Prevalent sequences in the human genome can form mini i-motif structures at physiological pH.

Bartomeu Mir[‡,§], Israel Serrano[†], Diana Buitrago[&], Modesto Orozco[&,%], Núria Escaja[‡,§]* and Carlos González[†,§]*.

[†]Instituto de Química Física 'Rocasolano', CSIC, Serrano 119, 28006 Madrid, Spain

[‡] Inorganic and Organic Chemistry Department, Organic Chemistry Section, and IBUB, University of Barcelona, Martí i Franquès 1-11, 08028 Barcelona, Spain.

[&] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST). 08028 Barcelona, Spain

[%] Departament de Bioquímica i Biomedicina. Facultat de Biología. Universitat de Barcelona, 08028 Barcelona, Spain

[§] BIOESTRAN associated unit UB-CSIC.

Supporting Information Placeholder

ABSTRACT: We report here the solution structure of several repetitive DNA sequences containing d(TCGTTCCGT) and related repeats. At physiological pH, these sequences fold into an i-motif like quadruplexes in which every two repeats a globular structure is stabilized by two hemiprotonated C:C$^+$ base pairs, flanked by two minor groove tetrads resulting from the association of G:C or G:T base pairs. The interaction between the minor groove tetrads and the nearby C:C$^+$ base pairs affords a strong stabilization, which results in effective pH$_T$ values above 7.5. Longer sequences with more than two repeats are able to fold in tandem, forming a rosary bead-like structure. Bioinformatics analysis shows that these sequences are prevalent in the human genome, and are present in development-related genes.

Repetitive DNA sequences are attracting much scientific attention because of their ability to induce genetic instabilities, which ultimately can lead to human diseases. Although the molecular mechanism of these genetic instabilities is not well understood, it has been suggested that it can be related to the tendency of repetitive sequences to adopt non B-DNA structures,[1] such as hairpins, Z-DNA or quadruplexes. Repetitive sequences containing guanine tracts have been extensively studied because of their occurrence in telomeric regions and their ability to fold into G-quadruplexes. Under acidic conditions, their complementary C-rich strands can form i-motif structures.[2] Although a number of repetitive sequences fold into this motif,[3] and recent studies reveal the active role of i-motifs in gene transcription regulation,[4] this class of non-canonical structures has been less studied than others due to its low stability at physiological conditions.

The i-motif is a four-stranded structure formed by the association of two parallel-stranded duplex through hemi-protonated C:C$^+$ base pairs. The two duplexes are intercalated in opposite orientations. Since i-motif formation requires cytosine protonation, these structures are usually stable only at acidic pH. However, recent studies have shown that some particular sequences may be stable at neutral conditions.[5] I-motif structures can be stabilized by external agents, such as molecular crowding agents[6] or by the appropriate chemical modifications.[7] Capping interactions at the ends of the C:C$^+$ stack also play an important role in i-motif stability. In a previous study, we showed that minor groove G:T:G:T tetrads are excellent capping elements of C:C$^+$ stacks.[8] These tetrads result from the association of two G:T mismatches through their minor groove side, and have been found in i-motifs[9] and in other structures.[10] This family of tetrads have been also observed with Watson-Crick G:C[11] and A:T[12] base pairs (G:C:G:C and A:T:A:T), as well as with combinations of G:C, A:T and G:T base pairs.[10, 13]

In this paper, we study several repetitive sequences designed to form small i-motifs based on interactions between C:C$^+$ base pairs and minor groove tetrads. We found that the combination of these two secondary structural elements renders a very unique i-motif like structure of an extraordinary stability under physiological conditions.

Table 1. Melting temperature and pH$_T$ values of different oligonucleotides studied in this paper. Experimental conditions: [oligonucleotide]=3.5 μM, 10 mM sodium phosphate, pH 7.

Name	Sequence	pH$_T$	T$_m$ (°C)
LL1	d(L-T-L)	n.d.	24.3
LL2	d(L-T$_2$-L)	n.d.	27.9
LL3	d(L-T$_3$-L)	7.8	32.1
LL4	d(L-T$_4$-L)	7.9	28.4
LL5	d(L-T$_5$-L)	n.d.	25.4
LL6	d(L-T$_6$-L)	n.d.	23.3
LL7	d(L-T$_7$-L)	n.d.	21.7
LL3rep	d(L-T$_3$-L-T$_3$-L-T$_3$-L)	7.6	23.7
LL3long	d(L-T$_3$-L-T$_6$-L-T$_3$-L)	7.5	23.0
L: TCGTTCCGT			

