

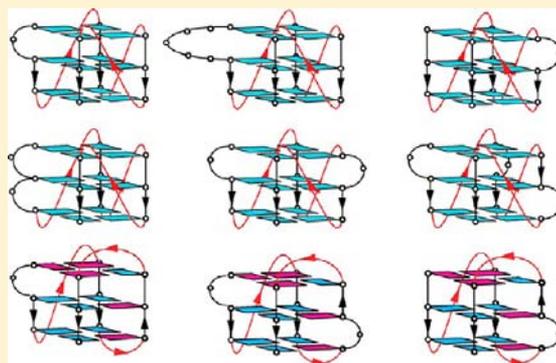
Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences

Vineeth Thachappilly Mukundan and Anh Tuân Phan*

School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore

S Supporting Information

ABSTRACT: We report on the first solution structure of an intramolecular G-quadruplex containing a single bulge and present evidence for extensive occurrence of bulges in different G-quadruplex contexts. The NMR solution structure of the d-(TTGTGGTGGGTGGGTGGGT) sequence reveals a propeller-type parallel-stranded G-quadruplex containing three G-tetrad layers, three double-chain-reversal loops, and a bulge. All guanines participate in the formation of the G-tetrad core, despite the interruption between the first guanine and the next two guanines by a thymine, which forms a single-residue bulge and is projected out of the G-tetrad core. To provide a more general understanding about the formation of bulges within G-quadruplexes, we systematically investigated the effects of the residue type, the size, the position, and the number of bulges on the structure and stability of G-quadruplexes. The formation of bulges has also been observed in two different G-quadruplex scaffolds with different strand orientations and folding topologies. Our results show that bulges can be formed in many different situations within G-quadruplexes. While many sequences tested in this study can form stable G-quadruplex structures, all of them defy the description of sequences $G_{3+N_{L1}}G_{3+N_{L2}}G_{3+N_{L3}}G_{3+}$, currently used in most bioinformatics searches for identifying potential G-quadruplex-forming sequences in the genomes. Broadening of this description to include the possibilities of bulge formation should allow the identification of more G-quadruplex-forming sequences which went unnoticed in the earlier searches. This study could also open the possibilities of exploiting bulges as recognition elements for interactions between G-quadruplexes and other molecules.



INTRODUCTION

G-quadruplexes are four-stranded structures formed by guanine-rich DNA or RNA sequences.^{1–3} There is growing evidence that such structures can form in the genome of many organisms and play important regulatory roles in gene expression and genome stability due to their involvement in the key cellular processes, such as transcription, recombination, and replication.^{4–9} The anticancer effect of their formation in the telomeres and promoter regions of several oncogenes makes them attractive targets in cancer research.^{10–12} On the other hand, some G-quadruplex-forming sequences have been shown to possess anticancer and anti-HIV activities.^{13–20}

The basic unit of a G-quadruplex structure is a G-tetrad, which is a cyclic planar alignment of four guanines each forming hydrogen bonds with two neighbors.²¹ A G-quadruplex is formed by stacking of G-tetrads on top of each other and further stabilized by cations located at the center of the structure.²² G-quadruplexes are polymorphic depending on the nucleic acid sequence and experimental conditions:^{1–3,23} the four strands serving as columns supporting the G-tetrad core can be oriented in the same or opposite direction with respect to each other; and loops connecting these columns can also adopt different configurations. G-quadruplex structures could be intra- or intermolecular forming from single or separate strands.^{1–3}

Intramolecular G-quadruplexes have attracted most attention as they can readily form in the genome in the context of single strands under cellular conditions.

An intramolecular G-quadruplex is often considered to be formed by a sequence containing four tracts of three or more continuous guanines connected by linkers, in which the G-tracts would form continuous columns supporting the G-tetrad core, while the linkers would form loops connecting the corners of the G-tetrad core. However, exceptions to this rule have been observed by NMR and X-ray crystallography. The sequences Pu24 from the human *c-myc* promoter²⁴ and *c-kit87up* from the human *c-kit* promoter²⁵ are two notable examples found in the human genome. Both these sequences exhibit discontinuous arrangement of guanines in one column of the G-tetrad core, despite the presence of four G-tracts each having at least three continuous guanines.^{24,25} Additionally, an isolated guanine participates in the G-tetrad core of the structure formed by *c-kit87up*.²⁵

Like loops, bulges are projections of bases from the G-tetrad core. However, while a loop connects two corners of the G-tetrad core, a bulge connects two adjacent guanines of the same column

Received: October 17, 2012

Published: March 22, 2013

of the G-tetrad core.^{26–28} Bulges are commonly observed in RNA duplex contexts with mismatches²⁹ and are involved in interactions with other nucleic acids or proteins.³⁰ The crystal structure of a tetrameric parallel G-quadruplex formed by the r(U)d(^{Br}G)r(UGGU) or r(UGGUGU) sequence revealed four bulges of uracil residues (from four different strands) projected outside of the G-tetrad core.^{26,27} Recently, we have shown by NMR that the d(GTGGTGGGTGGGTGGGT) sequence, named T30177, forms a six-layer G-quadruplex structure through the stacking of two parallel G-quadruplex subunits each containing a single bulge adopted by the thymine T2 residue.²⁸ This is also an example of a nonstandard sequence which involves an isolated guanine (G1) in the G-quadruplex formation. However, the high-resolution structure of this G-quadruplex could not be determined due to spectral broadening at the stacking interface.²⁸

Here, with the introduction of two thymines at the 5' end of the T30177 sequence, we could greatly improve the quality of the NMR spectra by destabilizing the stacking interaction between two G-quadruplex subunits, thus, favoring the formation of a single intramolecular G-quadruplex unit. This has allowed us to determine the NMR solution structure of this intramolecular G-quadruplex containing a single bulge. To provide a more general understanding about the formation of bulges within G-quadruplexes, we systematically investigated the effects of the sequence, the size, the position, and the number of bulges on the structure and stability of G-quadruplexes using UV absorption, CD, and NMR spectroscopy. This study revealed that bulges can be formed in many different situations within G-quadruplex structures.

This result alters the common view on the ability of many sequences to form G-quadruplexes. All the sequences tested defy the standard notion that continuous G-tracts are needed for quadruplex formation. Some sequences, with as many as three isolated guanines, still form stable G-quadruplex structures, in which these bases can participate in the G-tetrad core formation. Based on these findings, the sequence description used in most current bioinformatics searches for identifying putative G-quadruplex-forming sequences could be reformulated, and the number of putative G-quadruplexes in the genome is expected to be larger than previously reported.

RESULTS AND DISCUSSION

Conversion from a Stacked Dimer to a Monomeric G-Quadruplex. Recently, the T30177 oligonucleotide with the sequence d(GTGGTGGGTGGGTGGGT) was shown to form a dimeric structure involving the 5'-end stacking of two parallel-stranded G-quadruplex subunits each containing a bulge between the first and the second guanines.²⁸ Addition of non-guanine bases at the 5' end could break the stacking between the two G-quadruplex monomers.²⁸ Figure 1 displays NMR imino proton spectra of T30177 and sequences containing one (T30177-T) and two (T30177-TT) additional thymine bases at the 5' end: T30177 shows broad imino proton peaks corresponding to the stacked dimeric G-quadruplex;²⁸ T30177-T shows similar broad peaks at low abundance along with 12 major sharp peaks, indicating the coexistence of both stacked dimeric and monomeric G-quadruplexes; and T30177-TT (also called TB-1 in this paper) shows only 12 sharp well-resolved peaks corresponding to the formation of the monomeric G-quadruplex only (Table 1). The solution structure determination of TB-1 will be presented in the next section. The relative population between the stacked dimer and the monomer was also shown to

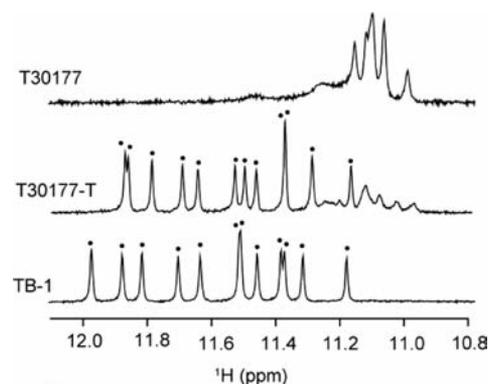


Figure 1. Imino proton spectra of T30177, T30177-T, and TB-1 (or T30177-TT) in K⁺ solution at 25 °C. Black dots indicate 12 imino protons of the monomeric G-quadruplex form.

