kit2 to assume a G-quadruplex arrangement.

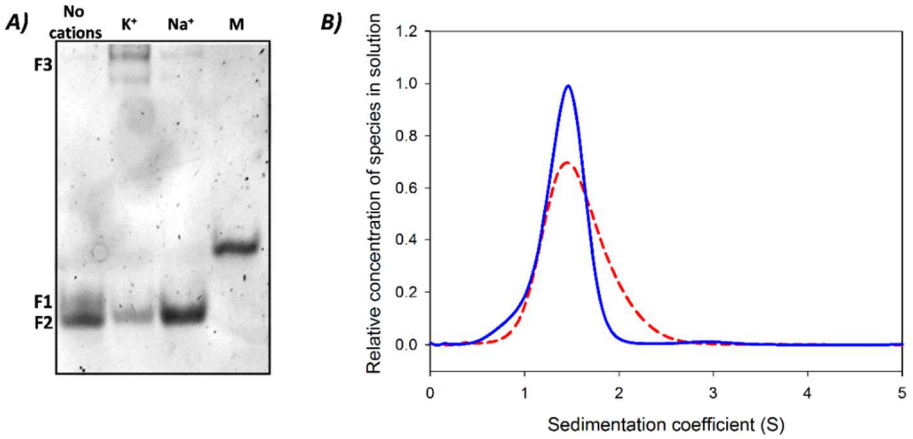


Figure 5. A) Electrophoretic resolution of kit2 annealed under different ionic conditions (no cations, 150 mM KCl, 150 mM NaCl) performed on 15% acrylamide native PAGE in 1x TBE using a scrambled oligonucleotide of the same molecular weight as marker (M), B) relative distribution of kit2 species in solution as function of their sedimentation coefficient with/without 50 mM NaCl (respectively, dashed and solid lines) in 10 mM TRIS pH 7.5.

Electrophoretic resolution showed that when kit2 was annealed in the presence of sodium, it formed one fast electrophoretic band corresponding to the monomeric G4 characterized in K^+ but no multimeric species (Figure 5A). Analytical ultracentrifugation results were in agreement: they confirmed that NaCl changes the hydrodynamic volume of the oligonucleotide and leads to a distribution of the sedimentation coefficients compatible with the monomeric oligonucleotide (Figure 5B). Nevertheless, the shape of

the curve appeared to broaden, thus suggesting the presence of a mixed population that includes more than one monomeric form of kit2.

In order to determine if this profile is comprised of a mixture of more than one G4 form or of unfolded/folded species, the CD spectrum of kit2 was recorded in the presence of increasing concentration of sodium chloride until a plateau was reached.

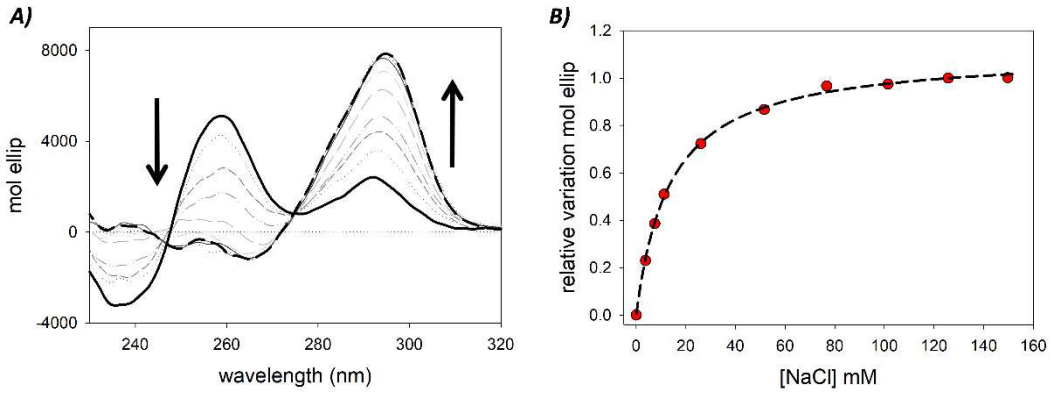


Figure 6. A) CD spectra representing the titration of 4 µM kit2 with increasing concentration of NaCl (0-150 mM) in 10 mM TRIS pH 7.5 at 25°C, and B) relative variation of molar ellipticity obtained monitoring the spectral changes induced by the addition of NaCl at 294 nm.

As reported in Figure 6A, the CD spectrum of the pre-folded kit2 changed upon NaCl addition and showed the formation of a positive band around 294 nm and a negative band at 264 nm. These are the expected features of an antiparallel G-quadruplex. During CD titrations, an isodichroic point was maintained at 275 nm. Accordingly, the relative variation of the signal recorded at 294 nm as function of the NaCl (Figure 6B) was well fitted using a "one binding site model". The derived KD_{app} value was 13.69 ± 0.47 mM, higher than the value reported for KCl (9.49 ± 0.83 mM) (22). Also the intensity of the final dichroic signal is lower in Na^+ than in K^+, thus suggesting that in sodium the G-quadruplex form of kit2 is less rigidly organized. This data is consistent with a general lower efficiency of Na^+ vs K^+ in the stabilization of G-quadruplex structures. When this observation is coupled with the evidence that in 50 mM NaCl kit2 is not completely folded into a G4 (approximately the 15% maintains the pre-folded form), it becomes easier to justify the low resolution of monomeric components derived by AUC.

Folding kinetics of kit2 in sodium

In order to assess how the presence of Na^+ also impacts the folding kinetics of kit2, a comparable spectroscopic approach (CD kinetics and stopped-flow UV analyses) was performed.

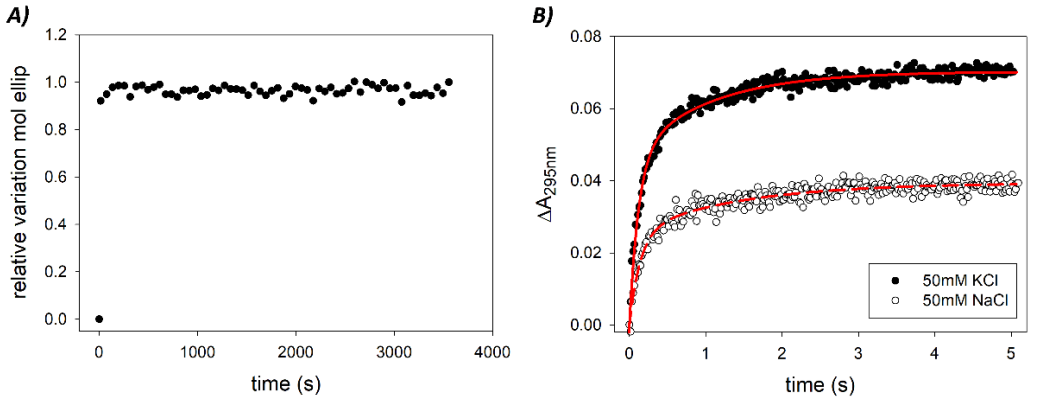


Figure 7. A) Kinetic profile of kit2 folding after the addition of NaCl (final concentration 50 mM) monitoring the variation of molar ellipticity at 294 nm, B) absorbance variation at 295nm of 6 µM kit2 induced by the addition of different metal ions (50 mM sodium or potassium chloride) plotted over time recorded in 10 mM TRIS, pH 7.5, 25°C.

When we explored the fast processes by stopped-flow UV, we observed that the addition of NaCl induced effects comparable to KCl (Figure 7B). Indeed, also in this case a "two kinetic processes model" (eq. (1)) properly fitted the absorbance variation at 295 nm as a function of time (Figure 7B). The kinetic parameters are summarized in Table 5.

Table 5. Kinetic constants determined applying a "two kinetic processes" (Eq. 1, see Materials and Methods) fitting model to the absorbance variation recorded at 295 nm for kit2 (6 µM) upon the addition of 50 mM NaCl, in 10 mM TRIS, pH 7.5, 25°C.

[NaCl]	k_1' (s^{-1})	k_2' (s^{-1})
50 mM	7.04 ± 0.49	0.71 ± 0.05

However, distinct from results in potassium-containing solution, in the presence of sodium no additional slow folding steps were detected, thus making the folding process

very fast and fully completed in the timescale of few seconds (Figure 7A). This means that in sodium a reduced number of folding intermediates (named I1' and I2') is required to reach the final antiparallel G-quadruplex.

Sodium - potassium folding competition

The final folded state of kit2 in NaCl consists of an antiparallel G-quadruplex that does not correspond to other species identified along the potassium driven folding pathway. The physiological environment contains a complex mixture of different ions, and even though cells present a smaller amount of Na^+ compared to the concentration of K^+, the first cation can affect the overall arrangement of the oligonucleotide as well as the intermediates required to interconvert the sodium vs the potassium folded forms. Therefore, after the characterization of the oligonucleotide conformations in each cation species, we considered the folding of kit2 in solutions containing both the metal ions. We started from a solution of kit2 equilibrated in 50 mM sodium chloride to which we added an equimolar KCl concentration and the changes of the dichroic spectrum were recorded over time.

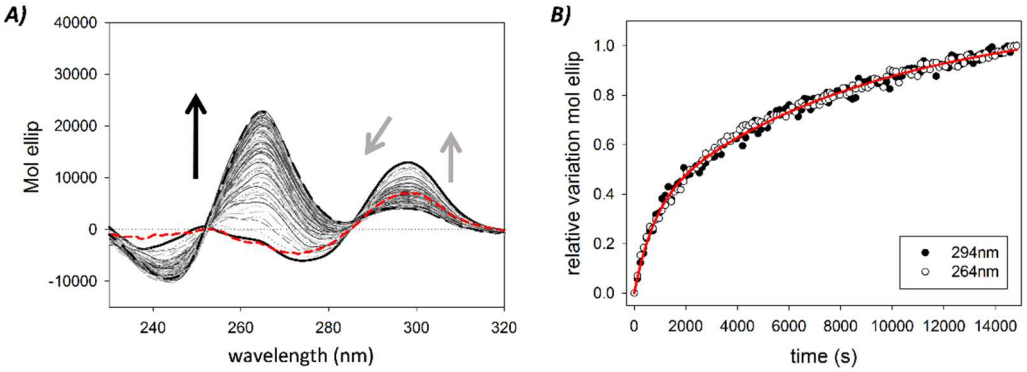


Figure 8. A) CD spectra acquired during the kinetic structural rearrangement of 6 µM kit2 after the addition of KCl to a final concentration of 50 mM in 10 mM TRIS pH 7.5, 50 mM NaCl at 25°C (data were collected for 4 h). Dashed red line, solid black line and the dashed black line represent, respectively, kit2 folded in 50 mM in NaCl, immediately after the addition of 50 mM KCl and at the thermodynamic equilibrium. The corresponding relative variations of the signals recorded at 294 nm and at 264 nm are reported in panel B.

Upon addition of KCl, the starting spectrum corresponding to the Na^+-stabilized antiparallel G-quadruplex (Figure 8A, red line), immediately changed (the timescale is

comparable to the mixing time) triggering an increase of the signal at 294 nm and a decrease of the band at 275 nm. This fast ion exchange has already been observed on the telomeric sequence (32). When compared to the fast-forming intermediate characterized in KCl-containing solution, the main difference rests in the lack of any contribution at 260 nm. Its presence suggests that in these experimental conditions potassium can transiently favor or stabilize an antiparallel arrangement of kit2 (named I3), not previously detected along the KCl folding pathway in the absence of Na^+. Also this intermediate then slowly interconverts to provide the expected spectrum for potassium containing solutions (a positive peak at 264 nm and a negative one at 245 nm). Again, this slower kinetic process was well described by a "first order + second order independent processes" (Eq.2).
Thus, the presence of sodium is not sufficient to impair the formation of the K^+-induced parallel forms although it modulates the conformational equilibria of the sequence in solution. Indeed, as reported in Table 6, the rates of the slower steps are slower in Na^+/K^+ than in K^+ alone.

Table 6. Kinetic constants obtained by a "first order + second order independent processes" fitting model applied to the variation of molar ellipticity recorded over time at 264 nm and 294 nm of kit2 in 50 mM NaCl and 10 mM TRIS upon the addition of 50 mM KCl at 25°C.

Salt conditions	k_3' ($s^{-1}M^{-1}$) ($\cdot 10^2$)	k_4' (s^{-1}) ($\cdot 10^2$)
50 mM NaCl-50 mM KCl	0.13 ± 0.01	0.011 ± 0.001

DISCUSSION

A key challenge in G4 ligand design concerns the polymorphic behavior of the selected targets. In fact, whereas high resolution techniques, like as NMR and X-ray are extensively applied to provide structural data related to the thermodynamically stable forms, information about G-quadruplex folding pathways are currently limited to a few examples (33). The folding of G4 structures in general is recognized to be kinetically complex and might be described by a number of theoretical concepts, including "folding funnels",

"kinetic partitioning" or more conventional sequential or parallel kinetic reaction mechanisms (5,33-35). These are difficult to test and distinguish experimentally, especially because of the differing timescales accessible to different experimental or computational methods. What is clear is that folding of G4 structures is not a simple two-state process and that a complex array of intermediate states is populated along the folding pathway. These transient intermediates may play physiologically important roles and should not be neglected.

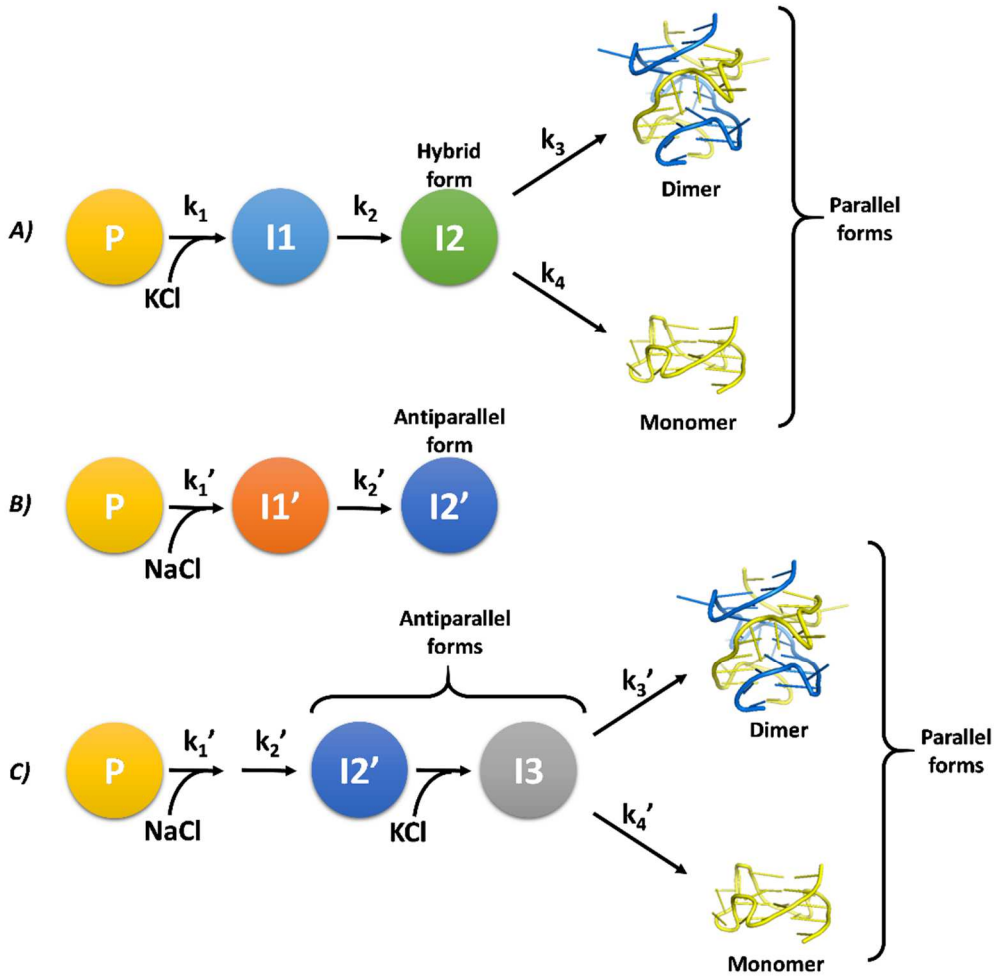


Figure 9. A) Potassium-induced folding pathway of kit2, B) sodium-induced folding pathway, C) sodium-potassium mixture folding pathway.

In this context, the results herein provide deep insights on kit2 folding equilibria.

The cation-induced folding pathways of kit2 were shown to be very complex and greatly affected by the environmental conditions. A schematic illustration of the folding processes based on data presented herein is shown in Figure 9.

The first observation is that in the absence of cations, the sequence appears to be arranged in a pre-folded form (P) with spectroscopic and electrophoretic features completely different from those expected for an unfolded oligonucleotide. This behavior has been reported earlier for other G-rich sequences (29,36). We inferred that our experimental starting conditions (10 mM TRIS) present a sufficient ionic strength to support the collapse of the unfolded oligonucleotide into a compact structure, e.g. hairpin (35,37,38). Upon the addition of cations to the solution, the electrostatic repulsion among the phosphate groups is strongly and quickly reduced. In accordance with electrolyte theory, this favors the formation of more structured species ultimately leading to the final G4 forms (39).

At the very early stage of cation-driven structural rearrangement of kit2, two fast conformational changes are detected either in the presence of sodium or potassium ions. UV stopped-flow measurements confirmed that these initial folding steps affect the absorbance signal at 295 nm and the derived kinetic constants are cation-concentration dependent. As far it concerns I1 and I1' (the first intermediates in K^+ and Na^+, respectively) we did not succeed in obtaining structural information, but these spectral features are compatible with G-quadruplex-like structures. They could represent intermediates with less than a full stoichiometric complement of cations as already reported in the literature for the initial folding step of other G-rich sequences or triplex intermediates (29).

The formation constants for these initial species are only slightly different when measured in Na^+ or K^+. The same occurs for I2 and I2', the first intermediates for which we succeeded in acquiring structural information. Under both experimental conditions, they derive from the full conversion of I1 or I1', respectively. This is supported by their half-life timescale and, limited to K^+ condition, by SVD analysis that excluded a significant presence of I1 in the slower part of the process. Despite these extensive similarities, I2 and I2' do not belong to overlapping folding pathways. Indeed, in sodium, I2' represents the thermodynamically stable structure and our data indicate it is a monomeric antiparallel G-quadruplex. This is in keeping with the behavior previously reported for the telomeric sequence (29). Conversely, the potassium-induced I2 is a short-lived intermediate corresponding to a hybrid G-quadruplex that slowly rearranges into the final forms.

Following the evolution of I2 on a time-scale of hours, SVD analysis showed that the Kit2 energetic landscape in potassium does not favor the stability of the hybrid form. Indeed, it converts into the final parallel monomeric and dimeric forms stably trapped into two

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

energetic minima. According to the data obtained by analytical ultracentrifugation, these two species correspond to the structures already reported in literature (21).

The plasticity of kit2 is further reinforced by exploring more complex ionic environments, where additional conformational rearrangements can be detected. Indeed, starting from the sodium-stabilized antiparallel G4 (I2'), the addition of potassium cations drives the process towards the formation of the same thermodynamically stable parallel G-quadruplexes. Interestingly, this process involves an unprecedented intermediate (I3), when compared to K^+-containing conditions. It fully retains the antiparallel arrangement typically observed in Na^+. This means that the oligonucleotide can follow multiple pathways to reach the final monomeric and the dimeric parallel structures.

As already reported for other sequences such as the telomeric one, we can propose that KCl efficiently replaces within the mixing time the sodium cations preserving the G-quadruplex core (40). Herein, the observed stabilization of the antiparallel form in potassium, underlined by the rapid increase of its dichroic signal after its addition, is the experimental evidence that this replacement occurs. Subsequently, this unique potassium-stabilized antiparallel G-quadruplex (I3) acts as short-lived intermediate that directly leads to the formation of the final monomer and dimer. Thus, in this experimental condition, the previous identified K^+-induced intermediate I2 is not involved in the oligonucleotide folding process.

It is worthwhile to compare our model to those derived for the widely studied telomeric sequence (5,33). A branched pathway is required to describe our data. A major difference is that we were not able to identify antiparallel intermediates in pure K^+ solutions. Nevertheless, we cannot exclude the involvement of this species since it appears to represent the main component in the Na^+/K^+ conformational switch (40). Such an highly polymorphic behavior for kit2 has been proposed to be enabled by the long and unstructured central loop that allows it to adopt either an all-parallel, propeller type conformation, or a mixed parallel/antiparallel conformation (20).

In conclusion, based on the data presented herein, it appears that the wild-type kit2 sequence also converges towards two shared main G-quadruplex folded forms. It emerges that the population of G-quadruplex or G-quadruplex-like conformations in solution is comprised of multiple components and that its composition significantly varies with time and environmental conditions. The presence of several relevant folding intermediates, some of them characterized by a half-life within the timescale of

physiological DNA processing events, definitely supports their potential role in gene regulatory processes. Thus, taking into account only the high-resolution structures of kit2 as the actual G4 elements working during the transcription, leads to a loss of information regarding the functional arrangement of the gene promoter.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

This work was supported by University of Padova [CPDA147272 to C.S., PhD fellowships to RR]. Additional support was provided by NIH grant GM077422 (to JBC).

CONFLICT OF INTEREST

Authors have nothing to declare.

ACKNOWLEDGEMENT

We acknowledge Del Villar-Guerra R. for support with data analysis.

REFERENCES

1. 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.
2. Davis, J.T. (2004) G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed. Engl.*, **43**, 668-698.
3. 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.
4. Hansel-Hertsch, R., Di Antonio, M. and Balasubramanian, S. (2017) DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell. Biol.*, **18**, 279-284.
5. Gray, R.D., Trent, J.O. and Chaires, J.B. (2014) Folding and unfolding pathways of the human telomeric G-quadruplex. *J. Mol. Biol.*, **426**, 1629-1650.
6. Maiuri, P., Knezevich, A., De Marco, A., Mazza, D., Kula, A., McNally, J.G. and Marcello, A. (2011) Fast transcription rates of RNA polymerase II in human cells. *EMBO Rep.*, **12**, 1280-1285.
7. Ben-Ari, Y., Brody, Y., Kinor, N., Mor, A., Tsukamoto, T., Spector, D.L., Singer, R.H. and Shav-Tal, Y. (2010) The life of an mRNA in space and time. *J. Cell. Sci.*, **123**, 1761-1774.
8. Morisaki, T., Muller, W.G., Golob, N., Mazza, D. and McNally, J.G. (2014) Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat. Commun.*, **5**, 4456-4464.

