Supporting Information

Insight into formation propensity of pseudocircular DNA G-hairpins.

Martina Lenarčič Živković[1,2,†], Martin Gajarský[1,†], Kateřina Beková[1,†], Petr Stadlbauer[3], Lukáš Vicherek[1], Magdalena Petrova[4], Radovan Fiala[1], Ivan Rosenberg[4], Jiří Šponer[1,3], Janez Plavec[2,5,6,*], and Lukáš Trantírek[1,*]

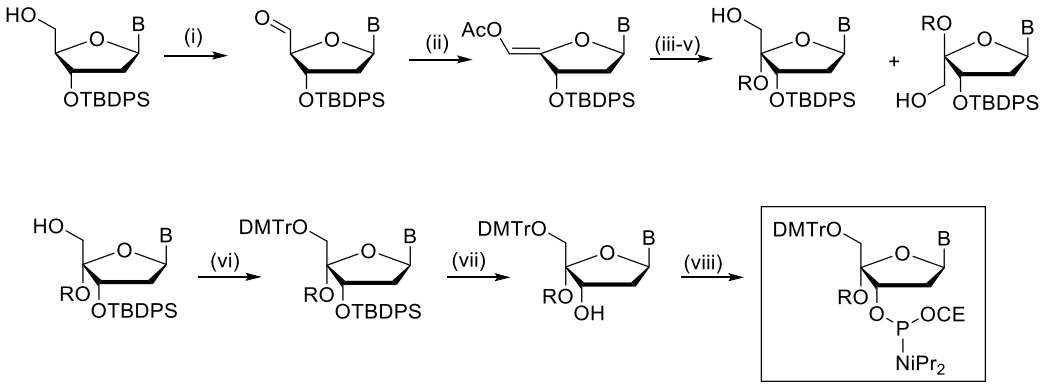
[1] Central European Institute of Technology, Masaryk University, Brno, 62500, Czech Republic
[2] Slovenian NMR Centre, National Institute of Chemistry, Ljubljana, SI-1000, Slovenia
[3] Institute of Biophysics of CAS, v.v.i., Brno, 61265, Czech Republic
[4] Institute of Organic Chemistry and Biochemistry of CAS, Prague, Czech Republic
[5] EN-FIST Centre of Excellence, Ljubljana, SI-1001, Slovenia
[6] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, SI-1000, Slovenia

* To whom correspondence should be addressed. Tel: +420 549 49 6476; Email: lukas.trantirek@ceitec.muni.cz
Correspondence may also be addressed to Janez Plavec. Tel: +386 1 4760 353; Email: janez.plavec@ki.si
† These authors contributed equally.

Table of Contents

1) Methods

Scheme 1. General procedure of synthesis of 4'-alkoxy phosphoramidite T and G monomers.(1, 2)



(i) DMSO, EDC, TFA, py, DMF, rt 4 h; (ii) Ac2O, K2CO3, ACN, 60°C 4 h; (iii) NIS, ROH in DCM (1:4), -20 °C to rt 16 h; (iv) TEAB, DMF, 0°C to rt 16 h; (v) NaBH4, MeOH, 0 °C 16 h; (vi) DMTr-Cl, py, rt 16 h; (vii) TBAF, THF, rt 16 h; (viii) ClP(OCE)NiPr2, DIPEA, THF, rt 1 h. R = CH3.

2) Tables

Table S1. Hydrogen bonds selected for the HBfix in unrestrained MD simulations of SC14.

Construct	Base-pair	HBfix
SC14	G^5:G^1	$N7_5$-$H2_1$ $O6_5$-$H1_1$ $H1_5$-$O6_1$ $H2_5$-$N7_1$
	G^7:G^{-II}	$N7_7$-$H2_{-II}$ $O6_7$-$H1_{-II}$ $H1_7$-$O6_{-II}$ $H2_7$-$N7_{-II}$
	G^6:G^9	$N7_6$-$H2_9$ $O6_6$-$H1_9$ $H1_6$-$O6_9$ $H2_6$-$N7_9$

Table S2. The list of organisms used in genome-wide search for sequences with potential to form PGHs (PGH sites) for *form I* **(A)** and *form II* **(B)**.

A

	organism	No. PGH sites[a]
viruses	*Canarypox virus*	0
	Ebola virus	0
	HIV virus	0
	Poliovirus	0
	Rubella virus	0
	Variola virus	0
bacteria	*Staphylococcus aureus*	2
	Salmonella enterica	4
	Klebsiella pneumoniae	6
	Escherichia coli	9
	Pseudomonas aeruginosa	10
	Mycobacterium tuberculosis	28
eukaryotes	*Saccharomyces cerevisiae*	114
	Arabidopsis thaliana	140
	Xenopus leavis	7261
	Danio rerio	7496
	Homo sapiens	13383
	Mus musculus	17544

[a]Query for bioinformatic search covered eleven sequences, namely parent *form I* SC11 sequence (5'-GTGTGGGTGTG-3') and its PGH-forming analogs (*formI*_T2C, *formI*_G3A, *formI*_G3C, *formI*_G3T, *formI*_T4A, *formI*_T4C, *formI*_T8A, *formI*_T8C, *formI*_T10A, and *formI*_T10C).

B

	organism	No. PGH sites[b]
viruses	*Canarypox virus*	0
	Ebola virus	0
	HIV virus	0
	Poliovirus	0
	Rubella virus	0
	Variola virus	0
bacteria	*Staphylococcus aureus*	2
	Salmonella enterica	0
	Klebsiella pneumoniae	3
	Escherichia coli	3
	Pseudomonas aeruginosa	6
	Mycobacterium tuberculosis	10
eukaryotes	*Saccharomyces cerevisiae*	77
	Arabidopsis thaliana	99
	Xenopus leavis	7653
	Danio rerio	18311
	Homo sapiens	10053
	Mus musculus	17208

[b]Query for bioinformatics search covered seven sequences displaying PGH formation potential: 5'-GTGTGTGGGTG-3' (corresponding to the *form II* minimal sequence) and its PGH-forming analogs (*form II*_T-IC, *form II*_G3C, *form II*_G3A, *form II*_G3T, *form II*_T4A, and *form II*_T4C).

Table S3. Lifetimes (ns) of PGHs formed by SC14 during the unrestrained MD simulations[a].

Construct	*Water model*								
	TIP3P			*SPC/E*			*OPC*		
Simulation #	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
SC14	780	5000	5000	5000	520	5000	5000	5000	5000

[a]The length of the individual simulation was 5000 ns, nine independent MD simulations were performed for the SC14 construct. SC14 remained stable over the whole simulation period in seven of the simulations.

Table S4. List of 11-nt long oligonucleotide constructs designed based on minimal sequence forming *form I* PGH (*i.e.* SC11) bearing individual point mutations used in this study.

name of the construct	sequence (5'→3')
form I	GTGTGGGTGTG
***form I*_G1A**	**A**TGTGGGTGTG
***form I*_G1C**	**C**TGTGGGTGTG
***form I*_G1T**	**T**TGTGGGTGTG
***form I*_T2A**	G**A**GTGGGTGTG
***form I*_T2C**	G**C**GTGGGTGTG
***form I*_T2G**	G**G**GTGGGTGTG
***form I*_G3A**	GT**A**TGGGTGTG
***form I*_G3C**	GT**C**TGGGTGTG
***form I*_G3T**	GT**T**TGGGTGTG
***form I*_T4A**	GTG**A**GGGTGTG
***form I*_T4C**	GTG**C**GGGTGTG
***form I*_T4C**	GTG**G**GGGTGTG
***form I*_G5A**	GTGT**A**GGTGTG
***form I*_G5C**	GTGT**C**GGTGTG
***form I*_G5T**	GTGT**T**GGTGTG
***form I*_G6A**	GTGTG**A**GTGTG
***form I*_G6C**	GTGTG**C**GTGTG
***form I*_G6T**	GTGTG**T**GTGTG
***form I*_G7A**	GTGTGG**A**TGTG
***form I*_G7C**	GTGTGG**C**TGTG
***form I*_G7T**	GTGTGG**T**TGTG
***form I*_T8A**	GTGTGGG**A**GTG
***form I*_T8C**	GTGTGGG**C**GTG
***form I*_T8G**	GTGTGGG**G**GTG
***form I*_G9A**	GTGTGGGT**A**TG
***form I*_G9C**	GTGTGGGT**C**TG
***form I*_G9T**	GTGTGGGT**T**TG
***form I*_T10A**	GTGTGGGTG**A**G
***form I*_T10C**	GTGTGGGTG**C**G
***form I*_T10G**	GTGTGGGTG**G**G
***form I*_G11A**	GTGTGGGTGT**A**
***form I*_G11C**	GTGTGGGTGT**C**
***form I*_G11T**	GTGTGGGTGT**T**

Table S5. List of 11-nt long oligonucleotide constructs designed based on the minimal sequence forming *form II* PGH (*i.e.* SC14 lacking last three 3'-residues) bearing individual mutations used in this study.

