Frustrated folding of guanine quadruplexes in telomeric DNA repeats

Simone Carrino, Department of Chemistry, McGill University, Montreal

March 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of M.Sc. in Chemistry

Table of contents

List of figures and tables

1 Abstract

G-Quadruplexes (GQs) are nucleic acids secondary structures ubiquitous in our genome, covering fundamental roles in regulating processes such as gene expression and ribosomal protein synthesis. Among them, GQs forming at the end of telomeres have a prominent contribution in controlling telomerase activity and therefore directly affect cell aging. In the ~200 nt long telomeric overhang, a maximum of ~8 GQs can be folded, opening the question of whether this is the most probable configuration. Through a combination of kinetic and thermodynamic measurements on relatively small-scale systems (Tel_{8ext} and Tel_{12}), we first obtained quantitative microscopic parameters on their dynamics. We then used these parameters in combination with statistical mechanical simulations to extend our analysis to very long telomeric sequences. We found that the telomeric overhang might be a "frustrated" system, where the maximum number of folded GQs is never reached, leaving unused TTAGGG tracts in-between them. This frustration effect is also proportional to telomeric DNA length and is further enhanced at the very end of the sequence.

G-Quadruplexes (GQs) sont structurés secondaires d'acides nucléiques omniprésent dans notre génome, en recouvrant des rôles fondamentaux pour le control des procès comme l'expression des gènes et la synthèse ribosomiaux des protéines. Parmi cette catégorie, le GQs qui forment à la fine des télomères ont des implications très importants pour le control de la télomérase, en régulant directement l'âge des cellules. Dans la région télomérique a filament singulier (~200 nt longue), un maximum de ~8 GQs peuvent être formés, en ouvrant la question si ça pourrait être la configuration la plus probable. À travers des mesurément kinétiques et thermodynamiques sur des systèmes relativement petits (Tel_{8ext} et Tel_{12}) nous avons premièrement obtenu des informations quantitatives sur leur dynamique. Ensuite, nous avons utilisé cette information en combination avec des simulations statistiques mécaniques pour analyser des séquences plus longues. Nous avons déterminé que la région télomérique a filament singulier pourrait essentiellement être un système "frustré", où le numéro maximum de GQs n'est jamais atteint, en laissant entre eux des tronçons TTAGGG inutilisés. Cet effet de frustration et aussi proportionnel à la longueur de l'ADN télomérique et en plus est augmenté à la fine de la séquence.

2 Acknowledgements

First, I'd like to express my most sincere thanks to Prof. Anthony Mittermaier, for his help and guidance during the whole duration of the degree program. As well, I'd like to acknowledge Prof. Masad J. Damha and Prof. Hanadi Sleiman for granting me access to their respective research laboratories and equipment.

3 Contribution of authors

The two authors of this work, Simone Carrino and Prof. Anthony Mittermaier, have equally contributed to the realization of each chapter.

4 Introduction

Guanine quadruplexes (GQs from now on) are four stranded nucleic acid secondary structures adopted by G-rich sequences. While most of the GQ related discoveries happened within the last 30 years, a first landmark was reached well before that. In 1960, guanines in a concentrated GMP solution were found to interact via Hoogsteen base pairs; a crystal structure was experimentally determined, where planes composed by four Gs (later called G-tetrads) formed linear aggregates via stacking interactions[1]. Then in the late 1980s, it was determined that telomeric DNA regions had the tendency to form GQs in vitro[2]. More recently, a selective antibody allowed GQ imaging in human cells, and was used to demonstrate that GQs do not just fold in vitro but are indeed biologically relevant[3]. This

introduction is meant to give the reader a broad picture of 1) GQ structures 2) DNA synthesis and GQ characterization 3) GQs and their biological functions.

4.1 Elements of GQs structures

From a structural point of view, GQs are commonly formed by four stretches of sequential guanines (usually three) called G-tracts, separated by loops of variable length and composition[4]. Four Gs are required to form a G-tetrad, where these bases are arranged in a plane and interact via Hoogsteen base pairs; this means forming an overall four stranded structure composed by a stack of three G-tetrads[5]. Although this represents the most common situation, there are reports of unusual GQ DNA structures forming from shorter G-tracts; $(GGA)_8$ is a DNA sequence known to form two stacked GQ units each composed of a G tetrad, and a unique heptad made by the remaining Gs and As[6, 7]. Interestingly, GGA repeats are present in the c-myb promoter and GQ formation regulates gene expression[8]. On the other hand, G-tracts longer than three guanines are also very common in our genome[9] and can lead to isomer exchange as previously shown by our lab[10]. This highlights the intrinsic conformational polymorphism exhibited by GQs.

u[11][12]that [13]tand loop ,[14] Parallel GQs are composed of four strands all oriented in the same direction (5'-3' or vice versa). In an antiparallel GQ, strands alternate their orientation, whereas mixed GQs usually do not follow any criteria (for example three strands are 5'-3', one is 3'-5'). Guanines arranged in a G-tetrad and several GQs topologies are showed in

Figure 1. Stability andy[15]and cations and -[16] For instance, GQ formation is enhanced by the presence of metal cations, which coordinate either within one or between two G-tetrads, depending on the ionic radius. The most common are monovalent cations like Na^+ and K^+, whose relatively high physiological concentration is key to GQ folding in vivo[17], whereas Li^+ is too small and therefore does not enhance GQ stability[18]. Cations do not necessarily have to be monoatomic, for example NH_4^+ has been shown to fit within a G-tetrad and overall stabilize telomeric GQs[19]. In principle, this means that solution conditions need to be carefully adjusted, because in most cases this will directly affect GQ stability. The effect of divalent cations has also been extensively studied; for example, cations of the second group like Ca^{2+} and Mg^{2+} have also been shown to decrease GQ stability[20]. Transition metals cations like Co^{2+} and Mn^{2+} have been explored, and interestingly they can both induce GQ destabilization and changes in topology[21]; it must be noted that it's impossible to establish trends, as any change in stability and/or topology is highly dependent on the sequence and divalent cation under examination. When talking about GQs structures, another important aspect to mention is the use of crowding agents. In principle, *in vitro*

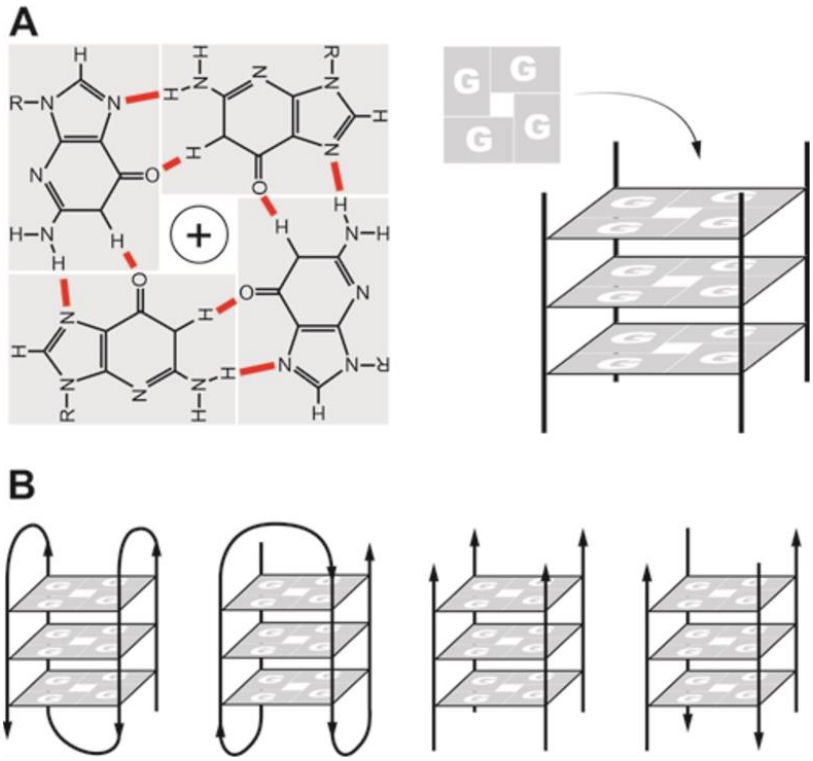

Figure 1. Structures of GQs. (A) four guanines interact via Hoogsteen base pairs and form G-tetrads. (B) GQs exhibit different topologies depending on strand orientation and molecularity

experimental setups for systems like proteins or nucleic acids, represent crude approximations of the crowded cellular environment. In order to simulate *in vivo* experimental conditions, a crowding agent like PEG_{200} can be added to the solution to make it more viscous. PEG_{200} has been shown to both induce changes in GQ topology and stability (both GQs and duplexes DNA)[22]. In the latter case, the effect can be rationalized as follows: GQ folding involves the burial of hydrophobic surface, meaning that hydrating H_2O molecules are now released in solution. Addition of PEG_{200} makes H_2O release and thus folding more favorable, producing an increase in the T_m value[23]. Interestingly, the opposite is sometimes true for duplex DNA sequences[24]; duplex folding is associated with H_2O hydration, meaning that addition of a crowding agent produces a negative ΔT_m. It follows that this effect has important biological implications; the double stranded region of human telomeres is formed by a G-rich strand $(TTAGGG)_n$ and a C- rich one $(AATCCC)_n$. This results in a competition between a duplex secondary structure, and the formation of GQs and i-motifs (G- rich and C-rich strands respectively) when the helix is unfolded. This means that a more crowded environment pushes equilibrium towards the latter[25]. In this regard, a very recent study showed that i-motifs favorably form in the double stranded regions of both promoters and telomeres[26].

4.2 Nucleic acids synthesis and characterization tools

The last 3-4 decades have seen the manufacture of progressively more capable nucleic acids synthesizers. This allowed scientists to: 1) bypass the necessity of isolating nucleic acids sequences from biological samples 2) design new and interesting artificial sequences.

Synthesizers are based on phosphoramidite chemistry[27], an automated synthesis process that in most cases only requires the user to supervise the process and make simple calibrations. Recent instruments are extremely high throughput and allow a relatively easy manufacture of up to ~100mer sequences, with acceptable yields for most applications. Each synthesis step (i.e. adding one base) is at most 99.5% efficient, meaning that some unreacted oligo carries over; the final product will be mixture of the target sequence (N total bases) and impurities (N-1, N-2…). Very good purity levels can then commonly be achieved via methods such as HPLC and PAGE. HPLC (or High Pressure Liquid Chromatography) usually separates impurities based on different hydrophobicity or charge, by using reverse phase and ion exchange columns respectively. PAGE (or Polyacrylamide Gel Electrophoresis) is a denaturing gel technique that easily differentiates between oligo fragments of different masses, it has by far the highest resolution. Nucleic acid folding and stability can then be characterized via many different tools. We will discuss three main examples of temperature-based methods: UV-Vis and CD (circular dichroism) spectroscopy, and differential scanning calorimetry (DSC). UV-Vis spectrophotometers are widely available in most laboratories and can often be upgraded with a Peltier block to change sample temperature in the ~2-100°C range. The small (μM) amount of sample required and the availability of multi-cell blocks make UV-Vis an extremely high throughput and convenient tool. For GQs, single wavelength (usually 295 nm) thermal melting experiments are based on the hyperchromic effect exhibited during GQ folding[28]. Usually, this gives a sigmoidal curve characterized by two baselines (folded and unfolded), and by a transition region of variable steepness. Thermal melts for two-state processes (only folded and unfolded states) can be easily analyzed to give key information, such as the melting temperature (50% folded) T_m, and the enthalpy,

entropy, and free energies of folding[28]: ΔH, ΔS and ΔG. In recent years, multidimensional techniques have also been developed to analyze more complex, multi-state GQ folding processes. Multidimensional means that for each temperature value, the absorbance is measured over a broad range of wavelengths instead of one. In practice, this means having a collection of temperature dependent absorbance spectra, one for each wavelength[29]. Mathematically, this dataset can be analyzed via SVD (singular value decomposition), a technique that finds the largest contributions to variations across the dataset. In terms of GQ folding, these are the spectral components directly related to the states in the multi-state process[29]. CD spectroscopy is another important tool; its application to GQs is based on the chirality originating from Gs stacking in different G-tetrads. There are two main applications: 1) collecting one- and multi-dimensional GQ thermal melts, much like UV-Vis 2) getting insight on GQ topology. In the second case, the different GQ topologies mentioned before give different CD spectra[30], usually acquired in the 200-300 nm range. Recently, a method based on spectral data clustering has made it possible to access more quantitative GQ structural information[31]; it involves calculating the spectral contribution of structural elements, such as different modes of base stacking and different loop arrangements to the total experimental spectrum. DSC represents another widely used tool that allows the quantitative thermodynamic characterization of GQs[32], other DNA secondary structures and other biomolecules. The instrument measures the heat necessary to raise the sample temperature at a constant rate (or in other words the temperature-dependent heat capacity), with respect to a reference solution, hence the term differential. Both the sample and reference cell are maintained at the same temperature, with extremely high accuracy. One drawback of this technique is the relatively high sample requirements; for DNA GQs, a

~100 μM, 1 ml solution is usually required to perform an experiment. Conversely, there are several advantages unique to this tool: 1) due to the sample cell being pressurized at ~3 atm, it's possible to reach temperatures up to 120-125°C and therefore analyze particularly stable biomolecules 2) the presence of spectroscopically active species can be ignored 3) an accurate estimate of the folding ΔC_p can be obtained from the dataset. DSC profiles that look symmetric usually indicate simple, two-state folding; conversely, asymmetry is usually a sign that a more complex folding process is present[33]. In the second case, model free deconvolution of DSC data gives populations of each folding intermediate[34].

4.3 The biological relevance of GQs

GQ forming sequences are widely present in the human genome, and they play a critical role in regulating various biological processes, from cell aging[35] to gene expression[36] (Figure 2). For instance, GQs can form in the promoter regions of genes, such as MYC or KRAS[37]; here, an intramolecular GQ folding within the double stranded DNA region acts like a roadblock, effectively causing RNA and DNA polymerases to stall during transcription and replication respectively[35]. A similarly inhibiting effect is produced by RNA GQs folding in mRNA, thereby regulating protein synthesis by ribosomes[38] (Figure 2). The last, perhaps

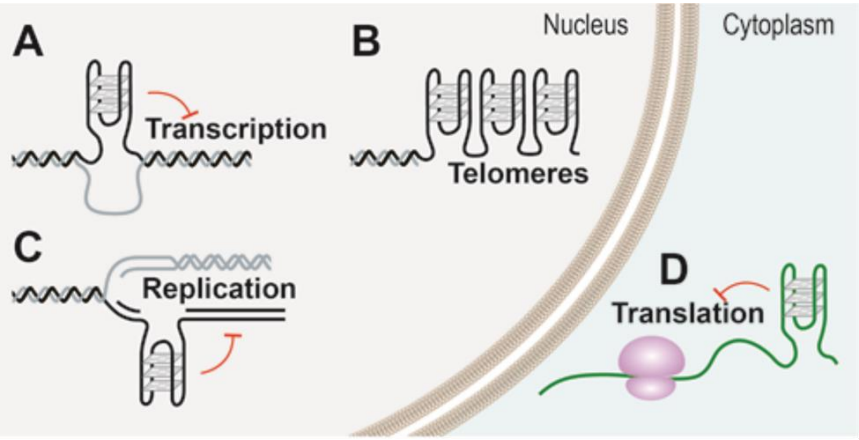

Figure 2. Roles of GQs in the human genome. (A) regulation of gene transcription. (B) GQ formation in telomeric overhangs. (C) Regulation of DNA replication. (D) Regulation of mRNA translation

most widely known example of physiologically occurring GQs is represented by telomeric GQs. Telomeres are repetitive DNA terminal regions of chromosomes, whose role is to protect the genetic material[39]. Telomere length is linked to cell aging: they are irreversibly shortened after each replication round, ultimately determining cell senescence. The enzyme telomerase can elongate telomeres and prevent cell aging; for example, cancer cells are known to overexpress telomerase, hence their proliferative potential[40]. Conversely, healthy cells are characterized by extremely low telomerase activity, hence the regular cell aging process associated with progressive telomere shortening. The ~200 nt single-stranded telomere overhang is composed of tandem TTAGGG (Tel) repeats, and can therefore fold into several adjacent GQs under physiological conditions[41]. GQ formation at the 3' end of telomeres has been shown to inhibit telomere extension by both telomerase[42] and the alternative lengthening of telomeres[43] (ALT). GQs may also control protein binding to the telomere overhang, notably helicases and telomerase activity regulators, such as POT1 and SSB1[44]. Figure 3 shows some of the processes regulated by the formation of GQs in telomeric overhangs. Telomeric GQs are therefore regarded as potential drug targets; many ligands have been designed, for the most part stacking on or between G-tetrads to make folding more favorable[45], thereby inhibiting telomerase and potentially reversing cancer cell

immortalization. Equally important are ligands designed towards stabilizing GQs in promoter regions of genes, and more recently in viruses such as HIV-1[46] and HSV-1. During the last two decades, many GQs ligands have been developed; some widely known examples are depicted in Figure 4. A complete survey is beyond the scope of this introduction; therefore, we will discuss two main categories of ligands as an example. The first category is represented by cationic porphyrins; these molecules are often characterized by a planar, aromatic surface. This allows them to favorably intercalate between two adjacent G- tetrads or on the top or bottom tetrad[47], via formation of π-π interactions. The binding free energy ΔG_b depends on the specific porphyrin and GQ forming sequence used; nevertheless, a rule of thumb can be established, because ΔH_b and $-T\Delta S_b$ almost always have negative values[48]. This can be rationalized in terms of the formation of new contacts (enthalpic contribution) and H_2O displacement from the DNA surface (entropic contribution). A second category of ligands is represented by those binding GQs grooves; within this category, the main examples are represented by dystamicin A and derivatives. These polyamides were originally known as duplex DNA binders, then found to exhibit high affinity for GQs and inhibit telomerase activity[49]. While trends cannot be established in terms of binding modes and stoichiometry,