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

9. Karpova, T.S., Kim, M.J., Spriet, C., Nalley, K., Stasevich, T.J., Kherrouche, Z., Heliot, L. and McNally, J.G. (2008) Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter. *Science*, **319**, 466-469.
10. Sharp, Z.D., Mancini, M.G., Hinojos, C.A., Dai, F., Berno, V., Szafran, A.T., Smith, K.P., Lele, T.P., Ingber, D.E. and Mancini, M.A. (2006) Estrogen-receptor-alpha exchange and chromatin dynamics are ligand- and domain-dependent. *J. Cell. Sci.*, **119**, 4101-4116.
11. Bosisio, D., Marazzi, I., Agresti, A., Shimizu, N., Bianchi, M.E. and Natoli, G. (2006) A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-kappaB-dependent gene activity. *EMBO J.*, **25**, 798-810.
12. McNally, J.G., Muller, W.G., Walker, D., Wolford, R. and Hager, G.L. (2000) The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science*, **287**, 1262-1265.
13. Gebhardt, J.C., Suter, D.M., Roy, R., Zhao, Z.W., Chapman, A.R., Basu, S., Maniatis, T. and Xie, X.S. (2013) Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat. Methods*, **10**, 421-426.
14. Metcalfe, D.D. (2008) Mast cells and mastocytosis. *Blood*, **112**, 946-956.
15. Gregory-Bryson, E., Bartlett, E., Kiupel, M., Hayes, S. and Yuzbasiyan-Gurkan, V. (2010) Canine and human gastrointestinal stromal tumors display similar mutations in c-KIT exon 11. *BMC Cancer*, **10**, 559-568.
16. Zorzan, E., Da Ros, S., Musetti, C., Shahidian, L.Z., Coelho, N.F., Bonsembiante, F., Letard, S., Gelain, M.E., Palumbo, M., Dubreuil, P. *et al.* (2016) Screening of candidate G-quadruplex ligands for the human c-KIT promotorial region and their effects in multiple in-vitro models. *Oncotarget*, **7**, 21658-21675.
17. 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.
18. 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.
19. Miller, M.C., Le, H.T., Dean, W.L., Holt, P.A., Chaires, J.B. and Trent, J.O. (2011) Polymorphism and resolution of oncogene promoter quadruplex-forming sequences. *Org. Biomol. Chem.*, **9**, 7633-7637.
20. 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.
21. Kuryavyi, 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.
22. Da Ros, S., Zorzan, E., Giantin, M., Zorro Shahidian, L., Palumbo, M., Dacasto, M. and Sissi, C. (2014) Sequencing and G-quadruplex folding of the canine proto-oncogene KIT promoter region: might dog be used as a model for human disease? *PLoS One*, **9**, e103876.
23. Greenfield, N.J. (2006) Analysis of the kinetics of folding of proteins and peptides using circular dichroism. *Nat. Protoc.*, **1**, 2891-2899.

24. Hendler, R.W. and Shrager, R.I. (1994) Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. *J. Biochem. Biophys. Methods*, **28**, 1-33.
25. Gray, R.D. and Chaires, J.B. (2011) Analysis of multidimensional G-quadruplex melting curves. *Curr. Protoc. Nucleic Acid Chem.*, **Chapter 17**, Unit17 14.
26. DeSa, R.J. and Matheson, I.B. (2004) A practical approach to interpretation of singular value decomposition results. *Methods Enzymol.*, **384**, 1-8.
27. Mergny, J.L., Li, J., Lacroix, L., Amrane, S. and Chaires, J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
28. Peat, J. and Barton, B. (2008) *Medical Statistics: A guide to data analysis and critical appraisal.* 1st ed. ed. Blackwell Publishing.
29. Gray, R.D. and Chaires, J.B. (2008) Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures. *Nucleic Acids Res.*, **36**, 4191-4203.
30. Ying, L., Green, J.J., Li, H., Klenerman, D. and Balasubramanian, S. (2003) Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U S A*, **100**, 14629-14634.
31. Tothova, P., Krafcikova, P. and Viglasky, V. (2014) Formation of highly ordered multimers in G-quadruplexes. *Biochemistry*, **53**, 7013-7027.
32. Gray, R.D., Petraccone, L., Trent, J.O. and Chaires, J.B. (2010) Characterization of a K+-induced conformational switch in a human telomeric DNA oligonucleotide using 2-aminopurine fluorescence. *Biochemistry*, **49**, 179-194.
33. Marchand, A. and Gabelica, V. (2016) Folding and misfolding pathways of G-quadruplex DNA. *Nucleic Acids Res.*, **44**, 10999-11012.
34. Bessi, I., Jonker, H.R., Richter, C. and Schwalbe, H. (2015) Involvement of Long-Lived Intermediate States in the Complex Folding Pathway of the Human Telomeric G-Quadruplex. *Angew. Chem. Int. Ed. Engl.*, **54**, 8444-8448.
35. Stadlbauer, P., Kuhrova, P., Banas, P., Koca, J., Bussi, G., Trantirek, L., Otyepka, M. and Sponer, J. (2015) Hairpins participating in folding of human telomeric sequence quadruplexes studied by standard and T-REMD simulations. *Nucleic Acids Res.*, **43**, 9626-9644.
36. Ceru, S., Sket, P., Prislan, I., Lah, J. and Plavec, J. (2014) A new pathway of DNA G-quadruplex formation. *Angew. Chem. Int. Ed. Engl.*, **53**, 4881-4884.
37. Murthy, V.L. and Rose, G.D. (2000) Is counterion delocalization responsible for collapse in RNA folding? *Biochemistry*, **39**, 14365-14370.
38. Thirumalai, D. and Hyeon, C. (2005) RNA and protein folding: common themes and variations. *Biochemistry*, **44**, 4957-4970.
39. Manning, G.S. (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys.*, **11**, 179-246.
40. Gray, R.D., Li, J. and Chaires, J.B. (2009) Energetics and kinetics of a conformational switch in G-quadruplex DNA. *J. Phys. Chem. B*, **113**, 2676-2683.

Published – Rigo, R.; Dean, W. L.; Grey, R. D.; Chaires, J. B.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (22), 13056-13067

SUPPLEMENTARY INFORMATION

Conformational profiling of a G-rich sequence within the *c-KIT* promoter

Riccardo Rigo[1], William L. Dean[2], Robert D. Gray[2], Jonathan B. Chaires[2] and Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

[2] James Graham Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA.

* To whom correspondence should be addressed. Tel: +39-049-827-5711; Fax: +39-049-827-5366; Email: claudia.sissi@unipd.it

Present Address: Claudia Sissi, Department of Pharmaceutical and Pharmacological Sciences, University of Padova, 35131 Padova, Italy

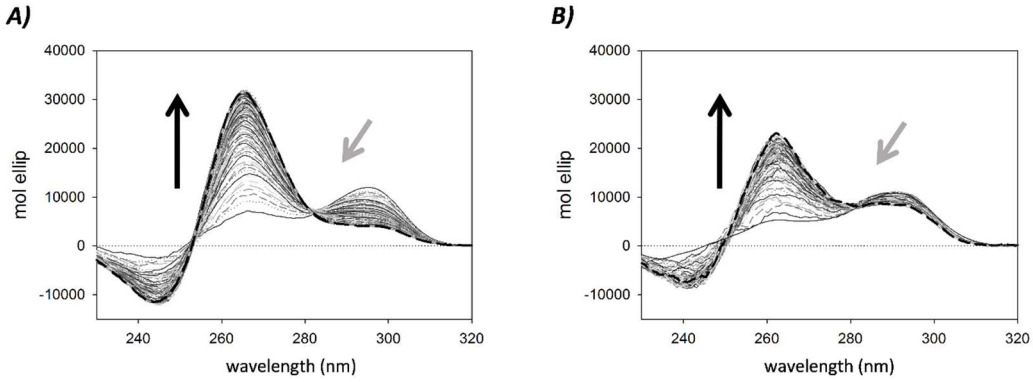


Figure S1. CD spectra acquired along the time-dependent folding of kit2 at different concentration - A) 20 µM and B) 6 µM DNA - in 10 mM TRIS pH 7.5 induced by the addition of 50 mM KCl at 25°C. Data were recorded for 4 h.

Characterization of G4-G4 crosstalk in *c-KIT* promoter region

Riccardo Rigo[1], Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Via Marzolo 5, 35131 Padova, Italy

ABSTRACT

The proximal promoter of *c-KIT* contains a peculiar domain that comprise three closely located short G-rich sequences able to fold into non-canonical DNA secondary structures called G-quadruplexes (G4). Here, we focused on a sequence corresponding to two consecutive G4 forming (kit2 and kit*) and, by electrophoretic, SPR and spectroscopic techniques, we demonstrated that they retain the ability to fold into G4 when inserted in the extended sequence. Noteworthy, we highlighted the occurrence of a crosstalk between the two forming units. This previously unexplored G4-G4 interaction modulates both the con-formation and the stability of the overall arrangement of *c-KIT* promoter. It is not supported by stacking of single nucleotides, but it refers to a G4-G4 interaction surface surrounded by a 2-nucleotides loop that might represent a reliable unprecedented target for anticancer therapy.

G quadruplexes (G4) are polymorphic DNA arrangements forming in G-rich sequences. They are characterized by planar arrays of four guanines paired by Hoogsteen hydrogen bonds (G-quartets) that interact one to each other through π-π stacking interactions. The overall stability of these arrangements is further improved by monovalent cations, such as potassium or sodium.[1] The promoters of many oncogenes contain G-rich sequences that are able to adopt this tetra-helical conformation.[2,3] A meaningful example is the proto-oncogene *c-KIT*. It encodes for a tyrosine kinase receptor (c-kit) which expression and activation are increased in many different cancer types where it sustains cell proliferation, migration, maturation and survival.[4] Within the *c-KIT* proximal promoter three short G-rich sequences have been identified, namely kit1, kit2 and kit*. Their ability to fold into G4 has been proved and, for kit1 and kit2, high resolution structures are available.[5-9] Experimental data suggest that upon stabilization of these G4 structures in

Published – Rigo, R.; Sissi, C. *Biochemistry.* **2017**, 56 (33), 4309-4312

promoter, the transcriptional process of *c-KIT* is impaired whereas the opposite occurs in cell treated with ligands able to reduce G4 formation.[10,11]

Within the promoter, the three G4-forming domains are closely clustered and spaced few nucleotides one from each other (Figure 1A). Thus, in analogy with the reported behavior of long telomeric repeats as well as for hTERT promoter, their potential crosstalk through G4-G4 interactions must be considered.[12,13] This is relevant since it can lead to a rearrangement of the over-all geometry of the participating DNA structural domains and it can generate a G4-G4 interaction surface that might represent a reliable unprecedented pharmaceutical target. Clearly, to properly dissect the functional role of this G-rich domain in DNA processing, the characterization of this higher-order arrangement is necessary. Indeed, its structural features are expected to regulate protein recruitment, in particular transcription factors for *c-KIT* transcription. Worth of mentioned, the promoter portion comprising kit2 and kit*contains validate and putative binding sites for Sp1 and AP2, respectively.[7,14]

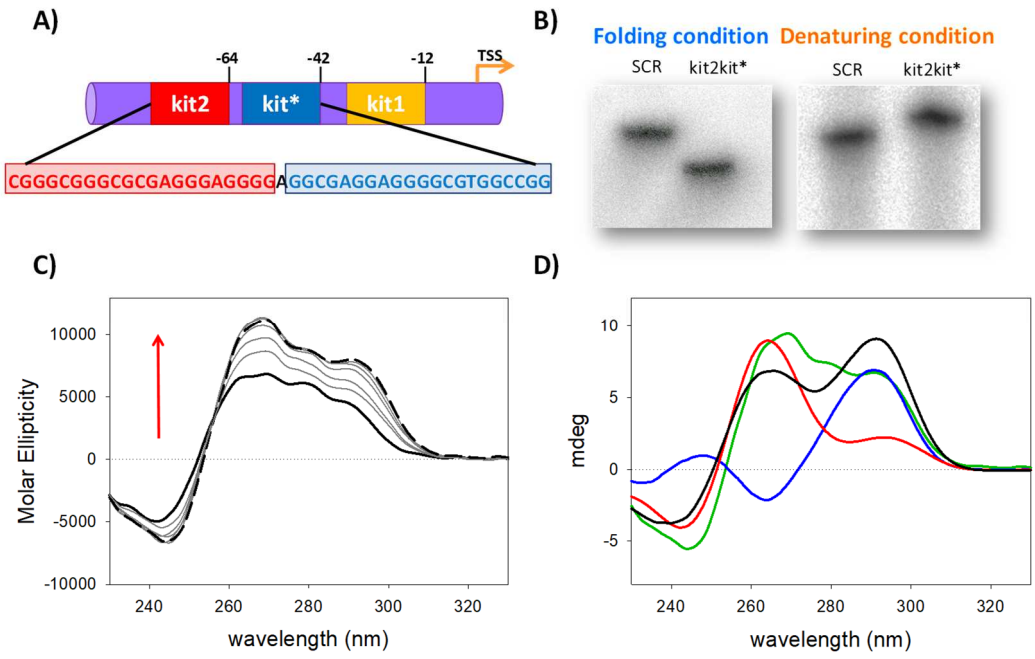


Figure 1. (A) Schematic representation of *c-KIT* proximal promoter, reporting the kit2kit* sequence; kit2 and kit* sequences are highlighted in red and blue, respectively. (B) Electrophoretic mobility of kit2kit* with reference to a 38 residues-long marker (SCR) performed under folding and denaturing conditions. (C) CD titration of kit2kit* with KCl in 10 mM TRIS, pH 7.5, 25 °C. Solid and dashed black lines refer to the sequence in the absence and in the presence of an excess of metal ion, respectively. (D) CD spectra of 2

μM kit* (blue line), kit2 (red line) and kit2kit* (green line) acquired in 10 mM TRIS, 150 mM KCl pH 7.5, 25 °C. The arithmetic sum of the dichroic contributions from kit2 and kit* is represented by the black line.

For these reasons, as first step, we investigated if kit2 and kit* could reciprocally interact and modify their conformations and stabilities.

To preliminary assess this issue, we performed electrophoretic mobility shift experiments on a sequence comprising both the kit2 and kit* G-rich segments, called kit2kit* (Figure 1A). As expected, in denaturing conditions the electrophoretic mobility of kit2kit* well parallels the one expected with reference to a 38-residue long oligonucleotide unable to intra-molecularly fold (SCR). Conversely, in the presence of 150 mM KCl (Figure 1, panel B, native condition) kit2kit* runs much faster, thus sup-porting its folding into an intra-molecular arrangement. This structural change was also followed by means of circular dichroism titrations of kit2kit* with KCl (Figure 1C). A potassium dependent folding of the oligonucleotide emerged as indicated by the increment of the intensity of the dichroic signal. On the so obtained K^+-induced folded form, we acquired the thermal difference spectrum (TDS) since its bands intensity and localization represent a specific signature for each single DNA structural motif.[15] The TDS of kit2kit* showed a negative band around 295 nm and two positive contributions at 273 nm and 255 nm with a shoulder at 243 nm (Figure S1), compatible with the expected result for a G4. Thus, we ruled out other possible arrangements[16,17] and concluded that the longer kit2kit* domain retains the ability to fold into intramolecular G4.

Interestingly, the response of kit2kit* to KCl proved to be largely different from the one obtained in the presence of equimolar concentration of kit2 and kit* (Figure S2). As a result, the spectrum of kit2kit* in 150 mM KCl (Figure 1, panel D, solid green line) does not overlap the one derived from the sum of the dichroic contribution of kit2 and kit* (Figure 1, panel D, solid black line) acquired in the same experimental conditions. In particular, the recorded CD spectrum of the longer sequence presents a positive peak at 268 nm, with two shoulders at 282 nm and 295 nm whereas the two isolated G4 are expected to generate just two positive peaks, the major one at 295 nm and a smaller one at 264 nm. The same results were acquired considering kit2-A, which was used along with kit* to obtain in solution a residue combination perfectly superimposable to the longer sequence (Figure S2). This observation supports that the tested G-rich units behave

differently when analyzed as isolated blocks or when inserted within a single long sequence.

Once assessed the occurrence of a distinctive G4 arrangement by kit2kit*, its melting profile was studied and compared to those of its building blocks. As reported in Figure S3, the CD melting profile of kit* corresponds to a single reversible process that is consistent with a direct conversion of the folded G4 into the unfolded oligonucleotide without significant contributions of intermediate species. By using a single-transition model based on van't Hoff formalism, we determined kit* melting temperature and unfolding enthalpy (Table 1).[18] We could not apply the same analytical protocol to kit2 because, as reported in the literature, in solution it distributes between a monomeric and a dimeric parallel G4 and this condition hampered a correct data analysis.[5,19] We solved this bias by taking advantage of the elongated kit2-A sequence. Indeed, it preserves the ability to fold into a parallel G4, but it does not form multimeric species (Figure S4).

Table 1. Thermodynamic parameters of unfolding processes of kit*, kit2-A and kit2kit* in 150 mM KCl, derived by CD titrations. T1-T3 refer to the three thermal transitions for kit2kit*.

Sequence		**T_m (°C)**	**ΔH (kcal/mol)**
kit*		60.5 ± 0.1[#]	-25.5 ± 0.4[#]
kit2-A		76.0 ± 0.3[#]	-37.2 ± 1.0[##]
kit2kit*	T1	42.5 ± 0.5[&]	-28.0 ± 2.5[&]
	T2	58.0 ± 1.4[&]	-32.8 ± 3.9[&]
	T3	75.3 ± 2.2[&]	-38.9 ± 4.6[&]

(#) from melting profile acquired at 295 nm
(##) from melting profile acquired at 264 nm
(&) Singular Value Decomposition analysis

Accordingly, also the melting of this sequence in 150 mM KCl was nicely fitted by a single-transition process allowing us to derive the corresponding thermodynamic parameters summarized in Table 1.

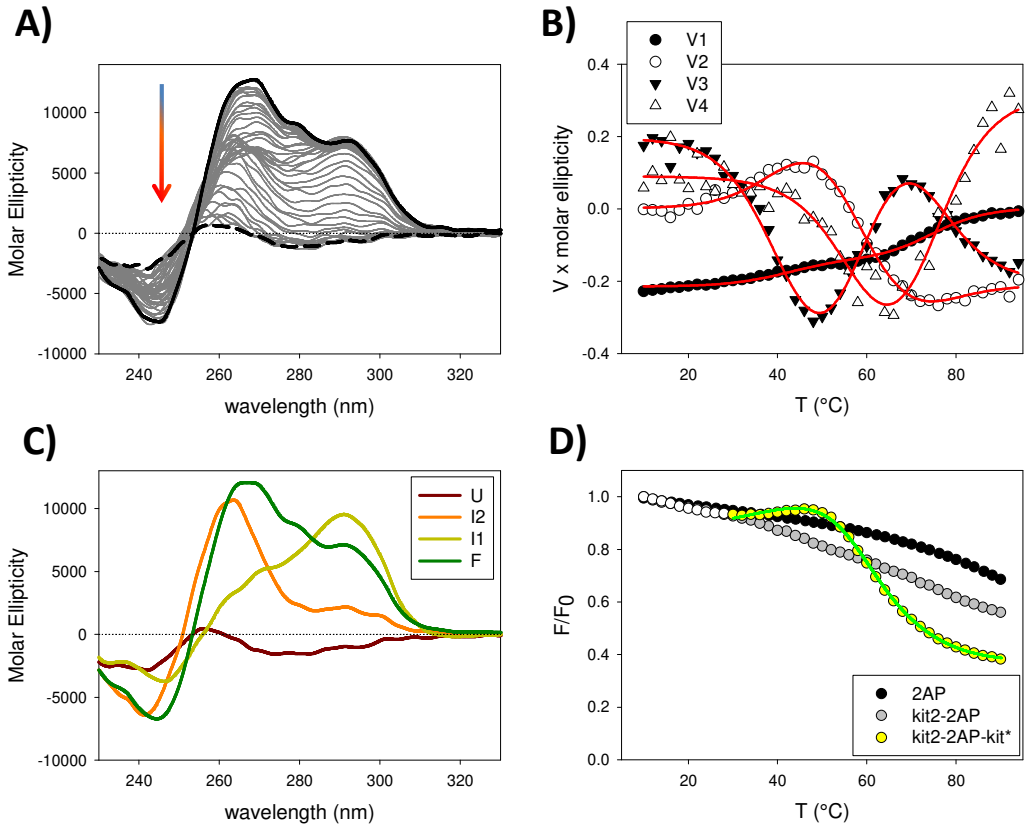


Figure 2. (A) CD spectra of kit2kit* in 150 mM KCl recorded at increasing temperature. Solid and dashed black lines refer to the spectra acquired at 10 and 94 °C, respectively (B) Significant V eigenvectors (V1-V4) fitted according to the "four states-three transitions" melting process model. (C) Derived CD spectra of the species contributing to the dichroic signal variations during the melting process. (D) Melting profile of 0.4 µM kit2-2AP-kit* previously folded in 150 mM KCl monitored following the emission signal at 370 nm (white and yellow dots). The green line represents the fitting of the experimental data (yellow dots) according to a “three transitions” unfolding model. Black and grey dots refer to the temperature dependence of the fluorescence signal of 2AP and kit2-2AP, respectively, in the same experimental conditions.

Conversely, in the same experimental conditions, the unfolding profile of kit2kit* was composed by multiple steps (Figure 2A). It is not easy to describe it by means of common protocols and for this reason we moved on the Singular Value Decomposition (SVD) analysis, an analytical tool very useful to characterize such complex events since it takes into account the optical contributions in the whole wavelengths range (see Supporting Information).[20-22]

The singular values contained in the S matrix as well as the U and V autocorrelation coefficients (Figure S5, Table S1, meaningful values are those above 0.75) highlighted that four main species in solution account for the whole dichroic signal variation. The significant V eigenvectors were fitted using several models and the "three transitions mechanism" provided the best description of the unfolding process of kit2kit* (Figure 2B).[15] The resulting fitting parameters (u, a, b and l from eq. (2) in Supporting Information) were used to derive the actual shapes of the CD spectra of each significant species in solution, as reported in Figure 2C: they nicely fit with those experimentally determined for kit2kit*, kit2 and kit* summarized in Figure 1D. Furthermore, SVD analysis provided the melting temperatures of each single species as well as their unfolding enthalpies (Table 1). The first species shows the same dichroic shape of the fully-folded sequence and it is characterized by the lowest melting temperature (42.5 °C). After the disruption of this form, the CD spectrum of kit2kit* well resembles the expected sum of the independent contributions of the parallel kit2 and the antiparallel kit* (Figure 2C, yellow line). A second thermal transition occurs at 58 °C. After this step, the dichroic signal indicates the presence in solution of only one parallel G4, in agreement with the loss of the antiparallel component (Figure 2C, orange line). Finally, the third and last transition takes place at 75 °C and it leads to the complete unfolding of the long oligonucleotide (Figure 2C, brown line). By comparing the structural rearrangement of kit2kit* with the data derived from the isolated G4 units, it clearly appears that the second and the third steps (T2 and T3 in Table 1) well correspond to the melting of the kit* and kit2, respectively. Conversely, the first one (T1) likely represents a novel conformational rearrangement, associated to a reduced enthalpy change. This initial transition perfectly fits with the disruption of a cross-talking interaction between kit2 and kit* G4s.