name of the construct	sequence (5'→3')
form II	GTGTGTGGGTG
***form II*_G-IIA**	**A**TGTGTGGGTG
***form II*_G-IIC**	**C**TGTGTGGGTG
***form II*_G-IIT**	**T**TGTGTGGGTG
***form II*_T-IA**	G**A**GTGTGGGTG
***form II*_T-IC**	G**C**GTGTGGGTG
***form II*_T-IG**	G**G**GTGTGGGTG
***form II*_G1A**	GT**A**TGTGGGTG
***form II*_G1C**	GT**C**TGTGGGTG
***form II*_G1T**	GT**T**TGTGGGTG
***form II*_T2A**	GTG**A**GTGGGTG
***form II*_T2C**	GTG**C**GTGGGTG
***form II*_T2G**	GTG**G**GTGGGTG
***form II*_G3A**	GTGT**A**TGGGTG
***form II*_G3C**	GTGT**C**TGGGTG
***form II*_G3T**	GTGT**T**TGGGTG
***form II*_T4A**	GTGTG**A**GGGTG
***form II*_T4C**	GTGTG**C**GGGTG
***form II*_T4C**	GTGTG**G**GGGTG
***form II*_G5A**	GTGTGT**A**GGTG
***form II*_G5C**	GTGTGT**C**GGTG
form II_G5T	GTGTGT**T**GGTG
form II_G6A	GTGTGTG**A**GTG
form II_G6C	GTGTGTG**C**GTG
form II_G6T	GTGTGTG**T**GTG
form II_G7A	GTGTGTGG**A**TG
form II_G7C	GTGTGTGG**C**TG
form II_G7T	GTGTGTGG**T**TG
form II_T8A	GTGTGTGGG**A**G
form II_T8C	GTGTGTGGG**C**G
form II_T8G	GTGTGTGGG**G**G
form II_G9A	GTGTGTGGGT**A**
form II_G9C	GTGTGTGGGT**C**
form II_G9T	GTGTGTGGGT**T**

Table S6. Occurences of putative PGH sites in introns of human genes for *form I* **(A)** and *form II* **(B)**.

A

ID	Sequence	No. of occurrences in unique genes
form I	GTGTGGGTGTG	1638
*form I*_T2C	GCGTGGGTGTG	248
*form I*_G3A	GTATGGGTGTG	395
*form I*_G3C	GTCTGGGTGTG	649
*form I*_G3T	GTTTGGGTGTG	508
*form I*_T4A	GTGAGGGTGTG	696
*form I*_T4C	GTGCGGGTGTG	161
*form I*_T8A	GTGTGGGAGTG	694
*form I*_T8C	GTGTGGGCGTG	233
*form I*_T10A	GTGTGGGTGAG	684
*form I*_T10C	GTGTGGGTGCG	161

B

ID	Sequence	No. of occurrences in unique genes
form II	GTGTGTGGGTG	1554
*form II*_T-IC	GCGTGTGGGTG	245
*form II*_G3A	GTGTATGGGTG	388
*form II*_G3C	GTGTCTGGGTG	638
*form II*_G3T	GTGTTTGGGTG	533
*form II*_T4A	GTGTGAGGGTG	587
*form II*_T4C	GTGTGCGGGTG	170

1) Figures

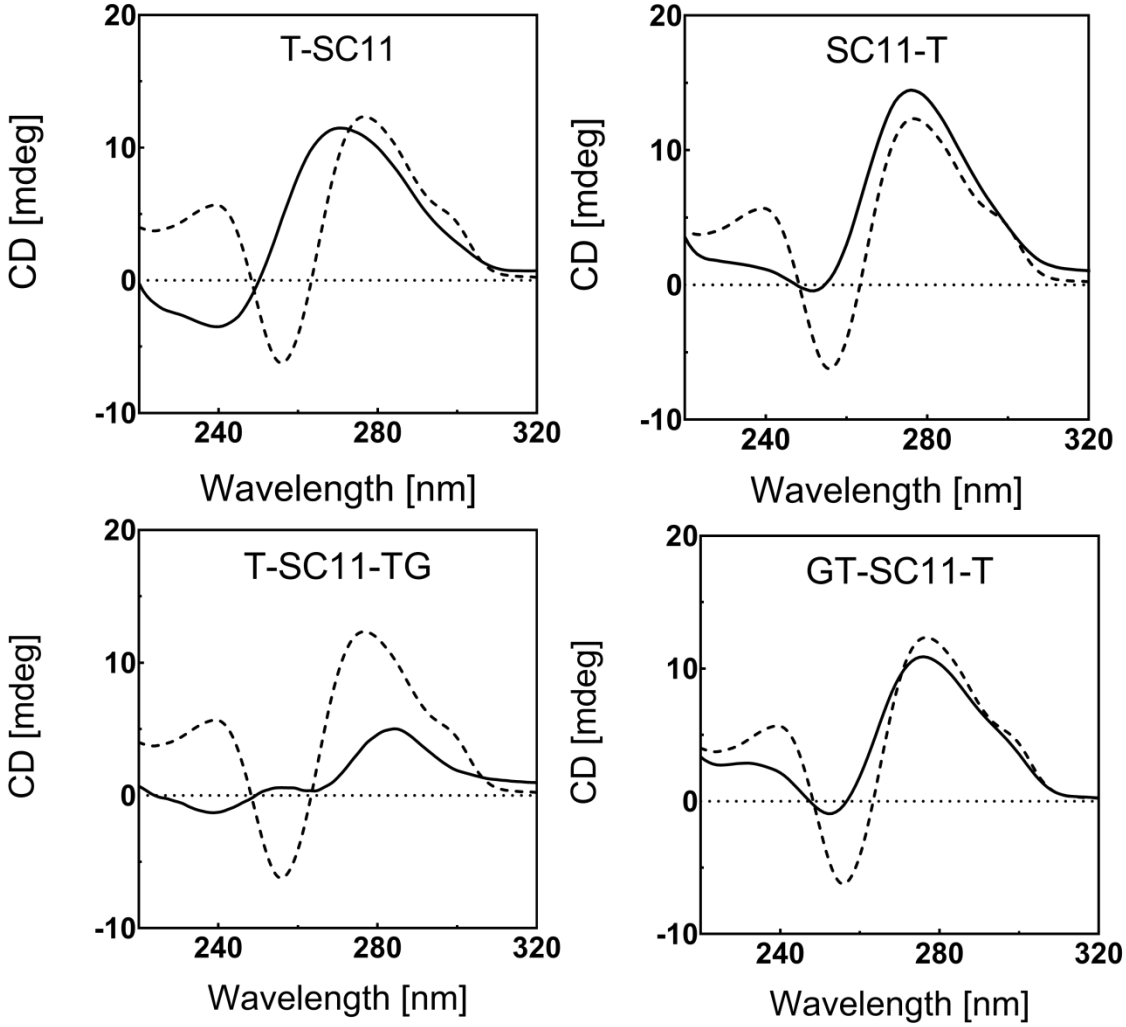


Figure S1. CD spectra of the parent SC11 (dashed line) and extended sequences based on SC11 (solid line) (cf. Figure 2).