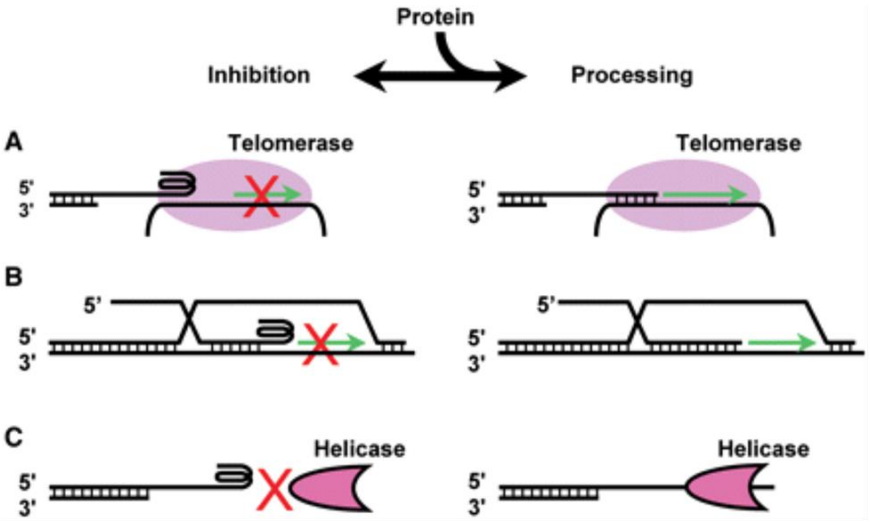

Figure 3. BIological roles of GQs in telomeric overhangs. (A) Telomerase inhibition. (B) ALT inhibition. (C) Inhibition of unwinding by helicases.

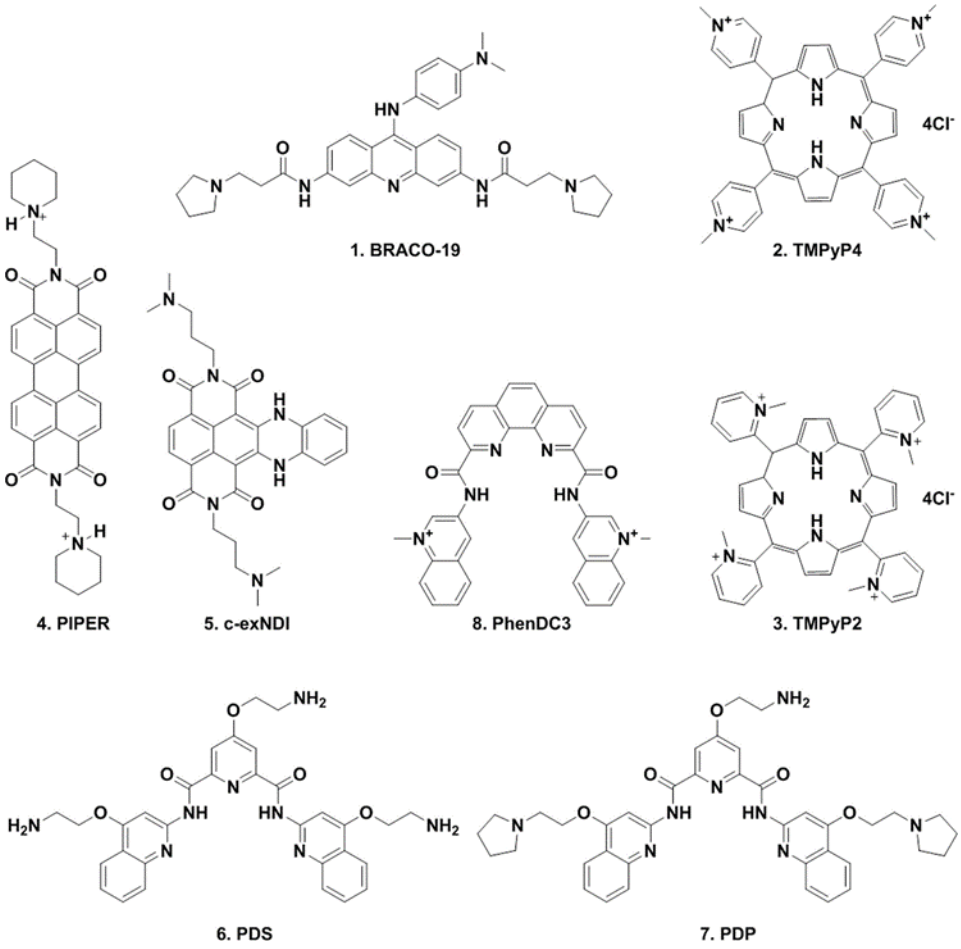

Figure 4. Examples of GQs ligands[1]

thermodynamic measurements are almost always consistent with an entropically driven process[50]. Again, this suggests favorable H_2O displacement from the grooves of the hydrated GQ. Ligand binding to GQs can be analyzed via both spectroscopic and calorimetric techniques. In the second category, ITC[51, 52] (isothermal titration calorimetry) is perhaps the most efficient and informative tool. Small (5-10 µl) volumes of a ligand solution are successively injected from a syringe into a cell containing the macromolecule, until saturation is reached. Heat can either be released or absorbed during the binding event, and the signal is measured as the heater power necessary to keep the cell temperature constant (hence isothermal). A good binding isotherm can be obtained with relatively low amounts of sample, then subsequently examined to accurately estimate

[1] G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy
Nucleic Acids Res. 2018;46(7):3270-3283. No copyright clearance required.

binding parameters (ΔH_b, ΔS_b, ΔG_b) and stoichiometry. When going from *in vitro* to *in vivo* conditions however, a major obstacle in designing these ligands is represented by their promiscuity, meaning that they bind GQs with high affinity but low specificity. Having previously mentioned how GQs are ubiquitous in the human genome, it's clear how difficult it is to target a specific GQ while at the same time preventing unwanted collateral effects. This is one of the main reasons why as of today, no ligand has made it past Phase II in clinical trials[53].

4.4 The dynamics of long telomeric DNA strands

A great deal is now known about how nucleotide sequence and solution conditions govern the stability[17, 18], topology[16], and folding pathways of DNA strands containing four G-tracts. For instance, it has been shown that telomeric GQs at the extreme 5' and 3' ends of DNA strands are more stable than those in the middle, with important implications for telomerase inhibition[54]. However, much less is known about the highly repetitive nature of the single-stranded 3' terminus with roughly 32 Tel repeats and how it regulates DNA folding and dynamics. There are indications that GQ folding in the context of the full telomeric 3' overhang may differ fundamentally from folding of individual Tel_4 GQs. For example, evidences suggest that multiple telomeric GQs form a beads-on-a-string structure[55], in opposition to a more common wire-like structure where GQs are vertically stacked. Petraccone et al. applied a deconvolution analysis to differential scanning calorimetric folding data for a Tel_{12} DNA sequence, showing that consecutive folded GQs are less stable than GQs formed in isolation. An unfavorable coupling energy associated with the folding of

adjacent GQs[56] was measured, although the strength of this interaction for GQs internal to the telomere was unclear. In addition, the multiplicity of different partly folded states available to long TTAGGG repeats may itself have a strong influence on GQ folding. TGGE (Temperature Gradient Gel Electrophoresis) experiments qualitatively showed that a Tel_8 DNA sequence can form partially folded states where only one GQ is folded[57]. Furthermore, a combination of AFM imaging and statistical calculations showed that a Tel_{16} DNA sequence rarely forms the maximum number of folded GQs (4) and instead populates states with a mixture of unfolded G-tracts and 2 to 3 folded GQs at various positions along its length[58]. That said, the extrapolation of these arguments to longer sequences (for example Tel_{32}) is not straightforward as the number of possible partly-folded states exponentially increases with the length of the sequence.

In order to better understand the intrinsic dynamics of the telomeric 3' DNA overhang, we have determined both thermodynamic and kinetic coupling effects for the formation of adjacent GQs. We employed a mutational trapping approach previously developed by our lab[10] in which we globally fit folding data for sets of mutants (Tel_{8ext}-Tel_{12}) where individual GGG tracts were systematically replaced with GTG. The extracted parameters were then used to perform closed-form statistical mechanical calculations, Monte Carlo, and kinetic simulations in order to predict the distribution of folded GQs throughout the 3' overhang and describe how this distribution changes over time. Interestingly, our calculations predicted that the overwhelming abundance of partly-folded states for long telomeric sequences results in an increase of about 9-fold in the probability that an individual G-tract will be unfolded in the middle of a telomere, compared to G-tracts in a short Tel_4 DNA sequence under identical conditions. Furthermore, this effect was highly

position-dependent, such that an oscillating pattern of more and less folded G-tracts appeared in about the terminal 20 *Tel* repeats at the 3' end. This natural tendency to expose unfolded Tel at certain positions near the 3' end of the telomere could have consequences on the targeting of single-stranded DNA binding proteins that stabilize telomeric structures[59].

5 Materials and methods

5.1 Sample preparation

DNA samples were produced on a Mermade 6 synthesizer (Bioautomation, USA) using reagents from Chemgenes Corporation (USA), then cleaved from the CPG with AMA (1:1 ammonium hydroxide and methylamine). $(TGGG)_8T$ samples were purified with Glen-Pak columns (Glen Research, USA); *Tel* samples were purified via ion exchange HPLC on an Agilent 1200 Infinity Series (Agilent Technologies), then desalted with Glen Gel-Pak columns (Glen Research, USA). Purity of all oligos was estimated via LC-ESI-MS on a Bruker Maxis Impact mass spectrometer (Bruker, USA). Oligos were then resuspended in milliQ water and their concentration measured using a NanoDrop Lite (Thermo Fisher Scientific, USA). All sequences are listed in Table 1, with a code that helps their visual recognition.

5.2 Thermal hysteresis UV-Vis melts

All UV-Vis experiments were performed on a Cary 100 Bio spectrophotometer (Agilent, USA), with the wavelength set to 295 nm. $(TGGG)_8T$ and mutants at 5 μM in Buffer B (5 mM

LiH_2PO_4, 5 mM Li_2HPO_4 and 1 mM KCl, pH=7.00) were scanned at 2 and 3 °C/min, between 20 and 95 °C. Tel_{8ext} and mutants at 3 μM in buffer C (5 mM KH_2PO_4, 5 mM K_2HPO_4 and 60 mM KCl, pH=7.00) were scanned at 2.5 and 4 °C/min, between 10 and 80 °C. An equilibration period was introduced right before the first scan for complete unfolding, and then at the end of each one. In all cases, a layer of mineral oil was applied to each sample to minimize evaporation, and when necessary a flow of nitrogen was used to prevent condensation.

5.3 Isothermal Folding UV-Vis

Isothermal folding experiments were performed at four different temperatures (10, 15, 20, 25 °C). Tel_{8ext} at 3 μM in buffer C was incubated for 5 minutes at 90 °C for complete unfolding. Temperature was monitored with a Cary Series II (Agilent, USA) probe and then changed to the target value at the fastest rate possible, calculated by taking timed screenshots. Measurement, the absorbance at 295 nm, was monitored as soon as the target temperature was reached.

5.4 Thermal equilibrium UV-Vis melts

Tel12 and mutants at 3 μM in Buffer D (10 mM KH_2PO_4, 10 mM K_2HPO_4 and 110 mM KCl, pH=7.00) were pre-unfolded at 90 °C and then scanned between 20 and 90 °C at 1 °C/min, or at 0.2 °C/min in the case of sequences forming two or three GQs to minimize thermal hysteresis.

5.5 Thermal hysteresis and isothermal folding global fitting (Tel_{8ext} and mutants)

The Tel_{8ext} (xx|||||||xx) kinetic pathway was modeled as follows (see Figure S1 for the $(TGGG)_8T$ case and Table 1 for sequence nomenclature):

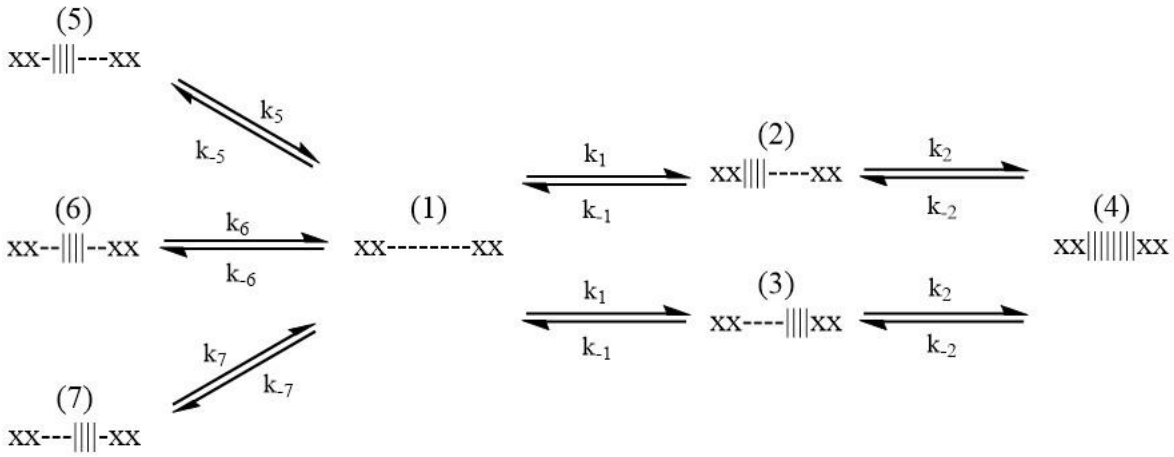

Figure 5. xx|||||||xx folding pathway

Each state is described by folding and unfolding rate constants k_F and k_U, and their temperature dependence described by activation energies E_F and E_U and the Arrhenius equation, such that

$$k_F = k_{0F} e^{\left(\frac{-E_F}{R(T_0 - T)}\right)} \quad (1)$$

$$k_U = k_{0U} e^{\left(\frac{-E_U}{R(T_0 - T)}\right)} \quad (2)$$

where T_0 is an arbitrary reference temperature. In order to individually represent each partially folded state, we mutate each unused G-tract from "GGG" to "GTG" (for example, xx--||||--xx becomes xxxx||||xxxx); the latter will be called a tract knockout mutant, and we assume its kinetics to be identical to the same GQ folding in the WT (wild-type) sequence.

Three types of data are analyzed here: 1) thermal hysteresis curves for the five knockout mutants, 2) thermal hysteresis curves for xx|||||||xx 3) isothermal folding curves for xx|||||||xx. Each set of data is fitted to the same set of microscopic rate constants, but simulated curves are calculated differently in each case.