Table 1. DNA Sequences Used to Demonstrate the Dimer–Monomer Conversion^a

name	sequence (5'–3')
T30177	GTGGTGGGTGGGTGGGT
T30177-T	TGTGGTGGGTGGGTGGGT
T30177-TT (or TB-1)	TTGTGGTGGGTGGGTGGGT

^aResidues expected to form a bulge are shown in boldface.

depend on the temperature, the DNA concentration, and the ionic strength for these and other related sequences. The monomeric form was favored at higher temperature (e.g., see Figure S1) or at lower DNA concentration and ionic strength.³¹

Solution Structure of a G-Quadruplex Containing a Single Bulge. The predominant formation of a monomeric G-quadruplex for TB-1 and its high-quality NMR spectra provided necessary conditions for the first high-resolution structure determination of a G-quadruplex containing a single bulge. Resonances of TB-1 were unambiguously assigned using site-specific labeling and through-bond correlation approaches^{32–34} (Figure 2). The guanine H8 protons were assigned to their respective positions by site-specific proton-to-deuterium substitutions³⁴ for each guanine in the sequence. Guanine imino protons were assigned using the through-bond correlations via ¹³C5 at natural abundance with the already assigned H8 protons.³³ Some imino proton assignments were independently confirmed by the site-specific low-enrichment approach,³² using 2% ¹⁵N-labeled samples. The assignments of other protons (Figure 3) were completed by a combination various through-bond and through-space experiments (COSY, TOCSY, ¹³C–¹H HSQC, and NOESY).³⁵

The folding topology of TB-1 was determined from NOEs observed between imino and H8 protons (Figure 4). Different colors are used to distinguish NOEs between protons of three different G-tetrads: G3·G8·G12·G16 (red), G5·G9·G13·G17 (green), and G6·G10·G14·G18 (purple). The strands in the G-tetrad core are parallel to each other, and the thymine between the first two guanines forms a bulge between two G-tetrads. Moderate intensities of the cross-peaks between H8 and H1' protons (Figure 3) indicate *anti* glycosidic conformations for all guanines, consistent with the formation of a parallel-stranded G-quadruplex. The folding topology of TB-1 was further supported by a solvent exchange experiment (Figure S2). After dissolving the sample in D₂O solution, the imino protons of the central G-tetrad layer remained protected, while other imino protons

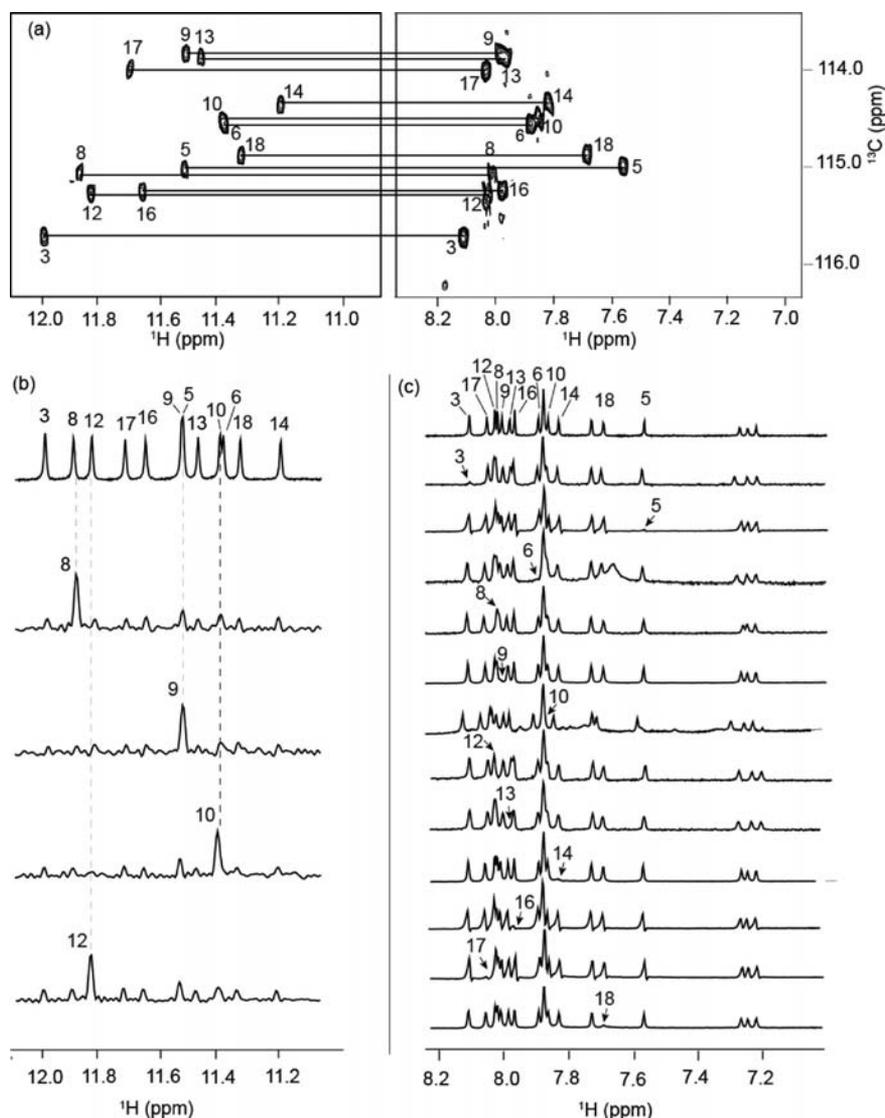


Figure 2. NMR spectral assignments of guanine imino and H8 protons of TB-1 in K^+ solution. (a) Through-bond correlations between guanine imino and H8 protons via $^{13}\text{C}_5$ at natural abundance. (b) Guanine imino proton assignments using site-specific 2% ^{15}N -labeled samples. (c) Guanine H8 proton assignments using site-specific deuterated samples.

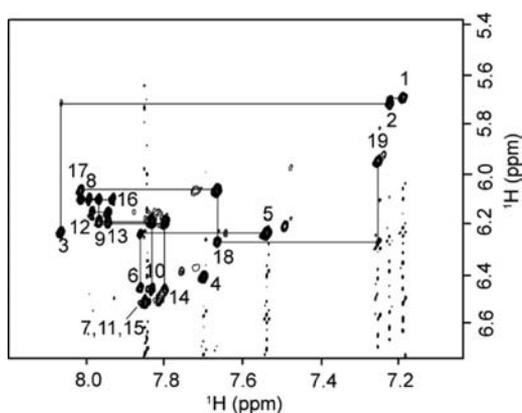


Figure 3. NOESY spectrum (mixing time, 200 ms) showing H8/H6–H1' NOE sequential connectivities of TB-1 in K^+ solution. Intraresidue cross-peaks are labeled with the corresponding residue numbers.

disappeared quickly. This contrasts the solvent exchange data for the stacked dimer of T30177, where imino protons from the

stacking interface were also well protected from the exchange with solvent.²⁸ The CD spectrum of TB-1 (Figure S3) gives a positive peak at 264 nm and a negative peak 240 nm, characteristic of parallel-stranded G-quadruplexes.

The structure of TB-1 (Figure 5) was determined on the basis of NMR restraints (Table 2) using the X-PLOR program.³⁶ This structure is a three-layer propeller-type parallel-stranded G-quadruplex containing three single-residue double-chain-reversal loops (T7, T11, and T15) and a bulge (T4). The first two thymines T1 and T2 are positioned on top of the upper G-tetrad with T2 being well stacked on the G-tetrad and T1 slightly tilted. The thymine at the last position (T19) is also found to be partially stack at the bottom of the lower G-tetrad layer. Thymines capping at both ends of the G-tetrad core might be responsible for preventing further stacking between different G-quadruplex blocks. The bulge T4 base is projected out with N3 and O2 atoms pointing away from the G-tetrad core. The guanine G3 preceding the bulge is tilted downward from the plane of the G-tetrad, and the corresponding sugar shows a near C3'-endo conformation