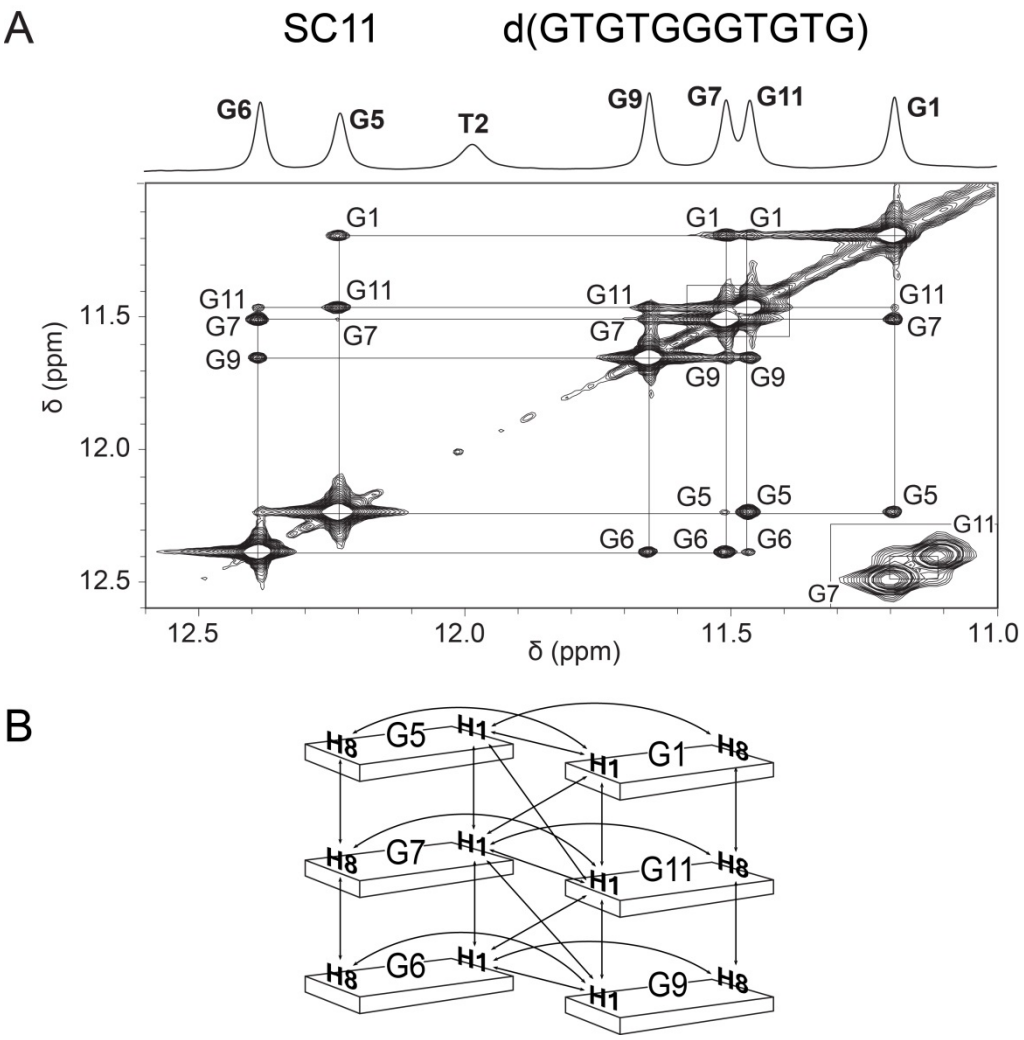


Figure S2. (A) Top: The sequence and assigned imino region of the 1D ^{1}H NMR spectrum of SC11. Bottom: The imino-imino region of 2D NOESY spectrum (τ_m = 150 ms) in $^{1}H_2O/^{2}H_2O$ (90:10). **(B)** Schematic representation of NOE connectivities (indicated by arrows) between imino and aromatic protons for core guanine residues observed in the NOESY spectrum of SC11. Figure S2 was adapted from (3).

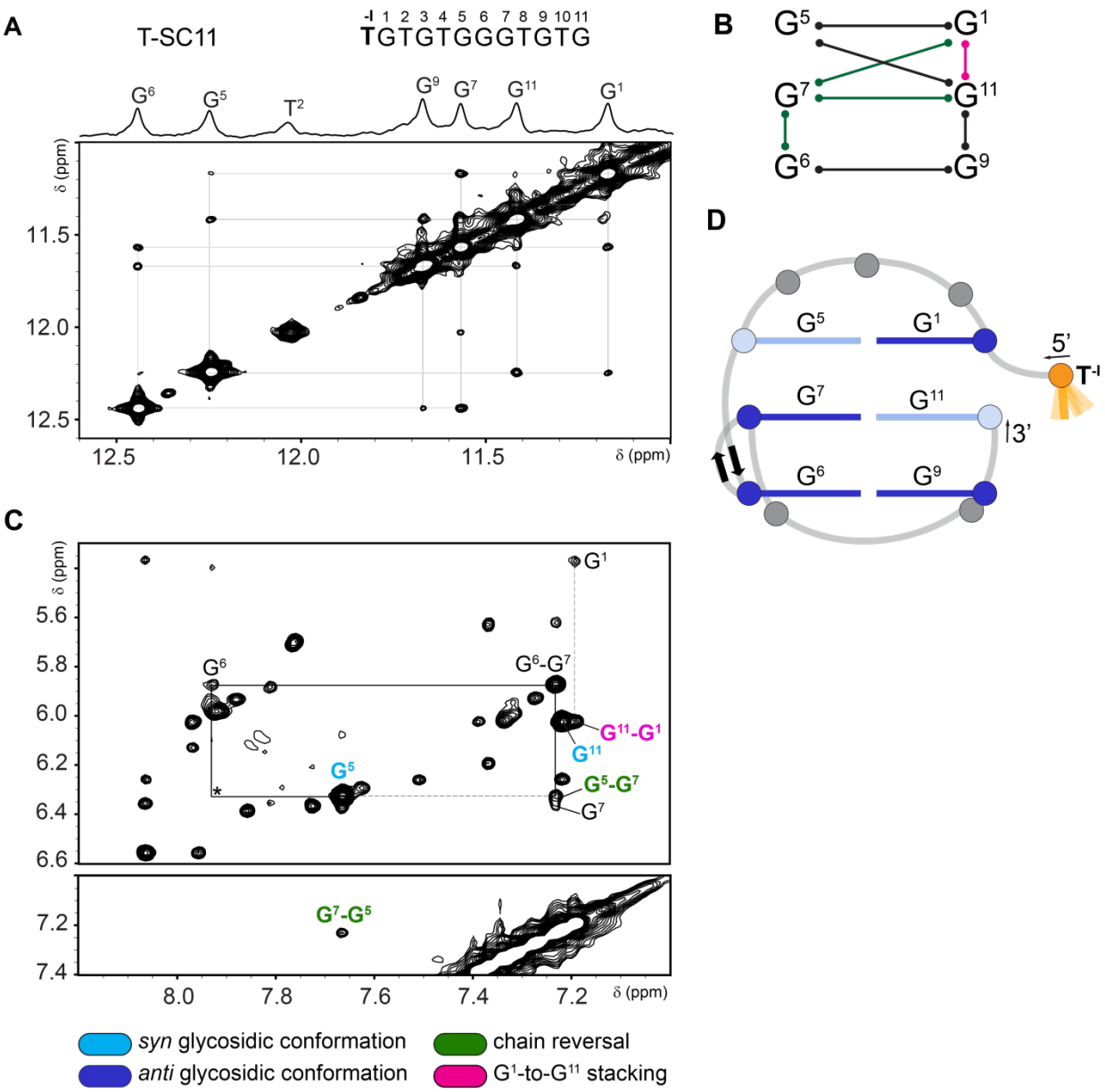


Figure S3. (A) T-SC11 sequence with the corresponding imino region of 1D ^{1}H NMR spectrum (top) and imino-imino region of 2D NOESY spectrum (τ_m = 150 ms) (bottom). **(B)** Schematic presentation of imino-imino NOE connectivities of T-SC11 observed in the NOESY spectrum. **(C)** Anomeric-aromatic (top) and aromatic-aromatic (bottom) regions of the 2D NOESY spectrum (τ_m = 150 ms) of T-SC11. NOE connectivities characteristic of chain reversal involving residues G^5, G^6, and G^7 (dark green) and G^1-to-G^{11} stacking (magenta) are highlighted. Residues G^5 and G^{11} that occupy *syn* glycosidic conformations are colored light blue. **(D)** Schematic of PGH topology adopted by T-SC11 as deduced from the NOE data.

***Commentary/Notes*:**

All the main structural elements characteristic for a PGH fold are preserved in T-SC11: three G:G base pair core, chain reversal and stacking of G^1 and G^{11}. Complex network of imino-imino NOE connectivities indicates that the core of the structure involves guanines G^1, G^5, G^6, G^7, G^9, and G^{11} that form G^5:G^1, G^7:G^{11} and G^6:G^9 base pairs (Figure S3A and S3B). Although, no G^5-G^7 and G^6-G^{11} imino-imino NOE contacts were observed, the position of G^7 between G^5 and G^6 is evidenced mainly by the presence of G^1-G^7 H1-H1, G^5-G^7 H1'-H8 and G^7-G^5 H8-H8 NOE connectivities (Figure S3A and S3C). G^1-to-G^{11} stacking is supported by the observation of G^{11}-G^1 (weak) H1-H1 and H1'-H8 NOE contacts (Figure S3C). H8-H8 NOE cross-peak between G^1 and G^{11} could not be observed due to similar chemical shifts of their aromatic protons. Intense intraresidual NOE cross-peaks in the anomeric-aromatic region of the NOESY spectrum (Figure S3C) indicate *syn* glycosidic conformations for G^5 and G^{11}, which corresponds to *syn* guanines at the equivalent positions as in parent SC11 structure. The lack of NOE connectivities between 5'-end residue T^{-l} and other residues indicates that 5'-T^{-l} is, in contrast to well-defined PGH unit, disordered (Figure S3D).