First, thermal hysteresis curves for each tract knockout mutant (xx|||xxxxxx, xxx|||xxxxx, xxxx||||xxxx, xxxxx||||xxx and xxxxxx||||xx) are calculated, according to a two-state model[28]:

$$S^{fit}(T) = A_F \cdot f_F + A_F \cdot f_U \text{ (3)}$$

$$\text{where } A_F = m_F \cdot T + b_F \text{ (4)}$$

$$\text{and } A_U = m_U \cdot T + b_U \text{ (5)}$$

are the folded and unfolded baselines. Each baseline is linearly dependent on temperature and have individual intercepts, whereas the slopes are constrained to a single value for all the folding and unfolding baselines respectively. Being based on a two-state model, the time dependent concentration for each mutant is calculated via the following ODE:

$$\frac{d[Folded]}{dt} = k_F[Unfolded] - k_U[Folded] \text{ (6)}$$

The MATLAB solver *ode15s* was always used here. For example, for xxxx||||xxxx $[Folded]$ and $[Unfolded]$ are replaced by xxxx||||xxxx and xxxx----xxxx respectively. Multiplying the equation by the scan rate dT/dt then transforms d/dt into d/dT. Lastly, fractions (f) are obtained as follows (again for xxxx||||xxxx):

$$f_{xxxx||||xxxx} = \frac{[xxxx||||xxxx]}{[xxxx||||xxxx]+[xxxx----xxxx]} \text{ (7)}$$

$$f_{xxxx----xxxx} = 1 - f_{xxxx||||xxxx} \text{ (8)}$$

We then use a multi-state model is used to calculate xx|||||||xx thermal hysteresis curves:

$$S^{fit}_{xx|||||||xx}(T) = \frac{1}{C^T}\big(A_{(4)}[xx|||||||xx] + A_{(1)}[xx - - - - - - - - -xx] + A_{(2)}[xx||||- - - - xx] + A_{(3)}[xx - - - -|||xx] + A_{(5)}[xx - ||||- - -xx] + A_{(6)}[xx - -||||- - xx] + A_{(7)}[xx - - - ||||-xx]\big) \text{ (9)}$$

Where C^T is the total concentration, $A_{(i)}$ is the temperature dependent absorbance for each state. This time, a system of 7 ODEs is solved to give the concentration of each of the seven states:

$$\frac{d[xx|||||||xx]}{dT} = k_2([xx|||| - - - -xx] + [xx - - - -||||xx]) - 2k_{-2}[xx|||||||xx] \text{ (10)}$$

$$\frac{d[xx||||----xx]}{dT} = k_1[xx - - - - - - - - -xx] + k_{-2}[xx|||||||xx] - [xx|||| - - - -xx](k_2 + k_{-1}) \text{ (11)}$$

$$\frac{d[xx----||||xx]}{dT} = k_1[xx - - - - - - - - -xx] + k_{-2}[xx|||||||xx] - [xx - - - -||||](k_2 + k_{-1}) \text{ (12)}$$

$$\frac{d[xx-||||---xx]}{dT} = k_5[xx - - - - - - - - -xx] - k_{-5}[xx - |||| - - - xx] \text{ (13)}$$

$$\frac{d[xx--||||--xx]}{dT} = k_6[xx - - - - - - - - -xx] - k_{-6}[xx - -|||| - -xx] \text{ (14)}$$

$$\frac{d[xx---||||-xx]}{dT} = k_7[xx - - - - - - - - -xx] - k_{-7}[xx - - - |||| - xx] \text{ (16)}$$

$$\frac{d[xx--------xx]}{dT} = k_{-6}[xx--||||--xx] + k_{-5}[xx-||||---xx] + k_{-7}[xx---||||-xx] + k_{-1}([xx||||-----xx] + [xx----|||xx]) - [xx---------xx](k_6 + k_5 + k_7 + 2k_1) \quad (17)$$

Once obtained the concentrations, f values are calculated accordingly.

Third, using the same kinetic parameters, a time dependent version of (10) gives the $f_{xx|||||||xx}$ values over time that are fitted to the four normalized isothermal folds: first, xx|||||||xx cooling from 90 °C to the target temperatures is simulated. Then, $f_{xx|||||||xx}$ corresponding to each target temperature is used as the initial value for the isothermal folding simulation. Time span, target temperatures and cooling rates were the same as in the isothermal folding experiment. To sum it up, thermal hysteresis curves for each trapped state are calculated as (3), xx|||||||xx curves as (9) and isothermal folds using a time dependent version of (10). In the global fit, all these sets of data are analyzed simultaneously. Parameters are then varied to minimize the global RSS function, which is calculated as follows:

$$RSS_{melt} = \sum_{q=1}^{M}\left[\sum_{j=1}^{4}\sum_{i=1}^{6}(S_{ijq}^{exp}(T) - S_{ijq}^{fit}(T))_{heat}^2 + \sum_{j=1}^{4}\sum_{i=1}^{6}(S_{ijq}^{exp}(T) - S_{ijq}^{fit}(T))_{cool}^2\right] \quad (18)$$

$$RSS_{iso} = \sum_{p=1}^{N}\sum_{k=1}^{4}(S_{kp}^{exp}(t) - S_{kp}^{fit}(t))^2 \quad (19)$$

$$RSS_{total} = RSS_{melt} + \frac{RSS_{iso}}{10000} \quad (20)$$

Where a suitable scaling factor was applied to RSS_{iso}, to adjust its weight with respect to RSS_{melt}. The subscript i, runs over the wild-type and all tract-knockout mutants, j, runs over all scan rates, k runs over the four temperatures used, p and q are referred to the number of time and temperature points for each curve respectively.

5.6 Thermal equilibrium global fitting (Tel_{12} and mutants)

Tel_{12} can in principle form many partially folded states (Chapter 3.1). In order to individually represents them, we used the same mutational approach seen for the thermal hysteresis global fitting; each tract knockout mutant is now assumed to have the same thermodynamics as the corresponding partially folded state in the wild-type sequence. This means that for example, $\frac{[-||||-------]}{[----------]} = \frac{[x||||xxxxxxx]}{[x----xxxxxxx]}$. Again, several sets of data (sequences) are analyzed, differing in the number of folded GQs (one, two or three). Their calculated melting curves are described by different models sharing the same set of thermodynamic parameters.

First, a two-state model (3) was used to calculate thermal melts of knockout mutants forming a single GQ, with temperature independent folding parameters ΔH_F and ΔS_F, where

$$\Delta G_F = \Delta H_F - T\Delta S_F \quad (21)$$

$$K_F = e^{\frac{-\Delta G_F}{RT}} \quad (22)$$

$$ff = \frac{K_F}{K_F + 1} \quad (23)$$

$$f_U = 1 - f_F \text{ (24)}.$$

Again, folded and unfolded baseline are described as (4) and (5).

Then, calculated thermal melts of sequences forming multiple GQs are dependent on the parameters from the individual partially folded states forming them; additionally, they depend on a single set of cooperativity parameters ΔH_c and ΔS_c ($\Delta G_c = \Delta H_c - T\Delta S_c$) if GQs fold adjacently. In the latter case, a partition function Z for neighboring GQs is then written, based on the stability of each contributing partially folded state. For example,

$$Z_{--|||||||--} = 1 + K_{--||||------} + K_{---||||-----} + K_{----||||----} + K_{-----||||---} + K_{------||||--} + K_{--|||||||--} \text{ (25)}$$

$$\text{where } K_{--|||||||--} = e^{\frac{-(\Delta G_x)}{RT}} \text{ (26)}$$

$$\Delta G_x = \Delta G_{--||||------} + \Delta G_{------||||--} + \Delta G_c \text{ (27)}$$

$$f_{--|||||||--} = \frac{K_{--|||||||--}}{Z_{--|||||||--}} \text{ (28)}$$

$$\text{and } f_{----------} = \frac{1}{Z_{--|||||||--}} \text{ (29)}$$

The various Ks are folding equilibrium constants; again, $S^{fit}_{--|||||||--}$ was calculated as (9). The same reasoning was followed for the other configurations whose experimental melts form two nearest neighbor GQs, where Z is again a combination of the K values for each partially folded state.

The partition function Z for Tel_{12} (also $(|)_{12}$) is then described as follows:

$$Z_{(|)_{12}} = 1 + K_{(|)_{12}} + \sum_{i=1}^{N} K_i \quad (30)$$

where N is the total number of partially folded states,

$$K_{(|)_{12}} = e^{\frac{-(\Delta G_x)}{RT}} \quad (31)$$

$$\Delta G_x = \Delta G_{||||--------} + \Delta G_{----||||----} + +\Delta G_{--------||||} + 2\Delta G_c \quad (32)$$

For Tel_{12}, the thermal melting curve is calculated as

$$S^{fit}_{(|)_{12}}(T) = A_{(|)_{12}} \cdot f_{(|)_{12}} + A_{(-)_{12}} \cdot f_{(-)_{12}} + \sum_{i=1}^{9} \left(A_{Fi} - A_{Ui} + A_{(-)_{12}}\right)_{1GQ} \cdot f_i$$

$$+ \sum_{j=1}^{15} \left(A_{Fj} - A_{Uj} + A_{(-)_{12}}\right)_{2GQ} \cdot f_j \quad (33),$$

where i and j refer to partially folded states forming one or two GQs respectively; A_F and A_U are their folded and unfolded baselines calculated as (4) and (5), $(|)_{12}$ and $(-)_{12}$ refer to a fully folded and unfolded Tel_{12} respectively. $f_{(|)_{12}}$, $f_{(-)_{12}}$, f_i and f_j are calculated with respect to $Z_{(|)_{12}}$. Again, simultaneous fit to the whole dataset is performed, as parameters are optimized to minimize the RSS function. The RSS for the global fit is calculated as:

$$RSS = \sum_{j=1}^{N} \sum_{i=1}^{q} (S_i^{exp}(T) - S_i^{fit}(T))^2 \quad (34)$$

with q being the total number of experimental melting curves, N the number of temperature points.

Not every possible partially folded state is represented by an experimental melting curve; ----------|||| and -------|||- are assumed to be mirror images of ||||-------- and -||||------- respectively. When two GQs are forming, their curve is based on the thermodynamic parameters and absorbance of their respective components. For example, $\Delta G_{\text{-|||||||---}}$ (two sequential GQs starting with the second G-tract at the 5' end) is calculated as $\Delta G_{\text{-||||------}} + \Delta G_{\text{----||||---}} + \Delta G_c$ and its contribution to $S^{fit}_{(|)_{12}}$ as

$$((A_{Fm} - A_{Um}) + (A_{Fn} - A_{Un}) + A_{(-)_{12}}) \cdot \frac{K_{\text{-|||||||---}}}{Z_{(|)_{12}}} \quad (35)$$

Where m and n refer to -||||------- and -----||||--- respectively.

5.7 Monte Carlo simulation

The simulation is based on the Metropolis Hastings model. For every step, a random G-tract between 1 and L-3 is chosen, where L is the total number of G-tracts (32 in this case). If the 5' end of a folded GQ, or a 5' end of a 4 contiguous unfolded G-tracts is selected then a probability test is performed, and a change (folded to unfolded or vice versa) made if $rand \leq e^{\frac{-\Delta G_{test}}{RT}}$, where $rand$ is a random number between 0 and 1 and ΔG_{test} (at T=310K) is the difference in free energy between the final and initial states. $\Delta G_{test} = \pm(\Delta G_{uf} + n\Delta G_c)$, where $\Delta G_{uf} = \Delta G_{||||xxxxxxxx}$ for GQs ||||(.)$_{\text{L-4}}$, and (.)$_{\text{L-4}}$||||, $\Delta G_{uf} = \Delta G_{x||||xxxxxxx}$ for -||||(.)$_{\text{L-5}}$

and $(.)_{L\text{-}5}||||\text{-}$, $\Delta G_{uf} = \Delta G_{mid}$ for $(.)_{M}||||(.)_{L\text{-}4\text{-}M}$ where 1<M<L-6 and (.) is (-) or (|) and ΔG_{mid} is the average of $\Delta G_{xx||||xxxxxx}$, $\Delta G_{xxx||||xxxxx}$, $\Delta G_{xxxx||||xxxx}$, $\Delta G_{xxxxx||||xxx}$, and $\Delta G_{xxxxxx||||xx}$. For an isolated GQ with no folded G-tracts on either side, n=0. For a singly contiguous GQ with another folded GQ immediately to one side, n=1. For a doubly contiguous GQ with folded GQs to both sides, n=2.

5.8 Kinetics simulation

For every iteration, the sequence of length L was analyzed from one end to the other in a sequential manner, and a probability test was executed each time the 5' end of a folded GQ or set of four unfolded G-tracts was selected. A change is accepted if $rand \leq k_{test} \cdot \Delta t$, where Δt is a sufficiently low time step, such that $k_{test} \cdot \Delta t \ll 1$. The rate constants for folding and unfolding were set to $R_{cf}k_f$ and $R_{cu}k_u$, where $k_f = (k_5 + k_6 + k_7)/3$ and $k_u = (k_{-5} + k_{-6} + k_{-7})/3$; $R_{cf} = 1,\ k_2/k_1$, and $2k_2/k_1$ and $R_{cu} = 1,\ k_{-2}/k_{-1}$, and $2k_{-2}/k_{-1}$ for isolated, singly contiguous, and double contiguous GQs, respectively. This is only valid for Tel_{8ext}; for $(TGGG)_8T$, the difference in stability between ||||xxxx and xxxx|||| requires the introduction of additional parameters k_2^*, k_{-2}^*, k_1^* and related activation energies. k_{-1}^* is then calculated from them instead of being independently optimized, ensuring that the two possible transitions from -------- to |||||||| (and viceversa) are thermodynamically identical. Compared to Tel_{8ext}, this means having two R_{cf} and two R_{cu} values.

5.9 McGhee and Von Hippel model (cooperative case)

The statistical mechanical model of McGhee and von Hippel (MGVH) describes ligand binding to a one-dimensional lattice, where each ligand occupies multiple consecutive lattice points[60]. This situation is formally identical to GQ folding in telomeric repeats, since a folded GQ occupies multiple (4) consecutive G-tracts. The MGVH model accounts for the fact that inter-GQ gaps consisting of fewer than 4 G-tracts cannot fold and are therefore trapped in an unfolded state. At equilibrium there are 3 classes of folded GQ: doubly contiguous GQs that are immediately adjacent to folded GQs on both 5' and 3' sides, singly contiguous GQs with a folded GQ on one side but not the other, and isolated GQs that are not immediately adjacent to folded GQs on either side. The folding behavior of the lattice is governed by the equilibrium constant $K = (\#\ isolated\ GQs)/(\#\ isolated\ sites)$, where an isolated site is a stretch of 4 unfolded G-tracts which, if folded, would produce an isolated GQ. In addition, folding is modulated by the cooperativity parameter ω, which describes the increase or decrease in the tendency for GQ to form adjacently to another GQ, compared to at an isolated location. Positive and negative cooperativity are indicated by $\omega > 1$ and $\omega < 1$, respectively, such that $K_{sc} = \omega K$ and $K_{dc} = \omega^2 K$, where K_{sc} and K_{dc} are the folding equilibrium constants for doubly and singly contiguous GQs, respectively. The MGVH model considers a region containing N G-tracts that is located within an infinite lattice of G-tracts. Given an average number of folded GQs (B) within the region of N G-tracts, the model allows one to calculate the number of sites (4 contiguous unfolded G-tracts) that would produce doubly, singly, and isolated GQs, if folded. In order to predict the folding behavior of tandem G-tracts, one must find the value of B that is consistent with the user-defined values of K and ω. This is

accomplished as follows: According to the MGVH model, the average numbers of isolated, singly- and doubly-contiguous sites are given by:

$$s_{isol} = \left(\frac{b_n f \cdot ff^{(n+1)}}{fb_1}\right) \cdot B \ (36)$$

$$s_{sc} = (2 \cdot b_n f \cdot ff^n) \cdot B \ (37)$$

$$s_{dc} = \left(b_n f \cdot ff^{(n-1)} \cdot fb_1\right) \cdot B \ (38)$$

where

$$b_n f = \frac{(n-1) \cdot v - 1 + R}{2 \cdot v \cdot (\omega - 1)} \cdot B \ (39)$$

$$fb_1 = \frac{(n-1) \cdot v - 1 + R}{2 \cdot (\omega - 1) \cdot (1 - n - v)} \cdot B \ (40)$$

$$ff = \frac{(2 \cdot \omega - 1) \cdot (1 - n \cdot v) + v - R}{2 \cdot (\omega - 1) \cdot (1 - n \cdot v)} \ (41)$$

$$R = \sqrt{(1 - n(n+1) \cdot v)^2 + 4\omega v \cdot (1 - n \cdot v))} \ (42)$$

n=4 is the number of G-tracts occupied by a folded GQ, and ν is the folding density, B/N. It follows that the average numbers of isolated, singly- and doubly-contiguous folded GQs are related to the average numbers of the corresponding unfolded sites by the equilibrium constant K and cooperativity parameter ω according to:

$$B_{isol} = K \cdot s_{isol} \ (43)$$

$$B_{sc} = K \cdot \omega \cdot s_{sc} \ (44)$$

$$B_{dc} = K \cdot \omega^2 \cdot s_{dc} \ (45)$$

Finally, self-consistency requires that $B = B_{isol} + B_{sc} + B_{dc}$. We used non-linear least-squares optimization, varying the value of B to minimize the target function $(B_{isol} + B_{sc} + B_{dc} - B)^2$. Thus, for any values of K and ω, the MHVH model yields the average number of isolated, singly-, and doubly-contiguous GQs that would be found in a region of N G-tracts. Additionally, $1/b_n f$ gives the average number of sequential folded GQs.