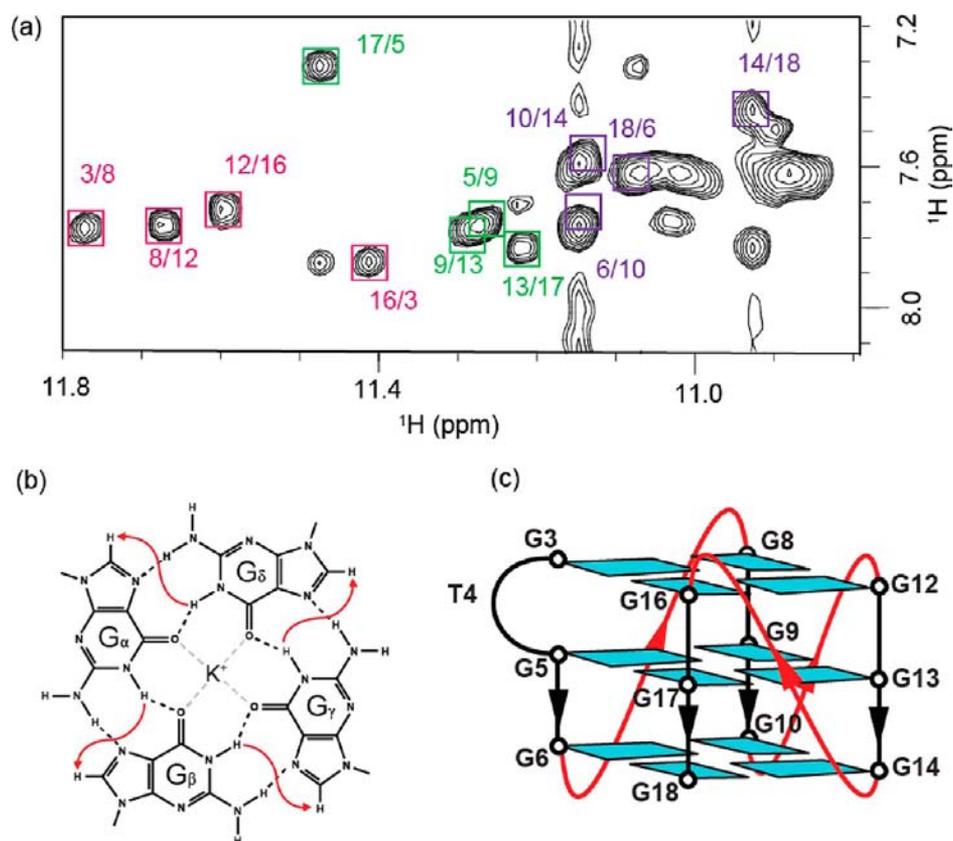


Figure 4. G-quadruplex fold determination for TB-1. (a) NOESY spectrum (mixing time, 200 ms), showing the imino-H8 connectivities around different G-tetrads. The characteristic guanine imino-H8 cross-peaks for G-tetrads are framed and labeled with the imino proton assignment in the first position and that of the H8 proton in the second position. The residues in different G-tetrads are indicated by different colors. (b) Specific imino-H8 connectivity pattern around a G_α - G_β - G_γ - G_δ tetrad indicated with arrows. (c) Schematic structure of the G-quadruplex adopted by TB-1 in K^+ solution.

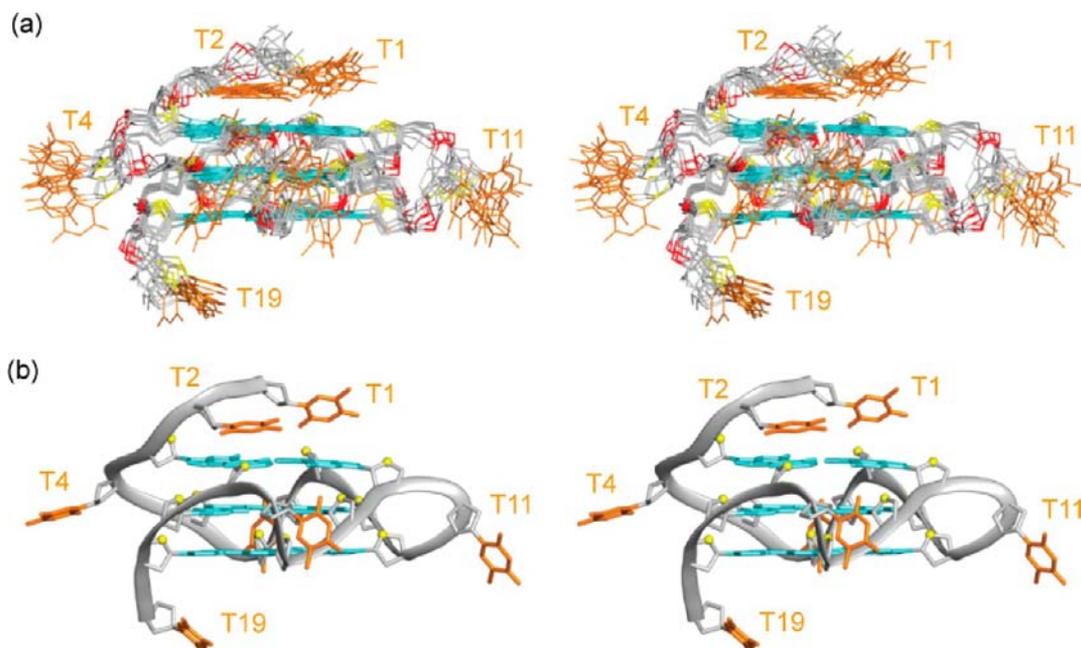


Figure 5. Stereoviews of the G-quadruplex structure formed by TB-1 in K^+ solution. (a) Ten superimposed refined structures. (b) Ribbon view of a representative structure. Guanines are colored in cyan; thymines, orange; backbone and sugar, gray; $\text{O}4'$ atoms, yellow; phosphorus atoms, red.

consistent with a strong NOE cross peak between $\text{H}3'$ and $\text{H}8$ protons of G3.

This is the first solution structure of a G-quadruplex with a single bulge in contrast to the previously reported crystal structures^{26,27} of tetrameric parallel G-quadruplexes of r(U)d-

Table 2. Statistics of the Computed Structures of the TB-1 G-Quadruplex

	D ₂ O	H ₂ O
A. NMR Restraints		
distance restraints		
intraresidue	238	0
sequential (<i>i</i> , <i>i</i> + 1)	68	2
long-range (<i>i</i> , <i>i</i> + 2)	6	26
other restraints		
hydrogen-bond	48	
dihedral	30	
repulsive	18	
B. Structure Statistics		
NOE violations		
number (>0.2 Å)		0
maximum violation (Å)	0.134 ± 0.039	
rmsd of violations (Å)	0.013 ± 0.002	
deviations from the ideal covalent geometry		
bond lengths (Å)	0.003 ± 0.000	
bond angles (°)	0.669 ± 0.006	
impropers (°)	0.349 ± 0.005	
pairwise all heavy atom rmsd (Å)		
all heavy atoms except T1, T2, T4, T7, T11, T15, and T19	0.84 ± 0.10	
all heavy atoms	2.10 ± 0.34	

(^{Br}G)_r(UGGU) and r(UGGUGU), where four uracil residues from four different strands form bulges simultaneously. However, the bulges in these structures share some similarities. In all cases, the presence of a bulge does not alter the global structure of the G-quadruplex. Potential hydrogen-bond acceptors O2 and N3 of the bulge bases (thymine and uracil) are pointed away from the G-tetrad core and available for recognition. However, the bulge bases were observed in different orientations with respect to the G-tetrad core in the three structures. The bulge residues adopt a *syn* conformation in r(U)(^{Br}dG)_r(UGGU) and an *anti* conformation in r(UGGUGU) and the current structure.

Effect of the Bulge Residue Type. In genomic sequences, G-tracts could be interrupted not only by thymines but also by other bases, such as cytosines or adenines. The size differences between purines and pyrimidines and the ability of these bases to be involved in secondary interactions could have an effect on the structure adopted by these sequences. In this section, we replace the thymine T4 by a cytosine and adenine, resulting in sequences CB-1 and AB-1 (Table 3). Guanine was not chosen, since it would result in a continuous G-tract, leading to the formation of a competing G-quadruplex structure. Figure 6 summarizes the experimental results. NMR spectra of the three sequences (TB-1, CB-1, and AB-1) are nearly identical, indicating the formation of the same G-quadruplex fold with a single bulge. CD spectra of the three sequences are also similar, supporting the formation of the same G-quadruplex fold. The UV-melting data show very similar melting temperatures for thymine and cytosine but a slightly lower value for adenine (Table 3).