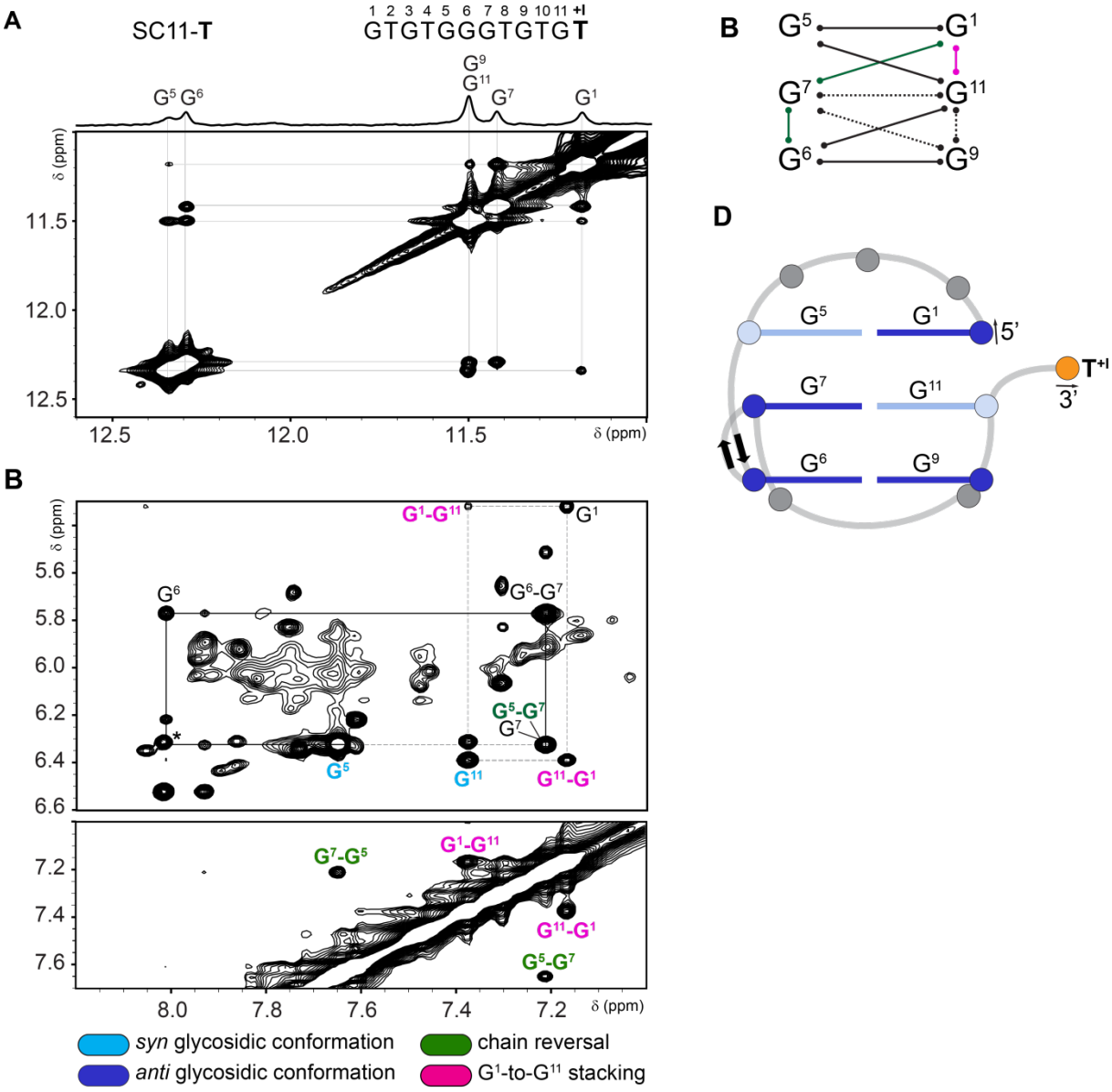


Figure S4. (A) SC11-T sequence with the corresponding imino region of 1D ^{1}H NMR spectrum (top) and imino-imino region of 2D NOESY spectrum (τ_m = 150 ms) (bottom). **(B)** Schematic presentation of imino-imino NOE connectivities of SC11-T observed in the NOESY spectrum. **(C)** Anomeric-aromatic (top) and aromatic-aromatic (bottom) regions of the 2D NOESY spectrum (τ_m = 150 ms) of SC11-T. NOE connectivities characteristic of chain reversal involving residues G^5, G^6, and G^7 (dark green) and G^1-to-G^{11} stacking (magenta) are highlighted. Residues G^5 and G^{11} that occupy *syn* glycosidic conformations are colored light blue. **(D)** Schematic of PGH topology adopted by SC11-T as deduced from NOE data.

***Commentary/Notes*:**

All the main structural elements characteristic for a PGH fold are preserved in SC11-T: three G:G base pair core, chain reversal and stacking of G^1 and G^{11}. Complex network of imino-imino NOE connectivities indicates that the core of the structure involves guanines G^1, G^5, G^6, G^7, G^9, and G^{11} that form G^5:G^1, G^7:G^{11} and G^6:G^9 base pairs (Figure S4A and S4B). Although, no G^5-G^7 imino-imino NOE contacts was observed, the position of G^7 between G^5 and G^6 is evidenced mainly by the presence of G^1-G^7 H1-H1, G^5-G^7 H1'-H8 and H8-H8 NOE connectivities (Figure S4A and S4C). G^1-to-G^{11} stacking is supported by the observation of G^{11}-G^1 H1-H1, H8-H8 and H1'-H8 NOE contacts (Figure S4C). Intense intraresidual NOE cross-peaks in the anomeric-aromatic region of the NOESY spectrum (Figure S4C) indicate *syn* glycosidic conformations for G^5 and G^{11}, which corresponds to *syn* guanines at the equivalent positions as in parent SC11 structure.

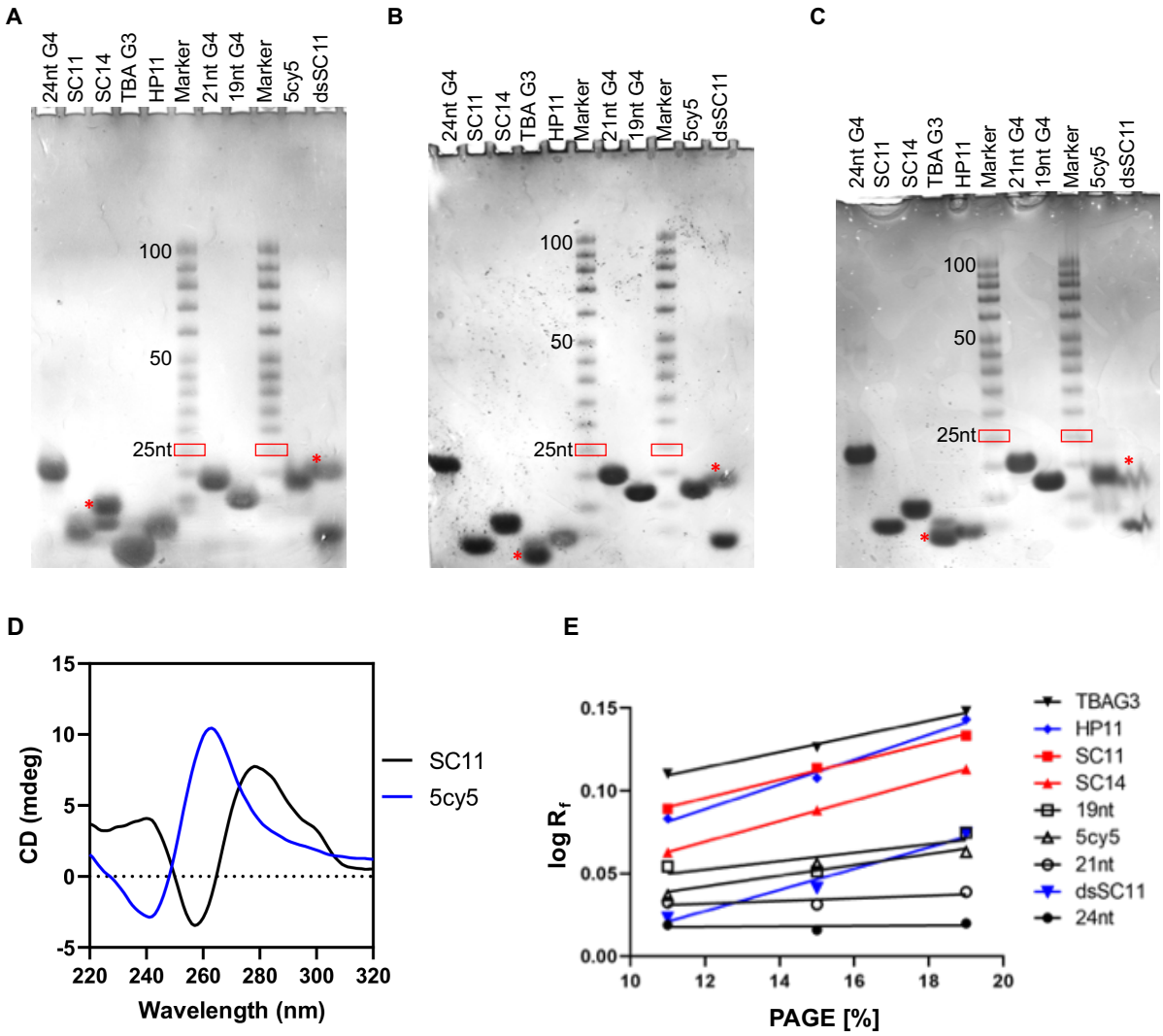


Figure S5. Native polyacrylamide gel electrophoresis (PAGE) run in 10mM KPOi and 100 mM KCl, pH=7 at 6°C. The acrylamide concentration in the gel was set to 11 **(A)**, 15 **(B)**, and 19% **(C)**, respectively. The 24nt, 21nt, and 19nt G4 correspond to the monomolecular G-quadruplexes formed by d[TT(GGGTTA)$_3$GGGA] (4), d(GTAGGTGGTTGGTGTGGTTGG) (5), and d(GGTTTGGTTGTTGGTTTGG) (5), respectively. The HP11 and dsSC11 corresponds to the extended ultra-stable mini-hairpin d(CGGCGAAGCCG) (6) and heteroduplex d(GTGTGGGTGTG).d(CACACCCACAC), respectively. The TBAG3 corresponds to the 11-nt d(GGTTGGTGTGG) forming a monomeric G-triplex structure (7,8). Note: TBAG3 has an identical nucleotide composition as SC11. The 5cy5 corresponds to the covalently tagged SC11 with a conventional fluorescence probe (cy5) (Sigma-Aldrich, USA). Note: The attachment of cy5 to the 5'-terminus of SC11 sequence was noted to remodel its conformational behavior as evidenced from the comparison of SC11 and 5cy5 CD spectra. While the CD spectrum of SC11 shows pattern typical for PGH, the CD spectrum of 5cy5 (see panel D) displays characteristic shape for a parallel G-quadruplex. The "*" marks the bands employed for the construction of the Ferguson plot (cf. panel E). The individual bands of the "Marker" corresponds (from the top) to 100, 90, 80, 70, 60, 50, 45, 40, 35, 30, 25, 20, 15, and 10nt. E) The (Ferguson) plot: The logarithm of relative mobility (with respect to the 25 nt band from the marker - red box in panels A, B, and C is plotted against the concentration of the polyacrylamide gel.