6 Results

6.1 Tel_{12} thermodynamics

We first examined the folding thermodynamics of Tel_{12} (or $(|)_{12}$), a twelve G-tract-containing telomeric sequence in order to analyze the folding of three sequential GQs. Our goal was to determine how the stability of a GQ depends on its position relative to the 5' and 3' termini and to characterize the interaction energies between neighboring GQs. We performed a thermal melt analysis of Tel_{12}, monitoring the spectroscopic absorbance at 295 nm as the temperature was varied in the 373-293 K range. As shown in Figures S1 and S2, we obtained a sigmoidal decrease in absorbance as the temperature was raised, as expected for thermal unfolding, given the hyperchromicity of folded GQs[61]. The relative simplicity of this denaturation curve belies a complex multi-state folding equilibrium. In what follows, we will employ a nomenclature where each G-tract is represented by "|" in the folded state and "-" in the unfolded state. For instance, "-||||-------" refers to a DNA sequence with 12 telomeric repeats in which G-tracts 2-5 have folded into a GQ. In principle, there are 9 ways of forming a single GQ (||||--------, -||||------- etc.), 5 ways of forming two neighboring GQs (||||||||----, -

||||||||--- etc.), 10 ways of forming two non-adjacent GQs (||||-||||---, ||||--||||-- etc.) and one way of forming three GQs, ||||||||||||, for Tel_{12}. The unfolding curve of Tel_{12} alone does not provide enough information to unambiguously characterize the populations of all these partly-folded intermediates as a function of temperature. Nevertheless, the relative populations of these intermediates are of great interest, as they reveal positional and cooperative effects on GQ folding. In order to proceed, we employed a mutational trapping approach our lab previously used to investigate conformational dynamics within individual GQs[10]. A set of mutants was used to probe the individual two-state equilibria that comprise the complex multi-state folding landscape of Tel_{12}. The mutants were characterized individually and the data globally fit, yielding the populations of all Tel_{12} folding intermediates as a function of temperature. Different sets of 8 of the 12 G-tracts in Tel_{12} were mutated from GGG to GTG, leaving four contiguous GGG tracts capable of folding into a single GQ in a two-state manner. In our nomenclature, "x" corresponds to a telomeric repeat in which GGG has been replaced with GTG. DNA molecules containing these substitutions will be referred to as "G-tract knockout mutants" in what follows. The key assumption of this approach is that the stability of a G-tract knockout mutant is identical to that of the corresponding wild-type (WT) configuration. For instance, we assumed that $\frac{[-||||-------]}{[----------]} = \frac{[x|||xxxxxxx]}{[x----xxxxxxx]}$, where square brackets indicate concentrations. We measured the unfolding profiles of 7 two-state G-tract knockout mutants, as well as 4 G-tract knockout mutants that could form tandem GQs, (Table 1 for reference). Note that all sequences employed here contained a flanking 5' TTA and 3' TT, as these were shown to promote two-state folding for a simple four G-tract telomeric sequence[56]. The data for each two-state G-tract knockout mutant was analyzed to yield the enthalpy (ΔH_F) and entropy

(ΔS_F) of folding, assuming a heat capacity change, $\Delta C_p = 0$, revealing how the stabilities of individual GQs vary as a function of position. Data for mutants with tandem GQs provided information on folding cooperativity. For instance, the folding free energy of ||||||||xxxx is expected to be sum of the folding free energies of ||||xxxxxxxx and xxxx||||xxxx plus ΔG_c, the cooperative interaction energy, where $\Delta G_c < 0$ implies that a GQ stabilizes adjacent GQs and $\Delta G_c > 0$ implies that a GQ destabilizes adjacent GQs. We assumed that the interaction energy is additive, i.e. the folding free energy of |||||||||||| is equal to the sum of the energies for ||||xxxxxxxx, xxxx||||xxxx, xxxxxxxx|||| plus $2\Delta G_c$. Thus, the thermodynamic parameters obtained from fits of the G-tract knockout mutants provide sufficient information to

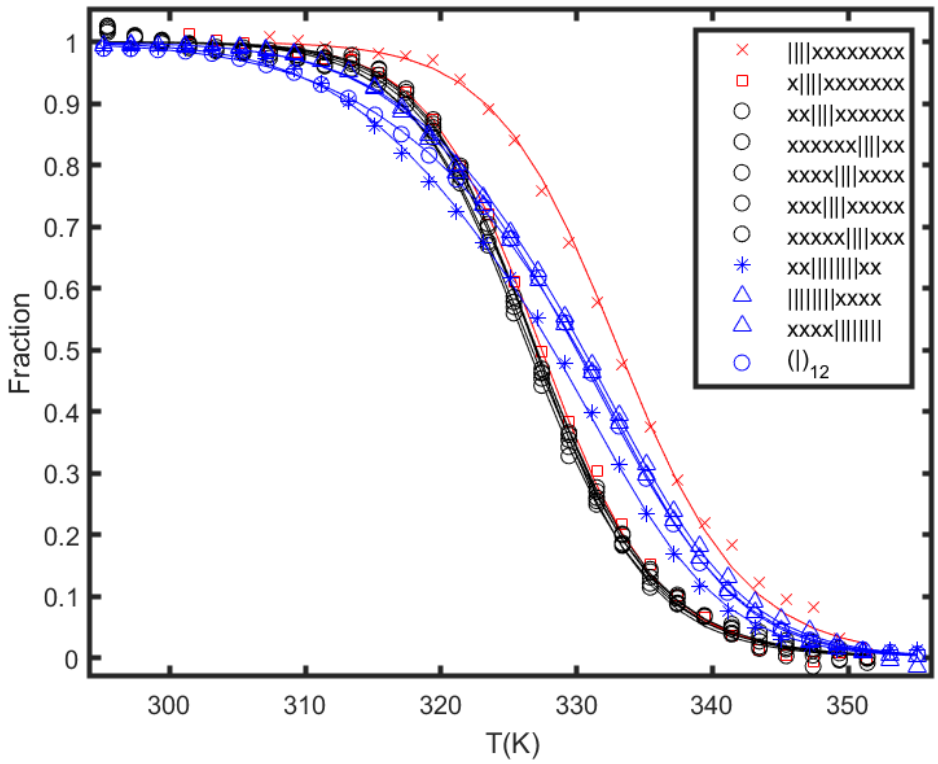

Figure 6. Raw data (converted to fraction) and fitted curves for the Tel_{12} system. Sequences in red and black form one GQ, sequences in blue form multiple GQs

reconstruct the unfolding UV-Vis profile of the WT molecule while mapping out the populations of all partly-folded intermediates (Figure S4) throughout the melting transition.

We performed a global analysis of all mutant and WT melting profiles, to extract position-specific ΔH_F and ΔS_F values for folding of individual quadruplexes located immediately at the termini, or separated from the nearest terminus by 1, 2, 3, or 4 unfolded G-tracts (i.e. 5 different stabilities). As well, we extracted the thermodynamic cooperativity parameters ΔH_c and ΔS_c, where $\Delta G_c = \Delta H_c - T\Delta S_c$. The extracted parameters are listed in Table 2. Importantly, the global fits (Figure 6) gave excellent agreement with all data sets, which validates the assumptions of the model, namely that folding cooperativity is position-independent and additive, and that the G-tract knockout mutations do not affect the folding stability of the remaining GQ. Violations of these assumptions would be expected to produce sets of mutants and of WT data that are mutually inconsistent [10]. Furthermore, we wanted to validate the assumption of ignoring the formation of long loop (3+1) GQs. These GQs readily fold in physiologically relevant conditions[62], however due to the entropic penalty caused by the presence of a long loop they are significantly less stable than a regular telomeric GQ. Figure S5 shows the melting profiles of sequences xx||||xx and xx|||x|xx in 100 mM K^+; while the latter exhibits significant hysteresis, its T_m is about 15 °C below the T_m of xx||||xx, therefore making it dramatically less stable.

Folding of all two-state G-tract knockout mutants was enthalpically driven with $\Delta H_F \approx -204\ kJ\ mol^{-1}$ and entropically unfavorable with $\Delta S_F \approx -622\ J\ mol^{-1}\ K^{-1}$, as expected for a disorder-to-order transition. It has been previously reported that the melting temperatures of terminal GQs are higher than those of internal GQs[54]. This was borne out in our extracted folding parameters. The melting temperature ($T_m = \Delta H_F/\Delta S_F$) of the Tel_{12} mutant containing a terminal GQ was about 5 degrees higher than those of the other mutants forming internal GQs. The enhanced stability of the terminal GQs is particularly evident in a

comparison of the melting profiles of the Tel_{12} G-tract knockout mutants containing internal GQs with the equivalent Tel_8 mutants, i.e. (xx||||xxxxxx, xxx||||xxxxx, xxxx||||xxxx, xxxxx||||xxx, xxxxxx||||xx) versus (||||xxxx, x||||xxx, xx||||xx, xxx||||x, xxxx||||). The Tel_8 terminal GQs melted at a higher temperature than did GQs separated from the terminus by one G-tract (Figure 7A). GQs two or more tracts away from the termini were the least stable, exhibiting melting curves essentially identical to those of the Tel_{12} internal GQs (Figure 7B).

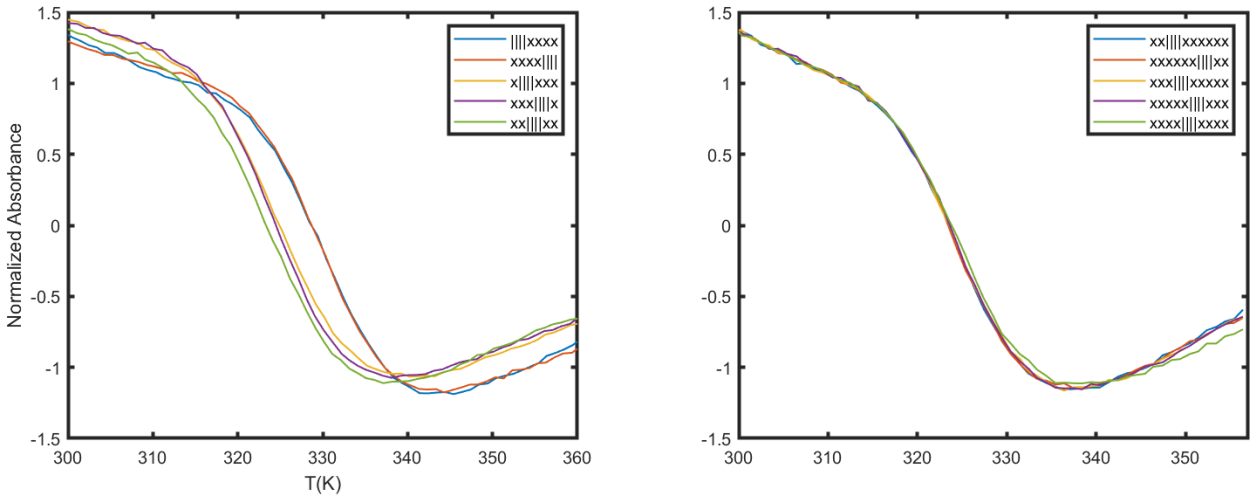

Figure 7. Sequences forming one GQ in Tel_8 (A) and Tel_{8ext} (B)

Their estimated T_m values are reported in Table 3. As previously shown[63], terminal GQs of vertebrate telomeric overhangs are indeed more stable than those further away from the 3' end. Interestingly, the G-tract knockout mutants that formed tandem GQs (||||||||xxxx, x||||||||xxx, xx||||||||xx, xxx||||||||x, and xxxx|||||||||) showed much broader melting transitions than did the two-state mutants. This is due to the formation of partly-folded intermediates near the midpoints of the transitions. For instance, in addition to the two-GQ fully folded state, ||||||||xxxx can adopt several one-GQ partly folded forms, such as ||||----xxxx, -||||---xxxx, --||||--xxxx, etc. We note that this broadening effect on the melting profiles was captured quantitatively by our folding model.

As well, our analysis showed that folded GQs have a weak tendency to destabilize their immediate neighbors at physiological temperatures. Our extracted parameters of $\Delta H_c = 44\ kJ\ mol^{-1}$ and $\Delta S_c = 139\ J\ mol^{-1}\ K^{-1}$ imply that at 37 °C, four contiguous G-tracts are only 64% as likely to fold adjacent to an already folded GQ as they would be in an isolated location. Interestingly this effect disappears at about 50 °C and is predicted to become slightly positively cooperative at higher temperatures. Similar weak negative folding cooperativity was previously observed in a differential scanning calorimetric study of Tel_8 and Tel_{12} folding[56]. Deconvolution of the DSC data revealed two- and three-step folding processes. In Tel_8 this presumably corresponded to transitions from the fully folded two-GQ form to the manifold of one-GQ partly folded forms to the fully unfolded state. In Tel_{12} this presumably corresponded to transitions from fully folded to the two-GQ manifold to the one-GQ manifold to the fully unfolded state. In both cases, the first unfolding transition from the fully folded state had the lowest energy, which was attributed to the presence of one or more adjacent GQs and an unfavorable coupling energy, $\Delta G_{coupling}$. For Tel_{12}, $\Delta G_{coupling}$ was similar (roughly double) to the $2\Delta G_c$ value we extrapolated at the same temperature for the microscopic pairwise energy. We believe this level of agreement is quite good, given the differences between the UV-Vis versus DSC methodologies and macroscopic vs microscopic analyses.

6.2 Tel_{8ext} kinetics

We then analyzed the Tel_{8ext} (xx|||||||xx) kinetic pathway. This sequence forms two neighboring GQs, meaning we can look at how a folded GQ affects the folding/unfolding kinetics of the next one. Retrieving kinetics information from UV-Vis thermal melts is possible thanks to the thermal hysteresis effect: a sufficiently fast (relative to the GQ kinetics) scan rate will cause the cooling/heating sigmoids to not overlay, instead pushing the apparent T_m towards lower and higher temperatures respectively. Temperature shifting is then proportional to the magnitude of the scan rate[64]; this is commonly referred to as an out of equilibrium situation. xx|||||||xx and trapped states were analyzed in a relatively low salt environment, in order to observe sufficiently high hysteresis. As previously observed[65], a high amount of K^+ increases the T_m value of a GQ, as well as making folding progressively faster and kinetics measurements impossible. When looking at xx|||||||xx folding there are two on-pathway isomers (xx|||----xx and xx----|||xx), whereas xx-|||---xx, xx--|||--xx and xx---|||-xx are off-pathway. On-pathway guarantees that once the first GQ is folded, there are still 4 sequential free G-tracts to fold the second one; conversely, off-pathway states create stretches of <4 free G-tracts. Importantly, once a molecule adopts an off-pathway configuration, it must fully unfold again before it can follow the pathway to the fully folded state. Thus the off-pathway intermediates are potentially deep kinetic traps. These states all related according to the kinetics model shown in Figure 5.

In order to individually measure the kinetics of xx|||||||xx and each partially folded state, we again mutate each unused G-tract such that "-" is now "x". For example, the folding kinetics of the xx--|||--xx partly folded isomer are probed with a xxxx|||xxxx G-tract knockout

mutant, etc. We assume the kinetics of each partly folded Tel_{8ext} isomer to be identical to that of the corresponding tract knockout mutant. This allows calculating the amount of each isomer present at any temperature. Assuming that the total absorbance of Tel_{8ext} at a given temperature and scan rate is proportional to the number of folded GQs, we can reconstruct its thermal hysteresis profile. Parameters are simultaneously adjusted to give good agreement between the tract knockout mutants and Tel_{8ext} wild-type profile. In the previous paragraph, we showed how single GQs folding at least 2 G-tracts away from each end have essentially the same stability. It follows that this effect should also be present in terms of thermal hysteresis; as shown in Figure 8A, the same tract knockout mutants from Figure 6B now give thermal hysteresis profiles that look alike.

In terms of normalized absorbance, we see the Tel_{8ext} signal going only up to ~0.8 instead of 1 at low temperatures (Figure 8A), therefore indicating incomplete folding; we rationalize this in terms of trapping, because folding an off-pathway isomer makes it impossible to fold the second one. Folding of both GQs requires that the off-pathway isomer first unfold and that on-pathway folding then proceeds. GQ unfolding is quite slow at these temperatures so Tel_{8ext} is trapped in a partly-folded state for a considerable length of time. Trapping is also kinetically dependent, because as shown in Figure 8C, the f_{8ext} value at low temperatures is inversely proportional to the cooling rate. In other words, more rapid cooling rates trap more molecules in the off-pathway partly folded states. In principle, this means that an ideally slow cooling process should allow $f_{8ext} = 1$ at low temperatures, because the kinetic trapping effect would be less noticeable.