The same melting temperature observed for TB-1 and CB-1 suggests that bulges in these contexts are not involved in any secondary interactions with the rest of the structure. The lower melting temperature observed for adenine could result from an increase in solvation entropy associated with the exposure of a bigger-size aromatic base. This result is comparable with a previous observation, in which the propeller loops made of a

thymine or cytosine resulted in similar melting temperatures, while loops with an adenine led to a lower melting temperature.³⁷

Effect of the Bulge Size. In this section, we study the effect of increasing the number of thymines (from one to seven) between the first two guanines. This has the effect of pushing these two guanines further apart in the sequence (Table 3). Figure 7 shows the results for all the seven sequences tested (TB-1, T2B-1, T3B-1, T4B-1, T5B-1, T6B-1, and T7B-1) with different number of thymines at the position that is expected to form a bulge. One-dimensional NMR imino proton spectra of these sequences show significant similarities, with all having 12 peaks, indicating a structure involving all the guanines in the G-tetrad core. Gradual increase in the number of thymines resulted in small changes in the spectrum, with some peaks coalescing while others separating from one another. The changes are dominant for peaks corresponding to the guanines G3 and G5 of the reference sequence, which are situated next the bulge. With four or more thymines at the bulge position, the spectrum of the major form remains unchanged, while additional small peaks appear. The associated CD spectra of the tested sequences are similar and characteristic of parallel-stranded G-quadruplexes.

The melting behavior of the tested sequences in the presence of 12 mM K⁺ is shown in Figure 7c. The melting temperature gradually decreases for the first four sequences with a 10 °C difference for each additional thymine base. The melting temperature remains nearly the same for the fourth, fifth and sixth sequences, while decreasing further for the last sequence T7B-1. The same trend is observed in the presence of 60 mM K⁺ but with higher melting temperatures (Table 3).

Taken all together, these results strongly suggest that all these sequences adopt the same folding topology as the reference structure but with bulges of different sizes. The stability of the structures decreases when the bulge size increases. This can be compared with the previous observations that the stability of G-quadruplexes decreases when the size of propeller loops increases.^{37–42}

Effect of the Bulge Position. G-quadruplex structures are sensitive to the sequence of bases, with even small modifications having a dramatic effect on the structure.^{43–45} Even though a bulge has been shown to form between the first two guanines of the propeller G-quadruplex scaffold, it is not clear whether bulges can form in other parts of the structure. To address this question we inserted a thymine between each of the eight pair of successive guanines, once at a time. NMR spectra of the eight sequences (TB-1–8, Table 3) are shown in Figure 8a. All the eight NMR spectra are similar and retain characteristic features of the reference spectrum. Most of the peaks are distinct and well resolved from each other. We could count 12 imino proton peaks in each spectrum indicating the formation a single three-layer G-quadruplex structure where all guanines participate in the G-tetrad core. CD spectra of all the sequences are almost the same and characteristic of parallel-stranded G-quadruplexes.

Melting curves in Figure 8c show the behavior of all the sequences in the presence of 12 mM K⁺. The reference sequence TB-1 has the highest melting temperature, while all the other sequences have nearly the same melting temperature, which is about 10 °C lower than that of TB-1.

These results suggest that bulges can be located at many positions in a G-quadruplex structure. In other words, presence of an isolated guanine at any position does not inhibit the formation of a G-quadruplex structure. However, we found that the bulge position in the sequence TB-1 is unique, as shown by the higher melting temperature of TB-1, compared to that of the

Table 3. DNA Sequences Used to Investigate the Effect of Bulges on the Structure and Stability of G-Quadruplexes^a

name	sequence (5'-3')	melting temperature (°C)	
		in 12 mM K ⁺	in 60 mM K ⁺
TB-1 ^b	TTGTGGTGGGTGGGTGGGT	77	>85
CB-1	TTGCGGTGGGTGGGTGGGT	77	>85
AB-1	TTGAGGTGGGTGGGTGGGT	73	>82
T2B-1	TTGTTGGTGGGTGGGTGGGT	63	71
T3B-1	TTGTTTGGTGGGTGGGTGGGT	53	62
T4B-1	TTGTTTTGGTGGGTGGGTGGGT	43	52
T5B-1	TTGTTTTTGGTGGGTGGGTGGGT	43	52
T6B-1	TTGTTTTTTGGTGGGTGGGTGGGT	43	52
T7B-1	TTGTTTTTTTGGTGGGTGGGTGGGT	<36	48
TB-2	TTGGTGTGGGTGGGTGGGT	67	77
TB-3	TTGGGTGTGGGTGGGTGGGT	69	78
TB-4	TTGGGTGGTGTGGGTGGGT	68	77
TB-5	TTGGGTGGGTGTGGTGGGT	68	77
TB-6	TTGGGTGGGTGGTGTGGGT	66	75
TB-7	TTGGGTGGGTGGGTGTGGT	67	76
TB-8	TTGGGTGGGTGGGTGGTGT	68	76
TB-1-2	TTGTGTGTGGGTGGGTGGGT	46	56
TB-1-4	TTGTGGTGGTGTGGGTGGGT	54	63
TB-1-5	TTGTGGTGGGTGTGGTGGGT	56	65
TB-1-8	TTGTGGTGGGTGGGTGGTGT	53	63
TB-3-4	TTGGGTGTGTGTGGGTGGGT	<36	46
TB-3-5	TTGGGTGTGGTGTGGGTGGGT	50	59
TB-3-6	TTGGGTGTGGTGGTGTGGGT	46	56
TB-4-5	TTGGGTGGTGTGTGGTGGGT	45	55
TB-5-6	TTGGGTGGGTGTGTGGGT	<33	45
TB-7-8	TTGGGTGGGTGGGTGTGTGT	<38	47
TB-1-2-3	TTGTGTGTGTGGGTGGGTGGGT	<27	<35
TB-1-3-5	TTGTGGTGTGGTGTGGTGGGT	<35	48
TB-1-5-7	TTGTGGTGGGTGTGGTGTGGT	<33	46
TB-2-3-8	TTGGTGTGTGGTGGGTGGTGT	<22	<29
TB-4-5-8	TTGGGTGGTGTGTGGTGGTGT	<24	<33
HT	TTGGGTTAGGGTTAGGGTTAGGGA	–	64
HT-T1	TTGTGGTTAGGGTTAGGGTTAGGGA	–	48
HT-T5	TTGGGTTAGGGTTAGTGGTTAGGGA	–	56
HT-T1-5	TTGTGGTTAGGGTTAGTGGTTAGGGA	–	45
TB-1-2-3-4	TTGTGTGTGTGTGGGTGGGT	–	NOG4 ^c
TB-1-3-5-7	TTGTGGTGTGGTGTGGTGTGGT	–	NOG4 ^c
TB-1-4-7-8	TTGTGGTGGTGTGGGTGTGTGT	–	NOG4 ^c
TB-2-3-5-8	TTGGTGTGTGGTGGGTGTGG T GT	–	NOG4 ^c
TB-1-2-3-4-5	TTGTGTGTGTGTGTGGTGGGT	–	NOG4 ^c
TB-1-2-3-4-5-6	TTGTGTGTGTGTGTGTGGTGGGT	–	NOG4 ^c
TB-1-2-3-4-5-6-7-8	TTGTGTGTGTGTGTGTGTGGTGGGT	–	NOG4 ^c

^aResidues expected to form bulges are shown in boldface. ^bG-quadruplex structure formed by the d(TTGGGTGGGTGGGTGGGT) sequence containing no bulge³¹ was too stable at both tested K⁺ concentrations and could not be used as a reference. ^cNo G-quadruplex formation observed.

other sequences. This effect could depend on the context of the sequence and G-quadruplex topology.

Effect of the Number of Bulges. Previous sections have shown that a bulge, which varies in the residue type, the size, or the position, can exist in a G-quadruplex structure. However, all these sequences have only one bulge in the structure resulted from a single interruption in a G-tract. In this section, we explore the ability of sequences with G-tracts interrupted at more than one location to form G-quadruplex structures. Since bulges generally reduce the stability of a G-quadruplex structure, there could be a limit in the number of bulges that the structure could tolerate.

We have tested ten different sequences (TB-1-2, TB-1-4, TB-1-5, TB-1-8, TB-3-4, TB-3-5, TB-3-6, TB-4-5, TB-5-6, and TB-7-

8), in which single thymines are inserted between two pairs of adjacent guanines at different positions in the sequence (Table 3). NMR spectra of all 10 sequences are shown in Figure 9, together with the reference spectrum TB-1. All the spectra show 12 imino proton peaks and retain the characteristics of the reference spectrum. The CD spectra of all sequences are similar and characteristic of parallel-stranded G-quadruplexes.