Providentially, in *c-KIT* promoter kit2 and kit* are separated by an adenine (A22): its substitution with the fluorescent analogue, 2-aminopurine (2AP) allowed us to further investigate such a potential interaction mode of kit2 and kit* within the longer sequence. Indeed, the use of this fluorescent probe pro-vides the advantage to selectively monitor localized conformational rearrangements with no or limited interference from the global structural movements.[23] Addition of 150 mM KCl to the fluorescent oligonucleotide induced an increment in the fluorescence emission (Figure S6). Since base stacking on a G-tetrad strongly quenches the fluorescence of 2AP, this suggests that in the folded oligonucleotide this adenine is not engaged in G4 interactions and that it is relatively free

to explore the solvent. Nevertheless, it is worth to mention that, accordingly to the reported kit2 structures,[5] the guanine at position 21 is not involved in G-tetrad formation thus, it can still partly quench the fluorescence signal of the close by 2AP.[24,25]

Subsequently, we melted the kit2kit* folded in KCl to follow the conformational changes of the environment surrounding the fluorescent base analogue. As reported in Figure 2D, by increasing the temperature, a further modest increment of the fluorescence signal is observed between 30 and 50 °C, followed by a remarkable decrement of the signal that reaches a minimum around 90 °C. This behavior is specific of the kit2-2AP-kit* since 2AP and kit2-2AP show only the expected fluorescence decrement related to the heating of the solution. Also in this case the best fitting of the experimental data relative to kit2kit* in the 24-90 °C temperature range was obtained by applying a "three transitions" model. The output provided melting temperatures perfectly overlapping those above reported (see Table S2). We attributed the first transition, associated to the increment of the fluorescence signal, to the break of the G4-G4 interaction: it increases the flexibility of the system and allows the fluorescent nucleotide to move away from G21 in the loop. During the second and third transitions, the sharp decrease in fluorescence emission is caused by the insertion of 2AP in an unfolded environment. Consistently, at the higher tested temperature the F/F_0 of kit2-2AP is lower in comparison to the free 2AP.

These findings support the hypothesis that the GA loop does not participate to the G4-G4 interaction. This was further sustained by acrylamide quenching experiments. They revealed a linear dependency of the kit2-2AP-kit* fluorescence to the quencher concentration (Figure S7), thus meaning that the inserted 2AP fluorophore experiences only one microenvironment thus excluding its stacking within the G4-G4 interface.[23,26] By the Stern-Volmer model we compared the accessibility of 2AP within the oligonucleotide to the free 2AP. The lower value of the extrapolated K_{sv} of kit2-2AP-kit* showed that in this case the fluorophore is more shielded compared to the free 2AP (Table S3). All these findings fully sustain that the adenine within the loop between does not participate to the G4-G4 interaction, but it is involved in a stacking interaction with the neighbor unpaired guanine. All our evidences can thus easily be summarized in the scheme reported in Figure 3.

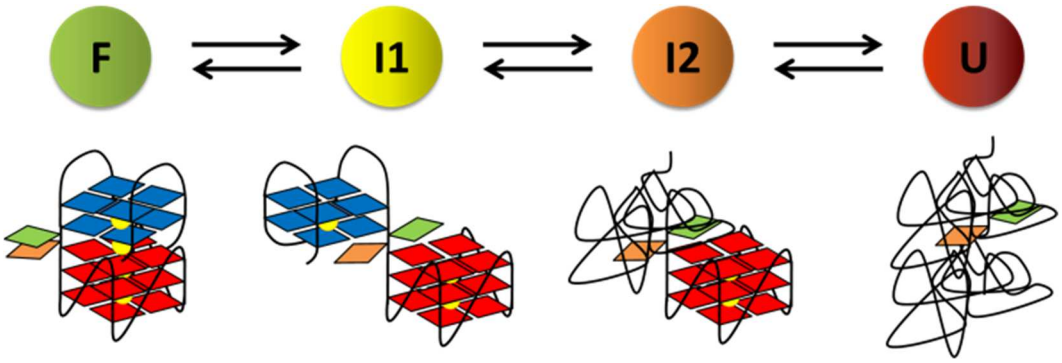


Figure 3. Graphical representation of the melting process of kit2kit*. Blue and red correspond to the kit* and kit2 G4 arrangements; in green and orange, A22 and G21 are highlighted, respectively.

It derives that one functional unit participating in the regulation of *c-KIT* expression might be represented by this novel G4-G4 module. Clearly, it is worth to keep in mind that in gene promoter this tandem G4 arrangement naturally competes with the duplex form. Nevertheless, from the data herein presented, we can expect that the higher order organization affects the equilibrium to a different extent in comparison to the isolated G4 modules. In order to assess this point, we monitor the duplex-G4 competition by SPR. In our assay, a biotinylated oligonucleotide corresponding to the cytosine-rich strand complementary to kit* sequence (biot-kit*-C) was immobilized on a streptavidin functionalized sensor-chip, a strategy that allows to avoid severe structural rearrangements on the chip surface. The G-rich strand (kit* or kit2kit*, previously folded in KCl) was then flushed at increasing concentrations, so the pairing process was monitored as variation of the SPR angle. From the sensor-grams reported in Figure 4, it is immediately evident that the efficiency of the surface hybridization of the short or the long sequences is completely different. Indeed, even if kit* has a lower molecular weight, it generates higher RU than kit2kit* at the same concentrations.

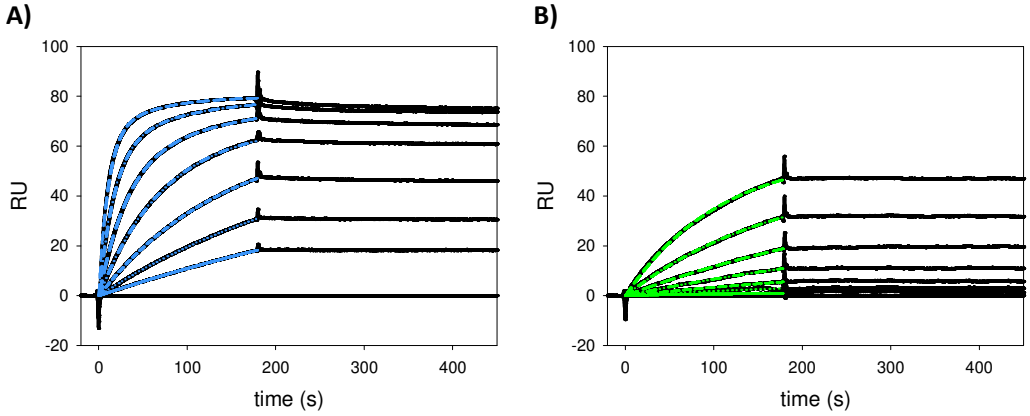


Figure 4. Differential pairing of 15-1000 nM kit* (PANEL A) and kit2kit* (PANEL B) to the complementary strand of kit* immobilized on a chip surface acquired by SPR kinetic analysis.

For data analysis we focused on the association step since the dissociation was almost negligible. This step was well fitted by a single exponential corresponding to a first-order Langmuir equation (eq. (4) in Supporting Information).[27-29] From each single sensorgram, we derived the apparent rate constants which were linearly related to the concentration of the flushing oligonucleotide thus allowing us to derive the association rate constant for the surface hybridization process (eq. (5) in Sup-porting Information and Figure S8). The results provide the association rate constant for the pairing of kit* (11.46·104 M-1s-1) that differs of about one order of magnitude in comparison to kit2kit* (0.91·104 M-1s-1). This is in line with all the evidences herein presented. Indeed, taking into account that the hybridization process requires the G4 unfolding, it derives that the G4-G4 interaction, which increases the overall stability of the fold-ed G-rich strand, negatively impacts on the double strand formation. Consistently, the same behaviour was detected when we immobilized the kit2 complementary strand (biot-kit2-C) and compared the pairing of kit2 to kit2kit* (Table S4). In this case, the reduced difference in the hybridization efficiency derives from the higher stability of the G4 arranged kit2 vs kit*.

To conclude, we provided evidences supporting an intra-molecular G4 folding for the sequence corresponding to the kit2kit* domain. It does not correspond to the sum of the individually structured kit2 and kit*, but it represents an unprecedented G4 form compatible with a G4-G4 tandem interaction. In addition to the unique structural significance of this new arrangement, our data indicate it as physiologically relevant since it efficiently shifts the equilibrium G4-duplex towards the tetrahelical form. Since other

oncogene promoters contain similar G4 repeats, (i.e. KRAS)[30] the control of these structural features might represent a shared mechanism of gene regulation worth to be further explored.

ASSOCIATED CONTENT

Supporting Information

Supporting Information contains: Experimental section; Figures S1-S8, Table S1-S4.

AUTHOR INFORMATION

Corresponding Author

* claudia.sissi@unipd.it

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

This work was founded by Universita' degli Studi di Padova (PRAT# CPDA147272/14, RR fellowship).

REFERENCES

1. Davis, J. T. *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 668-698.
2. Balasubramanian, S.; Hurley, L. H.; Neidle, S. *Nat. Rev. Drug Discov.* **2011**, 10, 261-275.
3. Rigo, R.; Palumbo, M.; Sissi, C. *Biochim. Biophys. Acta.* **2017**, *1861*, 1399-1413.
4. Lennartsson, J.; Ronnstrand, L. *Physiol. Rev.* **2012**, *92*, 1619-1649.
5. Kuryavyi, V.; Phan, A. T.; Patel, D. *J. Nucleic Acids Res.* **2010**, *38*, 6757-6773.
6. Hsui, S.D.; Varnai, P:; Bugaut, A.; Reszka, A. P:; Neidle, S.; Balasubramanian, S. *J. Nucleic Acids Res.* **2009**, *131*, 13399-13409.
7. Raiber, E. A.; Kranaster, R.; Lam, E.; Nikan, M.; Balasubramanian, S. *Nucleic Acids Res.* **2012**, *40*, 1499-1508.
8. Wei, D.; Parkinson, G. N.; Reszka, A. P: Neidle, S. *Nucleic Acids Res.* **2012**, *40*, 4691-4700.
9. Wei, D.; Husby, J.; Neidle, S. *Nucleic Acids Res.* **2015**, *43*, 629-644.
10. Paulo, A.; Francisco, A. P. *Curr. Med. Chem.* **2016**. *23*, 1-32.
11. Zorzan, E.; Da Ros, S.; Musetti, C.; Shahidian, L. Z.; Coelho, N. F.; Bonsembiante, F.; Letard, S.; Gelain, M. E.; Palumbo, M.; Dubreuil, P.; Giantin, M.; Sissi, C.; Dacasto, M. *Oncotarget* **2016**, *7*, 21658-21675.
12. Petraccone, L.; Trent, J. O.; Chaires, J. B. *J. Am. Chem. Soc.* **2008**, *130*, 16530-16530.
13. Palumbo, S. L.; Ebbinghaus, S. W.; Hurley, L. H. *J. Am. Chem. Soc.* 2009, 131, 10878–10891
14. Yamamoto, K.; Tojo, A.; Aoki, N.; Shibuya, M. *Jpn. J. Cancer Res.* **1993**, *84*, 1136-

1144.
15. Mergny, J. L.; Li, J.; Lacroix, L.; Amrane, S.; Chaires, J. B. *Nucleic Acids Res.* **2005**, *33*, e138.
16. Kocman, V.; Plavec, J. *Nat. Commun*. **2017**, *8*, 15355-15370.
17. Podbevsek, P.; Plavec, J. *Nucleic Acids Res.* **2016**, *44*, 917-925.
18. Gray, R. D.; Buscaglia, R.; Chaires, J. B. *J. Am. Chem. Soc.* **2012**, *134*, 16834-16844.
19. Da Ros, S.; Zorzan, E.; Giantin, M.; Zorro Shahidian, L.; Palumbo, M.; Dacasto, M.; Sissi, C. *PLoS One* **2014**, *9*, e103876.
20. Gray, R. D.; Chaires, J. B. *Curr. Protoc. Nucleic Acid Chem.* **2011**, Chapter 17, Unit17 4.
21. DeSa, R. J.; Matheson, I. B. *Methods Enzymol.* **2004**, *384*, 1-8.
22. Hendler, R. W.; Shrager, R. I. *J. Biochem. Biophys. Methods* **1994**, *28*, 1-33.
23. Gray, R. D.; Petraccone, L.; Buscaglia, R.; Chaires, J. B. *Methods Mol. Biol.* **2010**, *608*, 121-136.
24. Kawai, M.; Lee, M. J.; Evans, K. O.; Nordlund, T. M. *Journal of Fluorescence* **2001**, *11*, 23-32.
25. Narayanan, M.; Kodali, G.; Xing, Y.; Stanley, R. J. *The Journal of Physical Chemistry B* **2010**, 10573-1580.
26. Ballin, J. D.; Prevas, J. P.; Bharill, S.; Gryczynski, I.; Gryczynski, Z.; Wilson, G. M. *Biochemistry* **2008**, *47*, 7043-7052.
27. Gao, Y.; Wolf, L. K.; Georgiadis, R. M. *Nucleic Acids Res.* **2006**, *34*, 3370-3377.
28. Rapisarda, A.; Giamblanco, N.; Marletta, G. *J. Colloid Interface Sci.* **2017**, *487*, 141-148.
29. Halder, K.; Chowdhury, S. *Nucleic Acids Res.* **2005**, *33*, 4466-4474.
30. Kaiser, C. E.; Van Ert, N. A.; Agrawal, P.; Chawla, R.; Yang, D.; Hurley, L. H. *J. Am. Chem. Soc*. **2017**, *139*, 8522–8536.

Published – Rigo, R.; Sissi, C. *Biochemistry.* **2017**, 56 (33), 4309-4312

SUPPLEMENTARY INFORMATION

Characterization of G4-G4 crosstalk in *c-KIT* promoter region

Riccardo Rigo[1], Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Via Marzolo 5, 35131 Padova, Italy

Corresponding Author:

* claudia.sissi@unipd.it

Materials and Methods

Nucleic Acids

kit2kit*	5'-**CGGGCGGGCGCGAGGGAGGGGAGGCGAGGAGGGGCGTGGCCGGC**-3'
kit2	5'-**CGGGCGGGCGCGAGGGAGGGG**-3'
kit2-A	5'-**CGGGCGGGCGCGAGGGAGGGGA**-3'
kit*	5'-**GGCGAGGAGGGGCGTGGCCGGC**-3'
kit2-2AP-kit*	5'-**CGGGCGGGCGCGAGGGAGGGG-2AP-GGCGAGGAGGGGCGTGGCCGGC-**3'
kit2-2AP	5'-**CGGGCGGGCGCGAGGGAGGGG-2AP**-3'
biot-kit*-C	5'**-Biotin-GCCGGCCACGCCCCTCCTCGCC-**3'
biot-kit2-C	5'**-Biotin-CCCCTCCCTCGCGCCCGCCCG-**3'
SCR	5'**-CCTGCTTCTCGCCGAGCAATTGTCCAGGCGGATCCTCA-**3'
MW marker	5'**-GGATGTGAGTGTGAGTGTGAGG-**3'

Oligonucleotides were purchased from Eurogentec (Liège, Belgium) and were used without further purification. They were dissolved in 10 mM TRIS, pH 7.5, to prepare a 1 mM stock solution. Before use, each sample was heated at 95°C for 10 minutes in the required buffer and then slowly cooled down at room temperature to equilibrate the system.

Polyacrylamide gel electrophoresis

For the native PAGE, 200 ng of oligonucleotide were heated at 95 °C for 10 minutes in 10 mM TRIS, 150 mM KCl, pH 7.5 and let to cool down at room temperature overnight. Afterwards, samples were loaded on a native 15 % polyacrylamide (19:1 acrylamide: bisacrylamide) PAGE in 1X TBE (89 mM Tris, 89 mM boric acid and 2 mM EDTA). A scramble 38-residue long oligonucleotide (SCR) unable to assume any intramolecular folding was used as electrophoretic mobility marker.

For gel under denaturing conditions, before loading samples were added of in denaturing gel loading buffer (80% formamide, bromophenol blue), heated at 95°C for 10 minutes and immediately cool down in an ice-bath to avoid refolding processes. They were loaded

on a denaturing 20% polyacrylamide (7 M urea, acrylamide:bisacrylamide 19:1) PAGE in 1X TBE. Gels were stained with SYBR green II and the resolved bands were visualized on an image acquisition system (Geliance 600 Imaging system, Perkin-Elmer).

Thermal difference spectrum

The UV absorption spectra of the studied sequence were recorded at 95°C and at 25°C in 10 mM TRIS, 150 mM KCl, pH 7.5 using a Perkin-Elmer Lambda 20 Spectrophotometer. The spectrum acquired at 25°C was subtracted to the one at 95°C and the resulting thermal difference spectrum normalized to 1 at the maximal optical variation.

Circular Dichroism

Circular dichroism spectra were acquired on a JASCO J-810 spectropolarimeter equipped with a Peltier temperature controller. CD spectra were recorded from 230 nm to 320 nm, with the following parameters: scanning speed 100 nm/min; band width of 2 nm; data interval of 0.5 nm; response of 2 s. Measurements were performed using a 1 cm path length quartz cuvette and oligonucleotide concentration in cuvette ≈ 4 µM in 10 mM TRIS pH 7.5.

CD titration were performed at 25 °C by adding to the oligonucleotide solution increasing concentration of KCl. After each addition the system was leaved to equilibrate (2 h) before spectra acquisition.

CD melting studies were carried out between 10-94 °C in 150 mM KCl. The heating rate was 50°C/hour, each 2 °C the temperature was hold for 5 minutes and the corresponding CD spectrum was recorded.

After each acquisition, the contribution of the buffer was subtracted from the sample. Observed ellipticities (deg) were converted into molar ellipticity [Θ] = deg x cm2 x dmol-1 (Mol. Ellip.) calculated by the DNA residue concentration in solution determined by UV absorbance at 260 nm.

Data analysis of CD melting experiments

For melting profiles showing only one fully reversible melting transition, data analysis was performed by fitting the signal variation at a single wavelength according to a single-transition model based on van't Hoff formalism,[1]

$$y = \frac{u * e^{-\left(\frac{\Delta H}{RT} * \left(\left(\frac{T}{Tm}\right) - 1\right)\right)} + l}{e^{-\left(\frac{\Delta H}{RT} * \left(\left(\frac{T}{Tm}\right) - 1\right)\right)} + 1} \qquad \text{Eq.1}$$

where *u* and *l* are fitting parameters, *ΔH* is the enthalpy of the unfolding process, *R* is the ideal gas constant, *Tm* is the melting temperature of the folded oligonucleotide.

In order to describe unfolding processes with multiple transitions, such as the melting process of kit2kit* folding, SVD analysis was applied to the entire melting dataset, that correspond to a D matrix formed by the optical signals in the whole wavelengths range (rows) acquired at each single temperature (columns). This analytical tool splits the data matrix in three submatrices U, S, V, so that D = U x S x V. The singular values in S matrix and the autocorrelation coefficients of the U and V matrices give information about the number of species in solution that contribute to the signal variation.[2,3] For the U and V autocorrelation coefficients, significant values were considered those above 0.75.

Significant V eigenvectors in V matrix were globally fitted by applying different melting models and the best fitting was obtained by using a "three transitions" model,[1]

$$y = \frac{u * e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right) + \frac{\Delta H3}{RT} * \left(\left(\frac{T}{Tm3}\right) - 1\right)\right)} + b * e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right)\right)} + a * e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right)\right)} + l}{e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right) + \frac{\Delta H3}{RT} * \left(\left(\frac{T}{Tm3}\right) - 1\right)\right)} + e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right) + \frac{\Delta H2}{RT} * \left(\left(\frac{T}{Tm2}\right) - 1\right)\right)} + e^{-\left(\frac{\Delta H1}{RT} * \left(\left(\frac{T}{Tm1}\right) - 1\right)\right)} + 1} \qquad \text{Eq. 2}$$

where *u, a, b* and *l* are fitting parameters, *ΔH*1, *ΔH*2 and *ΔH*3 are the enthalpies associated to each melting step, *Tm*1, *Tm*2 and *Tm*3 are the melting temperatures. The fitting parameters formed the H matrix. By multiplying the H matrix for the U x S matrix, the actual shapes of the dichroic signals of species in solution were obtained.

Fluorescence spectroscopy

Fluorescence spectroscopy measurements were performed on kit2-2AP-kit* and kit2-2AP in which the adenine at position 22 was substituted with the 2-aminopurine fluorescent analogue (2AP). Measurements were performed on a JASCO-FP-6500 spectrofluorometer equipped with a Peltier Temperature controller. Emission spectra were recorded in the 320-460 nm range; with the excitation wavelength fixed at 305 nm. The scanning speed was 200 nm/min, the response 1 s and both the emission and the excitation band width were 3 nm. The fluorescently labeled oligonucleotides were diluted to a concentration of 0.4 µM in 10 mM TRIS, pH 7.5. The effect of addition 150 mM KCl on the fluorescence spectra was monitored over time at 25 °C.

Melting profiles of the sequences previously folded in 10 mM TRIS, 150 mM KCl, pH 7.5, were acquired in the 10-90°C range following the emission signal at 370 nm. Data points were recorded at each 2°C of temperature increment, after a 3 minutes of equilibration time. The melting profile of kit2-2AP-kit* was compared to those one of 2-AP alone and a kit2-2AP, as controls. For the longer sequence, data analysis was performed using a "three transitions" model, as reported for the CD melting (see Eq. 2).

Acrylamide quenching experiments were performed on the kit2-2AP-kit* previously folded in 150 mM KCl and on the free 2AP, used as control. The quenching efficiency was monitored at 370 nm, and plotted as function of the acrylamide concentration at 25°C. Stern-Volmer formalism was applied to obtain quenching parameters,[4,5]

$$\frac{F0}{F} = 1 + Ksv \cdot [acrylamide] \qquad \text{Eq. 3}$$

where F0 is the initial value of fluorescence and K_{sv} is the Stern-Volmer constant.