The DNA oligonucleotides studied here are shown in Table 1. Their sequences are similar to those that result from connecting the two subunits of the dimeric structure of d(TCGTTTCGT) with poly-thymidine linkers of different lengths.[8] The resulting construction may adopt different topologies that would be difficult to distinguish in a sequence with perfect TCGT repeats. Thus, to facilitate the assignment of the NMR spectra and to discriminate between possible topologies, we focused on related sequences containing d(TCGTTCCGT) repeats, named as **L**. These repeats can form different minor groove tetrads depending on the folding pattern, i.e. either two mixed G:C:G:T tetrads or one G:C:G:C and one G:T:G:T tetrad (Figure S1). Interestingly, the NMR spectra of all these sequences at pH 7 exhibit the distinctive imino signals of hemiprotonated C:C$^+$ base pairs (around 15 ppm), together with imino resonances characteristic of GC Watson-Crick base pairs (13 ppm) and additional narrow signals in the 10-11 ppm region, where GT mismatches are found (Figures 1D, S2 and S3). NMR spectra recorded at different temperatures indicate that the structures are stable at neutral pH (see Figure S2 and S3). This is confirmed by CD spectra, which, under the same neutral conditions, exhibit a positive band around 265 nm (Figure 1C) that disappears upon temperature increases (see Figure S2 and S3). The CD spectra are distinct from those seen in most i-motif structures with the characteristic band at 285 nm shifted to lower wavelength. These CD spectra resemble those observed in the dimeric structure of d<pTCGTTTCGT> and other structures stabilized by slipped minor groove tetrads.[8,10,11a] Melting experiments were conducted by following the decrease of the maximum ellipticity (Figure 1A). The resulting Tm values are shown in Table 1.

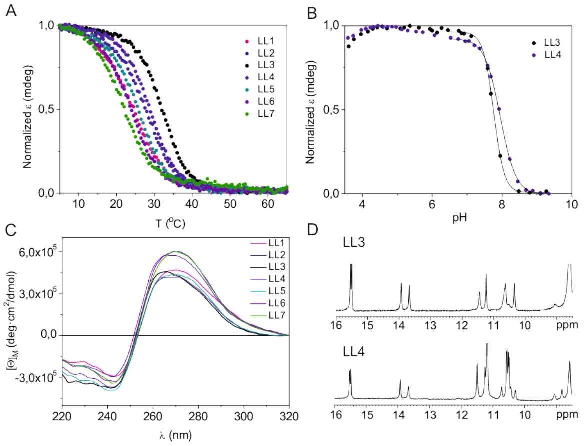


Figure 1. (A) CD melting experiments of LL1-7 at pH 7, B) pH titration of LL3 and LL4, C) CD spectra of the different sequences, and D) exchangeable protons region of the ^{1}H-NMR spectra of LL3 and LL4 at T=5°C, 10 mM phosphate buffer, pH7. [oligonucleotide]=3.5 μM for CD/UV and 1 mM for NMR.

The comparison of the apparent T_m and pH_T values for the different oligonucleotides reveals a clear dependence on the polythymidine loop length, with LL3 and LL4 being the most stable. Based on this, we focused on these two oligonucleotides for a more detailed structural study.

As expected for i-motif like structures, their thermal stability is pH-dependent. UV melting experiments recorded at two different pHs are shown in Figure S4. The effective pH_T values (midpoint of the pH transition) were estimated by carrying out a CD-monitored pH titration. As shown in Table 1 and S1, pH_T values are surprisingly high for structures containing hemiprotonated C:C$^+$ base pairs, although not completely unprecedented.[5]

To determine the molecularity of the structure, melting experiments were run at different oligonucleotide concentrations. The results shown in Figure S5 clearly indicate the formation of monomeric structures, since the T_m values do not depend on DNA concentration. This is confirmed by electrophoretic gel experiments shown in Figure S6. The electrophoretic mobility of sequences containing two repeats of the sequence **L** is compared with a sequence containing four repeats (LL3rep) in addition to poly-dT references. For all the LL sequences, the observed spots are in the range of dT15 and dT30 and migrate significantly faster than LL3rep, confirming the formation of unimolecular structures.