However, the melting temperatures are significantly different among different sequences (Table 3). Figure 9c shows the melting data in the presence of 60 mM K⁺. The 10 melting curves are arranged in three different groups. The sequences TB-1-4, TB-1-5 and TB-1-8 have melting temperatures in the range of 63–65 °C, followed by TB-1-2, TB-3-5 and TB-3-6 whose melting temperatures are in the range of 55–59 °C. The last

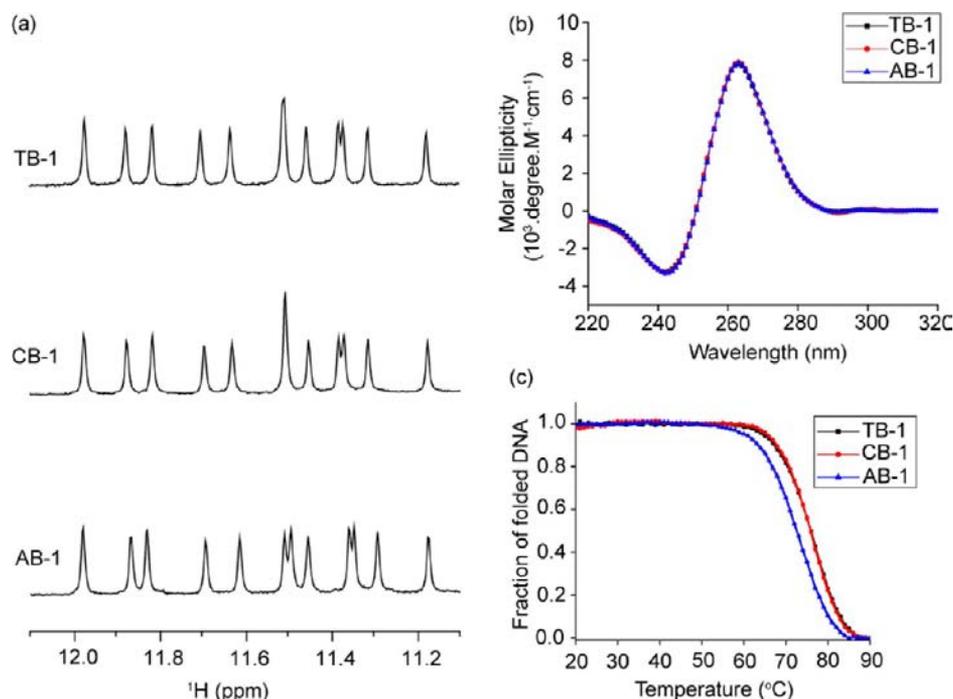


Figure 6. Effect of the bulge residue type on the structure and stability of a G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for G-quadruplex-forming sequences with a bulge of different residue types. UV melting experiments were performed in 12 mM K^+ solution.

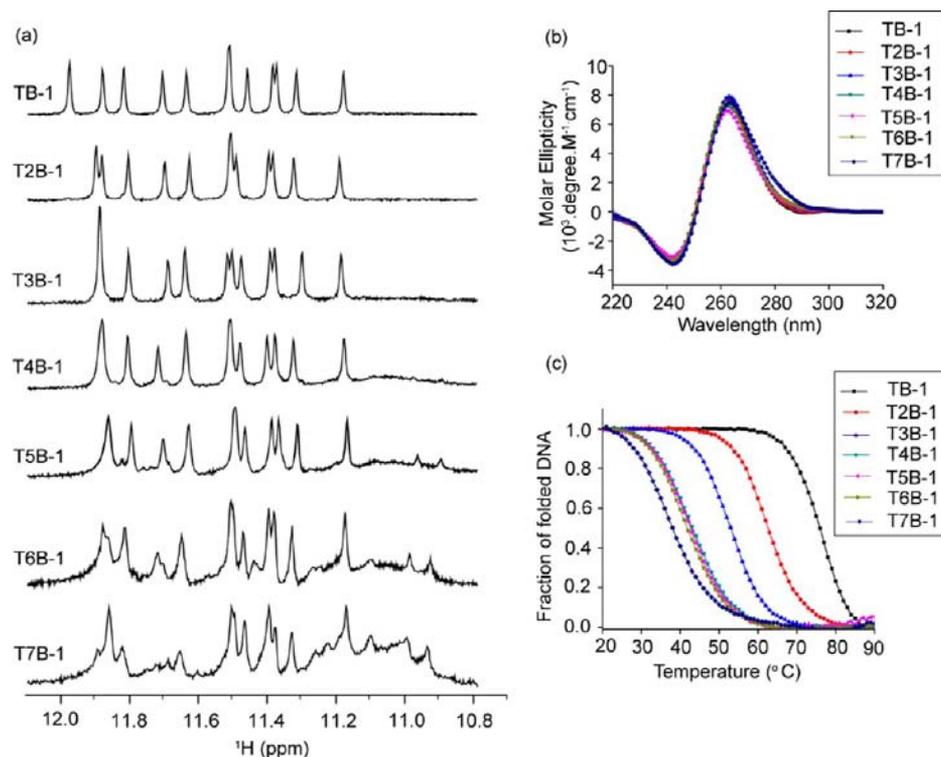


Figure 7. Effect of the bulge size on the structure and stability of a G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for G-quadruplex-forming sequences with a bulge of different sizes. UV melting experiments were performed in 12 mM K^+ solution.

group comprising TB-3-4, TB-5-6 and TB-7-8 has the lowest melting temperatures, which is in the range of 45–47 $^{\circ}\text{C}$.

Figure 10 summarizes the results from five sequences in which guanines are interrupted by thymines at three different locations (Table 3). Again, all the NMR spectra show 12 imino proton peaks and retain the features of the reference spectrum and CD

spectra are characteristic of parallel-stranded G-quadruplexes. The sequences TB-1-3-5 and TB-1-5-7 have the highest melting temperatures, followed by TB-1-2-3 in which thymines were inserted continuously between two successive pairs of guanines within the same G-tract. TB-2-3-8 and TB-4-5-8 have the lowest

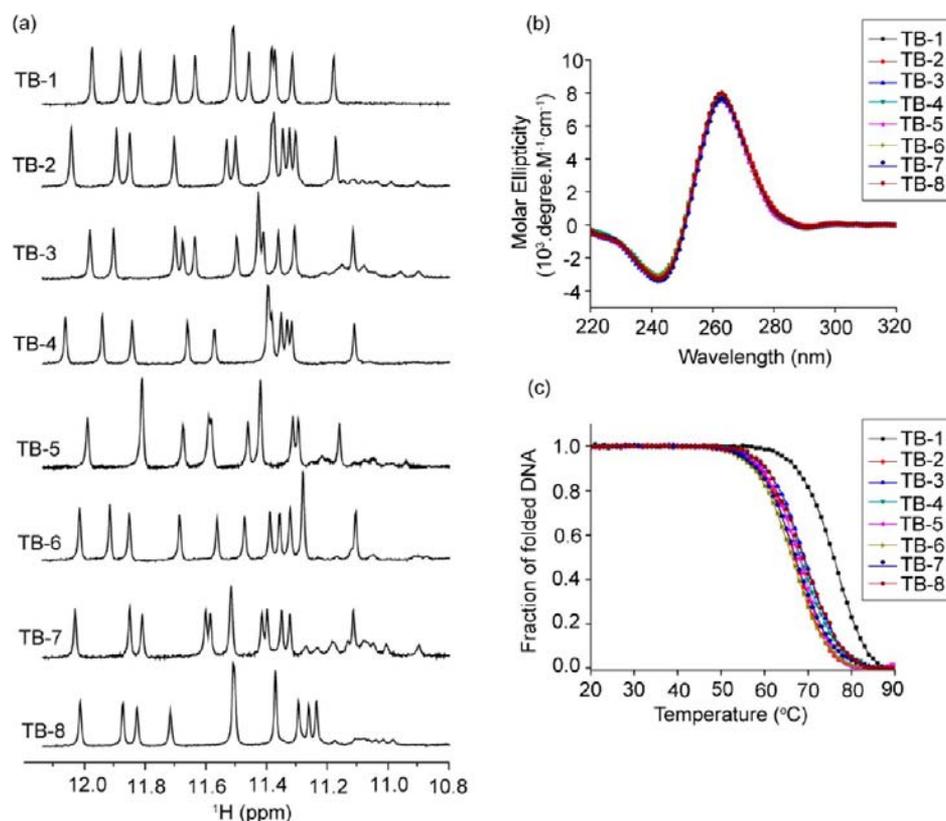


Figure 8. Effect of the bulge position on the structure and stability of a G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for G-quadruplex-forming sequences with a bulge at different positions. UV melting experiments were performed in 12 mM K^+ solution.

melting temperatures even though thymines were inserted in different G-tracts, in contrast to TB-1-2-3.

All sequences tested by us with thymines inserted in more than three locations (Table 3) did not show evidence for G-quadruplex formation in the NMR imino proton spectra in 60 mM K^+ solution (Figure S4).