Surface Plasmon Resonance (SPR)

Surface plasmon resonance measurements were performed using a Biocore X100 set up with streptavidin-coated sensor chips prepared for use by conditioning it with 1 min injections of 1 M NaCl, 50 mM NaOH in 50% isopropanol and finally extensively washed with a 0.22 µm filtered buffer (10 mM Tris, 150 mM KCl, 0.025% P20). Previously annealed biot-kit*-C or biot-kit2-C were then immobilized on flow-cell 2 of the chip surface by

adding a 50 nM DNA solution at a 1 µl min−1 flow rate until a 80 RU response was obtained. Flow-cell 1 was left blank as control. Sensorgrams were acquired using serial dilutions of kit* or kit2kit* (0–1000 nM) annealed in the running buffer. The G-rich sequences were injected at a 25 µl min−1 flow rate for 180 seconds, thereafter a dissociation time of 300 seconds were maintained. After each run, a 30 seconds regeneration step was performed with 1M NaCl and 50 mM NaOH followed by a 300 seconds stabilization period in running buffer.

In the applied experimental condition, dissociation processes were negligible, so it was studied only the association step. A mono-exponential kinetics was applied to the association curves in order to obtain the apparent rate constant of the process,[6,7]

$$RU = a \cdot (1 - e^{-k_{obs} \cdot t}) \qquad \text{Eq. 4}$$

where RU is the intensity of the signal at time t, a is a fitting parameter, k_{obs} is the apparent rate constant. It is dependent on the concentration of analyte in solution, according to this relationship:

$$k_{obs} = C \cdot ka \qquad \text{Eq. 5}$$

where C is the concentration of analyte in solution and ka is association rate constant of the hybridization process.

We plotted the k_{obs} as function of the analyte concentration, obtaining a linear relationship. This plot was used to derive the association rate constants by means of a simple linear regression.

For kit* hybridization on chip surface data analyses was limited to data points in 0-500 nM concentration range.

SELECTED REFERENCES

1. Gray, R. D.; Buscaglia, R.; Chaires, J. B. *J Am Chem Soc* **2012**, *134*, 16834.
2. DeSa, R. J.; Matheson, I. B. *Methods Enzymol* **2004**, *384*, 1.
3. Hendler, R. W.; Shrager, R. I. *J Biochem Biophys Methods* **1994**, *28*, 1.
4. Ballin, J. D.; Prevas, J. P.; Bharill, S.; Gryczynski, I.; Gryczynski, Z.; Wilson, G. M. *Biochemistry* **2008**, *47*, 7043.
5. Gray, R. D.; Petraccone, L.; Buscaglia, R.; Chaires, J. B. *Methods Mol Biol* **2010**, *608*, 121.
6. Halder, K.; Chowdhury, S. *Nucleic Acids Res* **2005**, *33*, 4466.
7. Rapisarda, A.; Giamblanco, N.; Marletta, G. *J Colloid Interface Sci* **2017**, *487*, 141.

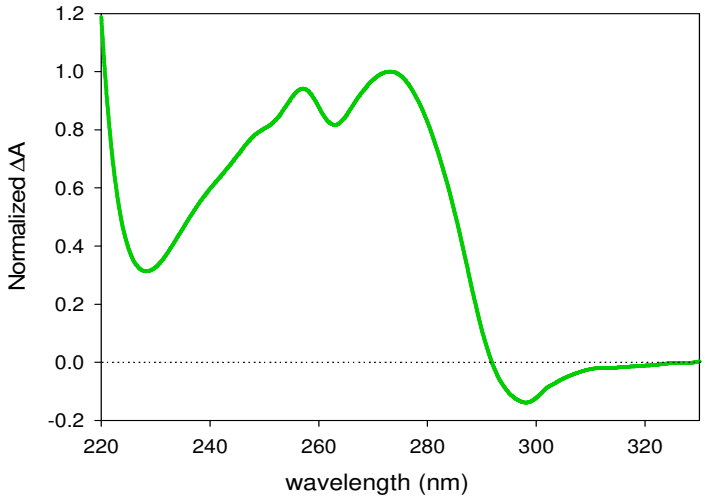


Figure S1. Thermal difference spectrum of kit2kit* in 10 mM TRIS, 150 mM KCl, pH 7.5

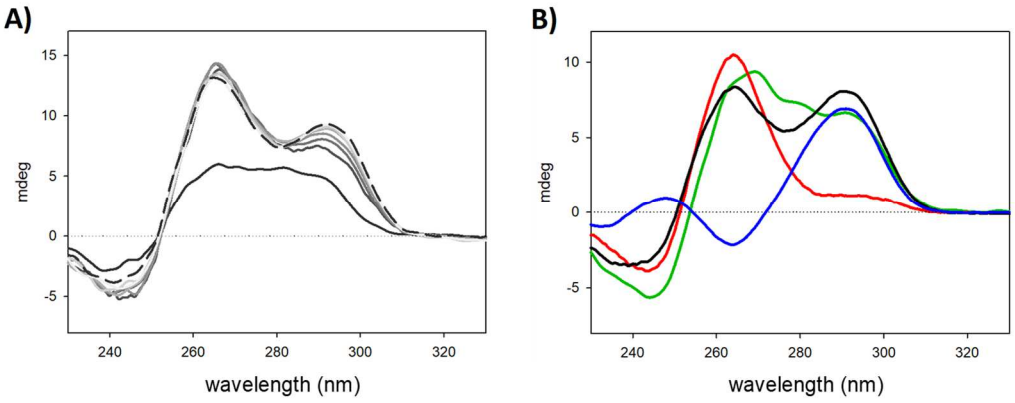


Figure S2. (A) CD titration with KCl of an equimolar mixture (2.8 µM) of kit2-A and kit* in 10 mM TRIS, pH 7.5, 25 °C. Solid and dashed black lines refer to the sequence in the absence and in the presence of an excess of metal ion, respectively. (B) CD spectra of 2 µM kit* (blue line), kit2-A (red line) and kit2kit* (green line) acquired in 10 mM TRIS,150 mM KCl pH 7.5, 25 °C. Black line corresponds to the spectra acquired by mixing equimolar concentration of kit* and kit2-A in 10 mM TRIS,150 mM KCl pH 7.5, 25 °C.

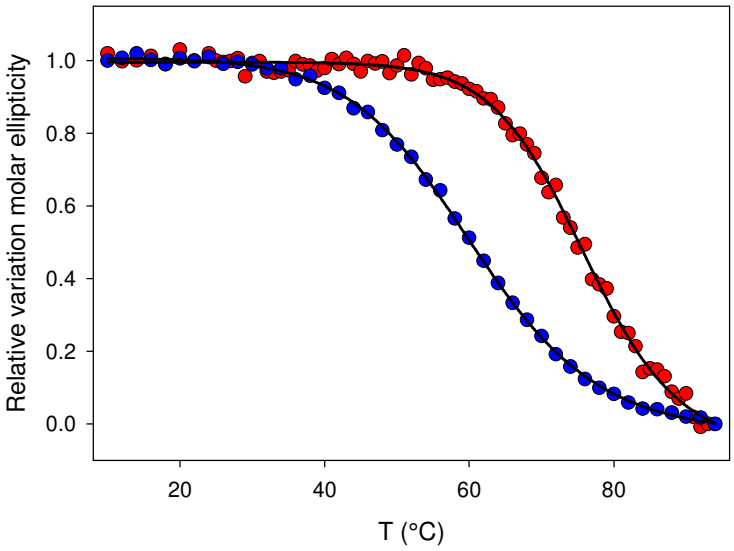


Figure S3. Melting profile of kit* (blue dots) and kit2-A (red dots) in 150 mM KCl, 10 mM TRIS pH 7.5. Data were fitted accordingly to a single-transition model based on van't Hoff formalism (black lines).

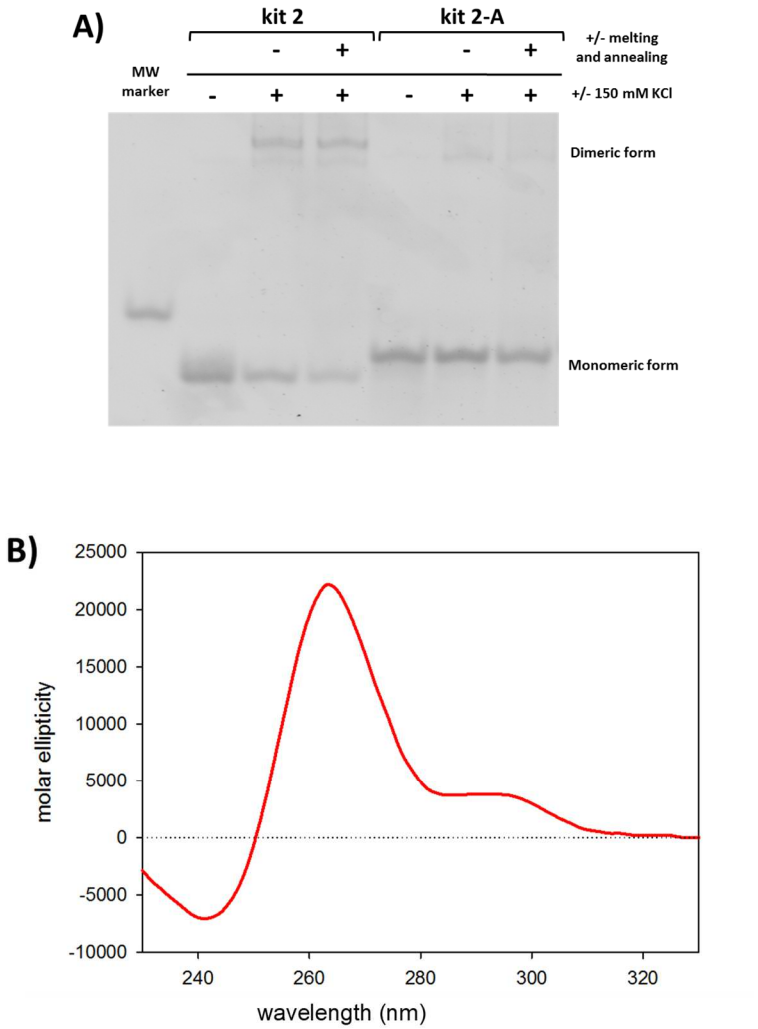


Figure S4. (A) Electrophoretic pattern of kit2 and kit2-A in absence/presence of 150 mM KCl with/without annealing resolved on 15% acrylamide native PAGE in 1x TBE using a 22-residues MW marker. (B) CD spectrum of kit2-A acquired in 150 mM KCl, 10 mM TRIS pH 7.5 at 25°C.

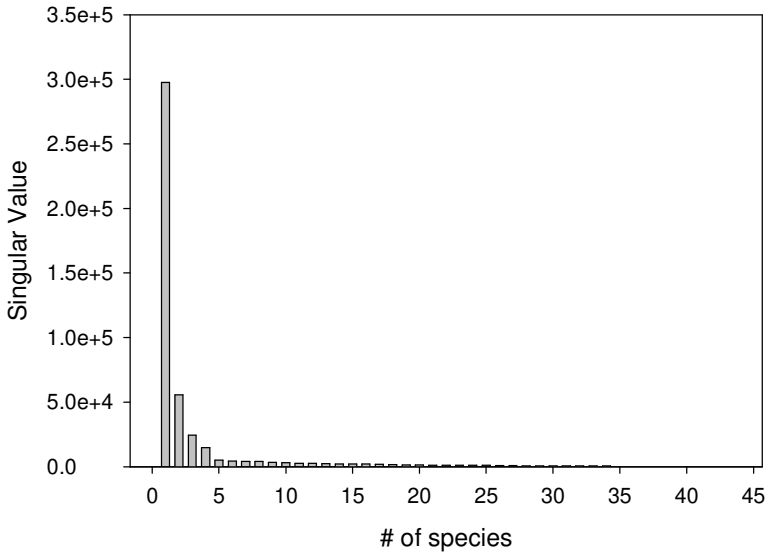


Figure S5. S matrix values derived from SVD analysis of the CD spectra of kit2kit* acquired at increasing temperature. Data indicate the relevance of species in solution participating to the overall dichroic signal variations.

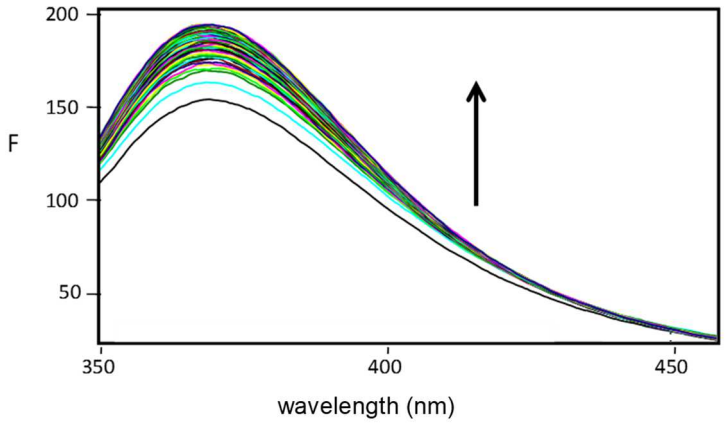


Figure S6. Fluorescence spectra of kit2-2AP-kit* acquired after the addition KCl up to 150 mM concentration. The arrow indicates the direction of spectra change.

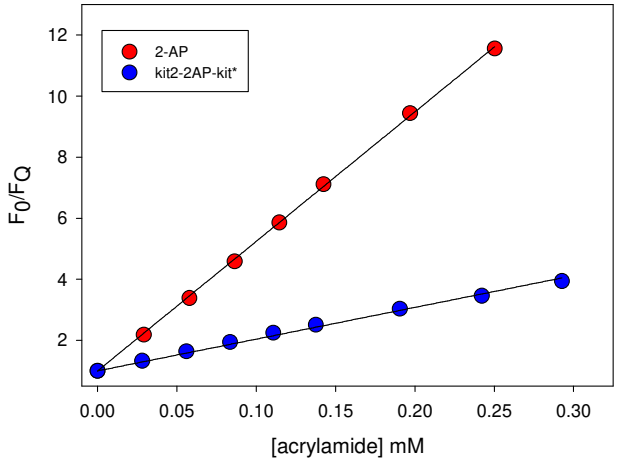


Figure S7. Fluorescence variation recorded at 370 nm of free 2-AP and kit2-2AP-kit* upon increasing concentration of acrylamide. Data were fitted according to the Stern-Volmer equation.

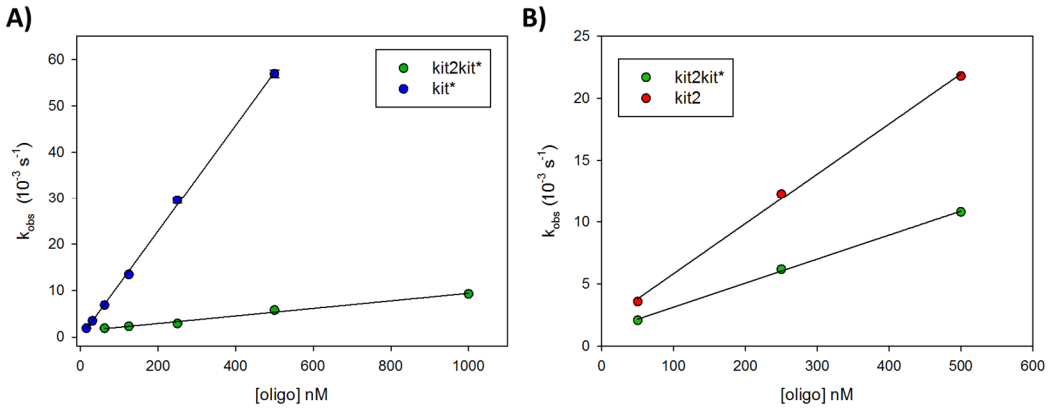


Figure S8. Apparent rate constants values derived from the SPR association phase (corresponding to the hybridization of kit*, kit2 or kit2kit* to biot-kit*-C (Panel A) or biot-kit2-C (Panel B)) vs flushing oligonucleotides concentration.

Table S1. U and V autocorrelation coefficients derived from SVD analysis indicating the relevance of species in solution participating to the signal variations. The not significant value is highlighted in red.

	Parameters	#	Values
Autocorrelation	U	1	9.95E-01
		2	9.91E-01
		3	9.86E-01
		4	9.87E-01
		5	8.90E-01
	V	1	9.73E-01
		2	9.69E-01
		3	9.42E-01
		4	8.60E-01
		5	-2.42E-03

Table S2. Thermodynamic parameters of the unfolding process of kit2-2AP-kit* in 150 mM KCl.

Transition	Tm (°C)	ΔH (kcal/mol)
T1	42.5 ± 1.0	-16.0 ± 3.0
T2	58.0 ± 1.2	-37.0 ± 4.1
T3	75.3 ± 2.5	-38.0 ± 10.81

Table S3. Summary of Stern-Volmer constants derived according eq. (3).

Transition	K_{sv} (M^{-1})
2-AP	42.43 ± 0.17
kit2-2AP-kit*	10.40 ± 0.16

Table S4. Summary of association rate constants for the hybridization processes of kit*, kit2 or kit2kit* on immobilized biot-kit*-C or biot-kit2-C.

Immobilized sequence	Analyte in solution	k_a($M^{-1}s^{-1}$)
biot-kit*-C	kit*	$(11.46 \pm 0.16) \cdot 10^4$
	kit2kit*	$(0.91 \pm 0.12) \cdot 10^4$
biot-kit2-C	kit2	$(4.03 \pm 0.16) \cdot 10^4$
	kit2kit*	$(1.93 \pm 0.06) \cdot 10^4$

Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter

Maria Laura Greco[1], Anita Kotar[2], Riccardo Rigo[1], Cristofari Camilla[1], Janez Plavec[2,3,4,*] and Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, v. Marzolo 5, Padova, 35131, ITALY

[2] Slovenian NMR Center, National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

[3] EN-FIST Center of Excellence, Trg OF 13, 1000 Ljubljana, Slovenia

[4] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Večna pot 113, Ljubljana

* To whom correspondence should be addressed. Tel: +39-049 8275711; Fax: +39-049 8275366; e-mail: claudia.sissi@unipd.it; janez.plavec@ki.si

ABSTRACT

EGFR is an oncogene which codifies for a tyrosine kinase receptor that represents an important target for anticancer therapy. Indeed, several human cancers showed an upregulation of the activity of this protein. The promoter of this gene contains some G-rich domains, thus representing a yet unexplored point of intervention to potentially silence this gene. Here we explore the conformational equilibria of a 30-nt long sequence located at position -272 (EGFR-272). By merging spectroscopic and electrophoretic analysis performed on the wild type sequence as well as on a wide panel of related mutants we were able to prove that in potassium ion containing solution this sequence folds into two main G-quadruplex structures one parallel and one hybrid. They show comparable thermal stabilities and affinities for the metal ion and, indeed, they are always co-present in solution. The folding process is driven by a hairpin occurring in the domain corresponding to the terminal loop which works as important stabilizing element for both identified G-quadruplex arrangements.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

INTRODUCTION

The cell surface receptor Epidermal Growth Factor Receptor (EGFR) is an important factor in the pathogenesis and progression of several human cancers (1). Along with ErbB2/HER-2, ErbB3/HER-3 and ErbB4/HER-4, it belongs to the ErbB family of protein kinase receptors. Up to seven physiological ligands (i. e. EGF, TGF-α, etc) have been identified that are able to activate EGFR by driving its homo- or hetero-dimerization. Physiologically, this activity is required in normal cell growth and proliferation comprising the maintenance of normal intestinal functions and homeostasis (2). However, EGFR overexpression or mutations that constitutively activate this receptor are known oncogenic drivers (3). Relevant examples of cancers connected to an aberrant activation of EGFR are non-small cell lung cancer, breast cancer and glioblastoma (4). Currently available therapeutic agents used to counteract the upregulation of EGFR are tyrosine kinases inhibitors and humanized monoclonal antibodies against the receptor extracellular domain. Both treatments are designed to switch off the kinase activity and, consequently, the signal transduction. Unfortunately, their efficacy is severely impaired by either intrinsic or acquired resistance, often deriving from a selection of pre-existing sub-clones (5). Several mechanisms can contribute to this phenomenon i. e. amplification of alternative pathways to activate common downstream factors able to promote cell proliferation or the expression of mutated forms of EGFR. Among them, worth of mention are the gatekeeper T790M mutation which increases the affinity of the receptor for ATP in its binding pocket (6).

The gene that encodes for EGFR is located on chromosome 7p12-13. Using in silico analytical tools it was discovered that sequences in EGFR gene at positions -37 and -272 from the transcriptional start site (TSS) potentially form G-quadruplex (G4) structures (7,8). G4s are four-stranded helical structures that can be formed by single-stranded guanine-rich DNA (and RNA) oligonucleotides. G4s arise from Hoogsteen hydrogen bonding of four guanines arranged in a planar G-quartet (9). Stacking of two or more G-quartets leads to formation of a G4 that is further stabilized by monovalent cations. The considerable structural diversity is characteristic for DNA G4s. It was shown that factors such as primary sequence, especially number of G-tracts and their length, number of assembling DNA molecules, length of loops, orientation of strands, type of cations and other external factors importantly influence the structure of G4 (10-15). G4s can adopt

parallel, antiparallel and hybrid (3+1) topologies characterized by different orientation of the four strands. It was demonstrated that proteins like DNA and RNA helicases can selectivity recognize the topology of G4 (16,17). The transition between different topologies could be induced by changing the type of cations, adding different co-solutes and crowding agents (11,18,19). Loops that connect guanine residues involved in G-quartets have an important role in overall folding and stability of G4s (12). Three basic types of loops are characteristic for G4s; propeller, lateral and diagonal loops (20). Orientation of loop depends on number and nature of nucleotides in loops as well strand orientation and number of G-quartets they traverse (12). Some long(er) loops have been found to adopt a well-defined structure. It has been shown that hairpin-like loop structures increase thermodynamic stability of G4s (12). It is noteworthy, that G4s are not the only structures that can be formed by G-rich sequences. According to recent reports they can adopt other noncanonical structures such as G-hairpin and AGCGA-quadruplexes (21-23).

The formation of G4 structures in human cells has been demonstrated clearly (24-29). G-rich regions are found in telomeric regions. In addition, bioinformatic genome analysis showed that they frequently cluster upstream of the TSS of many oncogenes (30). Compared to telomeres, G4-forming sequences found in promotor regions are more diverse with varying number and length of G-tracts and intersecting residues resulting in potential formation of multiple G4s. It was demonstrated that formation of G4 in promoter regions can be involved in regulation of gene transcription. Investigations into structural features of G4s enable structure-based design of ligands that would bind to and stabilize G4 structures. Indeed, in promotor regions, G4 stabilization by small-molecule ligands frequently results in the suppression of gene expression (31-34). The silencing of EGFR transcription by promoting G4 formation in its promotor could thus represent a powerful complementary therapeutic strategy to the currently available treatments.