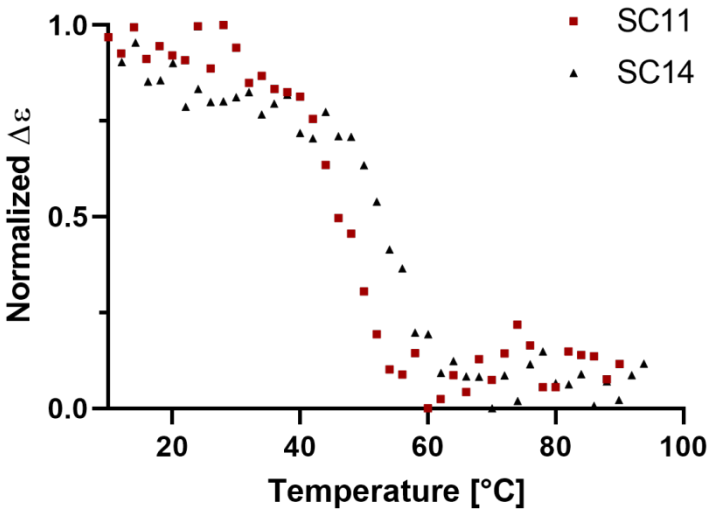
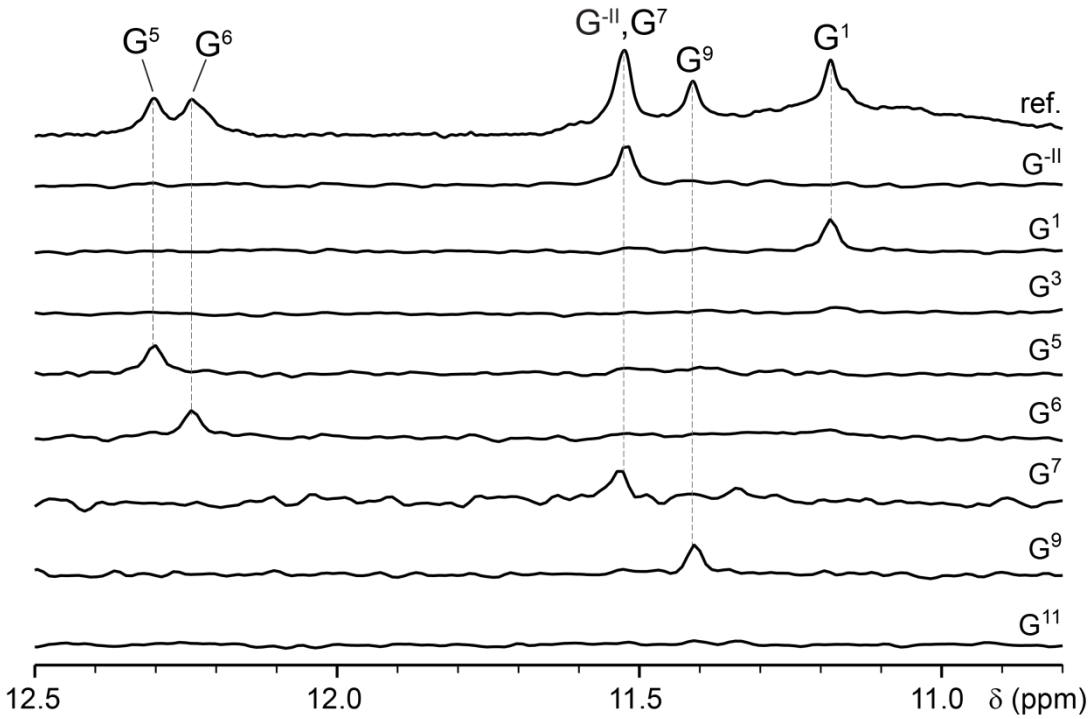


Figure S6. CD melting curves of SC11 and SC14 recorded at 272 nm in 10mM KPOi + 100 mM KCl, pH= 7.

A


G5
G6
G-II,G7
G9
G1
ref.
G-II
G1
G3
G5
G6
G7
G9
G11
12.5
12.0
11.5
11.0
δ (ppm)

B

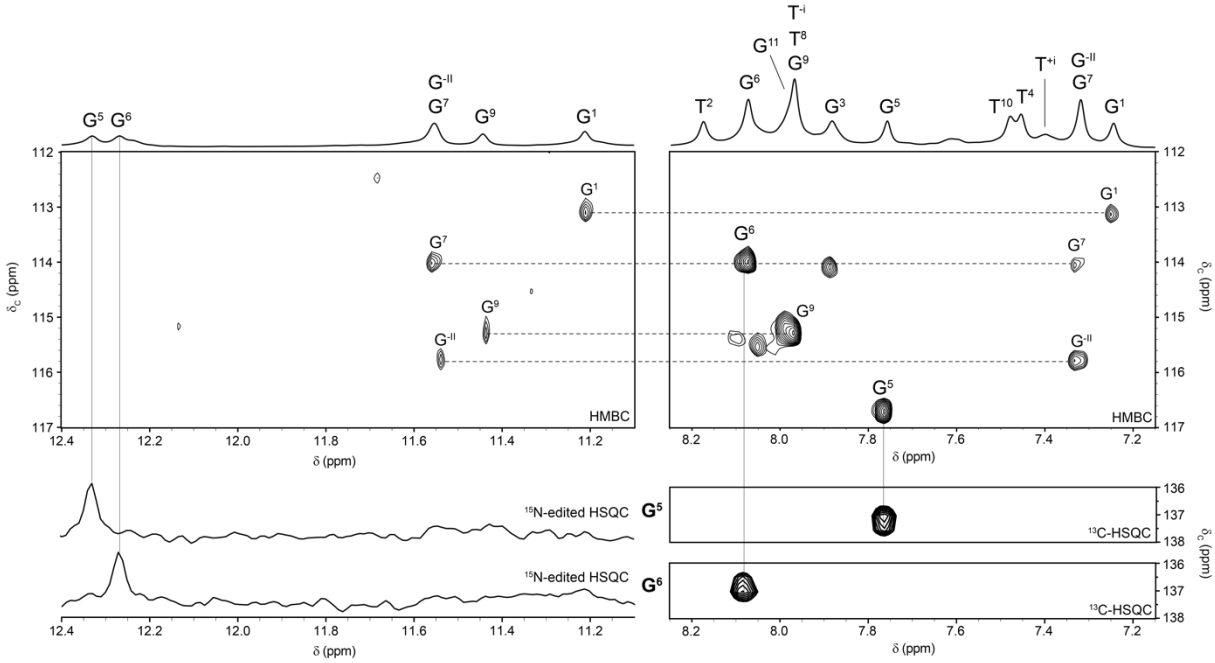

G5 G6
G-II G7
G9
G1
T2
G6
T-i T8 G9
G11
G3
G5
T10
T4
T+i
G-II G7
G1
HMBC
δ (ppm)
δC (ppm)
15N-edited HSQC
G5
G6
13C-HSQC

Figure S7. (A) Unambiguous assignment of imino proton resonances of SC14 was achieved by recording ^{15}N-edited HSQC spectra on partially (~6%) residue-specific ^{15}N/^{13}C-labeled oligonucleotides. **(B)** Unambiguous assignment of aromatic (H8) proton resonances of G:G core-forming residues (G^{-II}, G^{1}, G^{7}, and G^{9}) of SC14 was achieved by recording of 2D ^{1}H-^{13}C JR-HMBC spectrum at natural abundance of ^{13}C: Top-left and top-right panels display aromatic (^{13}C)-imino(^{1}H) and aromatic (^{13}C)-aromatic(^{1}H) regions of JR-HMBC spectrum, respectively. Unambiguous assignment of aromatic (H8) proton resonances of G:G core forming residues (G^{5} and G^{6}) of SC14 was achieved by recording ^{15}N-edited (bottom-left) and ^{13}C- (bottom-right) HSQC spectra on partially (~50%) residue specific ^{15}N/^{13}C-labeled G^{5} and G^{6} oligonucleotides. Imino and aromatic regions of 1D ^{1}H NMR spectrum of SC14 (GT-SC11-T) and assignment of corresponding resonances are shown on top. Spectra were recorded on Agilent DD2 600 MHz spectrometer at 10°C in 90% H_2O, 10% $^{2}H_2O$, 100 mM KCl, 10 mM phosphate buffer with pH 7.0. Oligonucleotide concentrations were 0.5 mM.