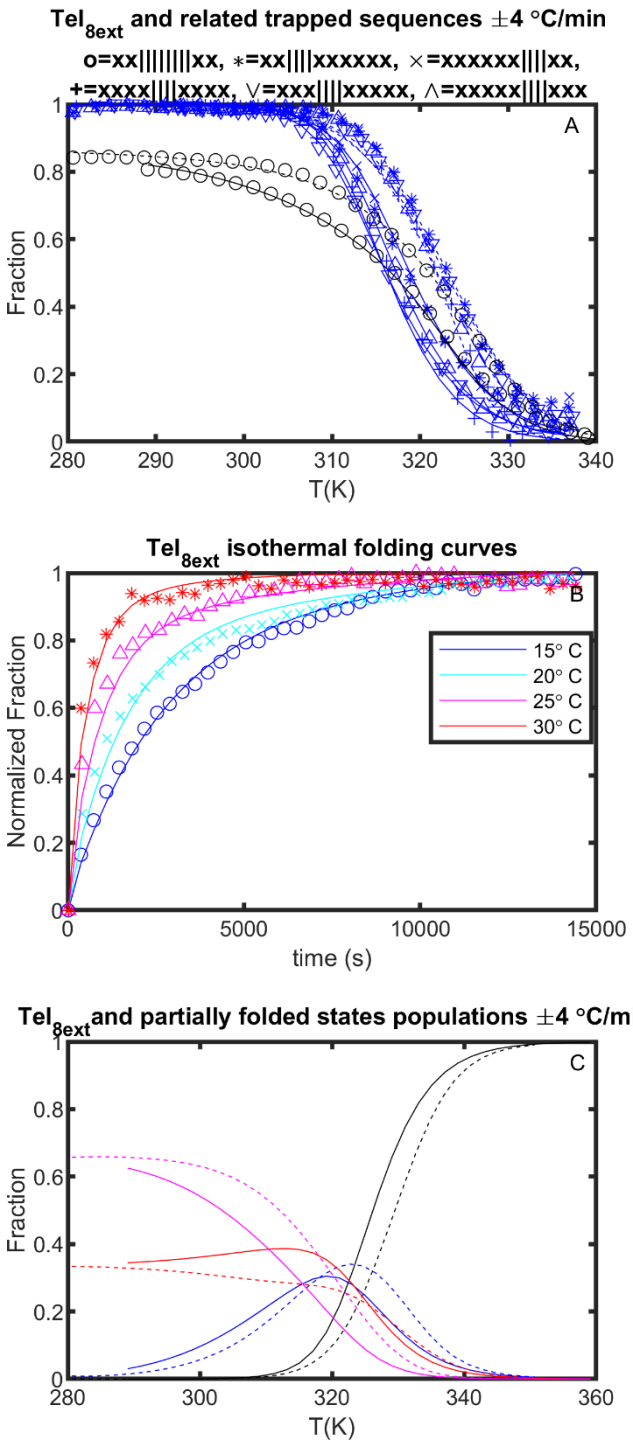

Figure 8. (A) Normalized absorbance data and fitted curves for Tel8ext and mutants. (B) Normalized isothermal folding data and fitted curves. (H) Temperature dependent populations of each state. Following the kinetic scheme (Figure 5), black is (1), blue is (2)+(3) (intermediates), magenta is (4), red is (5)+(6)+(7) (misfolded states). Solid lines refer to -4 °C/min and dashed lines to 4 °C/min (A and C)

In order to confirm the presence of this kinetic trapping effect, we performed a series of rapid cooling/isothermal folding experiments. Fast Tel_{8ext} cooling was expected to give populations of both the fully folded state and the off-pathway, partially folded ones. The off-pathway isomers were expected to slowly unfold over time and adopt the more stable fully folded state. It follows that gradual increases in absorbance over time should be visible, as the misfolded molecules containing one GQ convert to fully folded molecules containing two GQs. This was the case for the isothermal folding experiments shown in Figure 8B, performed at equilibration temperatures of 15 °C to 30 °C. Each of the normalized isothermal curves showed an increase in absorbance of about ~0.01 a.u. over 4 hours, and exhibited a roughly exponential shape, providing information on the rates of unfolding of the misfolded states. Equilibration was faster at higher temperatures, corresponding to a positive activation enthalpy for exiting the misfolded kinetic traps. As explained in Chapter 2.5, the three sets of data (thermal hysteresis curves for sequences forming one GQ, two GQs, and isothermal folding curves) were simultaneously analyzed and their respective fitted curves calculated using the same set of optimized parameters. For isothermal folding curves, the calculated change in $[Tel_{8ext}]$ is time dependent rather than being temperature dependent like the thermal hysteresis case. We also performed control experiments to test for instrument drift or time-dependent changes in the absorbance of the blank. The contributions of these effects were found to be negligible as summarized in Figure S6. Lastly, the kinetic parameters obtained from the Tel_{8ext} thermal hysteresis fit lead to the following considerations: 1) Consistently with a previous report[56], folding the second GQ is thermodynamically less favorable than the first one 2) Folding or unfolding the second GQ is overall slower than the first one (~30% and ~10% decrease of the respective rates at 315 K).

6.3 $(TGGG)_8T$ kinetics

We then used sequences based on TGGG tandem repeats as an additional validation for our kinetic model. Selective GQ folding along the sequence is again realized by mutating each unused G-tract from GGG to GTG. Like Tel_{8ext}, there are two on pathway and three off pathway isomers, meaning that a similar kinetic scheme applies (Figure S1). Figure S7 shows the global fit to the thermal hysteresis data. From a qualitative point of view, the stability of $(TGGG)_8T$ is much higher than any of the two intermediates, meaning that thermodynamic folding cooperativity is expected to be extremely positive. This agrees with a previous report on the similar sequence $(GGGT)_7GGG$, where folding the second GQ is enhanced by the formation of stacking interactions with the first one[66]. A similar, qualitative comparison to the same source can be made in terms of kinetics; at 300 K, folding (or unfolding) the first GQ is the slower step.

6.4 Simulating the behavior of longer tandem repeats

Analyzing sequences longer than Tel_{12} is not a simple task, because the number of partially folded states dramatically increases with length. For example, a telomeric overhang-like sequence of 32 G-tracts with 7 out of 8 GQs folded can adopt 330 unique configurations (see SI). Obviously, the full folding landscape contains many more states than this. It is not experimentally feasible to analyze each one of these states individually. Instead, we took the microscopic information obtained from the smaller scale systems and applied them to predict the behavior of the larger one. The success of the simple kinetic and thermodynamic

models in quantitatively reproducing the folding patterns of molecules containing up to twelve telomeric repeats gave us confidence that calculations for longer sequences would be realistic.

The technique of the McGhee von Hippel (MGVH) model can be used to quickly simulate long telomeric sequences. It was originally created to understand ligand binding on a 1D infinitely long, homogeneous lattice. Each ligand is perfectly symmetrical and occupies an equal number of identical residues on the lattice, meaning there's only one kind of nearest neighbor interaction. Due to telomeric DNA being composed of tandem TTAGGG repeats, this model can be used without any modification; each repeat represents a unit of the lattice and four of them are "occupied" when folding into a GQ. A rapid calculation not only gives the average number of folded GQs, but also how many of them are isolated (B_{isol}, no interactions), singly contiguous (B_{sc}, one interaction on either side) and doubly contiguous (B_{dc}, interactions on both sides) respectively. The main ideas are the following: 1) A sequence composed of L units within the infinitely long lattice is considered 2) The maximum number of foldable GQs is L/4, because four units are required to form one GQ 3) Due to the absence of end effects, GQ folding along this sequence is only dependent on ΔG_{mid} and ΔG_c 4) ΔG_{mid} (internal GQ folding free energy) is the average between xx||||xxxxxx, xxx||||xxxxx, xxxx||||xxxx, xxxxx||||xxx and xxxxxx||||xx folding ΔGs 5) ΔG_c is only used for nearest neighbors GQs 6) Folding can happen anywhere along the sequence, given four sequential free units (G-tracts).

The last consideration is especially important, because it leads to an overlap of the folding sites; this means that complete saturation of the sequence is difficult, unless highly positive

thermodynamic cooperativity is present. For example, for L=32 the calculated folding free energies at 310 K from the Tel_{12} fit ($\Delta G_{mid} = -10.51\ kJ/mol$ and $\Delta G_c = 1.14\ kJ/mol$) give $B_{isol} = 1.4$, $B_{sc} = 3.4$ and $B_{dc} = 2.1$; additionally, the average number of sequential folded GQs is ~2. This means that on average, ~7 GQs out of 8 are folded; in other words, compared to a sequence forming a single GQ whose fu~1.7% (derived from ΔG_{mid}), this leads to a ~9-fold increase of the unfolding probability. We refer to this as a frustration effect, meaning that the number of folded GQs will be lower than the maximum possible number. It makes sense from a statistical mechanics point of view, because there's only one way of fully saturating the lattice, compared to multiple ways of rearranging several folded GQs and free G-tracts; a more detailed description in the Materials and Methods section shows the math underlying the method. We then seek to account for the presence of end effects. While it's not a feature of this model, it's possible to introduce them in a Monte Carlo simulation; ΔG_1 (end GQ) and ΔG_2 (end-1 GQ) are based on ||||xxxxxxxx and x||||xxxxxxx parameters from the Tel_{12} fit. The McGhee and Von Hippel model proves useful here as well, to test the validity of the MC simulation. First, a very long sequence (L=1024 G-tracts) sequence is taken as our ideally infinitely long lattice for the MC simulation. A subset of variable length (M=32-128 G-tracts) in the middle of the sequence was considered, to ignore the influence of end effects; the different B_{isol}, B_{sc} and B_{dc} values were calculated and then compared to the ones obtained with the MGVH model. The two methods were shown to always agree within very reasonable boundaries, even upon changing ΔG_c. For example, for M=100 and using the parameters obtained from the Tel_{12} fit at 310 K, the MGVH model gave $B_{isol} = 4.3$, $B_{sc} = 10.5$ and $B_{dc} = 6.4$; the MC simulation gave $B_{isol} = 4.2$, $B_{sc} = 10.2$ and $B_{dc} = 6.1$. A similar simulation (L=1024) based on the same parameters and temperature, this time including

the 3’ end stabilizing effect, is depicted in Figure 9. While most of the lattice has the same unfolding probability (~15%), both ends are deviating by exhibiting a periodic pattern. The inset in Figure 8 showing the 3’ end allows several observations to be made. G-tracts close

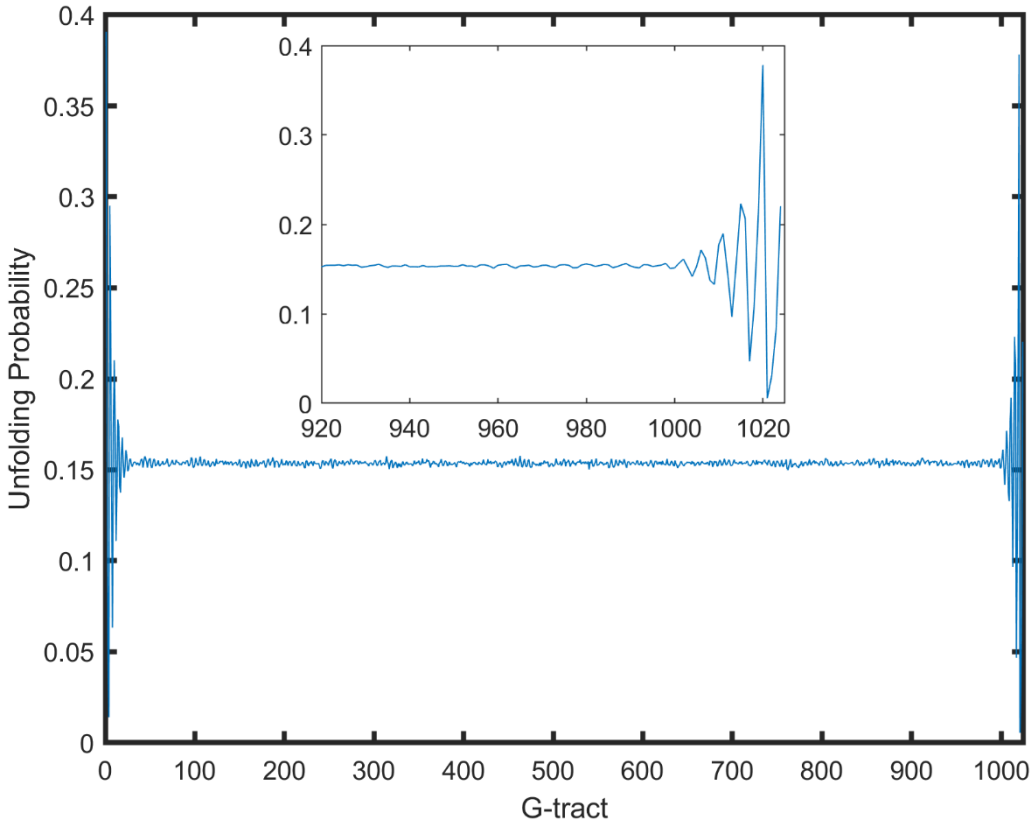

Figure 9. Monte Carlo simulation of the unfolding probability for a telomeric sequence of length L=1024 G-tracts

to the end are the most folded of all, due to the stabilizing effect being present; going towards the 5’, other G-tracts are expected to be more unfolded than the average, with an increase in probability of as much as ~2.5 times. It follows how this could have several biological implications that will be discussed later. We then wanted to ascertain whether this periodic pattern, or frustration end effect, was a consequence of the specific thermodynamic parameters used during the simulation. Therefore, we repeated the process with different

sets of parameters (ΔG_c <, >, = 0, presence or absence of end effects) and we found the frustration effect to exhibit small changes, but to overall persist in each case.

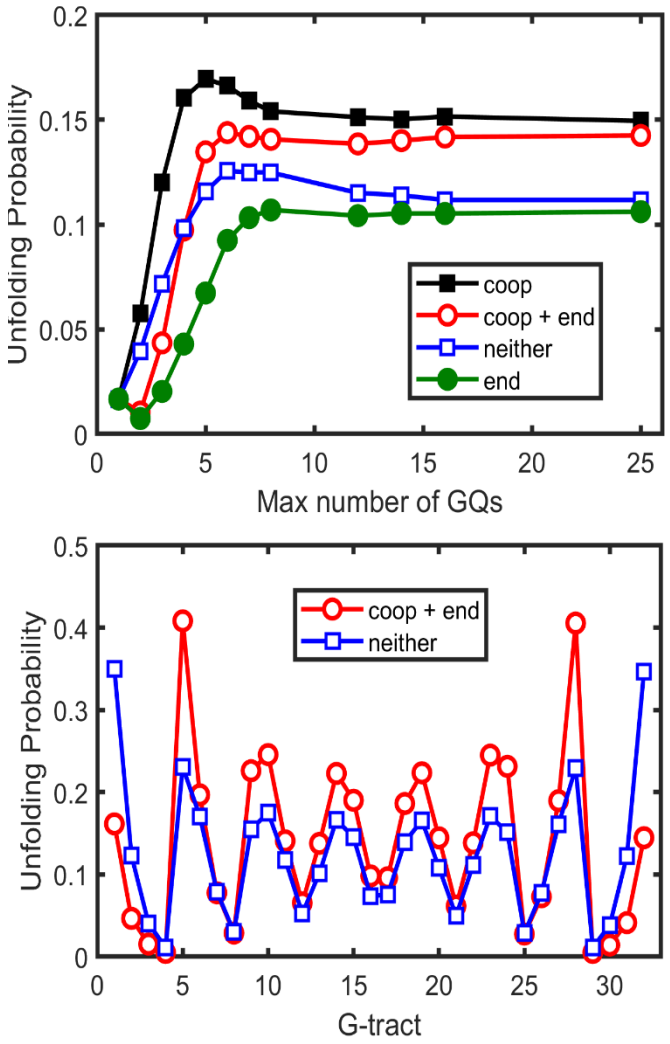

Figure 10. (Upper Panel) Monte Carlo simulation of the length dependent unfolding probability for telomeric sequences. (Lower Panel) Position dependent unfolding probability for a sequence of length L=32 G-tracts. Cooperativity values were the same in both cases.