The presence of 12 imino proton peaks in the NMR spectra for all tested sequences containing thymine insertions at two or three locations indicates the involvement of all the guanines in the G-tetrad core formation. Together with the similarities with the NMR and CD spectra of the reference structure, these results suggest a similar three-layer G-quadruplex structure with two or three bulges. The structures with two thymines inserted in the same G-tract are found to be less stable than those with two thymines in different G-tracts. With three thymine insertions, stability further decreases even when these thymines are situated far apart. Again, sequences involving a thymine at the bulge position 1 exhibit higher melting temperatures than the counterparts with an equivalent bulge at other positions.

Bulges in a Different G-Quadruplex Scaffold. We also examined the formation of bulges in a (3 + 1) G-quadruplex scaffold adopted by the human telomeric sequence d-(TTGGGTTAGGGTTAGGGTTAGGGGA), named HT (Table 3), in which three strands are oriented in one direction and the fourth in the opposite direction.⁴⁶ Guanines in this structure have both *anti* and *syn* glycosidic conformations. We modified this sequence by inserting thymines between guanines as described above. The sequences HT-T1 and HT-T5 have one thymine inserted at position 1 and 5 respectively, while the sequence HT-T1-5 has thymines inserted at both positions (Table 3). The experimental results of these sequences are shown in the Figure 11. The NMR imino proton spectra of the

three modified sequences show 12 peaks and are nearly identical to the spectrum of the HT sequence. The CD spectra of all sequences are also similar and characteristic of (3 + 1) G-quadruplexes. These results strongly suggest the formation of the same general G-quadruplex fold with bulges by these modified sequences. The melting temperatures of the modified sequences were lower than that of HT, which is 64 °C. Among the modified sequences, HT-T5 has the highest melting temperature of 56 °C in 60 mM K^+ solution, followed by HT-T1 with a melting temperature of 48 °C. HT-T1-5 with two interrupted G-tracts has the lowest melting temperature of 45 °C.

The position of the interrupting thymine with respect to the sequence is similar in HT-T1 and TB-1, as this is the first position of the first G-tract in both cases. However, the TB-1 G-quadruplex with a bulge at position 1 was found to be more stable than the counterparts with a bulge at other positions, while a bulge at position 1 in HT-T1 led to a lower melting temperature as compared to a bulge at position 5 in HT-T5. Hence, the stability of a G-quadruplex with bulges depends not only on their locations but also on the context of the sequence and overall G-quadruplex topology.

Extending the Definition of G-Quadruplex-Forming Sequences. Previously, a number of bioinformatics searches have been conducted to locate potential G-quadruplex-forming sequences in the genomes by using algorithms that look for sequences containing at least four tracts of consecutive guanines^{47–55} with the G-tract length being usually set to ≥ 3 .^{47,48,51,52,54,55} The algorithms used can be classified in two categories:⁵³ (i) those that search for individual specific G-quadruplex-forming sequences^{47–49} and (ii) those that aim to identify regions with a G-quadruplex-forming potential.^{51,54} With the sequence motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ used in

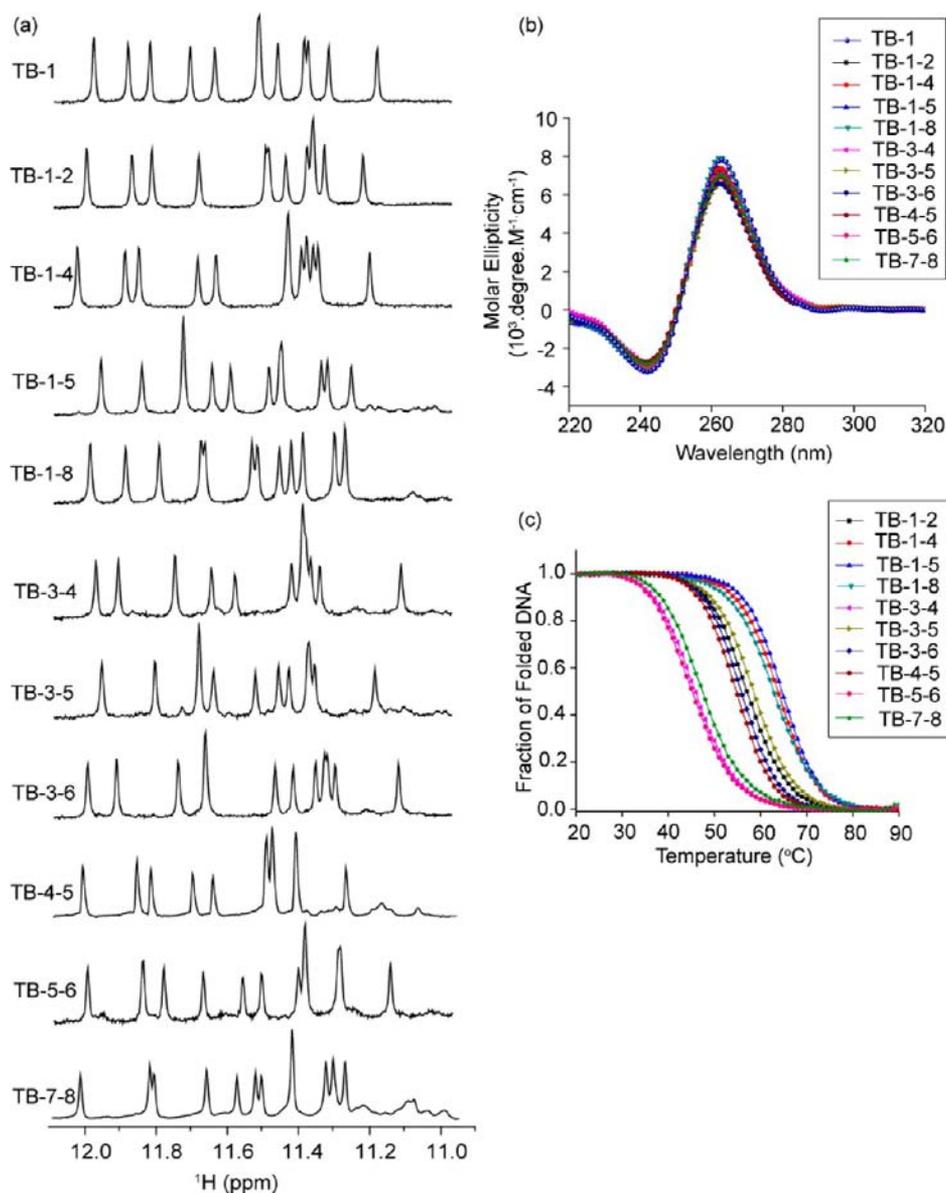


Figure 9. Effect of the number of bulges on the structure and stability of a G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for different G-quadruplex-forming sequences with two bulges. Spectra of TB-1 are shown as the reference. UV melting experiments were performed in 60 mM K^+ solution.

the putative quadruplex sequence approach,^{47,48} where G_{3+} represents a tract of three or more guanines and N_{1-7} represents a loop consisting of 1–7 nucleotides, the number of putative G-quadruplex-forming sequences in the human genome was estimated to be over 370 000. Our result alters this existing view about sequence–structure relationship for G-quadruplexes. The involvement of many isolated guanines in the G-tetrad core formation has been shown systematically here for the first time. This finding could lead to an increase in the number of potential G-quadruplex-forming sequences in many parts of the genome which earlier went unnoticed. Nevertheless, note that the notion of guanine frequency in a particular region (or G-density) has been discussed earlier in the literature^{50,53} and an algorithm relying on G-density,⁵⁶ rather than identifying G-tracts, might cover a part of bulge-containing G-quadruplex-forming sequences. Results of a search for potential G-quadruplex-forming sequences containing up to three bulges in the human genome will be reported elsewhere.

CONCLUSION

This is the first extensive study on the formation of bulges in G-quadruplexes. We have successfully shown that many different bulges, which vary in the sequence, the size, the position, or the number, can exist in G-quadruplex structures. Folding topology and high-resolution structure of one such structure have been determined, which are used as a reference. Melting temperatures of the resulting structures were found to be dependent on the nature of bulges. The formation of bulges has also been demonstrated in another G-quadruplex scaffold with different strand orientations and folding topology. While many of the sequences tested in this study can form stable G-quadruplex structures, all of them defy the description of sequences $G_{3+}N_{L1}G_{3+}N_{L2}G_{3+}N_{L3}G_{3+}$, currently used in most bioinformatics searches for predicting putative G-quadruplex-forming sequences in the genomes. Broadening of this description to include the possibilities of bulge formation should identify more potential G-quadruplex-forming sequences.