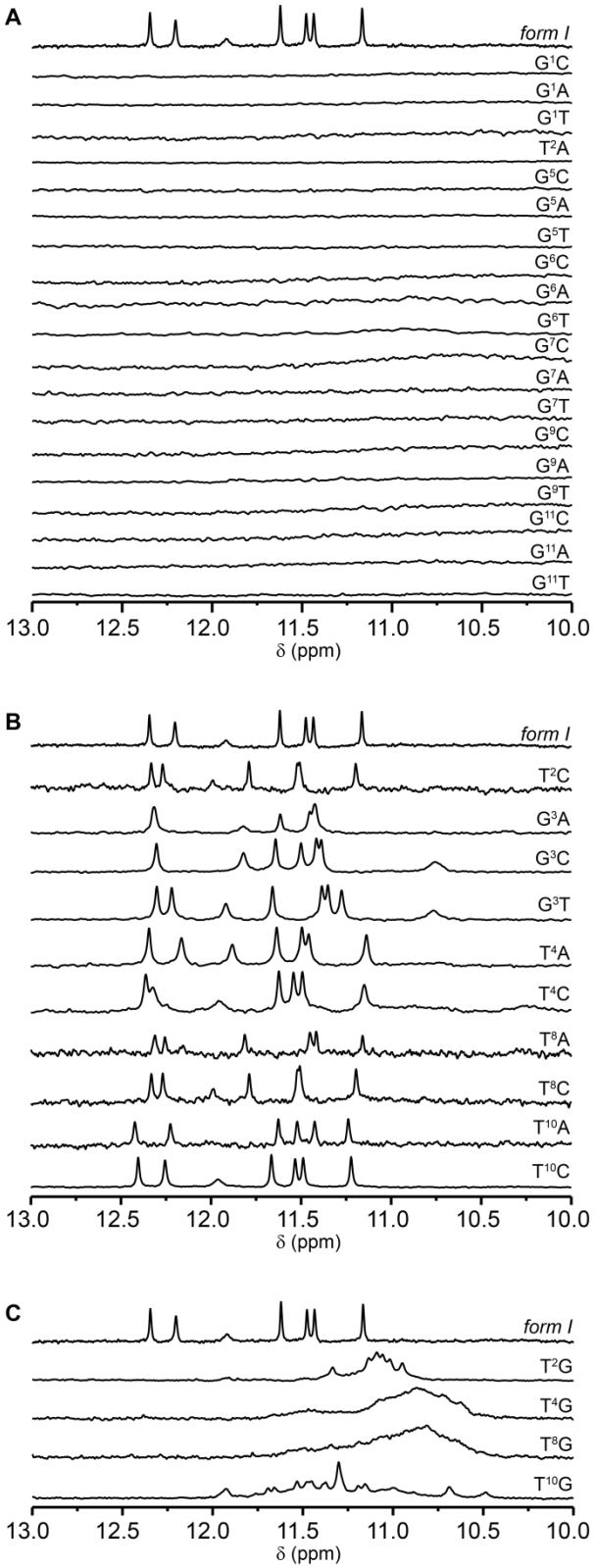


Figure S8. Imino regions of 1D ^{1}H NMR spectra of parent *form I* sequence (top) and its single-point mutations that prevent formation of a stable structure **(A)**, form PGH-like structures **(B)**, and lead to formation of G-quadruplex structures **(C)**.

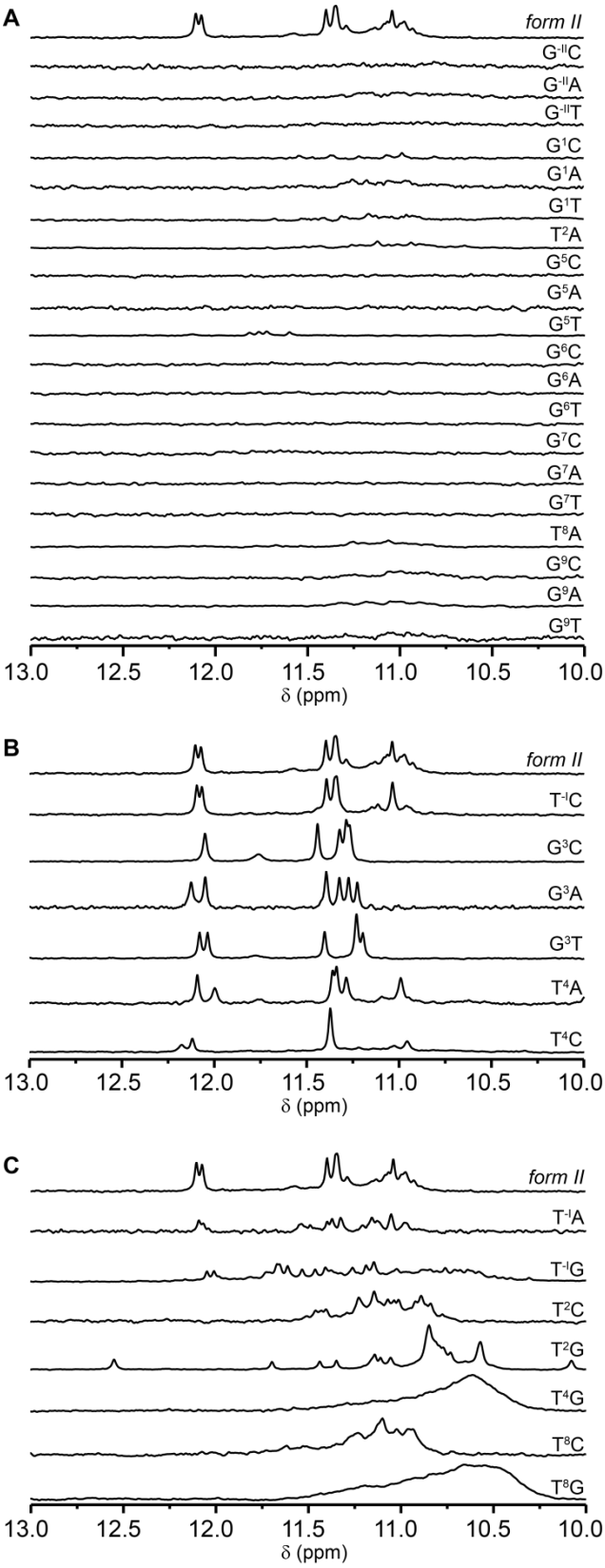


Figure S9. Imino regions of 1D ^{1}H NMR spectra of minimal *form II* sequence (top) and its single-point mutations that prevent formation of a stable structure **(A)**, form PGH-like structures **(B)**, and lead to formation of G-quadruplex structures **(C)**.