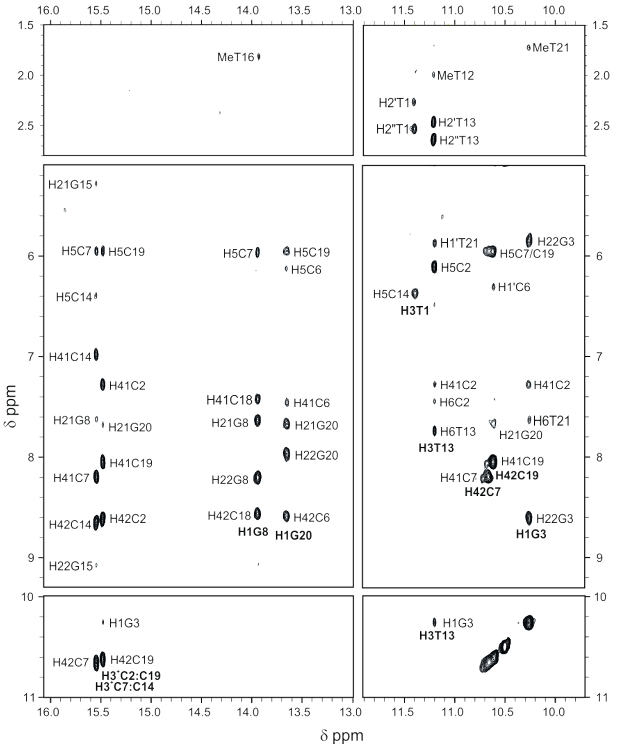


Figure 2. Exchangeable protons regions of the NOESY spectra (mixing time = 150 ms) of 1 mM LL3, T= 5°C, 10 mM phosphate buffer, pH 7.

Two-dimensional NMR spectra of LL3 and LL4 were recorded and fully assigned. Cytosine residues were assigned by analysing several constructions in which particular cytosines were substituted by 5-Me-dC (^{m}C) residues (see Figures S7-10 for details). The spectra are of excellent quality and show narrow lines and good dispersion of exchangeable signals (see Figures 2, S11 and S12). In both cases the main spectral features are identical. Two hemiprotonated imino signals are shown around 15.5 ppm. Each of these signals shows cross-peaks with two pairs of amino protons. Two guanine imino signals are found around 13-14 ppm, the characteristic region of Watson-Crick G:C base pairs, and exhibit cross-peaks with the amino protons of two cytosines. The other guanine imino signals are found upfield, (~10-12 ppm). These signals show cross-peaks with imino protons of two thymine residues, indicating the formation of G:T base pairs (see Figure 2 and S11, imino proton of G15 in LL3 is not observed). The formation of G:C:G:T tetrads is supported by a number of cross-peaks between amino and H1' protons of different guanines. Interestingly, significant differences are found in the chemical shift of amino protons of cytosine residues involved in the C:C$^+$ base pairs. Those belonging to the cytidine residues stacked with the G:C base pair of the tetrad exhibit downfield values. Full details on the assignment procedures are given in the supplementary material. The chemical shifts are listed in Tables S2-5.

The three-dimensional structure of LL3 was determined on the basis of approximately 125 experimental distance constraints by using restrained molecular dynamics methods (see Supplementary Information (SI). and Tables S6 and S7.). Except for some loop residues, the structures are very well-defined with an RMSD of 0.8 Å. The structures consist of a short stack of two hemiprotonated C:C$^+$ base pairs, surrounded by two minor groove G:C:G:T tetrads (see Figure 3). Adjacent guanines and cytosines in these tetrads are connected by two-thymine loops, being the first thymines in well-defined positions capping the tetrads, whereas the second ones are disordered. The two segments containing the d(TCGTTCCGT) repeats are oriented forming an X-shape when seen from the minor groove side (Figure 3). The loop connecting the two repeats is relatively disordered in both structures. As previously seen in i-motifs and other structures stabilized by minor groove tetrads, the minor grooves are extremely narrow with a number of inter-strand sugar-sugar contacts. These favorable contacts, together with the positive charge of the hemiprotonated base pairs, help alleviate the electrostatic repulsion arising between nearly phosphate groups on contiguous strands. Furthermore, the structure is stabilized by a number of hydrogen bonds, arising from two $C:C^+$, two G:C and two G:T base pairs, in addition to those from G:G interaction in the G:C:G:T tetrads.

These two guanines lay exactly on top of the $C:C^+$ base pair, as shown in Figures 3 and S13. These stacking interactions further stabilize the structure and, most probably, contribute to maintain the cytosines hemiprotonated at unusually high pH values. The key importance of this interaction is reflected in the temperature dependence of the exchangeable protons spectra. As shown in Figures S2 and S3, imino signals of the G:C:G:T tetrads and the $C:C^+$ base pairs disappear at the same temperature, indicating that these two structural elements unfold concomitantly.

The lateral two-residue loops are very similar to those found in the dimeric structure of d<pTCGTTTCGTT> and other related structures connecting the sides of minor grove tetrads. Previous studies showed that the optimal number of residues connecting the two sides of this kind of tetrads is two.[14] More interesting is the effect of the length of longer loop connecting the two repeats. The NMR spectra of the different oligonucleotides studied here clearly indicate that similar structures are formed regardless of the number of residues in this loop (Figures S2 and S3). Although the loop length affects the stability of the structure, as mentioned above, our results indicate that this motif is rather tolerant of different loop lengths.