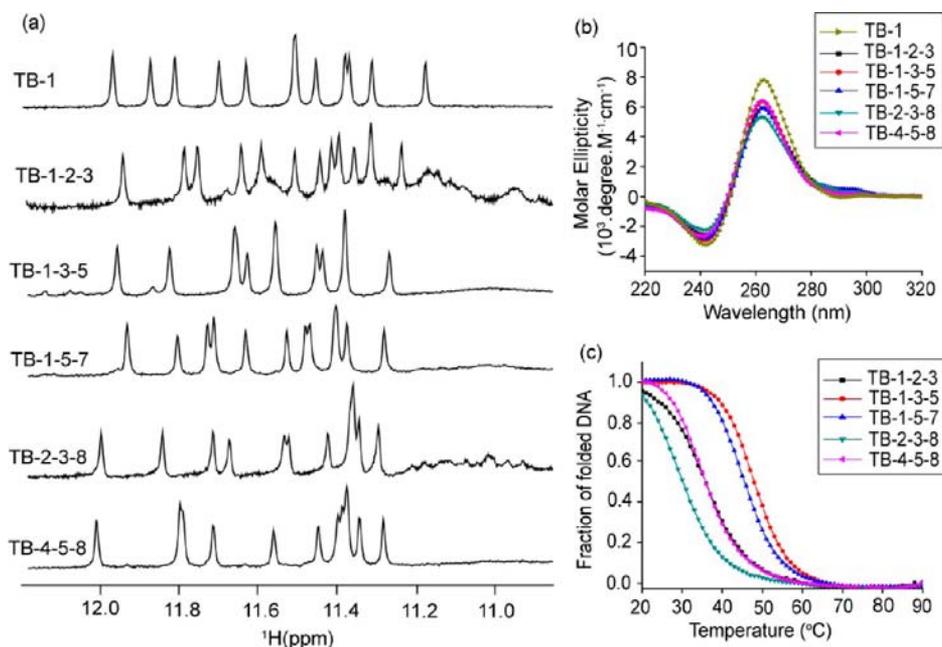


Figure 10. Effect of the number of bulges on the structure and stability of a G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for different G-quadruplex-forming sequences with three bulges. Spectra of TB-1 are shown as the reference. UV melting experiments were performed in 60 mM K^+ solution.

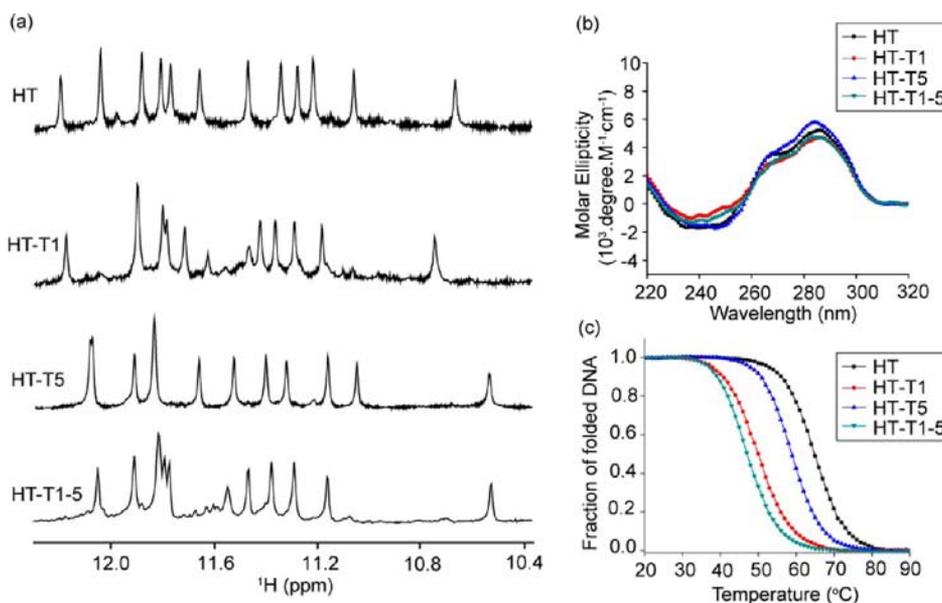


Figure 11. Bulges on a (3 + 1) G-quadruplex. (a) NMR imino proton spectra, (b) CD spectra, and (c) UV-melting curves for the human telomeric sequence HT and its derivatives with bulges.

METHODS

Sample Preparation. Unlabeled and site-specific (^{15}N or 2H) labeled DNA oligonucleotides were chemically prepared on an ABI 394 DNA/RNA synthesizer using products from Glen Research and Cambridge Isotope Laboratories. Samples were purified following Glen Research's protocol and then dialyzed successively against KCl solution and against water. Unless otherwise stated, DNA oligonucleotides were dissolved in solution containing 30 mM KCl, 20 mM potassium phosphate, pH 7.0. DNA concentration was expressed in strand molarity using a nearest-neighbor approximation for the absorption coefficients of the unfolded species.

UV Melting Experiments. The stability of G-quadruplexes was characterized in UV melting experiments conducted on a JASCO V-650 spectrophotometer. Experiments were performed with 1-cm path-

length quartz cuvettes. DNA concentration ranged from 4 to 6 μM . Solution contained 30 mM KCl and 20 mM potassium phosphate (pH 7.0) or 6 mM KCl and 4 mM potassium phosphate (pH 7.0). Samples were initially incubated at 90 °C for 10 min and then cooled down to 20 °C at a rate of 0.2 °C/min; after a delay of 10 min, they were heated back to 90 °C at the same rate of 0.2 °C/min. Absorbance at 295 nm was recorded as a function of temperature ranging from 20 to 90 °C. The linear pre- and post-transition regions of each absorbance-versus-temperature curve were taken as the baselines corresponding to the completely folded (low temperature) and completely unfolded (high temperature) states. The fractions of the folded and unfolded states were derived by taking the ratio of the differences between the baselines and the experimental curve at each temperature.⁵⁷

CD. CD spectra were recorded on a JASCO-815 spectropolarimeter using a 1-cm path-length quartz cuvette with a reaction volume of 600 μL at 20 °C. Scans from 220 to 320 nm were performed with a speed of 200 nm/min, 1 nm pitch, and 1 nm bandwidth. DNA concentration was 4–6 μM . All CD experiments were performed in solution containing 30 mM KCl and 20 mM potassium phosphate (pH 7.0).

NMR Spectroscopy. NMR experiments were performed on 600- and 700-MHz Bruker Avance spectrometers. Unless otherwise specified, all NMR experiments were performed at 25 °C in solution containing 30 mM KCl and 20 mM potassium phosphate (pH 7.0). Spectra in H_2O were recorded using jump- and return-type water suppression pulse sequences. Guanine resonances were assigned by using site-specific (^{15}N or ^2H) labeling and [^{13}C – ^1H] JRMBC through-bond correlation at natural abundance. Spectra assignments were completed by using NOESY, COSY, TOCSY, and [^{13}C – ^1H] HSQC spectra. Interproton distances were measured in NOESY experiments performed in H_2O and D_2O at different mixing times.

Structure Calculation. The NMR structure of the TB-1 sequence was determined using the X-PLOR program,³⁶ as described previously.³¹ In two general steps, 100 structures were calculated: (i) distance geometry simulated annealing and (ii) distance restrained molecular dynamics refinement. Ten lowest-energy structures were selected and analyzed by X-PLOR and viewed by the PyMOL program.⁵⁸

Data Deposition. The coordinates of 10 lowest-energy G-quadruplex structures of the TB-1 sequence d-[TTGTGGTGGGTGGGTGGGT] have been deposited in the Protein Data Bank (accession code 2M4P).

■ ASSOCIATED CONTENT

● Supporting Information

Additional experimental data (Figures S1–S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

phantuan@ntu.edu.sg

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Brahim Heddi and Kah Wai Lim for their assistance with NMR experiments and structure calculation. This research was supported by Singapore Ministry of Education (ARC33/12) and Nanyang Technological University (RG62/07 and RG72/10) grants to A.T.P.