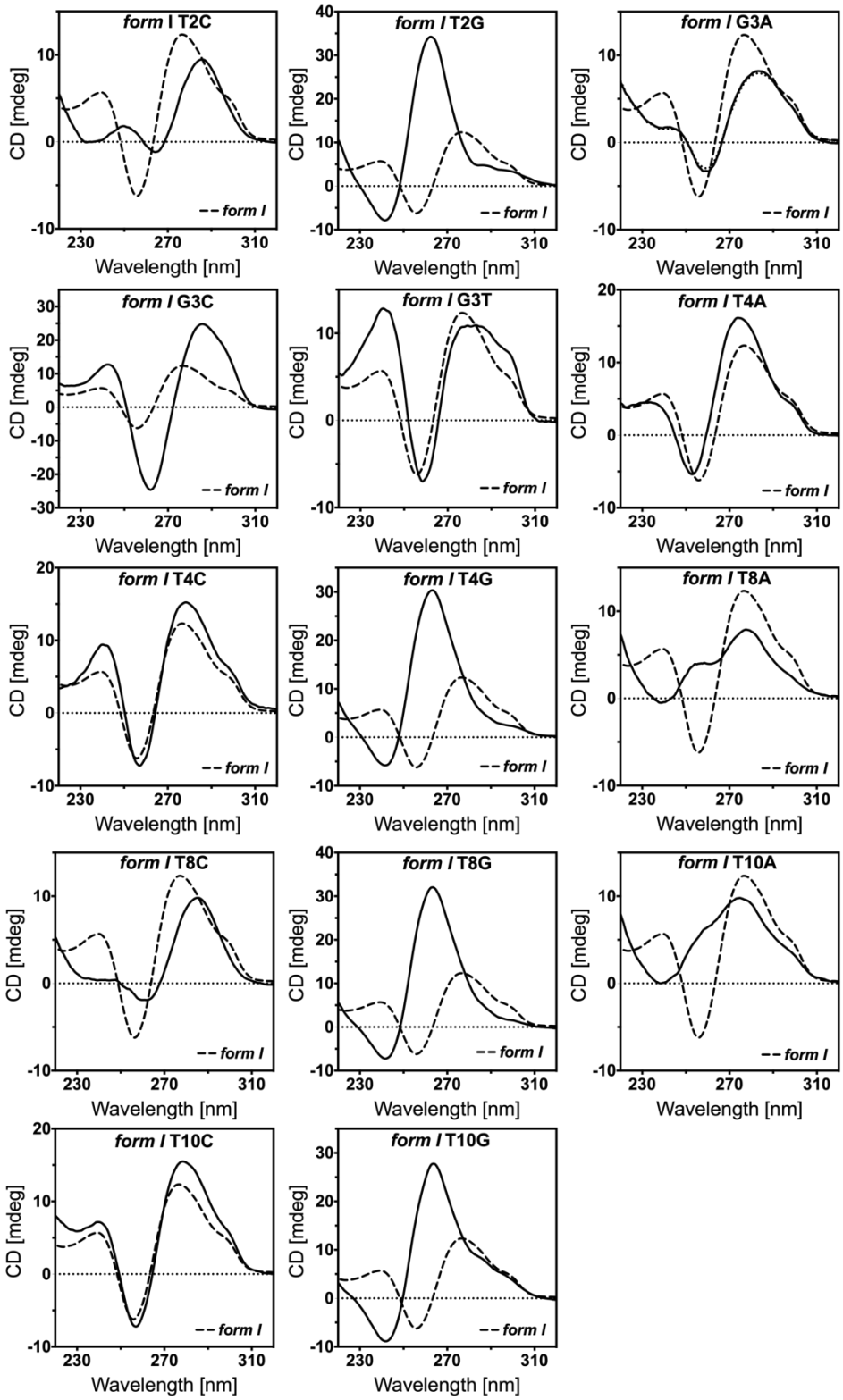


Figure S10. CD spectra of the parent *form I* sequence (dashed line) and selected single-point mutant variants (solid line). Please note that CD spectra of T2G, T4G, T8G, and T10G constructs display a typical shape of parallel G-quadruplex structure.

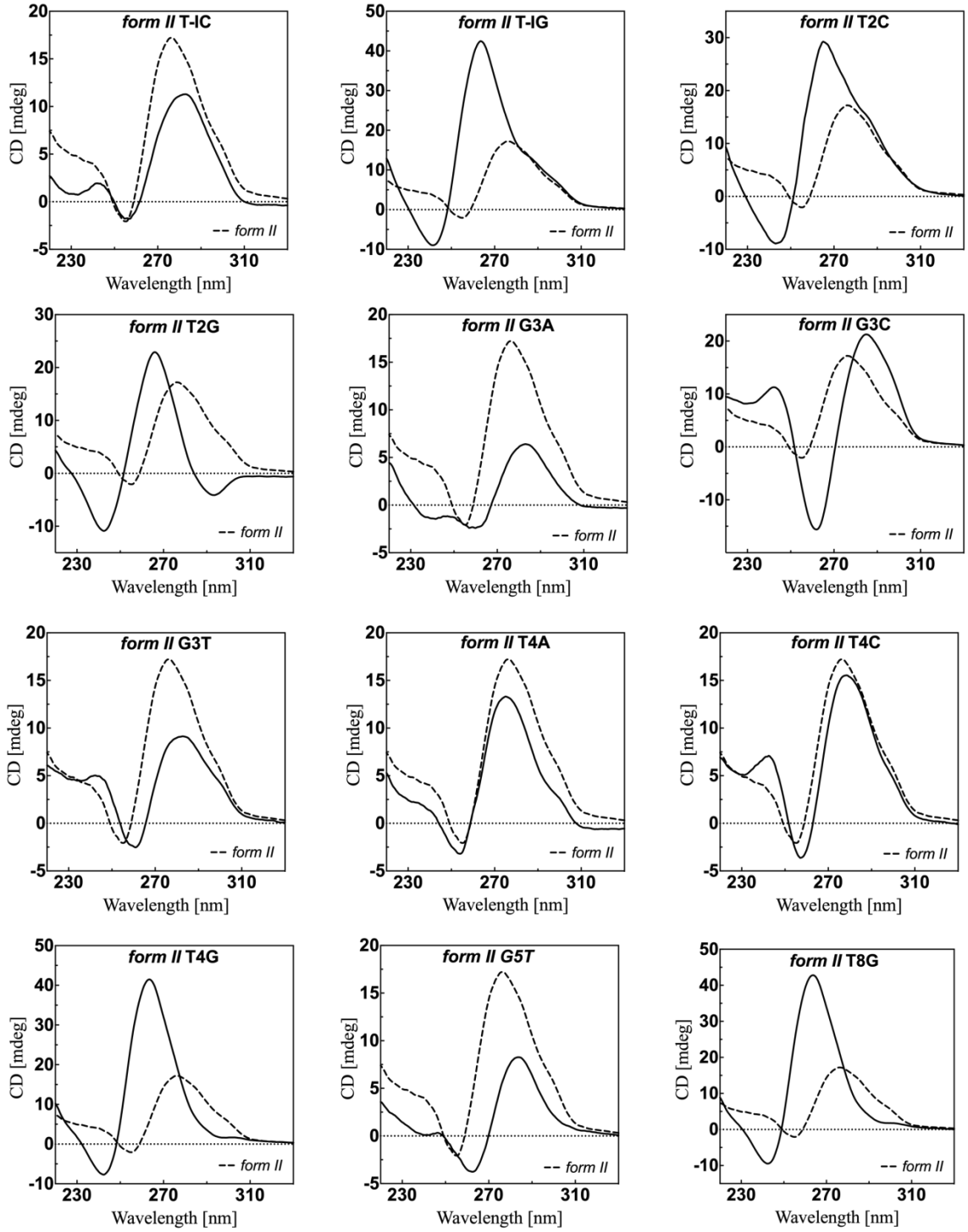


Figure S11. CD spectra of the minimal *form II* sequence (dashed line) and selected single-point mutant variants (solid line). Please note that CD spectra of T2G, T4G, T8G, and T10G constructs display a typical shape of parallel G-quadruplex structure.

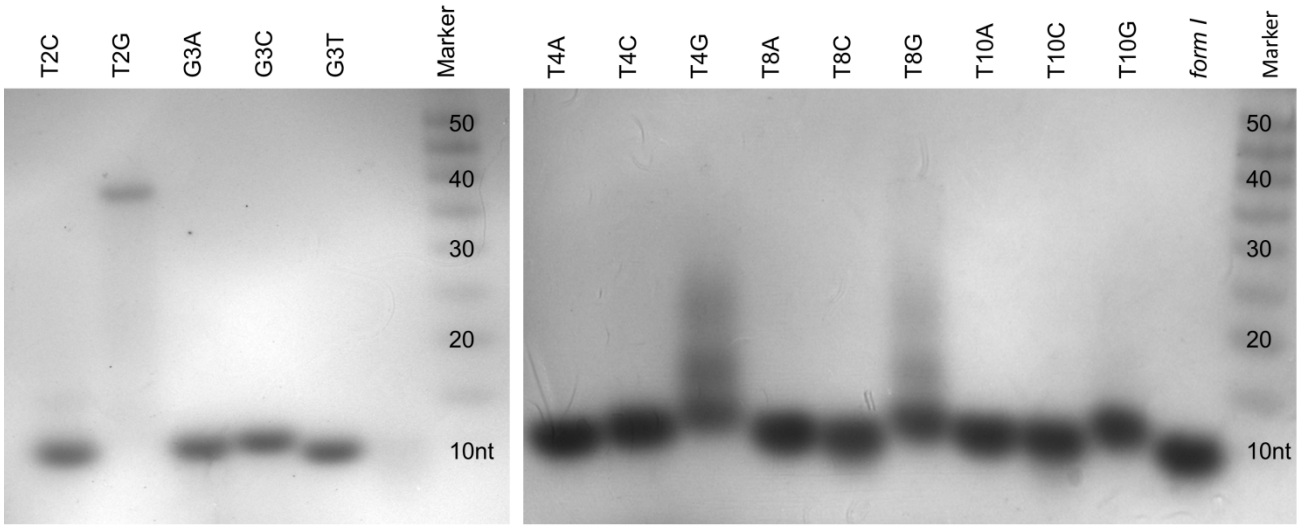


Figure S12. Native PAGE of the parent *form I* sequence and its single-point mutant variants that form stable secondary structures (cf. Figure S6 and S8).