A second, interesting frustration effect is found when looking at the change in unfolding probability upon increasing sequence length; due to previous studies mostly focusing on relatively short sequences, we feel that this effect has not been widely recognized as applicable to telomeric DNA. For a sequence forming one GQ the unfolding probability is simply calculated as $f_U = \frac{1}{K_{mid}+1}$; for sequences forming more GQs, the result is simulated (Figure 10, upper panel). Four different cases were analyzed (different cooperativity, presence or absence of more stable ends); they all exhibit the same trend of increasing the unfolding probability and then reaching an asymptote, with changes as big as 9 times the value for one GQ. The effect starts to become apparent for sequences forming at least 4 GQs, highlighting the importance of future studies focusing on long telomeric DNA strands. It follows that in principle, this

effect is solely dependent on the increase in sequence length: there is always one way of fully saturating the lattice, and multiple ways of folding several GQs with gaps. The lower panel shows the position dependent, unfolding probability for a telomeric overhang-like sequence (8 GQs).

Lastly, we used the parameters from the thermal hysteresis fit to compare thermodynamic and kinetic simulations of a 32 G-tracts long sequence. Figure 11 shows how the number of folded GQs over time rapidly reaches an asymptote; not only is the result independent on the initial configuration (fully folded or unfolded sequence), but also equal to the thermodynamic value. This suggests that in the case of telomeric DNA, GQ folding is a process

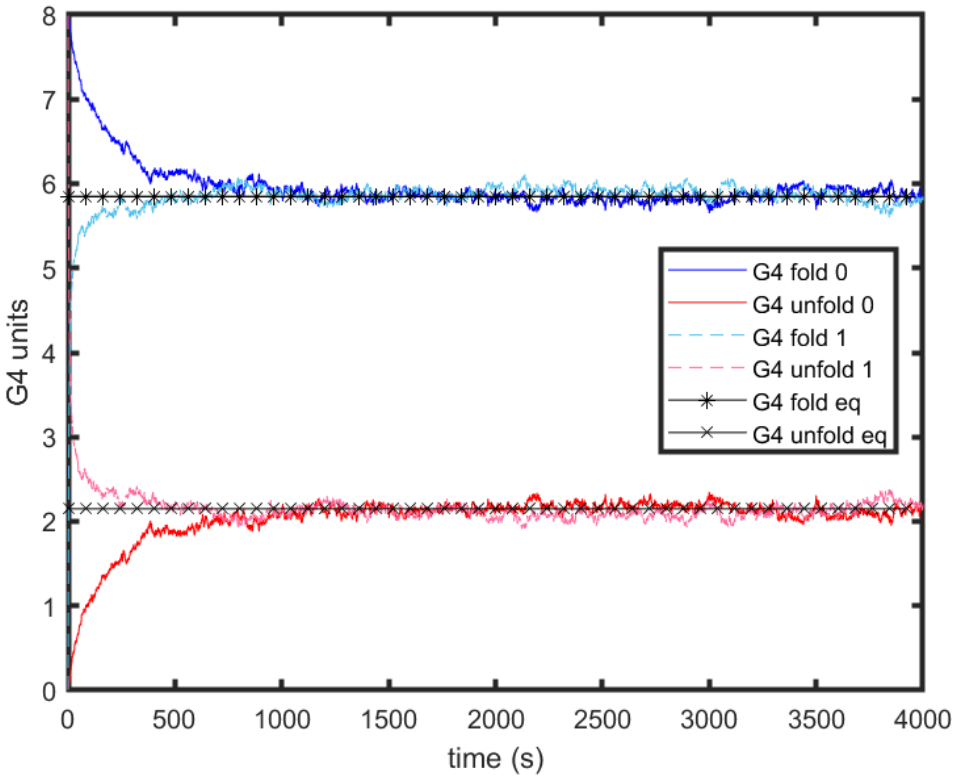

Figure 11. Simulated number of GQs folding/unfolding over time starting from either a fully folded (0) or fully unfolded (1) sequence. Black lines with markers refer to equilibrium values.

controlled by thermodynamics rather than kinetics. Again, the $(TGGG)_8T$ case represents an interesting comparison (Figure S8); in terms of kinetics, it now matters whether the

simulation starts with a fully folded or unfolded sequence. In the first case the number of folded GQs is always 8, whereas in the second case its asymptote is at 7. From a statistical mechanics point of view, it makes sense to assume that 8-GQ form is the most stable state. We mentioned before how there are 330 ways of distributing 7 GQs and 4 free G-tracts; even if we take the most stable among these states $((|)_{28}(x)_4$, which only occurs twice), calculate its Boltzmann's factor multiplied by 330 and compare it to $(|)_{32}$, we determined how the latter has a population about 170000 times higher. This last observation puts context into the big difference of having 7 folded GQs instead of 8 for such a cooperative GQ forming sequence. Overall, the difference between the kinetically and thermodynamically simulated values of folded GQs suggests the presence of significant kinetic trapping; this is contrast to what has been described above for the Tel case, where the folding process seems to be determined by thermodynamics.

6.5 Presence of gaps and biological implications

We found that for long telomeric repeats, the tendency for a G-tract to be unfolded can be 9-fold or higher than it would be in the context of an isolated, internal GQ. For negatively cooperative folding, this is a thermodynamic effect; the ensemble of states sampled at equilibrium contains unfolded gaps between the folded GQs. It must be emphasized that this tendency to unfold is dominated by the multiplicity of partly unfolded states, rather than by the unfavorable pairwise energy itself. An interesting study recently looked at the folding kinetics of telomeric sequences up to 48 G-tracts via force extension: while very different from our methodology, there is a qualitative agreement in observing an increase in the

number of vacant G-tracts as sequence length is increased. However, we do not ascribe this effect as sole consequence of an apparently unfavorable electrostatic interactions between neighboring GQs[67]; rather, we would like to introduce the idea of a length dependent frustration effects that is only further enhanced, but not caused, by negative thermodynamic cooperativity. For strongly cooperative GQ folding, where the thermodynamic minimum corresponds to nearly 100% folding, we found that a substantial percentage >10% of G-tracts likely remain unfolded nearly indefinitely. In this case it is a kinetic effect that dominates. The probability that 32 G-tracts transition to the fully folded state without encountering a kinetic trap is roughly 1 in 10^5. Correcting a misfolding event can require the coordinated unfolding and refolding of multiple GQs and may not occur on physiologically relevant timescales.

The existence of gaps between folded GQs in telomeres has likely biologically implications. Many proteins are known to interact with telomeric overhangs; the following are examples of proteins that do so more favorably when free G-tracts are present. As an example of proteins that take advantage of gaps present between telomeric GQs, POT1 represents one of the most important. It has been demonstrated to bind a 10 nucleotides single stranded sequence[68] and enhance telomerase activity even at modest concentration[69, 70]. In a previous study, it has been hypothesized that once finding a gap, POT1 drives steric unfolding of neighboring GQs, as observed for a Tel_{16} sequence[58]. Another study also pointed out how it is possible for POT1 to bind directly to a GQ and then unfold it, regardless of free single stranded regions being present[71]. This suggests that in the presence of both folded GQs and unfolded single stranded regions, POT1 would bind more favorably to the latter. Another example is represented by SSB1, which has been shown to preferentially bind

a single stranded G-rich telomeric sequence rather than a C-rich[72]. The same study points out how SSB1 facilitates hTERT recruiting and demonstrates its role in maintaining the G-overhang. Yet another example is represented by RPA, a different protein involved in telomere maintenance. RPA binding efficiency to telomeric G-rich ssDNA has been shown to be inversely proportional to GQ stability[73]. Considering our findings, we conclude that in general, any protein binding to free regions of the telomeric overhang would do so more favorably in proximity of the 3' end. Additionally, the relationship between sequence length and unfolding probability should play a significant role in binding, as telomere overhangs can be shortened and elongated. In the real context however, these proteins are not independent; interestingly, they are known to sometimes compete between each other[74] and interact with binding partners[75].

7 Conclusion

In this work, we sought to understand the dynamics of very long telomeric DNA sequences. The analysis of smaller scale systems allowed us to mainly determine that: 1) Sequences forming multiple GQs can be described in terms of stability of each individual GQ 2) A single cooperativity free energy parameter accounts for every nearest neighbor interaction. These results made it easy to approach the analysis of much longer sequences, without the necessity of making unreasonable approximations; consequently, MC simulations brought to our attention two different types of frustration effects, both of which could have real biological implications. Equally important, thermal hysteresis experiments allowed us to explore the kinetic aspect of GQ folding. Applied to two cases of tandem repeats ($TGGG_n$ and

TTAGGG$_n$) exhibiting drastically different thermodynamic cooperativity, this allowed us to the determine whether GQ folding is dictated by kinetics rather than thermodynamics. As well, despite the use of different experimental techniques and modeling approaches, we found our work to qualitatively agree and complement what has already been found in previous studies.

8 Supplemental information

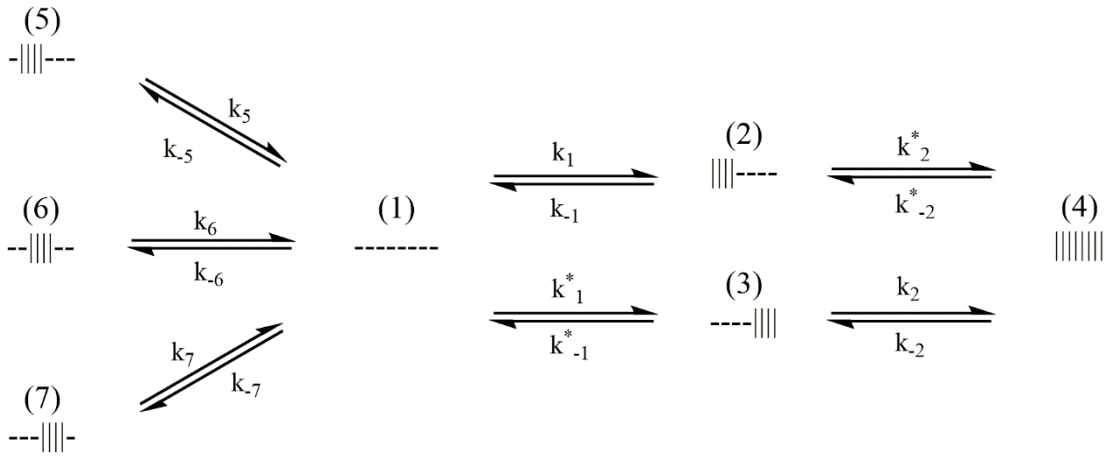

Figure S1. $(TGGG)_8T$ folding pathway

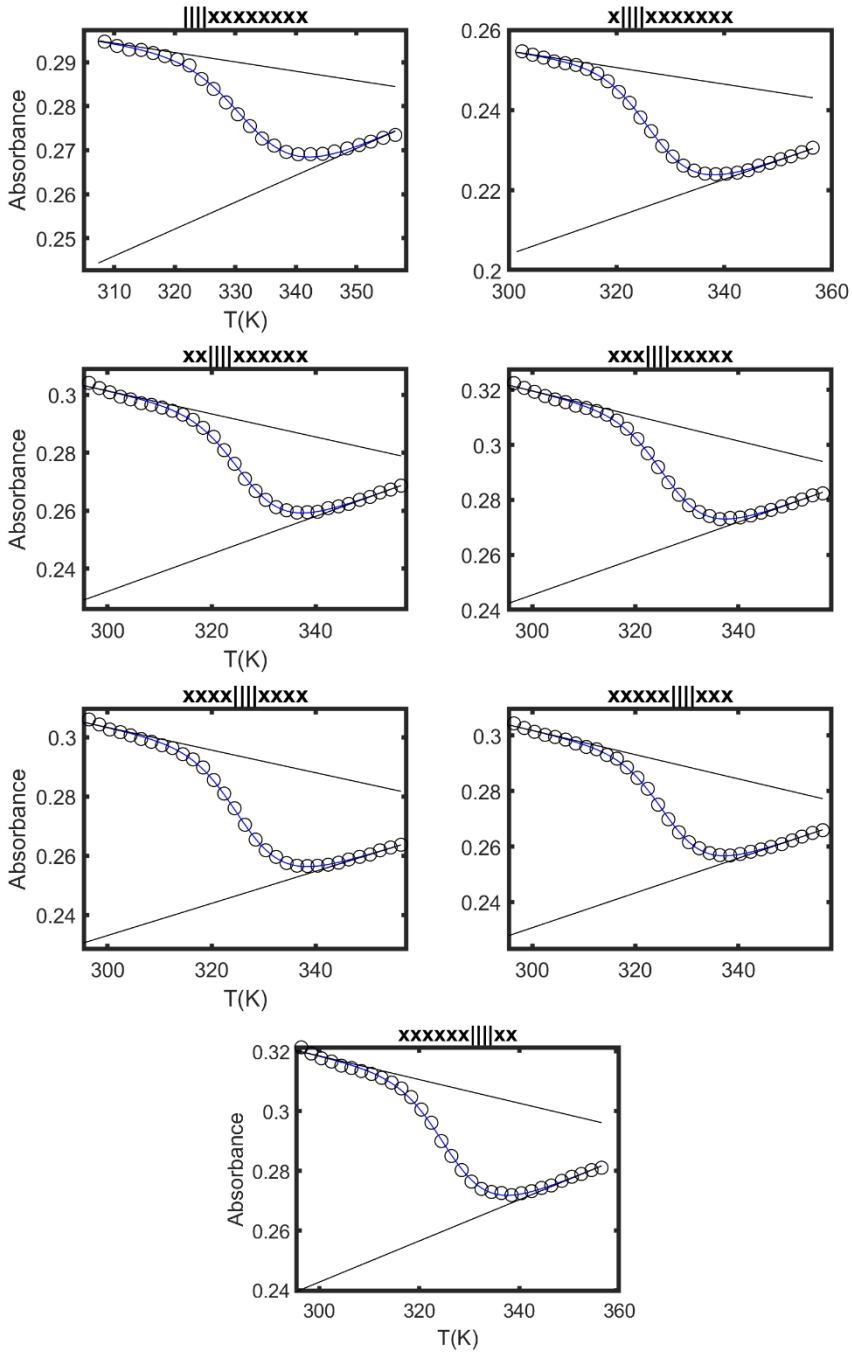

Figure S2. Raw absorbance data and fitted curves for Tel_{12} sequences forming two and three GQs

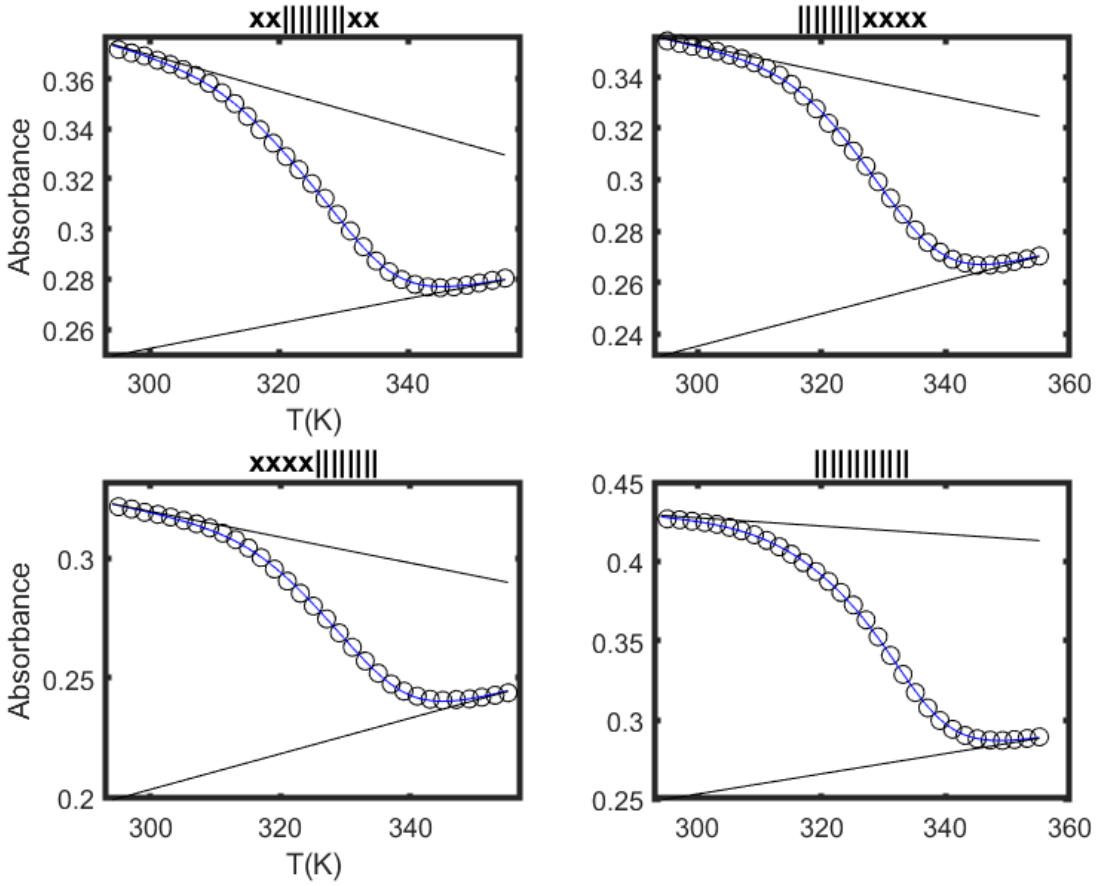

Figure S3. Raw absorbance data and fitted curves for Tel_{12} sequences forming two and three GQs

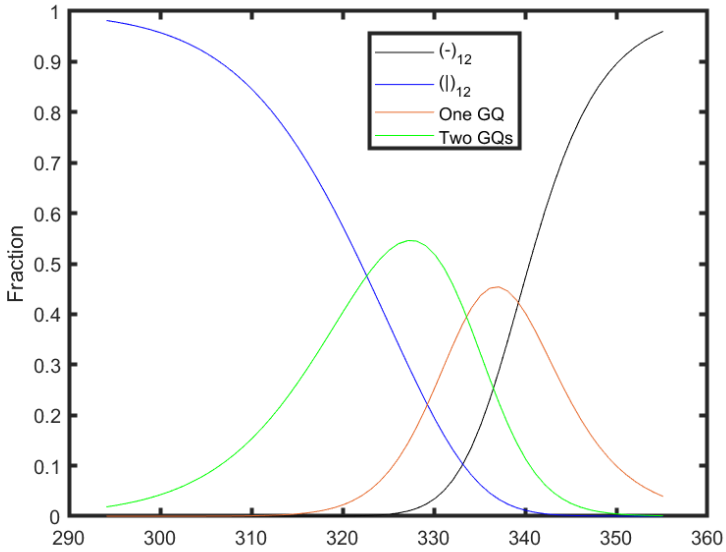

Figure S4. Populations of the various Tel_{12} states. For states forming either one or two GQs, the total sum is shown.