With the aim to explore the formation of G4 in the promotor region of the EGFR gene, we focused on the 30-nucleotide sequence named EGFR-272 d[GGGGACCGGGTCCAGAGGGGCAGTGCTGGG] that starts at positions -272 from the TSS. With the use of spectroscopic and electrophoretic techniques we describe its structural features in ionic conditions comparable to those found in the intracellular environment. The experimental data clearly established formation of G4 with additional Watson-Crick G-C base pairs within the loop regions. The effect of individual guanine and cytosine

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

residues on the folding of EGFR-272 in the presence of K+ ions has been tested rigorously with several spectroscopic methods including NMR. We identified two main G4 forms of EGFR-272, a kinetically favoured (3+1) hybrid and a slowly forming parallel one, both comprised of three stacked G-quartets.

MATERIAL AND METHODS

Oligonucleotides

Table 1. Sequences of oligonucleotides used in this work. In bold, mutated residues.

Oligonucleotide	Sequence											
				5		10		15		20		25		30	
EGFR-272		5'-	GGGG	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
EGFR-272-p	5'-	**phos**	GGGG	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
ΔG1		5'-	**/**GGG	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
G1T		5'-	**T**GGG	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
G2T		5'-	G**T**GG	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
G3T		5'-	GG**T**G	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
G4T		5'-	GGG**T**	ACC	GGG	TCC	AGA	GGGG	CAG	TGCT	GGG	-3'
G17T		5'-	GGGG	ACC	GGG	TCC	AGA	**T**GGG	CAG	TGCT	GGG	-3'
G18T		5'-	GGGG	ACC	GGG	TCC	AGA	G**T**GG	CAG	TGCT	GGG	-3'
G19T		5'-	GGGG	ACC	GGG	TCC	AGA	GG**T**G	CAG	TGCT	GGG	-3'
G20T		5'-	GGGG	ACC	GGG	TCC	AGA	GGG**T**	CAG	TGCT	GGG	-3'
ΔG1-G20T		5'-	**/**GGG	ACC	GGG	TCC	AGA	GGG**T**	CAG	TGCT	GGG	-3'
G1,20T		5'-	**T**GGG	ACC	GGG	TCC	AGA	GGG**T**	CAG	TGCT	GGG	-3'
G4,17T		5'-	GGG**T**	ACC	GGG	TCC	AGA	**T**GGG	CAG	TGCT	GGG	-3'
G4,18T		5'-	GGG**T**	ACC	GGG	TCC	AGA	G**T**GG	CAG	TGCT	GGG	-3'
G4,19T		5'-	GGG**T**	ACC	GGG	TCC	AGA	GG**T**G	CAG	TGCT	GGG	-3'
G4,20T		5'-	GGG**T**	ACC	GGG	TCC	AGA	GGG**T**	CAG	TGCT	GGG	-3'
C12,13T		5'-	GGGG	ACC	GGG	T**TT**	AGA	GGGG	CAG	TGCT	GGG	-3'
G25T-C26T		5'-	GGGG	ACC	GGG	TCC	AGA	GGGG	CAG	T**TT**T	GGG	-3'
ΔT27		5'-	GGGG	ACC	GGG	TCC	AGA	GGGG	CAG	TGC**/**	GGG	-3'
M_{22}		5'-	GGAT	GTG	AGT	GTG	AGT	GTGA	GG			-3'
M_{30}		5'-	GTTG	ACC	GTG	TCC	AGA	GTGG	CAG	TGCT	GGG	-3'

Oligonucleotides (Table 1) were purchased from Metabion International AG (German) and resuspended in milliQ water and then purified by electrophoretic technique (EGRF-272:

5'-GGG GAC CGG GTC CAG AGG GGC AGT GCT GGG-3'. The EGFR-272, ΔG1, G4T, G4,17T and G4,20T were also synthesized on K&A Laborgeraete GbR DNA/RNA Synthesizer. In all cases, the standard phosphoramidite chemistry was used. Deprotection and cleavage from the solid support was done with the use of aqueous ammonia at 55 °C for 12 h. The crude oligonucleotides were then purified by RP-HPLC and desalted, before use.

Electrophoretic Mobility Shift Assay (EMSA)

32P end-labelled single-stranded oligonucleotides were obtained by incubating the oligonucleotides with T4 polynucleotide Kinase (M-Medical S.r.l., Italy) and [γ-32P] ATP (Perkin Elmer S.p.a., Italy) for 30 min at 37 °C. The enzyme was then removed by extraction with phenol/chloroform/isoamylic alcohol (25:24:1). A mixture of purified labelled and unlabelled oligonucleotides (total final concentration 1 μM) was heated to 95 °C for 5 min in 10 mM Tris, 1 mM EDTA, pH 8.0 buffer at increasing KCl concentrations and let to cool overnight at room temperature. The folding of the starting material was monitored by native 20% polyacrylamide gel electrophoresis in 0.5X TBE (44.5 mM Tris base, 44.5 mM boric acid and 1 mM Na2EDTA) added of 10 mM KCl. Resolved bands were visualized and quantified on a Phosphor Imager (STORM 840, Pharmacia Biotech Amersham).

Thermal differential spectrum (TDS)

The thermal difference spectrum was obtained by subtracting of the oligonucleotide UV-spectra at 25 °C from the one recorded at 95 °C (below and above the oligo melting temperature respectively). The experiments were performed in 10 mM NaCacodilate, 150 mM KCl pH 7.0. The resulting thermal difference spectra have been normalized to the value of 1 at the maximal intensity (35). Before the thermal difference spectra, the CD signal of the oligo in the same buffer has been recorded.

Circular dichroism (CD) measurements

Circular dichroism spectra were acquired on a Jasco J 810 spectropolarimeter equipped with a Peltier temperature controller using 10 mm path length cells. Before data

acquisition, oligo solutions (ca. 4 µM in 10 mM Tris, pH 7.5) were heated at 95°C for 5 min and cooled overnight at room temperature. The reported spectrum of each sample represents the average of 3 scans recorded with 1-nm step resolution. For kinetic analyses, a single CD spectrum was acquired every 2 min. Thermal denaturation experiments were performed by heating the sample by 2 °C and allowing sample equilibration before spectra acquisition. Observed ellipticities were converted to mean residue ellipticity [Θ] = deg x cm2 x dmol-1 (Mol. Ellip.).

SVD analysis

Multiple wavelength CD experiments were analyzed using Singular Value Decomposition (SVD) analysis (36,37). The entire dataset forms a matrix, called D matrix, in which each row corresponds to a single wavelength and each column refers to an acquisition time. The applied analytical tool broke up the D matrix into three different submatrices according to the relation: D = U x S x V; where S matrix keeps information about the importance of every species contributing to the dichroic signal; U matrix contains information related to the spectral shapes of the significant species; V matrix indicates how the spectral changes occur over time. The S values and U and V autocorrelation coefficients allow determining the number of species that significantly contribute to the dichroic changes, thus selecting the significant V eigenvectors.
For kinetic CD experiment the significant V eigenvectors were globally fitted by applying different kinetic models and the best fitting was obtained using a first order mono-exponential kinetic model,

$$\theta_{t,\lambda} = \theta_{\infty,\lambda} + \theta_{\lambda} \cdot e^{(-k \cdot t)} \qquad \text{Eq. 1}$$

where θ t, λ is the value of signal at time t, $\theta \infty, \lambda$ is the final value of the signal, $\theta \lambda$ is amplitude factor for the exponential, k is the kinetic constant. The fitting parameters formed the so-called H matrix that allowed us to determine the spectral shapes of the species contributing to the dichroic signal.

Dimethyl sulfate (DMS) footprinting

For each reaction, 150000 cpm; ca. 300 ng of ^{32}P-labeled DNA were annealed 10 mM Tris, 1 mM EDTA, pH 8.0 in the presence/absence of 200 mM KCl. The samples were loaded onto a 20% polyacrylamide gel in 1xTBE and the solved bands were extracted by crushed and soak in the same buffer. The recovered DNA solutions were added of 0.4% DMS (25 µL EtOH, 5 µL DMS, 20 µL milliQ water) and 1 µM ctDNA (final concentration) in 50 µl total volume. After 5 min incubation at r.t. the reaction was stopped with 3.5 µL of β-mercaptoethanol and 10 µL of 40% glycerol. DNA samples were ethanol precipitated and further incubated for 30 min at 90 °C in 100 µl of diluted piperidine (1:10 in milliQ water). Finally, the samples were dried, washed two times with 20 µL of milliQ water and loaded on a 20% denaturing polyacrylamide sequencing gel along with Maxam and Gilbert purine marker.

Preparation of NMR samples

Oligonucleotides for NMR samples were dissolved in H2O with 10% of 2H2O and titrated with aliquots 3 M KCl solutions to a finale concentration ranging from 50 to 200 mM K+ ions. pH values of samples were adjusted to 7.0 with the 10 mM potassium phosphate buffer and to 8.0 with the 10 mM Tris buffer. Strand concentration in the samples was from 2.6 µM to 0.4 mM and was determined by UV absorption at 260 nm using UV/VIS Spectrophotometer Varian CARY-100 BIO UV-VIS.

NMR spectroscopy

NMR spectra were recorded on Agilent (Varian) NMR System 300, 600 and 800 MHz spectrometers at 25 °C in 90%/10% H2O/2H2O. 1H NMR spectra were recorded with the use of the DPFGSE solvent suppression method. NMR spectra were processed and analysed by using VNMRJ 4.2 (Varian Inc.) and the Sparky (UCSF) software. All NMR spectra were referenced to the TMSP.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

RESULTS

EGFR-272 shows distinct conformational features in the presence/absence of KCl

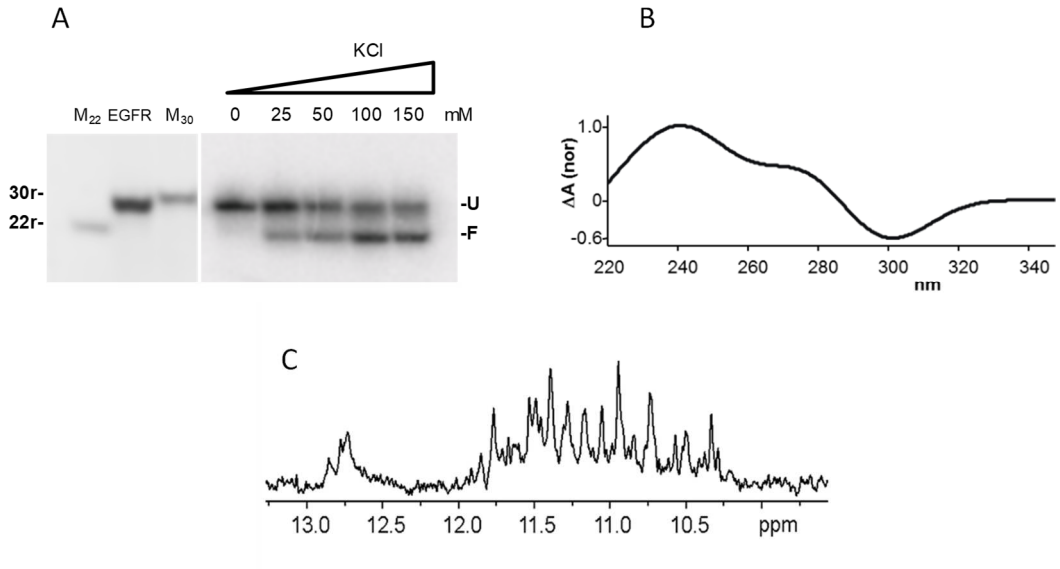


Figure 1. (A) EMSA of EGFR-272 annealed in the presence of increasing KCl concentrations. Lane 0 and EGFR refer to the oligonucleotide annealed in the absence of the metal ion; lanes M22 and M30 refer to random sequences 22 and 30 residues long; U and F to the unfolded and folded form of EGFR-272, respectively. (B): TDS of 4 μM EGFR-272 previously annealed in 150 mM KCl. (C): Imino region of 1H NMR spectrum of 0.1 mM EGFR-272 recorded at 100 mM KCl, pH 7.0, 25°C on a 800 MHz spectrometer.

To assess the potential of EGFR-272 to fold into a defined G4 secondary structure, we investigated the effect of KCl as a G4 stabilizer/inducer. When loaded on a native polyacrylamide gel, EGFR-272 runs essentially as a single band comprised between a 22 and a 30-mers (Figure 1A). Conversely, samples containing increasing concentrations of KCl showed a second band characterized by a higher electrophoretic mobility. Due to its reduced hydrodynamic volume, this form can be easily referred to an intramolecular folded form. Interestingly, a TDS analysis on the KCl-containing sample showed a negative pick at about 295-300 nm and of two positive bands at about 245 and 270 nm (Figure 1B) (35). These evidences sustain the folding of EGFR-272 into a G4 conformation in KCl concentrations (150 mM) comparable to those find in the intracellular environment. In full agreement, 1H NMR spectrum of EGFR-272 revealed signals between δ 10.35 and

12.00 ppm which are characteristic of imino protons of guanine residues involved in G-quartets as building blocks of G4 structures (Figure 1C). Additionally, the signals observed between δ 12.70 and 12.90 ppm indicated formation of G-C base pairs (38,39).

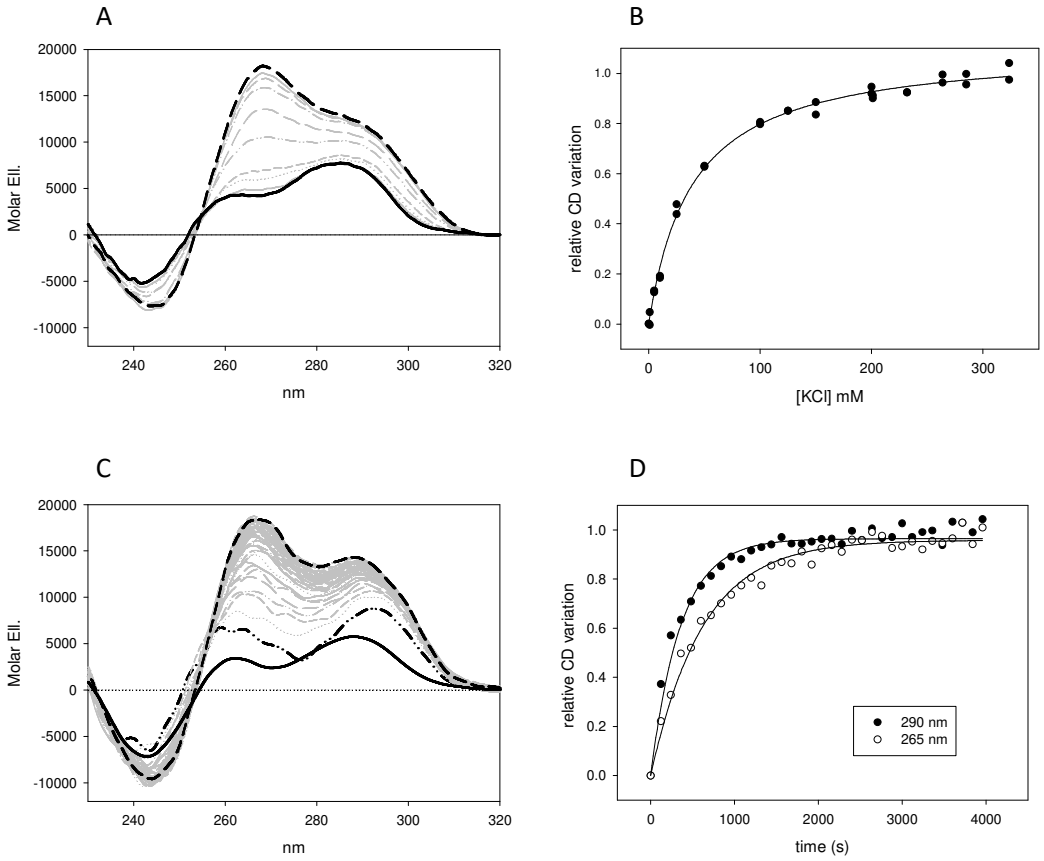


Figure 2. CD spectra of 4 μM EGFR-272 acquired in 10 mM Tris, pH 7.5, 25 °C in the absence (solid black line) and in the presence of increasing concentrations of KCl (dashed black line refers to 200 mM KCl) (A) and corresponding relative variation of the dichroic signal recorder at 265 nm as a function of metal ion concentration (B). In (C) the time-dependent variation of CD spectrum of 4 μM EGFR-272 in 10 mM Tris, pH 7.5, 25 °C upon addition of KCl is reported (1 acquisition/120 sec). Black line refers to the sample in the absence of KCl, dotted-dashed line to the one acquired immediately after addition of 200 mM KCl, dotted line to the one after equilibration. The derived relative variations of the dichroic signal recorder at 265 nm and 290 nm as a function of incubation time are reported in (D).

More than double number of imino signals in ^{1}H NMR spectrum than expected for a three G-quartet G4 indicated the formation of multiple G4 forms. This is not surprising since EGFR-272 sequence contains four guanine (G) tracts in which the guanine residues are

repeated 4-3-4-3 times, respectively, that can differently pair thus producing variable G4 arrangements. To get better insights on EGFR-272 conformational features, we performed circular dichroism studies (Figure 2). The CD spectrum of the oligonucleotide in the absence of KCl presents three main bands thus indicating the occurrence of a pre-folded state: this would actually explain the slightly higher electrophoretic mobility rate of EGFR-272 in comparison to a random 30-mer. In line, a theoretical prediction indicates the possible formation of different hairpins with melting temperatures comprised between 35 and 39 °C and repeated melting/annealing cycles monitored by recording the CD signal at 260 nm showed a reversible melting process within this temperature range although with a not well-defined thermal transition (Figure S1).

Addition of the monovalent cation extensively altered the oligonucleotide CD spectrum according to a process that, in agreement with PAGE data, was associated to a metal ion half maximal effective concentration EC50 = 38 ± 2 μM (Figure 2A and 2B). The observed spectral variations were not extremely fast and required about 30 min to reach equilibrium (Figure 2C and 2D). In these conditions, the metal ion induced a positive signal at about 265 nm and a negative one close to 240 nm that is reminiscent of a parallel G4. However, a significant shoulder at 290 nm is preserved and, even upon careful equilibration of the sample after each KCl addition, not well resolved isodichroic points are present. Noteworthy, the formation rate of the 290 nm contribution was much faster than the one at 265 nm (Figure 2C and 2D). This picture anticipates the formation of at least two G4 conformations comprising both parallel and antiparallel components.

This model was further supported by SVD (singular value decomposition) analysis of the oligonucleotide CD spectra acquired at different incubation time after addition of 200 mM KCl. As anticipated, the very first step was extremely fast, and this prevents us to kinetically analyse it according to the applied protocol. Thus, we moved to consider the slower part that comprises only spectra acquired after the addition of the metal ion thus removing from the analysis the contribution of the metal-free oligonucleotide (spectra ranging from the dashed-dotted line up to the dotted one in Figure 2C).

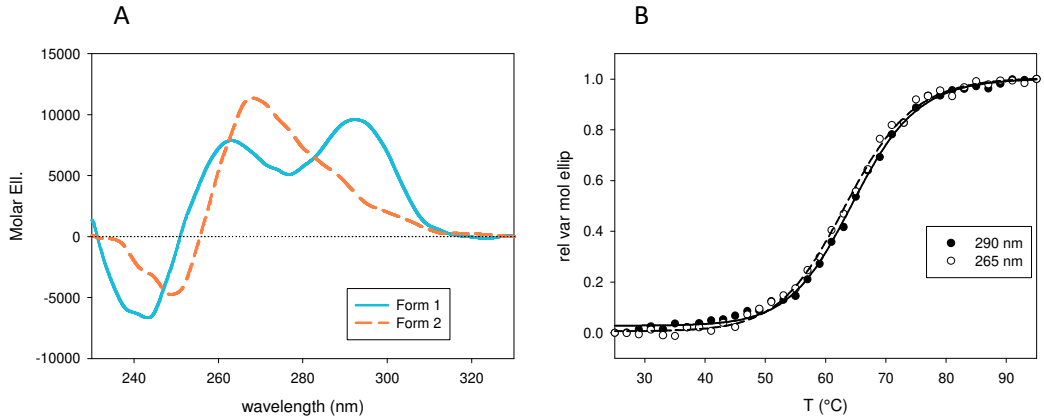


Figure 3. Generated CD spectra for the two folded species (Form 1 and Form 2 in blue and orange, respectively) derived from SVD analysis of the time-dependent titration of EGFR-272 in 200 mM KCl, 25 °C (A) and temperature-dependent relative variations of the dichroic signals of EGFR-272 recorded at 265 nm and 290 nm in 10 mM Tris, 200 mM KCl, pH 7.5 (B).

Data analysis indicated that two main species are sufficient to properly describe this second folding process (Figure S2A, S2B). Accordingly, the two significant derived V eigenvectors were satisfactorily described by a global mono-exponential process (Figure S2C, S2D). The derived fitting parameters allowed us to obtain the actual spectral shapes of the two folded forms in solution that are reported in Figure 3A. Among them, one can easily be referred to a parallel (Form 2) and the other one to a hybrid G4 (Form 1). The shape of the equilibrated EGFR-272 doesn't correspond to any of the two deconvoluted forms thus suggesting they coexist in solution.