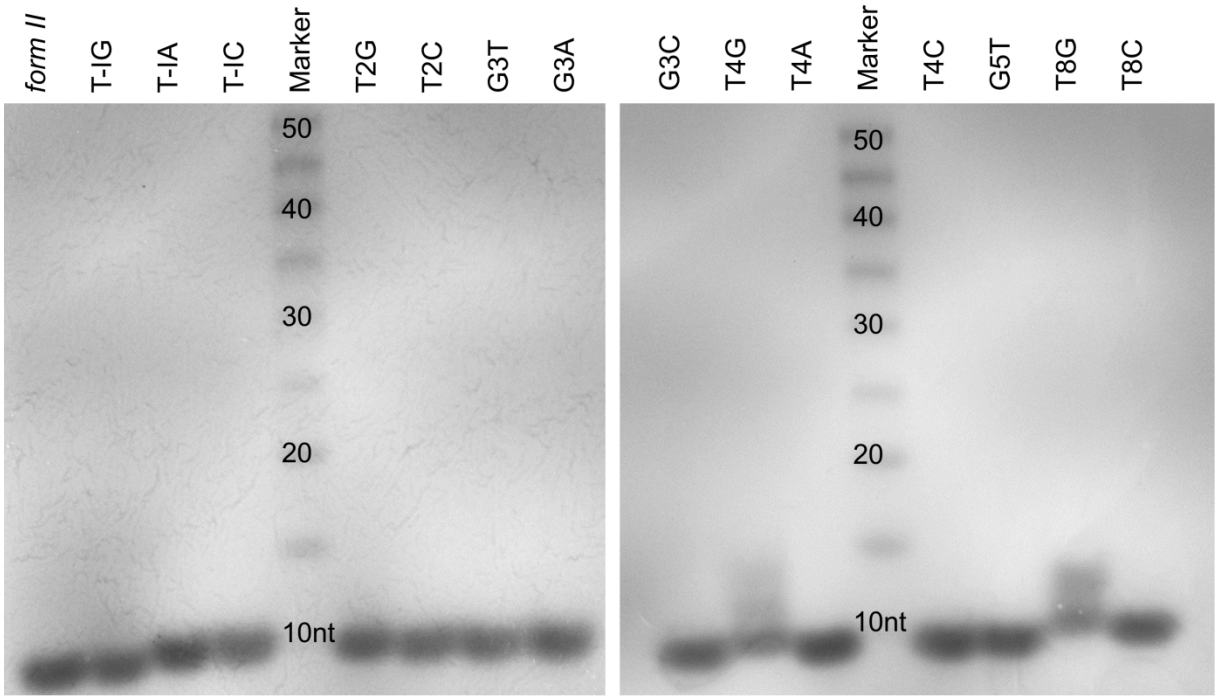


Figure S13. Native PAGE of the parent *form II* sequence and its single-point mutant variants that form stable secondary structures (cf. Figure S7 and S9).

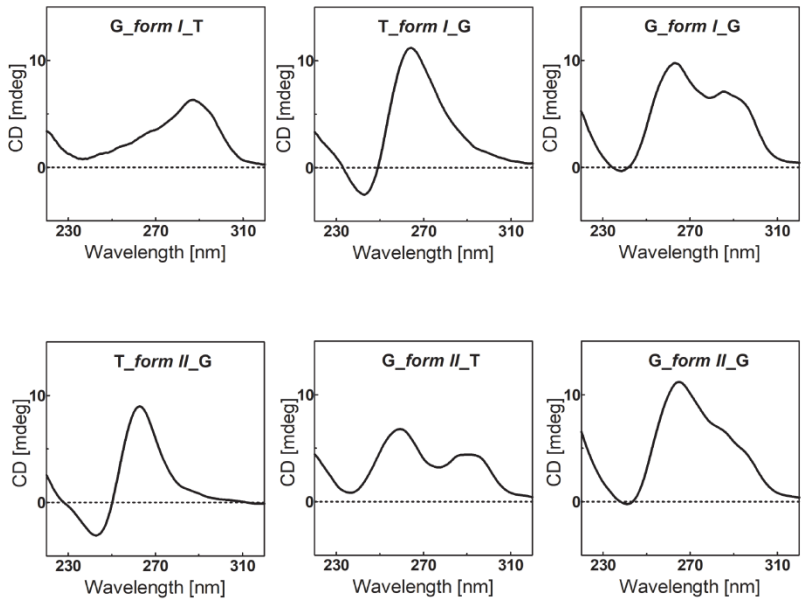


Figure S14. CD spectra of selected extended minimal constructs for *form I* and *form II*.

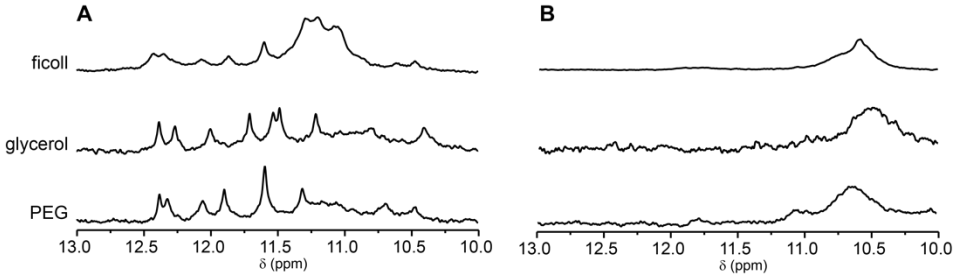


Figure S15. Imino regions of 1D ^{1}H NMR spectra of SC11 **(A)** and SC14 **(B)** acquired at 10 °C in the potassium based buffer (10 mM KPOi + 100 mM KCl, pH=7) supplemented with 20% w/v of Ficoll 70, PEG200, or glycerol.

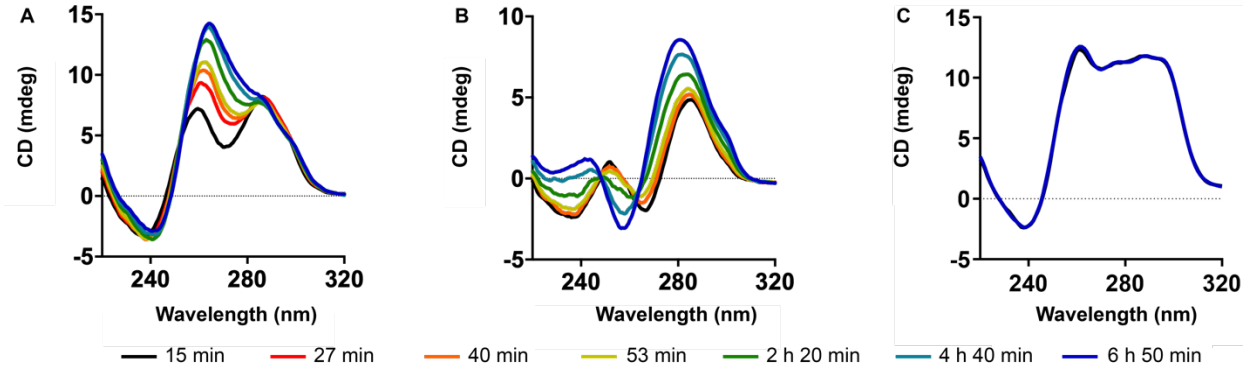


Figure S16. **(A)** and **(B)** CD spectra of SC11 acquired at 20 °C and 1 °C as a function of the time after annealing and quenching on ice in potassium phosphate based buffer (10 mM KPO_i, pH=7, 100 mM KCl), respectively. **(C)** CD spectra of SC11 acquired at 1 °C as a function of the time after annealing and quenching on ice in the sodium phosphate based buffer (10 mM $NaPO_i$, pH=7, 100 mM NaCl).

References

1. Liboska,R., Snášel,J., Barvík,I., Buděšínský,M., Pohl,R., Točík,Z., Páv,O., Rejman,D., Novák,P. and Rosenberg,I. (2011) 4'-Alkoxy oligodeoxynucleotides: a novel class of RNA mimics. *Org. Biomol. Chem.*, **9**, 8261–8267.
2. Petrová,M., Páv,O., Buděšínský,M., Zborníková,E., Novák,P., Rosenbergová,Š., Pačes,O., Liboska,R., Dvořáková,I., Šimák,O., et al. (2015) Straightforward Synthesis of Purine 4′-Alkoxy-2′-deoxynucleosides: First Report of Mixed Purine–Pyrimidine 4′-Alkoxyoligodeoxynucleotides as New RNA Mimics. *Org. Lett.*, **17,** 3426–3429.
3. Gajarský,M., Živković,M.L., Stadlbauer,P., Pagano,B., Fiala,R., Amato,J., Tomáška,L., Šponer,J., Plavec,J. and Trantírek,L. (2017) Structure of a Stable G-Hairpin. *J. Am. Chem. Soc.*, **139**, 3591–3594.
4. Ngoc Luu K., Phan A.-T., Kuryavyi V., Lacroix L, Patel D.J. (2006) Structure of the Human Telomere in K+ Solution: An Intramolecular (3 + 1) G-Quadruplex Scaffold *J. Am. Chem. Soc.* 128, 30, 9963–9970
5. Fialova M., Kypr J., Vorlickova M. (2006) The thrombin binding aptamer GGTTGGTGTGGTTGG forms a bimolecular guanine tetraplex. *Biochem Biophys Res Commun.* 344(1):50-4.
6. Padrta P., Stefl R., Králík L., Zídek L., Sklenar V. (2002) Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J Biomol NMR.* 24(1):1-14.

7. Cerofolini L., Amato J., Giachetti A., Limongelli V., Novellino E., Parrinello M., Fragai M., Randazzo A., Luchinat C. (2014) G-triplex structure and formation propensity. *Nucleic Acids Res*. 42(21):13393-404.
8. Limongelli V., De Tito S., Cerofolini L., Fragai M., Pagano B., Trotta R., Cosconati S., Marinelli L., Novellino E., Bertini I., Randazzo A., Luchinat C., Parrinello M. (2013) The G-triplex DNA. *Angew Chem Int Ed Engl*. 52(8):2269-73.