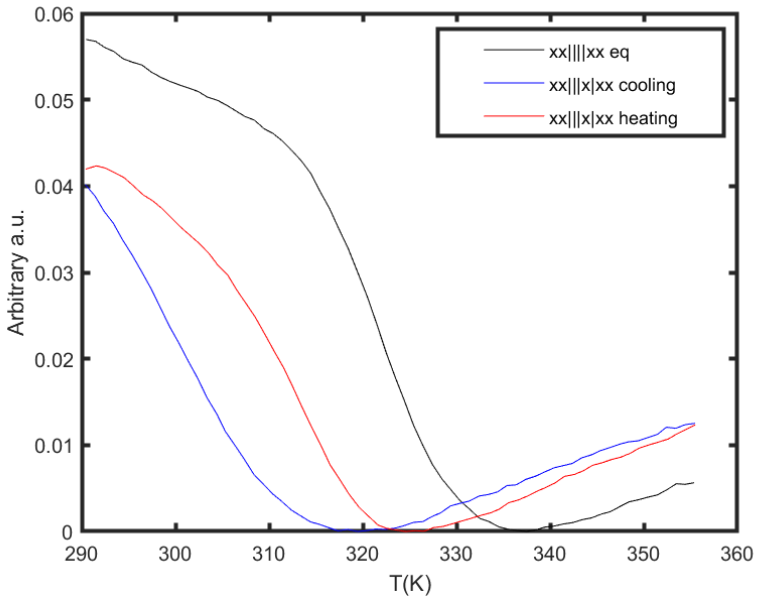

Figure S5. Comparison between the UV-Vis melting profiles of xx||||xx and its corresponding 3+1 GQ

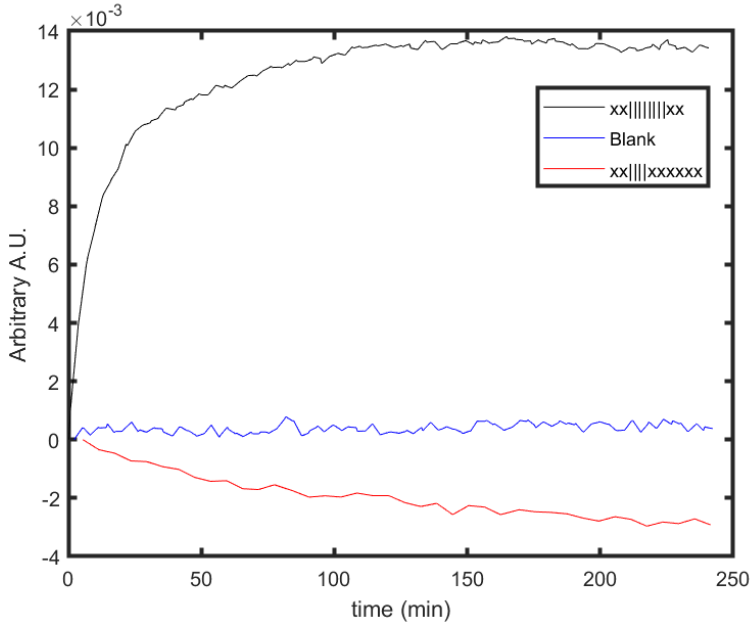

Figure S6. Comparison between the an isothermal folding curve for xx|||||||xx and possible absorbance drifting caused by the blank or by a GQ forming sequence with no kinetic trapping effect. All three time courses were acquired simultaneously in the same conditions.

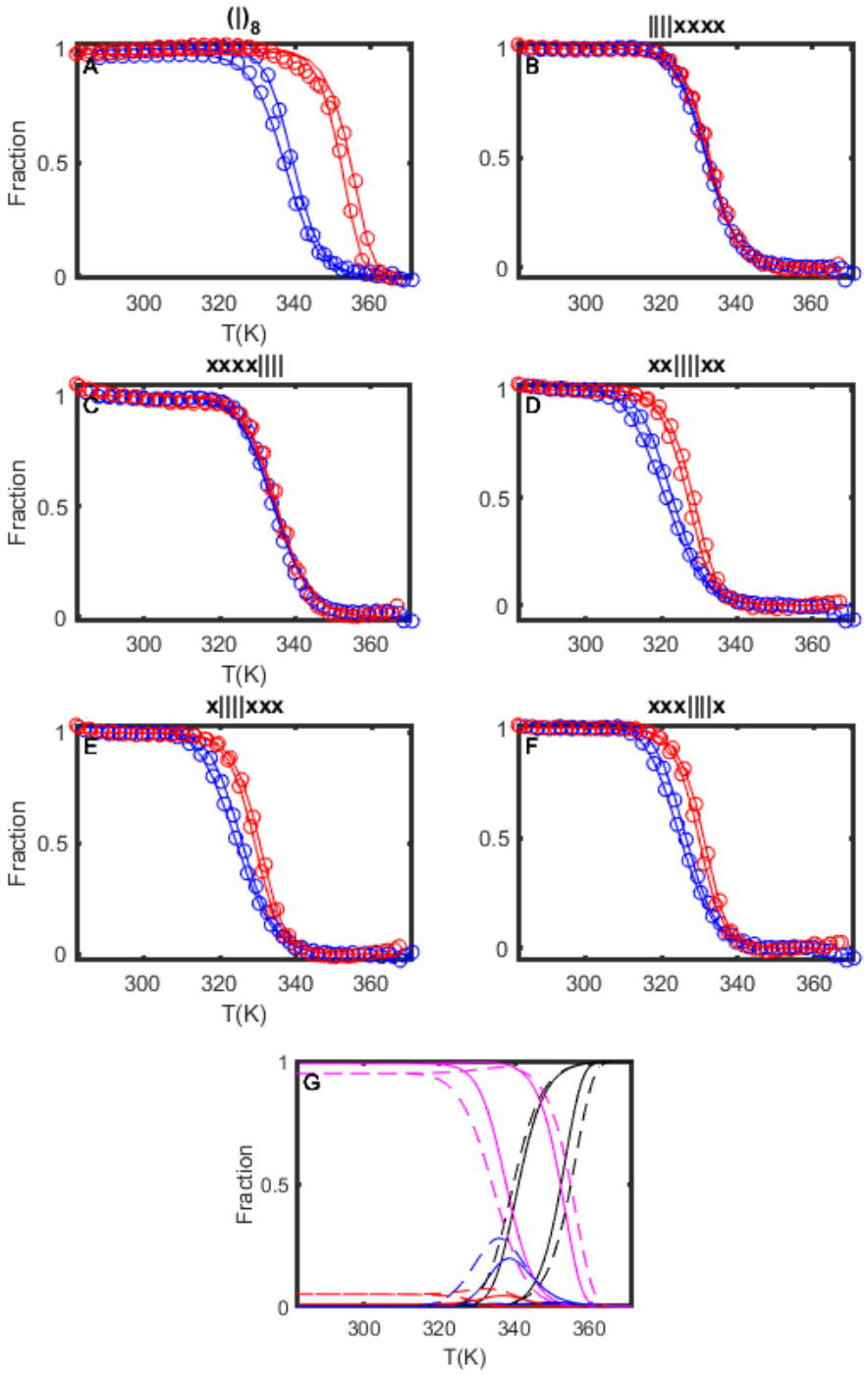

Figure S7. (A-F) Normalized absorbance data and fitted curves for $(TGGG)_8T$ and mutants. (H) Temperature dependent populations of each state. Following the kinetic scheme, black is (1), blue is (2)+(3) (intermediates), magenta is (4), red is (5)+(6)+(7) (misfolded states). Solid lines refer to 2 °C/min and dashed lines to 3 °C/min

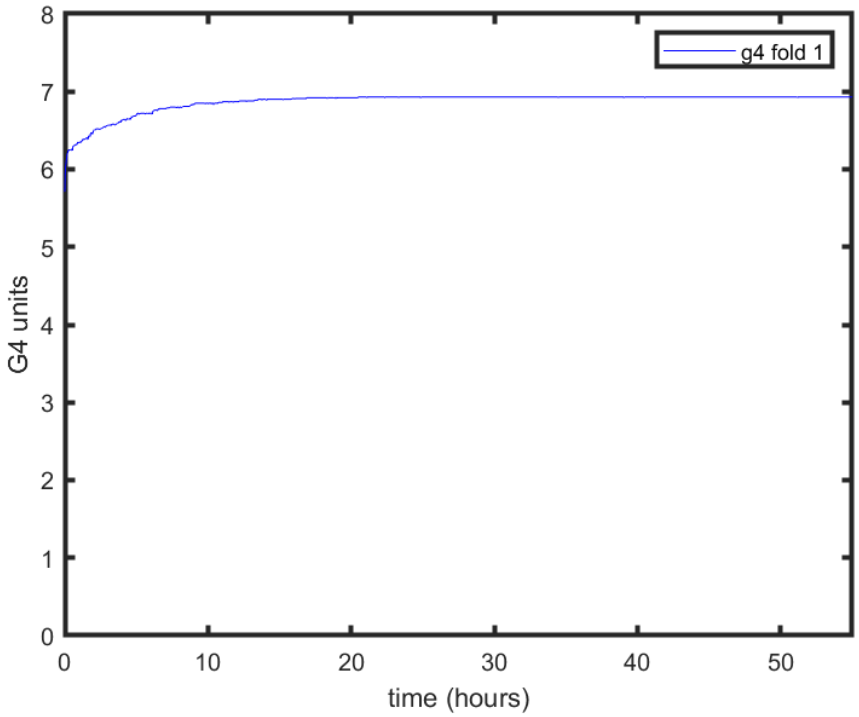

Figure S8. Monte Carlo simulation of the number of folded GQs (starting from fully unfolded) for a $TGGG_n$ sequence of length L=32 G-tracts.

Table 1. DNA sequences used in this study: |||| = 4 folded G-tracts (1 GQ), x = 1 mutated G-tract

Formula	Sequence code
$(TGGG)_8T$	||||||||
$(TGGG)_4(TGTG)_4T$	||||$xxxx$
$TGTG(TGGG)_4(TGTG)_3T$	x||||xxx
$(TGTG)_2(TGGG)_4(TGTG)_2T$	xx||||xx

$(TGTG)_3(TGGG)_4TGTGT$	$xxx\|\|\|\|x$
$(TGTG)_4(TGGG)_4T$	$xxxx\|\|\|\|$
$(TTAGGG)_{12}TT$	$\|\|\|\|\|\|\|\|\|\|\|\|$
$(TTAGTG)_2(TTAGGG)_8(TTAGTG)_2TT$	$xx\|\|\|\|\|\|\|xx$
$(TTAGGG)_8(TTAGTG)_4TT$	$\|\|\|\|\|\|\|xxxx$
$(TTAGTG)_8(TTAGGG)_4TT$	$xxxx\|\|\|\|\|\|\|$
$(TTAGGG)_4(TTAGTG)_8TT$	$\|\|\|\|xxxxxxxx$
$TTAGTG(TTAGGG)_4(TTAGTG)_7TT$	$x\|\|\|\|xxxxxxx$
$(TTAGTG)_2(TTAGGG)_4(TTAGTG)_6TT$	$xx\|\|\|\|xxxxxx$
$(TTAGTG)_3(TTAGGG)_4(TTAGTG)_5TT$	$xxx\|\|\|\|xxxxx$
$(TTAGTG)_4(TTAGGG)_4(TTAGTG)_4TT$	$xxxx\|\|\|\|xxxx$
$(TTAGTG)_5(TTAGGG)_4(TTAGTG)_3TT$	$xxxxx\|\|\|\|xxx$
$(TTAGTG)_6(TTAGGG)_4(TTAGTG)_2TT$	$xxxxxx\|\|\|\|xx$
$(TTAGGG)_4(TTAGTG)_4TT$	$\|\|\|\|xxxx$
$TTAGTG(TTAGGG)_4(TTAGTG)_3TT$	$x\|\|\|\|xxx$
$(TTAGTG)_2(TTAGGG)_4(TTAGTG)_2TT$	$xx\|\|\|\|xx$
$(TTAGTG)_3(TTAGGG)_4TTAGTGTT$	$xxx\|\|\|\|x$
$(TTAGTG)_4(TTAGGG)_4TT$	$xxxx\|\|\|\|$

Table 2. Tel_{12} fit parameters

Sequence code	Folding ΔH (kJ/mol)	Folding ΔS (J/(K·mol))
‖‖xxxxxxxx	-200 ± 8	-600 ± 25
x‖‖xxxxxxx	-210 ± 10	-641 ± 30
xx‖‖xxxxxx	-200 ± 6	-612 ± 18
xxx‖‖xxxxx	-208 ± 6	-635 ± 20
xxxx‖‖xxxx	-199 ± 5	-609 ± 15
xxxxx‖‖xxx	-215 ± 7	-659 ± 20
xxxxxx‖‖xx	-196 ± 5	-599 ± 16
$\Delta H_c = 44 \pm 5$ kJ/mol		
$\Delta S_c = 139 \pm 14$ J/(K·mol)		

Table 3. T_m values of sequences in figure S7A (blue) and S7B (black).