■ REFERENCES

- (1) Davis, J. T. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 668.
- (2) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. *Nucleic Acids Res.* **2006**, *34*, 5402.
- (3) Patel, D. J.; Phan, A. T.; Kuryavyi, V. *Nucleic Acids Res.* **2007**, *35*, 7429.
- (4) Maizels, N. *Nat. Struct. Mol. Biol.* **2006**, *13*, 1055.
- (5) Lipps, H. J.; Rhodes, D. *Trends Cell Biol.* **2009**, *19*, 414.
- (6) Cahoon, L. A.; Seifert, H. S. *Science* **2009**, *325*, 764.
- (7) Law, M. J.; Lower, K. M.; Voon, H. P. J.; Hughes, J. R.; Garrick, D.; Viprakasit, V.; Mitson, M.; De Gobbi, M.; Marra, M.; Morris, A.; Abbott, A.; Wilder, S. P.; Taylor, S.; Santos, G. M.; Cross, J.; Ayyub, H.; Jones, S.; Ragoussis, J.; Rhodes, D.; Dunham, I.; Higgs, D. R.; Gibbons, R. J. *Cell* **2010**, *143*, 367.
- (8) Lopes, J.; Piazza, A.; Bermejo, R.; Kriegsman, B.; Colosio, A.; Teulade-Fichou, M. P.; Foiani, M.; Nicolas, A. *EMBO J.* **2011**, *30*, 4033.
- (9) Paeschke, K.; Capra, J. A.; Zakian, V. A. *Cell* **2011**, *145*, 678.
- (10) Sun, D.; Thompson, B.; Cathers, B. E.; Salazar, M.; Kerwin, S. M.; Trent, J. O.; Jenkins, T. C.; Neidle, S.; Hurley, L. H. *J. Med. Chem.* **1997**, *40*, 2113.

- (11) Patel, D. J.; Phan, A. T.; Kuryavyi, V. *Nucleic Acids Res.* **2007**, *35*, 7429.
- (12) Balasubramanian, S.; Hurley, L. H.; Neidle, S. *Nat. Rev. Drug Discovery* **2011**, *10*, 261.
- (13) Wyatt, J. R.; Vickers, T. A.; Roberson, J. L.; Buckheit, J., R. W.; Klimkait, T.; DeBaets, E.; Davis, P. W.; Rayner, B.; Imbach, J. L.; Ecker, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 1356.
- (14) Bates, P. J.; Kahlon, J. B.; Thomas, S. D.; Trent, J. O.; Miller, D. M. *J. Biol. Chem.* **1999**, *274*, 26369.
- (15) Jing, N. J.; Marchand, C.; Liu, J.; Mitra, R.; Hogan, M. E.; Pommier, Y. *J. Biol. Chem.* **2000**, *275*, 21460.
- (16) Qi, H.; Lin, C.-P.; Fu, X.; Wood, L. M.; Liu, A. A.; Tsai, Y.-C.; Chen, Y.; Barbieri, C. M.; Pilch, D. S.; Liu, L. F. *Cancer Res.* **2006**, *66*, 11808.
- (17) Phan, A. T.; Kuryavyi, V.; Ma, J. B.; Faure, A.; Andreola, M. L.; Patel, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 634.
- (18) Jing, N. J.; Sha, W.; Li, Y. D.; Xiong, W. J.; Twardy, D. J. *Curr. Pharm. Des.* **2005**, *11*, 2841.
- (19) Do, N. Q.; Lim, K. W.; Teo, M. H.; Heddi, B.; Phan, A. T. *Nucleic Acids Res.* **2011**, *39*, 9448.
- (20) D'Atri, V.; Oliviero, G.; Amato, J.; Borbone, N.; D'Errico, S.; Mayol, L.; Piccialli, V.; Haider, S.; Hoorelbeke, B.; Balzarini, J.; Piccialli, G. *Chem. Commun.* **2012**, *48*, 9516.
- (21) Gellert, M.; Lipsett, M. N.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1962**, *48*, 2013.
- (22) Sen, D.; Gilbert, W. *Nature* **1990**, *344*, 410.
- (23) Phan, A. T. *FEBS J.* **2010**, *277*, 1107.
- (24) Phan, A. T.; Kuryavyi, V.; Gaw, H. Y.; Patel, D. J. *Nat. Chem. Biol.* **2005**, *1*, 167.
- (25) (a) Phan, A. T.; Kuryavyi, V.; Burge, S.; Neidle, S.; Patel, D. J. *J. Am. Chem. Soc.* **2007**, *129*, 4386. (b) Wei, D.; Parkinson, G. N.; Reszka, A. P.; Neidle, S. *Nucleic Acids Res.* **2012**, *40*, 4691.
- (26) Pan, B. C.; Xiong, Y.; Shi, K.; Sundaralingam, M. *Structure* **2003**, *11*, 1423.
- (27) Pan, B.; Shi, K.; Sundaralingam, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3130.
- (28) Mukundan, V. T.; Do, N. Q.; Phan, A. T. *Nucleic Acids Res.* **2011**, *39*, 8984.
- (29) Hermann, T.; Patel, D. J. *Structure* **2000**, *8*, R47.
- (30) Wu, H. N.; Uhlenbeck, O. C. *Biochemistry* **1987**, *26*, 8221.
- (31) Do, N. Q.; Phan, A. T. *Chem.—Eur. J.* **2012**, *18*, 14752.
- (32) Phan, A. T.; Patel, D. J. *J. Am. Chem. Soc.* **2002**, *124*, 1160.
- (33) Phan, A. T. *J. Biomol. NMR* **2000**, *16*, 175.
- (34) Huang, X. N.; Yu, P. L.; LeProust, E.; Gao, X. L. *Nucleic Acids Res.* **1997**, *25*, 4758.
- (35) Adrian, M.; Heddi, B.; Phan, A. T. *Methods* **2012**, *57*, 11.
- (36) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 65.
- (37) Rachwal, P. A.; Brown, T.; Fox, K. R. *FEBS Lett.* **2007**, *581*, 1657.
- (38) Phan, A. T.; Modi, Y. S.; Patel, D. J. *J. Am. Chem. Soc.* **2004**, *126*, 8710.
- (39) Hazel, P.; Huppert, J.; Balasubramanian, S.; Neidle, S. *J. Am. Chem. Soc.* **2004**, *126*, 16405.
- (40) Bugaut, A.; Balasubramanian, S. *Biochemistry* **2008**, *47*, 689.
- (41) Guédin, A.; De Cian, A.; Gros, J.; Lacroix, L.; Mergny, J. L. *Biochimie* **2008**, *90*, 686.
- (42) Guédin, A.; Gros, J.; Alberti, P.; Mergny, J. L. *Nucleic Acids Res.* **2010**, *38*, 7858.
- (43) Crnugelj, M.; Sket, P.; Plavec, J. *J. Am. Chem. Soc.* **2003**, *125*, 7866.
- (44) Hu, L.; Lim, K. W.; Bouaziz, S.; Phan, A. T. *J. Am. Chem. Soc.* **2009**, *131*, 16824.
- (45) Lim, K. W.; Alberti, P.; Guédin, A.; Lacroix, L.; Riou, J. F.; Royle, N. J.; Mergny, J. L.; Phan, A. T. *Nucleic Acids Res.* **2009**, *37*, 6239.
- (46) Luu, K. N.; Phan, A. T.; Kuryavyi, V.; Lacroix, L.; Patel, D. J. *J. Am. Chem. Soc.* **2006**, *128*, 9963.
- (47) Huppert, J. L.; Balasubramanian, S. *Nucleic Acids Res.* **2005**, *33*, 2908.
- (48) Todd, A. K.; Johnston, M.; Neidle, S. *Nucleic Acids Res.* **2005**, *33*, 2901.

- (49) Kikin, O.; D'Antonio, L.; Bagga, P. S. *Nucleic Acids Res.* **2006**, *34*, W676.
- (50) Rawal, P.; Kummarasetti, V. B. R.; Ravindran, J.; Kumar, N.; Halder, K.; Sharma, R.; Mukerji, M.; Das, S. K.; Chowdhury, S. *Genome Res.* **2006**, *16*, 644.
- (51) Eddy, J.; Maizels, N. *Nucleic Acids Res.* **2006**, *34*, 3887.
- (52) Scaria, V.; Hariharan, M.; Arora, A.; Maiti, S. *Nucleic Acids Res.* **2006**, *34*, W683.
- (53) Huppert, J. L. *Biochimie* **2008**, *90*, 1140.
- (54) Hershman, S. G.; Chen, Q.; Lee, J. Y.; Kozak, M. L.; Yue, P.; Wang, L. S.; Johnson, F. B. *Nucleic Acids Res.* **2008**, *36*, 144.
- (55) Du, Z.; Zhao, Y. Q.; Li, N. *Genome Res.* **2008**, *18*, 233.
- (56) Mergny, J. L. Third International Meeting on G-quadruplex and G-assembly, Sorrento, Italy, 2011.
- (57) Marky, L. A.; Breslauer, K. J. *Biopolymers* **1987**, *26*, 1601.
- (58) DeLano, W. L. *The PyMOL User's Manual*; DeLano Scientific: Palo Alto, CA, 2002.