Their stabilities were analysed by following the melting profile of EGFR-272 after proper equilibration in 200 mM KCl (Figure S3). In these conditions, EGFR-272 showed a fully reversible thermal denaturation profile. Analysis of the CD signal in the whole wavelength range derived from spectra acquired at increasing temperatures indicates that three species are significantly changing along the process. Comparison of the melting profile acquired at 265 vs 290 nm (the two wavelengths corresponding to the more intense dichroic contribution of the two main forms) showed a modest, although significant, difference in the melting temperature (Figure 3B). In particular, a Tm1 = 63.4 ± 0.2 °C and a Tm2 = 64.6 ± 0.2 °C were determined at 265 and 290 nm, respectively. This points towards the presence of two main forms with, unexpectedly, very similar thermal stabilities or to a large prevalence of one form over the other.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

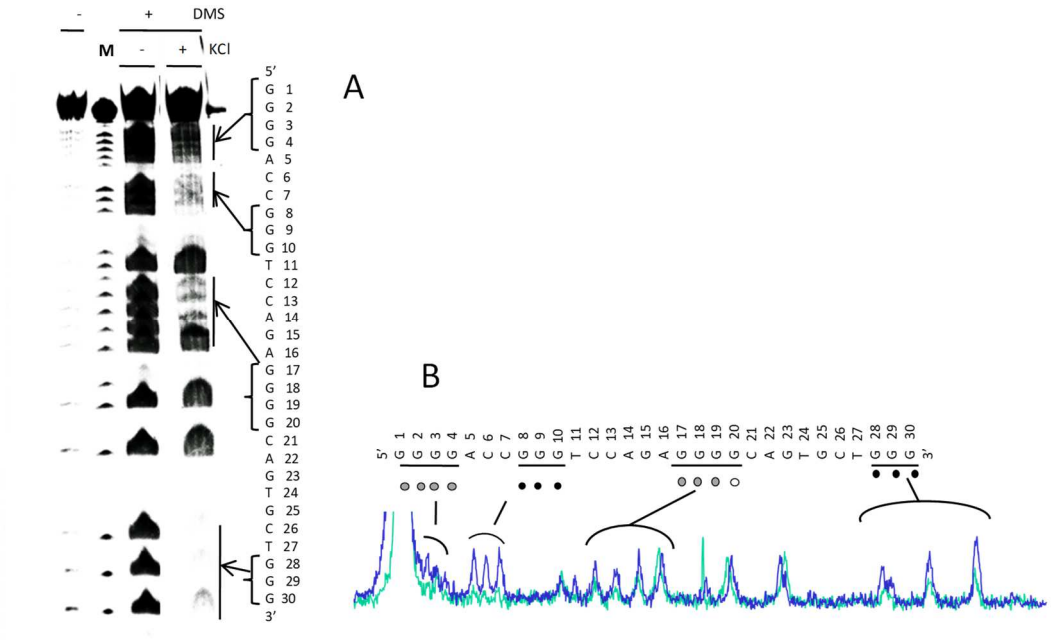


Figure 4. DMS footprinting of EGFR-272 in the absence/presence of 200 mM KCl (A). The densitometric analyses of the lanes corresponding to samples prepared in the absence (blue line) and in the presence of the metal ion (cyan line) are reported in (B).

This prompted us to investigate which were the most relevant guanines involved in the G-tetrads pairing. Thus, we isolated the high mobility band resolved by PAGE and we treated the extracted product(s) according to the DMS footprinting protocol in order to detect the guanines involved in the G4 tetrads as bases protected from the chemical modification (40). Results are summarized in Figure 4. Comparison of the DMS footprinting traces obtained in the presence and in the absence of KCl showed a clear metal-induced protection of the guanine triplets at positions 8-10 and 27-30. Within the G17-G20 tract G20 was always clearly exposed to the cleavage reaction as well as the isolated G15, G23 and G25. Conversely, it was less easy to unambiguously attribute G-quartets pairing within the G1-4 guanines repeat.

Mutation of selected guanines in EGFR-272 alters its conformational arrangement

To solve this lacking information we decided to analyze some oligonucleotides containing one or two G->T mutations within the two stretches formed by four consecutive guanines.

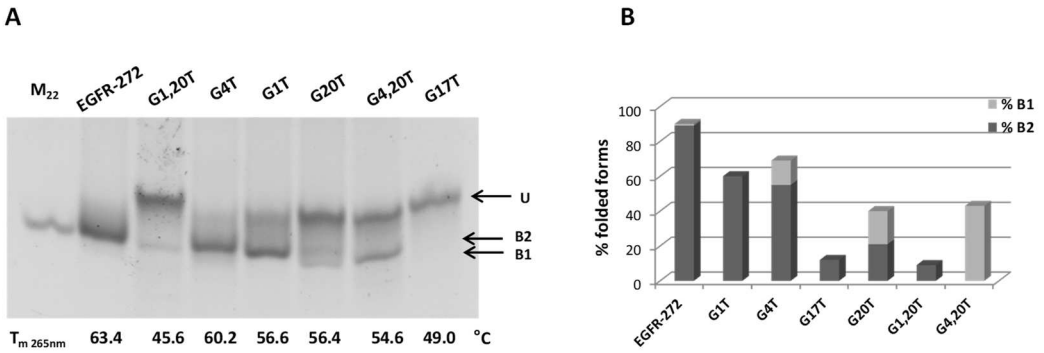


Figure 5. (A) EMSA of EGFR-272 and selected mutants annealed in 200 mM KCl. Lane M22 refers to a random sequence 22 residues long, U to the unfolded oligonucleotides and B1 and B2 to two high electrophoretic mobility forms. On the bottom, the melting temperatures derived by recording the CD signal at 265 nm in 200 mM KCl, is reported. In (B) the quantification of B1 and B2 with reference to the total DNA amount is reported.

EMSA of the oligonucleotides annealed in 200 mM KCl showed clear differences among the tested sequences (Figure 5). Mutation of either G1 or G4 poorly affected the extent of the folded fraction in comparison to the wt sequence thus suggesting that a shuffling of the guanines forming this G4 column is actually possible. Conversely, as expected, mutation of the G17 (G17T) almost completely prevents the formation of any high electrophoretic mobility form. Thus, in this third G4 column, no recruitment of other guanines is accepted. Nevertheless, peculiar data derived upon mutation of G20. Indeed, although footprinting results indicated that for the wt sequence G20 was not involved in G-quartet pairing, the G20T sequence remains largely unfolded (≈ 65%). Unexpectedly, the folded fraction was split into two well resolved bands (B1 and B2). Using them as markers it can be notice that whereas for EGFR-272 and G1T the fastest band is almost totally absent, with G4T the metal ion induces only the more compact form although to a minor percentage.

To get clearer insights on these folded forms, we acquired the CD spectrum of all the mutated sequences both in the absence and presence of 200 mM KCl as well as the corresponding melting profiles (Figures 5,6 and S4).

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

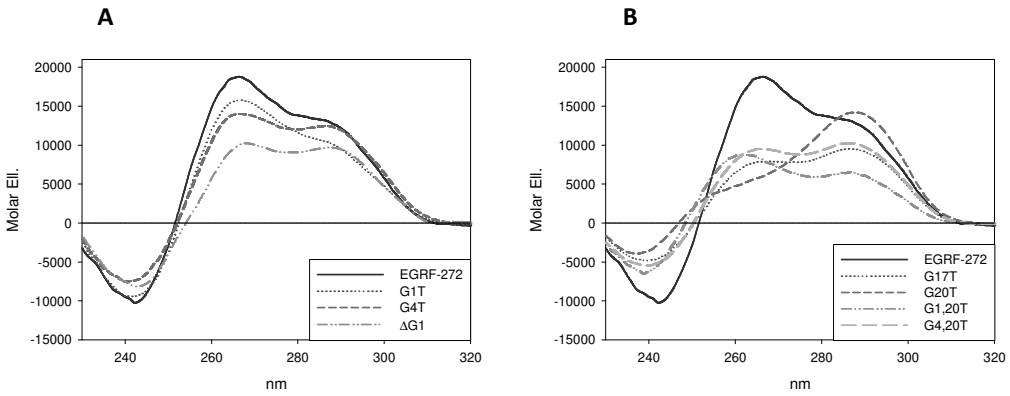


Figure 6. CD spectra of EGFR-272 single mutations in the first G-tract (A) or in the third one along with double mutations (B) acquired in 10 mM Tris, 200 mM KCl, pH 7.5, 25°C.

In the absence of KCl the spectral features of the tested sequences were variable. This is likely related to the aforementioned occurrence of hairpins formations as highlighted by the variation in the oligonucleotides electrophoretic mobility in native conditions which is actually elicited in denaturing one (Figure S5). After annealing in KCl, the acquired CD spectra present a picture that well parallels EMSA results (Figure 6).

Among oligonucleotides containing a single mutation, G17 was confirmed to be essential for G4 formation and stability. Indeed, upon its substitution with a T (G17T), a low dichroic signal was recorded which was lost at quite low temperature.

The mutations of G1 as well as G4 did not prevent the overall folding of the nucleic acid thus further supporting that a shuffling of the guanines forming the first column within the G4 is acceptable. The slight lower stability of the folded G1T (Tm 56.6 °C at 265 nm) derives from a limited reduction of the affinity for the metal ion and fits with the presence of a small fraction of unfolded DNA when resolved by PAGE. It is worth to underlines that the G4T mutant showed a CD signal overlapping the one corresponding to the wt EGRF-272 at 290 nm but a lower one at 265 nm, thus suggesting a redistribution of the relative amounts of the folded forms.

Unexpectedly, the removal of the first G1 (ΔG1) did not provide a CD profile comparable to the wt EGRF-272 or its closely related G1T. Conversely, it showed a profile with two well resolved maxima at 265 and 290 nm which relative ratio parallels the one obtained for G4T. Consistently, the melting temperature determined at 265 nm is notable (Tm 65.0 °C).

Striking result was obtained with the G20T mutant. G20 was expected to be not involved in G-quartet pairing. However, G20T had a CD signal at 265 nm (as well as the melting temperature herein detected corresponding to 56.4 °C) extremely reduced whereas the contribution at 290 nm was actually reinforced. When the G20T mutation was associated to G4T or G1T mutation the effect was even more dramatic: both sequences with two mutations showed CD intensities comparable (G4,20T) or even lower (G1,20T) with reference to G17T. Thus, we can suggest that G20 although not directly involved in G-quartet pairing might contribute to drive G4 formation, possibly by providing an external capping or through interactions with other bases in loop(s).

Small modifications in the first G-tract of EGFR-272 leads to two well-defined G4 forms

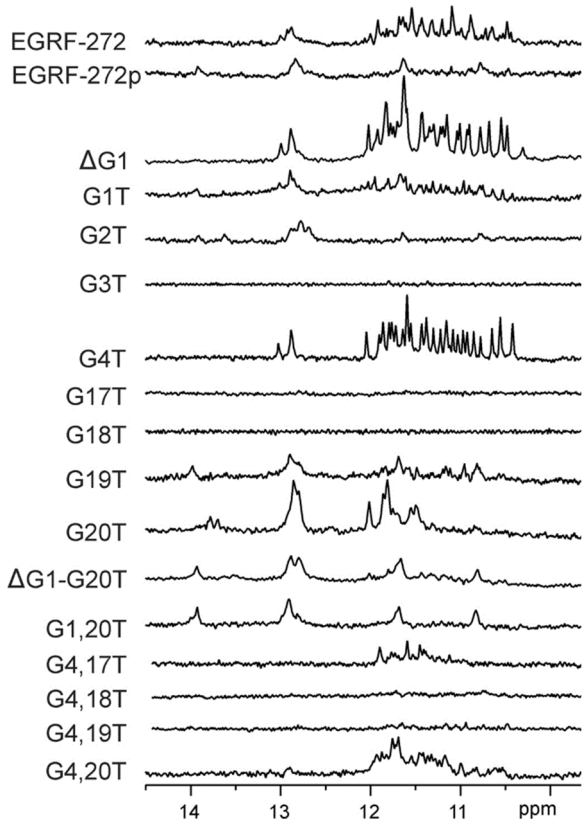


Figure 7. Imino region of ^{1}H NMR spectra of EGFR-272 and mutants. NMR spectra were recorded at 0.1 mM concentration of the oligonucleotide per strand, pH 7.0, 25°C on 800 MHz spectrometer. The concentration of K^{+} ions was 50 mM for T3, T17, T18, ΔG1, G20T; 70 mM for ΔG1, G4T, G4T,18 and 100 mM for all other samples. The sequences of mutants are presented in Table 1.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

Despite trying many different experimental conditions and annealing protocols, we were not able to reduce the structural diversity of EGFR-272 in ^{1}H NMR samples (Figure S6). Furthermore, in order to trap only one G4 fold in solution we enlarged the mutant library up to sixteen different mutated EGFR-272 oligonucleotides, which included an added phosphate group at the 5'-end (EGFR-272p), the removal of G1 residue (ΔG1) and fourteen oligonucleotides with one or two G->T mutations (Figure 7).

For ΔG1 and G4T mutants we observed twenty-four narrow and resolved signals in ^{1}H NMR spectra between δ 10.30 and 12.05 ppm indicating formation of well-defined G4 forms (vide infra). In the case of G1T and G4,20T mutants the high number of imino signals in ^{1}H NMR spectra demonstrated the presence of multiple G4 structures in the samples. These results suggested that removal of G1 in ΔG1 was more favorable to stabilize two major forms in NMR samples in comparison to its mutation with T (G1T). Interestingly, the phosphate group added at the 5'-end in the EGFR-272p mutant greatly influenced the folding of the oligonucleotide as it can derive from the observed four broad signals in imino region between δ 10.70 and 14.00 ppm. A similar pattern of imino signals in ^{1}H NMR spectra was detected also for G2T, ΔG1-G20T and G1,20T mutants. This pattern of signals does not support formation of G4 structures. While changes in the first G-tracts resulted in resolved signals in NMR spectra of two different major G4 forms, modifications in the third tract did not lead to well-defined structures. In the case of G17T, G18T, G4,18T and G4,19T mutants no or only minor signals were detected in the imino region of ^{1}H NMR spectra indicating that the modified guanine residues play a vital role in G-tetraplex folding and structural integrity. No imino signals were detected also for G3T mutant. For G19T and G20T we observed additional signal at around δ 13.8 ppm besides imino signals involved in Hoogsteen and Watson-Crick G-C base pairs. This signal could indicate an additional base paired thymine residue, or the presence of the form observed also in the NMR samples of some other mutants (e. g. G1,20T).

Interestingly, by comparing double mutants (ΔG1-G20T and G1,20T) we observed that they share a very similar pattern of signals in the imino region of ^{1}H NMR spectra (Figure 7). On the other hand, for double mutants of residue G4 and all four guanine residues in the third G-tract we observed no signals in imino region for G4,18T and G4,19T. Only imino signals between δ 11.10 and 11.90 ppm were observed for G4,17T mutant. The imino region of the ^{1}H NMR spectrum of G4,20T mutant was more similar to EGFR-272 with signals between δ 10.43 and 12.00 ppm and a very broad signal at δ 12.90 ppm (Figure 7).

It is noteworthy that similar spectral characteristics point to comparable structural elements, but not necessary to the same structure.

Due to the better peaks resolution provided by ΔG1 and G4T mutants, they were studied in more detail with the use of 2D NOESY spectra. In the case of both mutants we could distinguish between two sets of signals based on intensity in the imino region between δ 10.30 and 12.05 ppm of 1D ^{1}H and 2D NOESY spectra (Figure 7, Figures S7 and S8). Each set was comprised of twelve signals. The number and intensity of signals suggested the formation of two monomeric G4s each consisting of three G-quartets. The ratio between the two forms was ≈65:35 and ≈60:40 for ΔG1 and G4T, respectively and was not dependent on K^{+} ion concentration as shown by titration experiments (Figure S9). For both mutants we detected at least three broad signals in the region between δ 12.80 and 13.00 ppm that are characteristics for imino protons involved in Watson-Crick G-C base pairs. The aromatic-anomeric region of 2D NOESY NMR spectra revealed four very intense intermolecular NOE cross-peaks between H8 and H1' protons that indicated a syn orientation of at least four guanine residues in the case of ΔG1 as well as G4T (Figures S10 and S11). This is consistent with one of the G4 forms occupying a (3+1) hybrid strand orientation for each mutant. However, further investigation beyond the scope of this study is needed for more detailed high-resolution structural data.

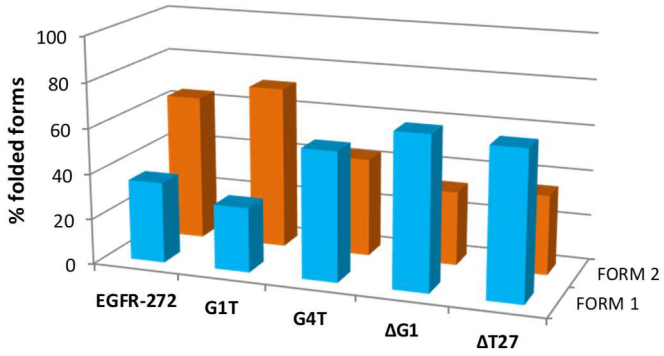


Figure 8. Prediction of the distribution of Form 1 (blue bars) and Form 2 (orange bars) in the folded population as derived from deconvolution of the experimental CD spectra of selected oligonucleotides based on the SVD-generated dataset reported in Figure 3A.

To asses if the predicted two folded forms identified for ΔG1 and G4T are adopted also by the wt oligonucleotide, we referred to the basic spectra obtained by deconvolution of CD profile of EGFR-272. We found that also the experimental acquired CD spectra of ΔG1,

G1T and G4T are well described by proper combination of these two datasets (Figure S12). Interestingly, the relative weights of the two forms cluster our oligonucleotides into two groups: the first one comprising EGFR-272 and G1T in which Form 2 counts for ≈ 70% of the entire population, the second one groups ΔG1 and G4T in which, in agreement with NMR data, the ratio is reversed to about 60% of Form 1 (Figure 8).

The third loop contributes in defining the relative distribution among the two main G4 forms

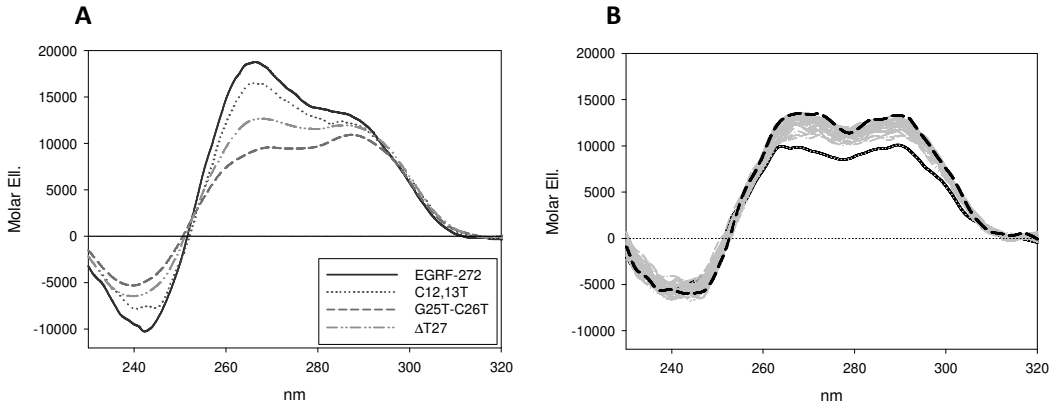


Figure 9. CD spectra of EGFR-272 related sequences containing mutations in the loops acquired in 10 mM Tris, 200 mM KCl, pH 7.5, 25 °C (A). Variation of the CD spectra acquired in 10 mM Tris, pH 7.5, 25 °C of 4 µM ΔT27 upon addition of 200 mM KCl as a function of incubation time (1 acquisition/120 sec); dashed black line correspond to the end of the titration (B).

The so far collected data confirmed that when EGFR-272 and its related mutants fold into G4 two forms are the most relevant. However, no explanation was derived for the critical contribution of G20 to the folding efficiency. In particular, the massive reduction of G4 formation by G20T suggests that both forms might benefit from interactions with G20. It is worth to remind that accordingly to NMR and footprinting data, G20 results inserted in a very long loop and any condition that freezes it in a fixed position is expected to be favourable for G4 stability. Actually, NMR of G4 folded systems always contains peaks attributable to GC pairs.

With this in mind we observed that the second loop contains two cytosines that eventually can pair with G20 to form an extended cupping element and thus strengthening its

contribution to G4 stability. Furthermore, also within the third loop two Watson-Crick GC pairs can occur between G20-C21 and G25-C26. This pairing is expected to support the formation of a stable hairpin within the loop thus decreasing its flexibility. To dissect among these two different models, we analysed by CD spectroscopy an additional set of EGFR-272 mutants in which we selectively modified these two loops (Figure 9).
Substitution of the guanines at position 12 and 13 (C12,13T) did not affect to a significant extent the G4 formation thus ruling out any relevant contribution from them to the G4 folding process.

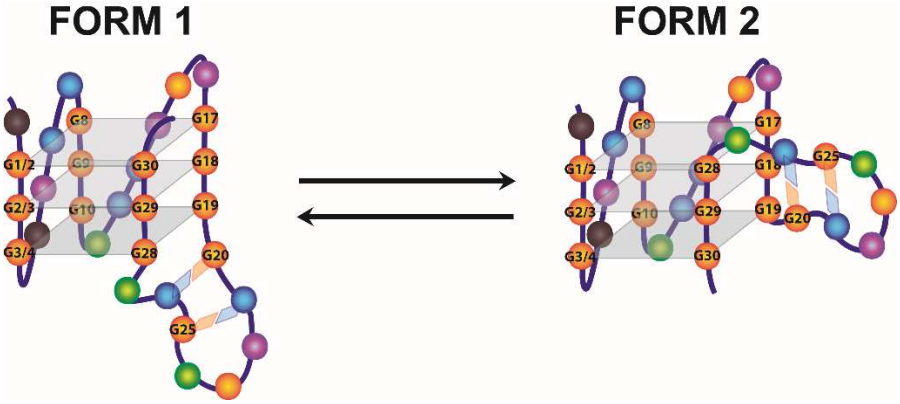


Figure 10. Schematic drawing of the main structural equilibrium occurring within EGFR-272. Guanines are shown in orange, adenines in magenta, cytosines in blue and thymines in green.

On the opposite, the G25T-C26T double mutant showed a remarkably reduced tendency to fold. The CD spectrum showed that this mutation mostly influenced the parallel form while the hybrid one was more preserved in comparison to EGFR-272. Unfavourable effects of G25 and C26 modifications on folding and structural integrity were evident also from 1H NMR spectrum (Figure S13). The hump together with many superimposed and overlapped imino signals between δ 10.70 and 12.12 ppm indicated formation of aggregates and several multimeric structures. Very broad imino signals characteristic for G-C base pairs were observed at δ 12.83 ppm. They probably do not belong to the same base pairs as in EGFR-272 due to their very different fingerprint of imino signals in comparison to G25T-C26T mutant (Figure S13). These results suggest that the third loop is organized in a hairpin-like structure stabilized by two G-C base pairs. We believe that hairpin-like structures drive the G4 formation, since the decreased folding of the G25T-

C26T mutant is most likely due to disruption of G-C base pairs. It is noteworthy that pre-folded hairpin structures were observed in the absence of KCl as well (Figure S1).

According to this model it derives that I n the wild type EGFR-272, an unpaired T (T27) connects the G-tetrad and the stem of the loop (Figure 10). This element appears to be required to allow the orthogonal positioning of the loop double helical domain across a medium wide groove as the one present in an all parallel structure as the one corresponding to Form 2 of EGFR-272 (41).

Conversely, from a wide groove, as the one occurring between two antiparallel-oriented strands as expected in the hybrid Form 1, the hairpin can directly exit in a coaxial orientation with reference to the G4 core without losing any base stacking and thus not impairing the overall stability of the system. Here, deletion of the connecting T27 on EGFR-272 (ΔT27) provided an oligonucleotide that preserves the ability to fold into G4. Additionally, it didn't show time dependent structural rearrangements in agreement with the indication that the slow forming parallel Form 2 is not expected to favourably accommodate the hairpin stem when the oligonucleotide contains this constrain. Consistently, Form 1 dominates vs Form 2 in the folded fraction (Figure 8).