Sequence code	Tm (°C)
xx‖‖xxxxxx	$\sim 54.4 \pm 0.3$
xxx‖‖xxxxx	$\sim 54.4 \pm 0.3$
xxxx‖‖xxxx	$\sim 54.4 \pm 0.3$
xxxxx‖‖xxx	$\sim 54.4 \pm 0.3$
xxxxxx‖‖xx	$\sim 54.4 \pm 0.3$
‖‖xxxx	$\sim 59.2 \pm 0.3$
xxxx‖‖	$\sim 59.2 \pm 0.3$
x‖‖xxx	$\sim 55.7 \pm 0.3$
xxx‖‖x	$\sim 55.1 \pm 0.3$
xx‖‖xx	$\sim 54.4 \pm 0.3$

Table 4. Tel_{8ext} fit parameters

Rate constants at 315 K (10^3 s^{-1})		Activation energies (kJ/mol)	
k_5	16.9 ± 0.5	E_5	-125 ± 5
k_{-5}	6.2 ± 0.2	E_{-5}	66 ± 3
k_6	16.1 ± 0.3	E_6	-100 ± 3
k_{-6}	4.8 ± 0.1	E_{-6}	154 ± 2
k_7	26.9 ± 1.3	E_7	-131 ± 7
k_{-7}	8.9 ± 0.4	E_{-7}	71 ± 5
k_1	33.3 ± 1.3	E_1	-119 ± 2
k_{-1}	8.8 ± 0.4	E_{-1}	67 ± 2
k_2	21.3 ± 1.7	E_2	2 ± 5
k_{-2}	7.8 ± 0.6	E_{-2}	100 ± 6

Table 5. $(TGGG)_8T$ fit parameters

Rate constants at 300 K (10^3 s^{-1})		Activation energies (kJ/mol)	
k_5	31 ± 3	E_5	-55 ± 3
k_{-5}	0.008 ± 0.0004	E_{-5}	189 ± 1
k_6	20 ± 1	E_6	-54 ± 2
k_{-6}	0.012 ± 0.0009	E_{-6}	187 ± 2
k_7	51 ± 5	E_7	-69 ± 3
k_{-7}	0.006 ± 0.0007	E_{-7}	195 ± 3
k_1	41 ± 5	E_1	-67 ± 3
k_{-1}	0.004 ± 0.0002	E_{-1}	232 ± 3

k_1^*	40 ± 3	E_1^*	-45 ± 1
k_{-1}^*	0.008 ± 0.0008	E_{-1}^*	202 ± 0.4
k_2^*	0.015 ± 0.002	E_2^*	~ 0
k_2	104 ± 67	E_2	-87 ± 20
k_{-2}	$7 \cdot 10^{-6} \pm 5 \cdot 10^{-6}$	E_{-2}	182 ± 22

8.1 Fit parameters errors calculation

Errors were estimated using the variance-covariance matrix, calculated as

$$\hat{V} = \frac{RSS}{DOF}(\hat{X}\widehat{W}\hat{X}^T)^{-1}$$

Where RSS (Materials and Methods for calculation) is the sum of squared differences between experimental and calculated data points; DOF is the degrees of freedom, equal to the difference between the total number of data points and the total number of parameters. $\widehat{W}$ is a diagonal matrix of fitting weights, with a default value of 1. Each element of the $\hat{X}$ matrix is described as follows

$$X_{ij} = \frac{\partial(C_j^{exp} - C_j^{fit})}{\partial\theta_i}$$

Where the partial derivative of the difference between the jth experimental and calculated data points is taken with respect to a small variation of the ith parameter. Mathematically,

$$X_{ij} = \frac{\left(C_j^{exp} - C_j^{fit}(+\Delta)\right) - (C_j^{exp} - C_j^{fit}(-\Delta))}{2\Delta}$$

Where $C_j^{fit}(+\Delta)$ and $C_j^{fit}(-\Delta)$ are calculated by using the optimized parameters obtained from the fit, except for the ith parameter, whose value is varied by $\pm\Delta$. The increment Δ is calculated as a percentage of each parameters. For an $\hat{X}$ matrix with p rows (number of parameters) and q columns (number of data points) it follows that

$$\hat{X} = \begin{pmatrix} \frac{\partial C_1}{\partial \theta_1} & \cdots & \frac{\partial C_q}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial C_1}{\partial \theta_p} & \cdots & \frac{\partial C_q}{\partial \theta_p} \end{pmatrix}$$

The diagonal elements of $\hat{V}$ are the variance values for each parameter, the square root gives the standard deviation.

8.2 Combinatorial calculation of the number of partially folded states

Given a sequence with $M-1$ folded GQs and N free G-tracts, the total number of different ways to rearrange them is given by

$$\frac{(N+M-1)!}{N!\,(M-1)!}$$

9 Bibliography

1. Gellert, M., M.N. Lipsett, and D.R. Davies, *HELIX FORMATION BY GUANYLIC ACID.* 1962. **48**(12): p. 2013-2018.

2. Henderson, E., et al., *Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs.* Cell, 1987. **51**(6): p. 899-908.

3. Biffi, G., et al., *Quantitative visualization of DNA G-quadruplex structures in human cells.* Nat Chem, 2013. **5**(3): p. 182-6.

4. Guedin, A., et al., *How long is too long? Effects of loop size on G-quadruplex stability.* Nucleic Acids Res, 2010. **38**(21): p. 7858-68.

5. Han, H. and L.H. Hurley, *G-quadruplex DNA: a potential target for anti-cancer drug design.* Trends Pharmacol Sci, 2000. **21**(4): p. 136-42.

6. Matsugami, A., et al., *New quadruplex structure of GGA triplet repeat DNA--an intramolecular quadruplex composed of a G:G:G:G tetrad and G(:A):G(:A):G(:A):G heptad, and its dimerization.* Nucleic Acids Res Suppl, 2001(1): p. 271-2.

7. Matsugami, A., et al., *Unique quadruplex structures of d(GGA)4 (12-mer) and d(GGA)8 (24-mer)--direct evidence of the formation of non-canonical base pairs and structural comparison.* Nucleic Acids Res Suppl, 2002(2): p. 49-50.

8. Palumbo, S.L., et al., *A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity.* Nucleic Acids Res, 2008. **36**(6): p. 1755-69.

9. Ambrus, A., et al., *Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization.* Biochemistry, 2005. **44**(6): p. 2048-58.

10. Harkness, R.W.t. and A.K. Mittermaier, *G-register exchange dynamics in guanine quadruplexes.* Nucleic Acids Res, 2016. **44**(8): p. 3481-94.

11. Burge, S., et al., *Quadruplex DNA: sequence, topology and structure.* Nucleic Acids Res, 2006. **34**(19): p. 5402-15.

12. Tran, P.L., et al., *Tetramolecular quadruplex stability and assembly.* Top Curr Chem, 2013. **330**: p. 243-73.

13. Ilc, T., et al., *Formation of G-Wires: The Role of G:C-Base Pairing and G-Quartet Stacking.* The Journal of Physical Chemistry C, 2013. **117**(44): p. 23208-23215.

14. Karsisiotis, A.I., et al., *Topological Characterization of Nucleic Acid G-Quadruplexes by UV Absorption and Circular Dichroism.* 2011. **123**(45): p. 10833-10836.

15. Risitano, A. and K.R. Fox, *Influence of loop size on the stability of intramolecular DNA quadruplexes.* Nucleic Acids Research, 2004. **32**(8): p. 2598-2606.

16. Fujii, T., et al., *Effects of metal ions and cosolutes on G-quadruplex topology.* J Inorg Biochem, 2017. **166**: p. 190-198.

17. Bhattacharyya, D., G. Mirihana Arachchilage, and S. Basu, *Metal Cations in G-Quadruplex Folding and Stability.* Front Chem, 2016. **4**: p. 38.

18. *The Role of Cations in Determining Quadruplex Structure and Stability*, in *Quadruplex Nucleic Acids*, S. Neidle and S. Balasubramanian, Editors. 2006, The Royal Society of Chemistry. p. 100-130.

19. Nagesh, N. and D. Chatterji, *Ammonium ion at low concentration stabilizes the G-quadruplex formation by telomeric sequence.* J Biochem Biophys Methods, 1995. **30**(1): p. 1-8.

20. Hardin, C.C., A.G. Perry, and K. White, *Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids.* Biopolymers, 2000. **56**(3): p. 147-94.

21. Miyoshi, D., et al., *Effect of divalent cations on antiparallel G-quartet structure of d(G4T4G4).* FEBS Letters, 2001. **496**(2): p. 128-133.

22. Trajkovski, M., et al., *Pursuing origins of (poly)ethylene glycol-induced G-quadruplex structural modulations.* Nucleic Acids Res, 2018. **46**(8): p. 4301-4315.

23. Miyoshi, D., H. Karimata, and N. Sugimoto, *Hydration regulates thermodynamics of G-quadruplex formation under molecular crowding conditions.* J Am Chem Soc, 2006. **128**(24): p. 7957-63.

24. Nakano, S.-i., et al., *The Effect of Molecular Crowding with Nucleotide Length and Cosolute Structure on DNA Duplex Stability.* Journal of the American Chemical Society, 2004. **126**(44): p. 14330-14331.

25. Miyoshi, D., et al., *Duplex Dissociation of Telomere DNAs Induced by Molecular Crowding.* Journal of the American Chemical Society, 2004. **126**(1): p. 165-169.

26. Zeraati, M., et al., *I-motif DNA structures are formed in the nuclei of human cells.* Nature Chemistry, 2018. **10**(6): p. 631-637.

27. Caruthers, M., *Gene synthesis machines: DNA chemistry and its uses.* 1985. **230**(4723): p. 281-285.

28. Mergny, J.-L. and L. Lacroix, *UV Melting of G-Quadruplexes.* 2009. **37**(1): p. 17.1.1-17.1.15.

29. Gray, R.D. and J.B. Chaires, *Analysis of multidimensional G-quadruplex melting curves.* Curr Protoc Nucleic Acid Chem, 2011. **Chapter 17**: p. Unit17 4.

30. Randazzo, A., G.P. Spada, and M.W. da Silva, *Circular dichroism of quadruplex structures.* Top Curr Chem, 2013. **330**: p. 67-86.

31. del Villar-Guerra, R., J.O. Trent, and J.B. Chaires, *G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy.* 2018. **57**(24): p. 7171-7175.

32. Pagano, B., et al., *Differential scanning calorimetry to investigate G-quadruplexes structural stability.* Methods, 2013. **64**(1): p. 43-51.

33. Dettler, Jamie M., et al., *DSC Deconvolution of the Structural Complexity of c-MYC P1 Promoter G-Quadruplexes.* Biophysical Journal, 2011. **100**(6): p. 1517-1525.

34. Ernesto Freire and R.L. Biltonen, *Statistical mechanical deconvolution of thermal transitions in macromolecules. I. Theory and application to homogeneous systems.* Biopolymers, 1978. **17**(2): p. 463-479.

35. Rhodes, D. and H.J. Lipps, *G-quadruplexes and their regulatory roles in biology.* Nucleic Acids Research, 2015. **43**(18): p. 8627-8637.

36. Huppert, J.L. and S. Balasubramanian, *G-quadruplexes in promoters throughout the human genome.* Nucleic Acids Research, 2007. **35**(2): p. 406-413.

37. Cogoi, S. and L.E. Xodo, *G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription.* Nucleic Acids Res, 2006. **34**(9): p. 2536-49.

38. Endoh, T. and N. Sugimoto, *Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells.* Sci Rep, 2016. **6**: p. 22719.

39. O'Sullivan, R.J. and J. Karlseder, *Telomeres: protecting chromosomes against genome instability.* Nature Reviews Molecular Cell Biology, 2010. **11**: p. 171.

40. Jafri, M.A., et al., *Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies.* 2016. **8**(1): p. 69.

41. Oganesian, L. and J. Karlseder, *Telomeric armor: the layers of end protection.* Journal of Cell Science, 2009. **122**(22): p. 4013.

42. Wang, Q., et al., *G-quadruplex formation at the 3' end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase.* Nucleic Acids Research, 2011. **39**(14): p. 6229-6237.

43. Min, J., W.E. Wright, and J.W. Shay, *Alternative lengthening of telomeres can be maintained by preferential elongation of lagging strands.* Nucleic acids research, 2017. **45**(5): p. 2615-2628.

44. de Lange, T., *Shelterin: the protein complex that shapes and safeguards human telomeres.* 2005. **19**(18): p. 2100-2110.

45. Neidle, S., *Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer.* FEBS J, 2010. **277**(5): p. 1118-25.

46. Perrone, R., et al., *A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter.* J Med Chem, 2013. **56**(16): p. 6521-30.

47. Haq, I., et al., *Intercalative G-Tetraplex Stabilization of Telomeric DNA by a Cationic Porphyrin1.* Journal of the American Chemical Society, 1999. **121**(9): p. 1768-1779.

48. Martino, L., et al., *Shedding Light on the Interaction between TMPyP4 and Human Telomeric Quadruplexes.* The Journal of Physical Chemistry B, 2009. **113**(44): p. 14779-14786.

49. Zaffaroni, N., et al., *Inhibition of telomerase activity by a distamycin derivative: effects on cell proliferation and induction of apoptosis in human cancer cells.* European Journal of Cancer, 2002. **38**(13): p. 1792-1801.

50. Pagano, B., et al., *Targeting DNA quadruplexes with distamycin A and its derivatives: An ITC and NMR study.* Biochimie, 2008. **90**(8): p. 1224-1232.

51. Freyer, M.W. and E.A. Lewis, *Isothermal Titration Calorimetry: Experimental Design, Data Analysis, and Probing Macromolecule/Ligand Binding and Kinetic Interactions*, in *Methods in Cell Biology*. 2008, Academic Press. p. 79-113.

52. Pagano, B., C.A. Mattia, and C. Giancola, *Applications of isothermal titration calorimetry in biophysical studies of G-quadruplexes.* Int J Mol Sci, 2009. **10**(7): p. 2935-57.

53. Giancola, C. and B. Pagano, *Energetics of Ligand Binding to G-Quadruplexes*, in *Quadruplex Nucleic Acids*, J.B. Chaires and D. Graves, Editors. 2013, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 211-242.

54. Bugaut, A. and P. Alberti, *Understanding the stability of DNA G-quadruplex units in long human telomeric strands.* Biochimie, 2015. **113**: p. 125-33.

55. Yu, H., et al., *Beads-on-a-String Structure of Long Telomeric DNAs under Molecular Crowding Conditions.* Journal of the American Chemical Society, 2012. **134**(49): p. 20060-20069.

56. Petraccone, L., et al., *Structure and Stability of Higher-Order Human Telomeric Quadruplexes.* Journal of the American Chemical Society, 2011. **133**(51): p. 20951-20961.

57. Bauer, L., et al., *G-quadruplex motifs arranged in tandem occurring in telomeric repeats and the insulin-linked polymorphic region.* Biochemistry, 2011. **50**(35): p. 7484-92.

58. Wang, H., et al., *Single molecule studies of physiologically relevant telomeric tails reveal POT1 mechanism for promoting G-quadruplex unfolding.* J Biol Chem, 2011. **286**(9): p. 7479-89.

59. Martínez, P. and M.A. Blasco, *Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins.* Nature Reviews Cancer, 2011. **11**: p. 161.

60. McGhee, J.D. and P.H. von Hippel, *Theoretical aspects of DNA-protein interactions: Co-operative and non-co-operative binding of large ligands to a one-dimensional homogeneous lattice.* Journal of Molecular Biology, 1974. **86**(2): p. 469-489.

61. Lane, A.N., et al., *Stability and kinetics of G-quadruplex structures.* Nucleic Acids Res, 2008. **36**(17): p. 5482-515.

62. Yue, D.J.E., K.W. Lim, and A.T. Phan, *Formation of (3+1) G-Quadruplexes with a Long Loop by Human Telomeric DNA Spanning Five or More Repeats.* Journal of the American Chemical Society, 2011. **133**(30): p. 11462-11465.

63. Tang, J., et al., *G-quadruplex preferentially forms at the very 3' end of vertebrate telomeric DNA.* Nucleic Acids Res, 2008. **36**(4): p. 1200-8.

64. Mergny, J.-L., et al., *Kinetics of tetramolecular quadruplexes.* Nucleic Acids Research, 2005. **33**(1): p. 81-94.

65. Bardin, C. and J.L. Leroy, *The formation pathway of tetramolecular G-quadruplexes.* Nucleic Acids Research, 2008. **36**(2): p. 477-488.

66. Kankia, B., et al., *Stable Domain Assembly of a Monomolecular DNA Quadruplex: Implications for DNA-Based Nanoswitches.* Biophys J, 2016. **110**(10): p. 2169-75.

67. Abraham Punnoose, J., et al., *Random Formation of G-Quadruplexes in the Full-Length Human Telomere Overhangs Leads to a Kinetic Folding Pattern with Targetable Vacant G-Tracts.* Biochemistry, 2018. **57**(51): p. 6946-6955.

68. Zaug, A.J., E.R. Podell, and T.R. Cech, *Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(31): p. 10864-10869.

69. Colgin, L.M., et al., *Human POT1 Facilitates Telomere Elongation by Telomerase.* Current Biology, 2003. **13**(11): p. 942-946.

70. Baumann, P. and C. Price, *Pot1 and telomere maintenance.* FEBS Letters, 2010. **584**(17): p. 3779-3784.

71. Hwang, H., et al., *POT1-TPP1 Regulates Telomeric Overhang Structural Dynamics.* Structure, 2012. **20**(11): p. 1872-1880.

72. Pandita, R.K., et al., *Single-Strand DNA-Binding Protein SSB1 Facilitates TERT Recruitment to Telomeres and Maintains Telomere G-Overhangs.* 2015. **75**(5): p. 858-869.

73. Safa, L., et al., *Binding polarity of RPA to telomeric sequences and influence of G-quadruplex stability.* Biochimie, 2014. **103**: p. 80-8.

74. Flynn, R.L., S. Chang, and L. Zou, *RPA and POT1: friends or foes at telomeres?* Cell cycle (Georgetown, Tex.), 2012. **11**(4): p. 652-657.

75. Rajavel, M., M.R. Mullins, and D.J. Taylor, *Multiple facets of TPP1 in telomere maintenance.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2014. **1844**(9): p. 1550-1559.