DISCUSSION

EGFR represents a highly valuable target for anticancer treatment but with high demanding need for novel strategies to suppress its activity. In this connection the presence of G-rich domains within its gene promoter is an attractive starting point which prompted us to explore its potential conformational rearrangements. Our data confirmed that the sequence located at position -272 from TSS can fold into G4 in the presence of physiological concentration of K+ ions. Interestingly, this sequence is not fully folded at lower metal ion concentrations thus suggesting a reliable structural plasticity within the cell which can be proficiently exploited to regulate gene expression by small ligands targeting.

Although it was not possible to solve a unique G4 structures interesting peculiar structural features emerged. First of all, in our experimental conditions, two main folded forms were identified in solution and we succeed to attribute them to a kinetically favoured hybrid G4 corresponding to a (3+1) arrangement and to a slower forming parallel one both comprising three stacked G-quartets. The easiest model would correspond to the full

conversion of the first structure into the second one. However, this was not the case since we always found the two of them simultaneously present in solution. The resolution of the two forms was further hampered by their unexpected comparable affinities for the metal ion as well as by their similar thermal stabilities. Since EGFR-272 is quite rich in guanines, this led us to hypothesize the formation of distinct G4 structures deriving from the recruitment of different residues in the G-tetrads. In particular, our evidences indicated that among the four guanines at 5'-end it exists the possibility of a shift of the residues involved in G4 pairing. Consistently, the recruitment of residues 1-3 and 2-4 to produce the two forms could be envisaged. Nevertheless, mutation studies elicited this hypothesis since both G1T and G4T containing only three guanines at 5'-end still contains a balanced contribution of the hybrid and parallel G4 folded forms. Only the absence of an unpaired nucleotide at 5'-modulates this ratio towards a preference for the hybrid form, as shown by the shared structures distribution between G4T and ΔG1.

By merging these data with the observation that G20 was not directly involved in G-tetrad formation, the emerging picture indicates that both G4 forms derive from pairing of the same guanines. This produce a general model in which three G-quartet stack one over the other and where the combination of 3/4, 6 and 8 nucleotide long loops fits with the quite relevant concentration of KCl required to fully fold the sequence.

Unexpectedly, highly detrimental for G4 formation resulted the mutation of G20 located in the longest terminal loop which affect both the parallel and the hybrid form (G20T). This foresees a clear role of G20 as stabilizing element. Insertion of proper mutations in the loops excluded the G20 pairing with a cytosine of the second loop. Conversely, it indicated that such a function is related to the formation of an hairpin within the third loop. It has been reported that G4 containing a long loop arranged in an hairpin can be joined to the duplex domain with different arrangement and energetic contribution according to the G4 geometry at the loop insertion points (41-43). By applying the proposed models to our sequence, we can propose that in the all parallel structure the stem can be orthogonally inserted in the G4 groove. In our case the energetic stabilization can be partly reduced by the presence of just one unpaired nucleotide at the junction (T27). Upon removal of this thymine (ΔT27), the stability of the overall system is not hampered. This suggests the direct exit of the stem from a wide groove as the one occurring between two antiparallel-oriented strands in a (3+1) structure. This corresponds

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

also to the kinetically favoured form which maximally benefits of the preformed arrangement of the sequence.

Thus, in conclusion the unique base composition of the long-third loop of EGFR-272 is suitable to provide the energetic contributions that are required to describe the fascinating structural equilibria of EGFR-272.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

This work was supported by University of Padova [CPDA147272 to C.S., PhD fellowships to C.C. and RR]; by Cariparo [PhD fellowships to M.L.G.] and by the Slovenian Research Agency (ARRS, research cores funding No. P1-242 and J1-6733).

CONFLICT OF INTEREST

Authors have nothing to declare.

ACKNOWLEDGEMENT

We acknowledge M. Folini for support with bioinformatics analysis.

REFERENCES

1. Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M.R., Carotenuto, A., De Feo, G., Caponigro, F. and Salomon, D.S. (2006) Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, **366**, 2-16.
2. Lemmon, M.A., Schlessinger, J. and Ferguson, K.M. (2014) The EGFR family: not so prototypical receptor tyrosine kinases. *Cold Spring Harbor perspectives in biology*, **6**, a020768.
3. Holowka, D. and Baird, B. (2016) Mechanisms of epidermal growth factor receptor signaling as characterized by patterned ligand activation and mutational analysis. *Biochimica et biophysica acta*.
4. Goffin, J.R. and Zbuk, K. (2013) Epidermal growth factor receptor: pathway, therapies, and pipeline. *Clinical therapeutics*, **35**, 1282-1303.
5. Hata, A.N., Niederst, M.J., Archibald, H.L., Gomez-Caraballo, M., Siddiqui, F.M., Mulvey, H.E., Maruvka, Y.E., Ji, F., Bhang, H.E., Krishnamurthy Radhakrishna, V. *et al.* (2016) Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.*, **22**, 262-269.
6. Morgillo, F., Della Corte, C.M., Fasano, M. and Ciardiello, F. (2016) Mechanisms of resistance to EGFR-targeted drugs: lung cancer. *ESMO open*, **1**, e000060.

7. Collie, G.W. and Parkinson, G.N. (2011) The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society reviews*, **40**, 5867-5892.
8. Wong, H.M., Stegle, O., Rodgers, S. and Huppert, J.L. (2010) A toolbox for predicting g-quadruplex formation and stability. *Journal of nucleic acids*, **2010**.
9. Hansel-Hertsch, R., Di Antonio, M. and Balasubramanian, S. (2017) DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Biol.*, **18**, 279-284.
10. Čeru, S., Šket, P., Prislan, I., Lah, J. and Plavec, J. (2014) A New Pathway of DNA G-Quadruplex Formation. *Angew. Chem. Int. Ed.*, **53**, 4881-4884.
11. Marušič, M. and Plavec, J. (2015) The Effect of DNA Sequence Directionality on G-Quadruplex Folding. *Angew. Chem. Int. Ed.*, **54**, 11716-11719.
12. Dolinnaya, N.G., Ogloblina, A.M. and Yakubovskaya, M.G. (2016) Structure, properties, and biological relevance of the DNA and RNA G-quadruplexes: Overview 50 years after their discovery. *Biochem. (Mosc.)*, **81**, 1602-1649.
13. Brčić, J. and Plavec, J. (2017) ALS and FTD linked GGGGCC-repeat containing DNA oligonucleotide folds into two distinct G-quadruplexes. *Biochimica et biophysica acta*, **1861**, 1237-1245.
14. Marušič, M., Hošnjak, L., Krafčikova, P., Poljak, M., Viglasky, V. and Plavec, J. (2017) The effect of single nucleotide polymorphisms in G-rich regions of high-risk human papillomaviruses on structural diversity of DNA. *Biochimica et biophysica acta*, **1861**, 1229-1236.
15. Trajkovski, M., Webba da Silva, M. and Plavec, J. (2012) Unique Structural Features of Interconverting Monomeric and Dimeric G-Quadruplexes Adopted by a Sequence from the Intron of the N-myc Gene. *J. Am. Chem. Soc.*, **134**, 4132-4141.
16. Moye, A.L., Porter, K.C., Cohen, S.B., Phan, T., Zyner, K.G., Sasaki, N.i., Lovrecz, G.O., Beck, J.L. and Bryan, T.M. (2015) Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat. Commun.*, **6**, 7643.
17. Ray, S., Bandaria, J.N., Qureshi, M.H., Yildiz, A. and Balci, H. (2014) G-quadruplex formation in telomeres enhances POT1/TPP1 protection against RPA binding. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 2990-2995.
18. Fujii, T., Podbevšek, P., Plavec, J. and Sugimoto, N. (2017) Effects of metal ions and cosolutes on G-quadruplex topology. *J. Inorg. Biochem.*, **166**, 190-198.
19. Miyoshi, D., Nakao, A. and Sugimoto, N. (2003) Structural transition from antiparallel to parallel G-quadruplex of d(G4T4G4) induced by Ca2+. *Nucleic Acids Res.*, **31**, 1156-1163.
20. Marušič, M., Šket, P., Bauer, L., Viglasky, V. and Plavec, J. (2012) Solution-state structure of an intramolecular G-quadruplex with propeller, diagonal and edgewise loops. *Nucleic Acids Res.*, **40**, 6946-6956.
21. Gajarský, M., Živković, M.L., Stadlbauer, P., Pagano, B., Fiala, R., Amato, J., Tomáška, L., Šponer, J., Plavec, J. and Trantírek, L. (2017) Structure of a Stable G-Hairpin. *J. Am. Chem. Soc.*, **139**, 3591-3594.
22. Kocman, V. and Plavec, J. (2014) A tetrahelical DNA fold adopted by tandem repeats of alternating GGG and GCG tracts. *Nat. Commun.*, **5**, 5831.
23. Kocman, V. and Plavec, J. (2017) Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. *Nat. Commun.*, **8**, 15355.
24. 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.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

25. Rhodes, D. and Giraldo, R. (1995) Telomere structure and function. *Current opinion in structural biology*, **5**, 311-322.
26. Laguerre, A., Hukezalie, K., Winckler, P., Katranji, F., Chanteloup, G., Pirrotta, M., Perrier-Cornet, J.M., Wong, J.M. and Monchaud, D. (2015) Visualization of RNA-Quadruplexes in Live Cells. *J. Am. Chem. Soc.*, **137**, 8521-8525.
27. Henderson, A., Wu, Y., Huang, Y.C., Chavez, E.A., Platt, J., Johnson, F.B., Brosh, R.M., Jr., Sen, D. and Lansdorp, P.M. (2014) Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.*, **42**, 860-869.
28. Shivalingam, A., Izquierdo, M.A., Marois, A.L., Vysniauskas, A., Suhling, K., Kuimova, M.K. and Vilar, R. (2015) The interactions between a small molecule and G-quadruplexes are visualized by fluorescence lifetime imaging microscopy. *Nat. Commun.*, **6** 9178.
29. Kotar, A., Wang, B., Shivalingam, A., Gonzalez-Garcia, J., Vilar, R. and Plavec, J. (2016) NMR Structure of a Triangulenium-Based Long-Lived Fluorescence Probe Bound to a G-Quadruplex. *Angew. Chem. Int. Ed. Engl.*, **55**, 12508-12511.
30. Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406-413.
31. Balasubramanian, S., Hurley, L.H. and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nature reviews. Drug discovery*, **10**, 261-275.
32. Grand, C.L., Han, H., Munoz, R.M., Weitman, S., Von Hoff, D.D., Hurley, L.H. and Bearss, D.J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo. *Molecular cancer therapeutics*, **1**, 565-573.
33. Howell, R.M., Woodford, K.J., Weitzmann, M.N. and Usdin, K. (1996) The chicken beta-globin gene promoter forms a novel "cinched" tetrahelical structure. *The Journal of biological chemistry*, **271**, 5208-5214.
34. Rigo, R., Palumbo, M. and Sissi, C. (2017) G-quadruplexes in human promoters: A challenge for therapeutic applications. *Biochimica et biophysica acta*, **1861**, 1399-1413.
35. Mergny, J.L., Li, J., Lacroix, L., Amrane, S. and Chaires, J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
36. DeSa, R. J., Matheson, I. B.(2004) A practical approach to interpretation of singular value decomposition results, *Methods Enzymol.*, **384**, 1-8.
37. Hendler, R. W., Shrager, R. I. (1994) Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners, *J. Biochem. Biophys. Methods*, 28**,** 1-33.
38. Webba da Silva, M. (2007) NMR methods for studying quadruplex nucleic acids. *Methods*, **43**, 264-277.
39. Adrian, M., Heddi, B. and Phan, A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, **57**, 11-24.
40. Sun, D. and Hurley, L.H. (2010) Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods Mol. Biol.*, **608**, 65-79.
41. Lim, K.W. and Phan, A.T. (2013) Structural Basis of DNA Quadruplex–Duplex Junction Formation. *Angew. Chem. Int. Ed.*, **52**, 8566-8569.
42. Lim, K.W., Jenjaroenpun, P., Low, Z.J., Khong, Z.J., Ng, Y.S., Kuznetsov, V.A. and Phan, A.T. (2015) Duplex stem-loop-containing quadruplex motifs in the human

genome: a combined genomic and structural study. *Nucleic Acids Res.*, **43**, 5630-5646.

43. Onel, B., Carver, M., Wu, G., Timonina, D., Kalarn, S., Larriva, M. and Yang, D. (2016) A New G-Quadruplex with Hairpin Loop Immediately Upstream of the Human BCL2 P1 Promoter Modulates Transcription. *J. Am. Chem. Soc.*, **138**, 2563-2570.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

SUPPLEMENTARY INFORMATION

Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter

Maria Laura Greco[1], Anita Kotar[2], Riccardo Rigo[1], Cristofari Camilla[1], Janez Plavec[2,3,4,*] and Claudia Sissi[1,*]

[1] Department of Pharmaceutical and Pharmacological Sciences, University of Padova, v. Marzolo 5, Padova, 35131, ITALY

[2] Slovenian NMR Center, National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

[3] EN-FIST Center of Excellence, Trg OF 13, 1000 Ljubljana, Slovenia

[4] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Večna pot 113, Ljubljana

* To whom correspondence should be addressed. Tel: +39-049 8275711; Fax: +39-049 8275366; e-mail: claudia.sissi@unipd.it; janez.plavec@ki.si

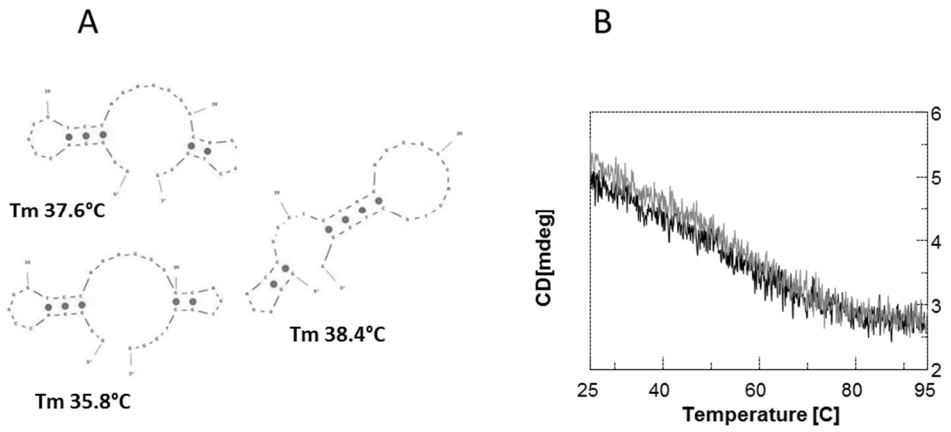
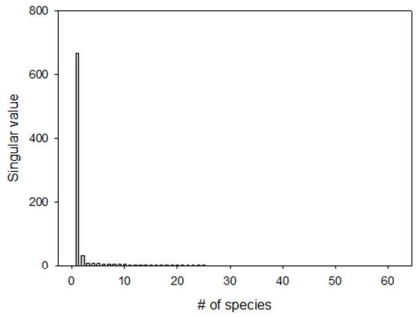
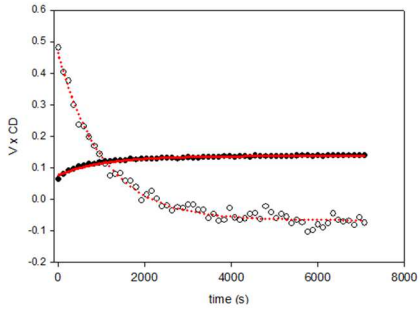


Figure S1. Theoretical intramolecular folding of EGFR-272 as derived by IDT Oligo-Analyzer tool (A) and melting profile of 4 µM EGFR-272 recorded at 260 nm in 10 mM Tris, pH 7.5, 25 °C (B). Black and gray lines refer to melting and annealing processes performed at 0.8°C/min.

A



C



B

	#	frequency
V	1	0.9879
	2	0.8628
	3	0.0439

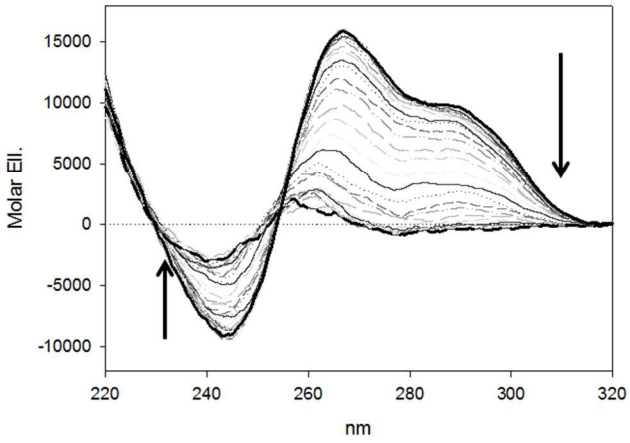
D

Best fitting equation	R^2	Rate constant (s^{-1})
$y(t)=y_0+a*(\exp(-k*t))$	0.989	$(9.54 \pm 0.24) \times 10^{-4}$

Figure S2. Data derived for the kinetic rearrangement of 4 µM EGFR-272 upon addition of 200 mM KCl in 10 mM Tris, pH 7.5, 25 °C: singular values (A), V matrix autocorrelation coefficients (B, the not significant coefficient is highlighted in grey), amplitudes of the first three *V* vectors (C) and final fitting model (D).

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

A



B

	#	frequency
V	1	0.9696
	2	0.9534
	3	0.6542
	4	0.2950

Figure S3. Variation of CD spectra of EGFR-272 acquired in in 10 mM Tris, 200 mM KCl, pH 7.5 upon increment of the temperature (A) as showed by the arrows. In (B), V matrix autocorrelation coefficients derived from this data set are reported; the not significant coefficient is highlighted in grey.

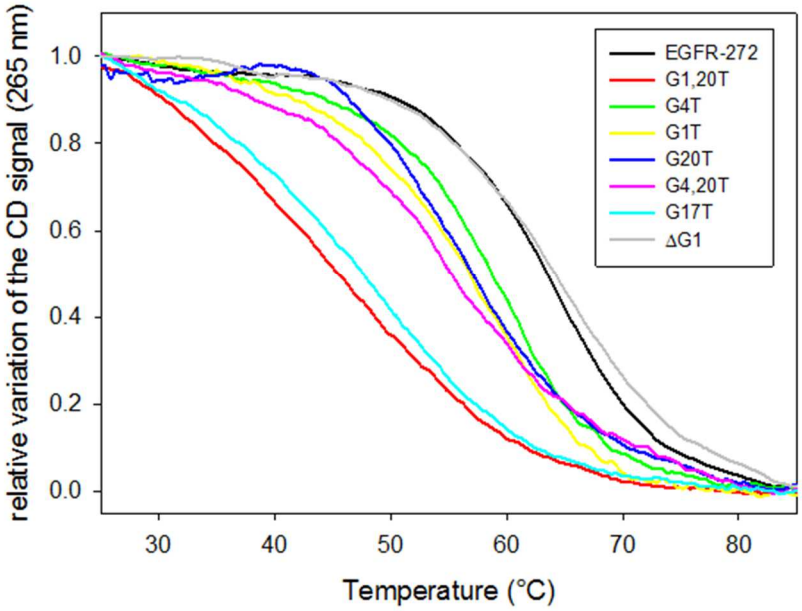


Figure S4. Thermal denaturation profiles of 4 µM EGFR-272 and related mutants previously folded in 200 mM KCl acquired by CD spectroscopy at 265 nm by applying a heating rate of 50 °C/h.

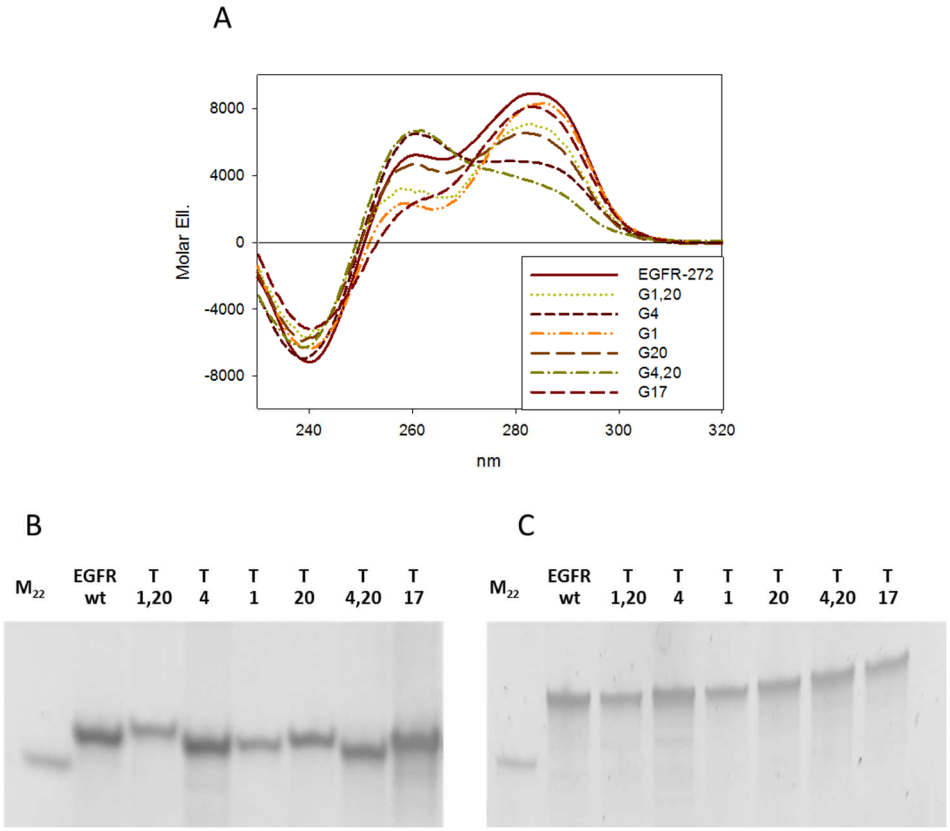


Figure S5. CD spectra of 4 μM EGFR-272 and related mutants in 10 mM Tris, pH 7.5, 25 °C (A) and PAGE resolution of the same samples run under native (B) or denaturing (C) conditions.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

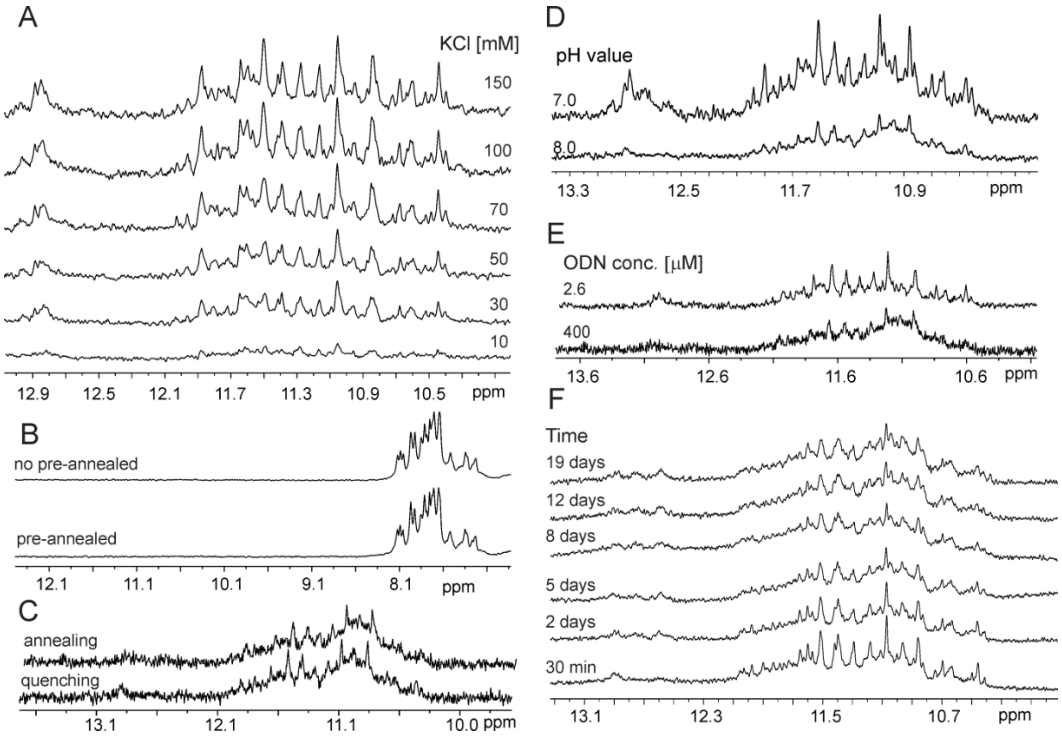


Figure S6. Imino and aromatic region of ^{1}H NMR spectra of EGFR-272 recorded at different experimental conditions. All NMR spectra were obtained in H_2O:$^{2}H_2O$=9:1 at 25°C. (A) NMR spectra of EGFR-272 as K^+ ions are titrated into solution (pH 7.0) at 0.1 mM concentration of the oligonucleotide per strand and at 800 MHz. The concentration of K^+ ions is indicated on the right. (B) The NMR spectra of the sample that was not annealed and of the sample that was annealed over night before adding the K^+ ions to solution (pH 6-7) at 0.4 mM concentration of the oligonucleotide per strand and at 300 MHz. (C) NMR spectra obtained after annealing (sample was heated to 95°C and slowly cooled down to room temperature) and quenching (sample was heated to 95°C and quickly cooled down on ice). The spectra were recorded in 10 mM Tris buffer (pH 8.0), at 200 mM concentration of KCl, 0.4 mM concentration of the oligonucleotide per strand and at 600 MHz. (D) The influence of pH value on NMR spectra that were recorded at pH values 7.0 (10 mM potassium phosphate buffer) and 8.0 (10 mM Tris buffer) after overnight annealing, at 200 mM concentration of KCl, 0.4 mM concentration of the oligonucleotide per strand and at 600 MHz. (E) NMR spectra at different concentration of oligonucleotide. The spectra were obtained at 2.6 and 400 μM concentration of the oligonucleotide per strand, after overnight annealing, at 200 mM concentration of KCl in 10 mM Tris buffer (pH 8.0) and at 600 MHz. (F) The influence of time from 30 min to 19 days on folding of EGFR-272 oligonucleotide. The sample was prepared in 10 mM Tris buffer (pH 8.0),

200 mM concentration of KCl, 0.35 mM concentration of the oligonucleotide per strand and recorded on 600 MHz spectrometer. The sample was stored in NMR tube at room temperature.

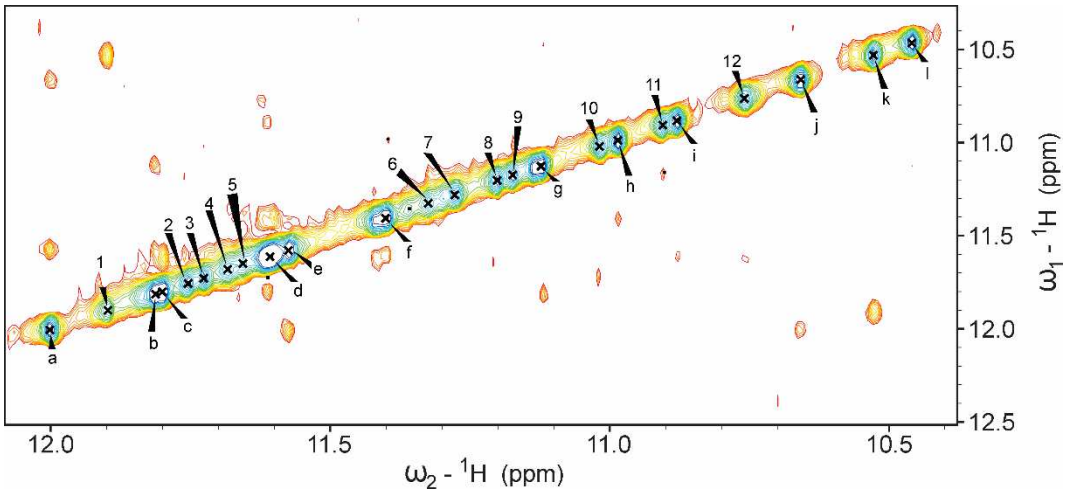


Figure S7. Imino-imino region of NOESY spectrum of ΔG1 mutant in the presence of 70 mM KCl at pH 7.0, 25°C in H_2O:^2H_2O=9:1 at 600 MHz. Mixing time was 300 ms. Diagonal cross peaks corresponding to two major forms of G4 structures are labeled with numbers and letters.

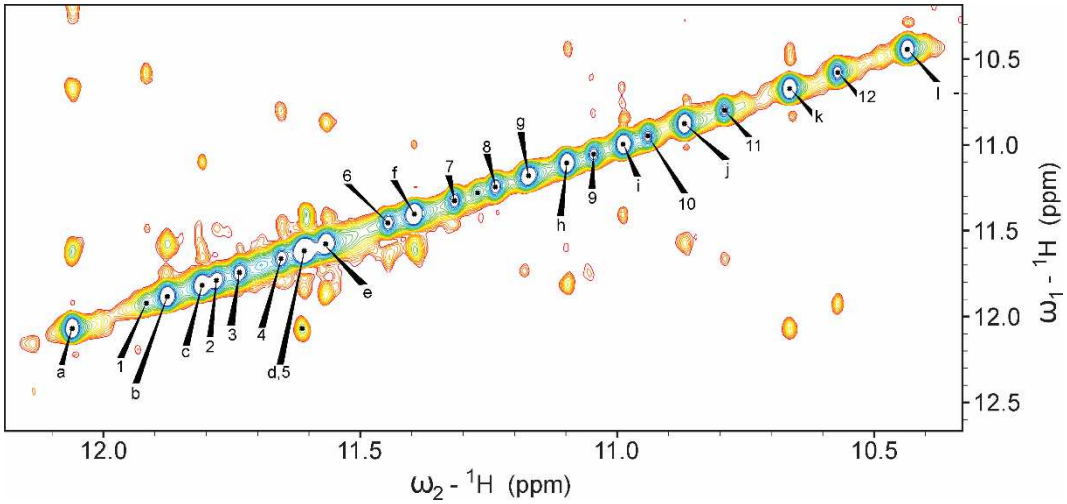


Figure S8. Imino-imino region of NOESY spectrum of T4 mutant in the presence of 70 mM KCl, at pH 7.0, 25°C in H_2O:^2H_2O=9:1 at 600 MHz. Mixing time was 300 ms. Diagonal cross peaks corresponding to two major forms of G4 structures are labeled with numbers and letters.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

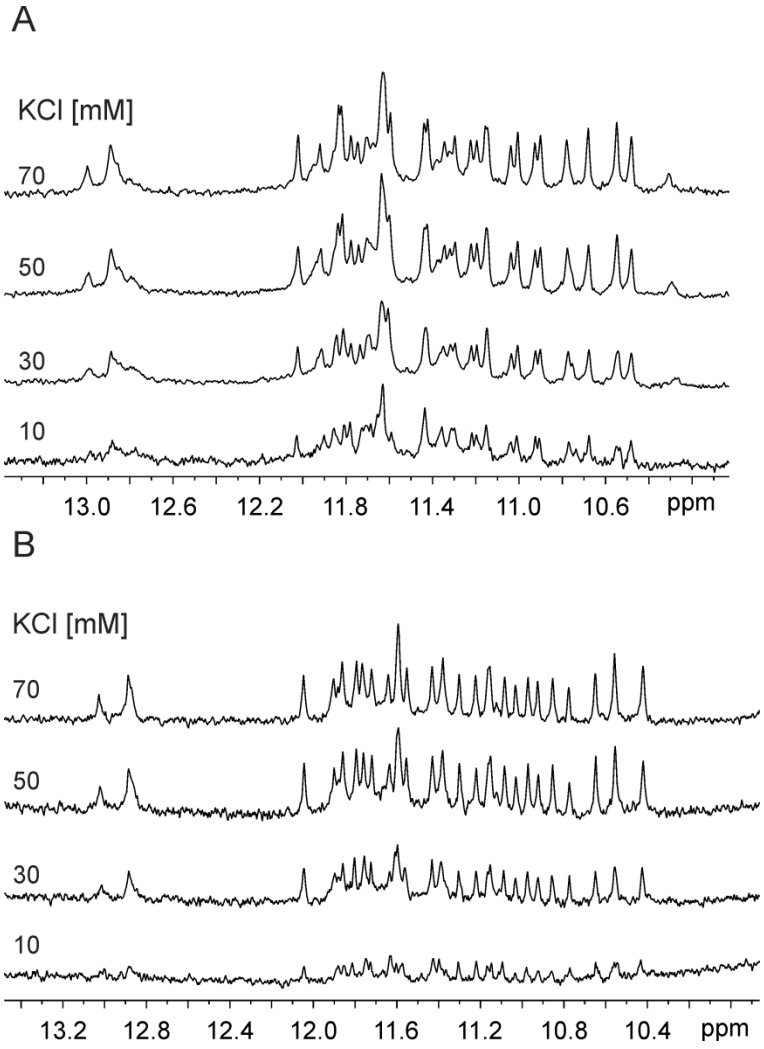


Figure S9. Imino region of ^{1}H NMR spectra of ΔG1 (A) and T4 (B) mutants at various concentrations of K^{+} ions as indicated on left. NMR spectra were recorded in H_2O:$^{2}H_2O$=9:1, at 0.1 mM concentration of the oligonucleotide per strand, pH 7.0, 25°C at 600 MHz.

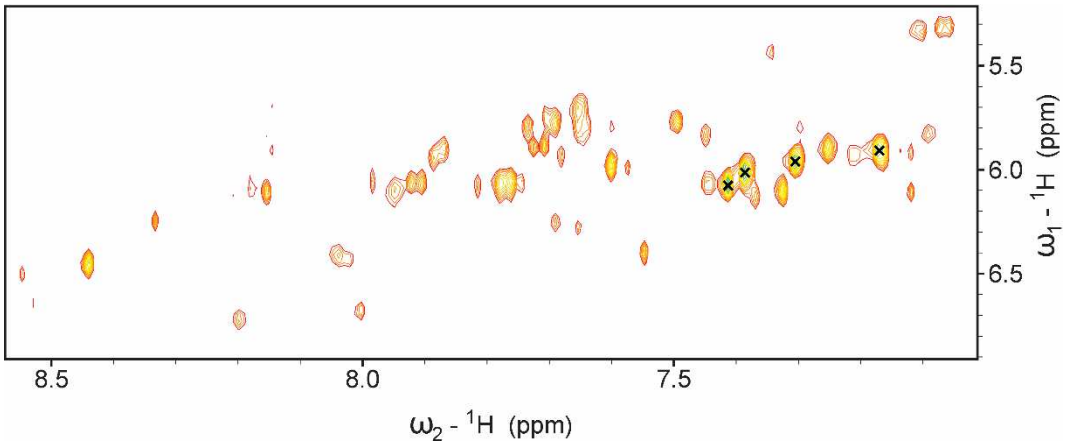


Figure S10. Aromatic-anomeric region of NOESY spectrum of ΔG1 mutant in the presence of 70 mM KCl at pH 7.0, 25°C in H_2O:^2H_2O=9:1 at 600 MHz. Mixing time was 150 ms. Very intense intermolecular NOE cross-peaks between H8 and H1' protons are marked with cross.

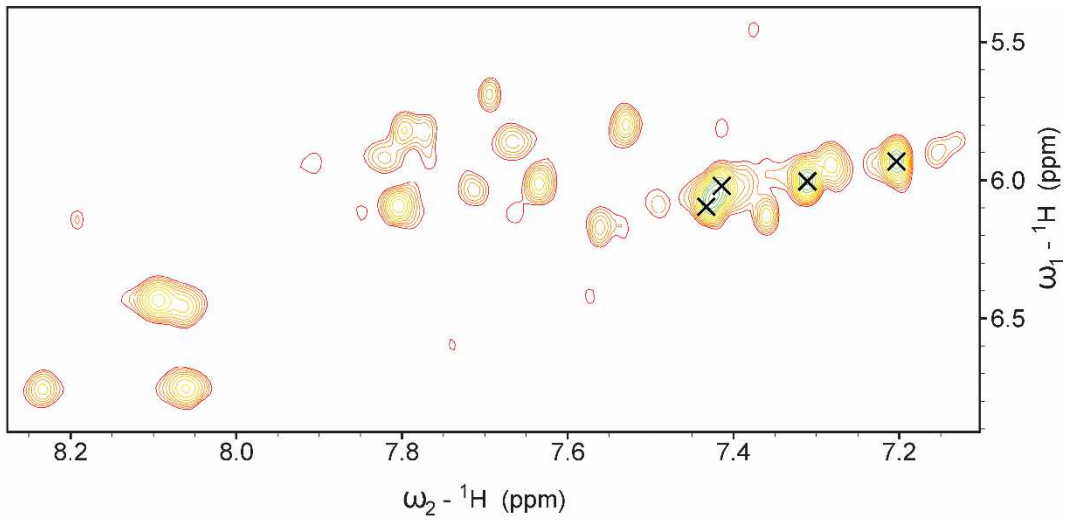


Figure S11. Aromatic-anomeric region of NOESY spectrum of T4 mutant in the presence of 70 mM KCl at pH 7.0, 25°C in H_2O:^2H_2O=9:1 at 600 MHz. Mixing time was 80 ms. Very intense intermolecular NOE cross-peaks between H8 and H1' protons are marked with cross.

Published – Greco, M. L.; Kotar, A.; Rigo, R.; Cristofari, C.; Plavec, J.; Sissi, C. *Nucleic Acids Res.* **2017**, 45 (17), 10132-10142

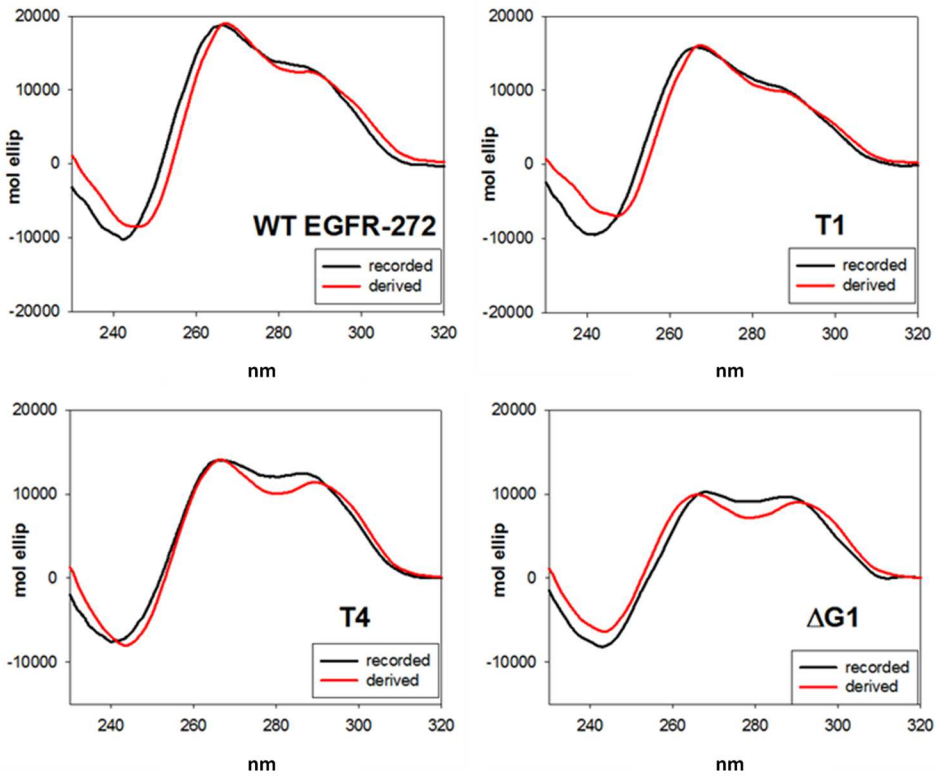


Figure S12. Description of experimentally acquired CD spectra (black lines) from analytical combination of the SVD derived spectra corresponding to folded form1 and form2 reported in Fig. (red lines).

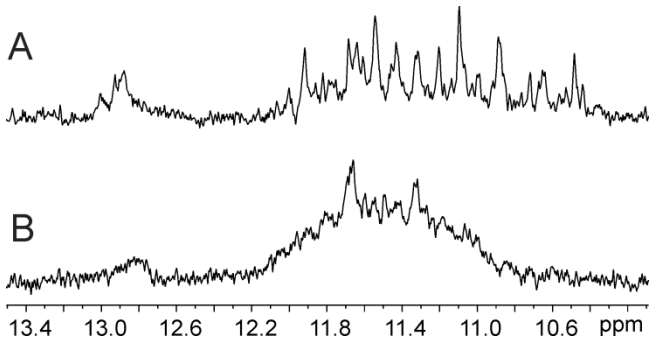


Figure S13. Imino region of ^{1}H NMR spectra of EGFR-272 (A) and C25T-G26T mutant (B). NMR spectra were recorded in H_2O:$^{2}H_2O$=9:1, at 0.1 mM concentration of the oligonucleotide per strand, at 100 mM KCl (EGFR-272) and 120 mM KCl (C25T-G26T), pH 7.0, 25°C at 600 MHz.

5. Conclusions

Structural characterization of G-quadruplexes is a tricky matter. This challenge has to be faced by applying complementary approaches in order to consider the multiple aspects of folding which are typical for these DNA arrangements. For instance, G-quadruplexes follow variegate folding pathways and, in general, present highly polymorphic behaviors. Even though all these features complicate the analysis of the system, the proper description of G-quadruplexes as pharmaceutical targets cannot be achieved by shelving these problems.

In this Ph.D.'s project, the spotlight has been focused on thermodynamic and kinetic issues of G-quadruplex folding considering c-KIT and EGFR promoters as case-studies.

The study of c-KIT promoter allowed to clear up many critical points in G-quadruplex folding and structures starting from the putative G-quadruplex forming sequences (PQS) search. PQS within the genome are found by applying bioinformatic tools which algorithms preferentially identify G-rich sequences containing at least three consecutive guanines in each G-run. Moreover, these approaches tend to divide sequences with more than four G-runs into many overlapping segments considering their tendency to form more than one G-quadruplex structure. The lesson learned by kit* G-quadruplex evidenced that both these assumptions are bias which could mislead a correct prediction. Indeed, kit* sequence nicely folds into a single stable G-quadruplex even though it contains six stretches of two consecutive guanines.

At the same time, the case of kit* G-quadruplex comes up with observations on the selection of the minimal G-quadruplex forming sequence. Usually, the shorter sequence containing all G-runs involved in the G-quadruplex is the studied model. kit* shows a 3'-tail that, even though not directly involved in G-quadruplex core, further stabilizes the overall arrangement via a pool of ancillary interactions and nicely drives the conformational selection towards only a single G-quadruplex structure. Removal of this 3'-tail leads to formation of multiple structures in solution. The emerging lesson is that behind the simple contribution of the 3'-tail in structure's stabilization through different interactions, the flanking element induces a marked structural rearrangement, even influencing the folding pathway of the oligonucleotide. Predominant effects of flanking bases have been identified even when the overhang is structured into a G-quadruplex, as occurs by elongating kit* at the 5'-end to include kit2 sequence. These closely clustered G-quadruplex units showed the ability to interact one each other shifting the quadruplex-duplex equilibrium towards the tetrahelical form. Since formation of multiple interacting

G-quadruplexes within the promoter has been observed for other oncogenes, as hTERT and c-Myb, this modular organization could be relevant in controlling transcriptional events. Thus, this arrangement might be an unprecedented suitable target for drug discovery and development. The reported data provide with essential tips to get a refinement of future search of putative G-quadruplex forming sequences, reinforcing the dignity of G-quadruplexes with two G-quartets which stability can be further improved by flanking bases or neighboring structures.

Looking at the interplay between kinetics and thermodynamics, kit2 and EGFR-272 G-quadruplexes provided with interesting insight. In both cases, the folding processes are complex events including many folded species. kit2 folding kinetics requires long timescale (about 2.5 hours) to achieve the thermodynamically stable forms. Its folding pathway comprises two kinetically favored folding intermediates which, even though transiently, could be involved in the regulation of c-KIT expression. Indeed, their half-life times are compatible with the timescale of transcriptional events, i.e. the rate of processing of RNA polymerase II or the residence time of transcription factors on their consensus sites. On the other side, EGFR-272 does not present any long-lived folding intermediate, but in potassium containing solution it is distributed into two forms, a kinetically-favored hybrid G-quadruplex and a thermodynamically-favored parallel one which are in dynamic equilibrium. As argued for kit2 G-quadruplexes, the hybrid form in EGFR-272 might be more physiologically relevant than the parallel G-quadruplex considering the timescale of folding. These two examples highlight the importance of the variable “time” which, sometimes, might help in discriminating between what is easy to study (states at thermodynamic equilibrium) and what is biologically relevant (the kinetically evolving systems).

To conclude, merging kinetics and thermodynamics, high-resolution data and complementary biophysical studies and placing the G-rich sequence in a more realistic context like the whole promoter is the correct strategy to improve the description of a G-quadruplex in terms of structure and function and, consequently, to understand the regulation of oncogene transcription. Only fulfilling this demanding task, we can support with a strong rationale the selection of effective G-quadruplex binders.