Slipped G:C:G:C tetrads are structurally very similar to G:C:G:T and G:T:G:T tetrads. To check whether these tetrads can also stabilize this minimal i-motif, two sequences, containing T←→C mutations at positions 1 and 6 of the **L** repeat were studied. The resulting repeats d(TCGTTTCGT) and d(CCGTTCCGT), named as **M** and **N**, respectively, were connected with a four thymine loop. The NMR spectra of these oligonucleotides (MM4 and NN4) are shown in Figure S14. In addition to the imino signals from hemiprotonated $C:C^+$ base pairs, imino proton resonances from G:C bases pair (in NN4) and from G:T mismatches (MM4) are observed. These NMR data clearly indicate that their overall structures exhibit the same features as LL3 and LL4.

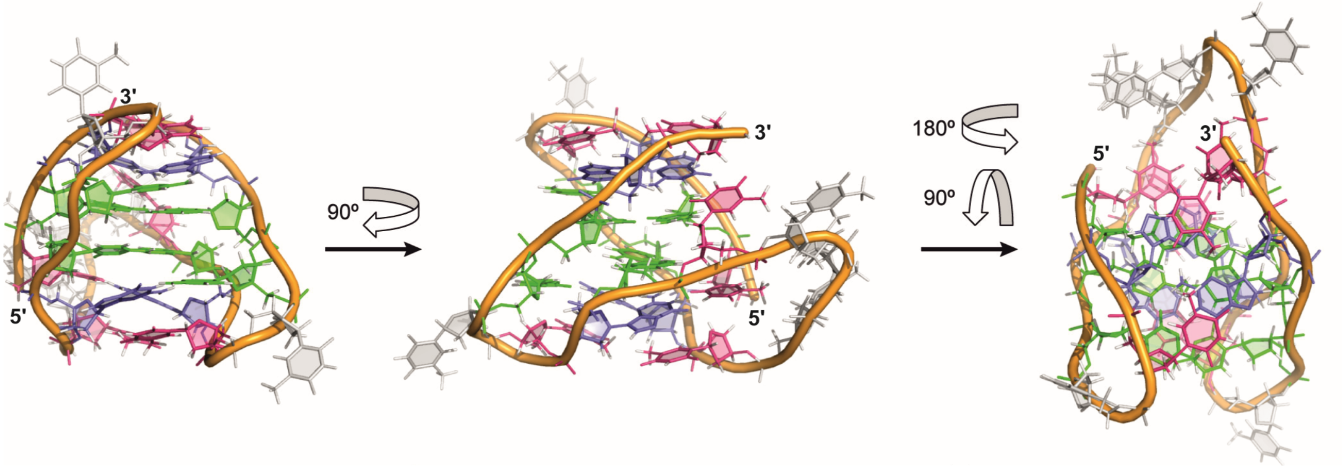


Figure 3. Different views of the calculated structure of LL3. Cytosines are shown in green; guanines in blue; thymines with a well-defined structure are in magenta and the rest in grey. PDB code 5OGA.

Finally, we tested whether this motif can fold in tandem. Two constructions (LL3rep and LL3long), consisting of four repeats of d(TCGTTCCGT) connected with poly-T loops of different lengths were prepared. Most interestingly, the NMR and CD spectra recorded at pH 7 show the same general features observed for LL3 (Figure 4C and 4D) and similar pH and thermal stability as that found for LL1-7 sequences (see Table 1 and Figure S15). We must conclude that this structural motif can occur in tandem, forming rosary bead-like structures as shown in Figure 4E. The model shown in this figure was built from a sequence with 6 L-repeats, forming a domain with the structure of LL3 every two repeats. As working hypothesis, we have assumed that the loop connecting each of the units is disordered, and no interaction occurs between each individual unit. Although this is consistent with our NMR data, more work is now in progress in our group to fully confirm these assumptions.

Although a more systematic study is necessary to fully assess the sequential requirements for the formation of these structures, the fact that most thymines in the loops are disordered suggests that their contribution to the stability is not significant, and they might be substituted by other nucleotides. C:G steps at the appropriate distance in the sequence to bring two G:C or G:T base pairs into register for a minor groove tetrad formation besides $C:C^+$ base pair is probably the necessary sequential requirement. Based on these premises, a consensus motif can be established as $d(YCG(XX)YCG(X_n)YCG(XX)YCG)$, where X can be any nucleotide.

We mapped this consensus motif with n from 4 to 10 bases (forcing exact matching) to the hg19 version (UCSC GRCh37, Feb/2009) of the human genome, finding 4971 hits, more than expected from a random model (p value $< 10^{-28}$). The most common length for the connection loop is 7 nucleotides (which appear twice more than expected, p-value $< 10^{-113}$). Very interestingly, regions susceptible to form mini i-motif structures are not randomly distributed in the genome, but are very localized in regulatory regions, very close to the transcription start site (TSS; see Figure S16), especially in promoters and 5'UTRs (see Table S8). Gene ontology analysis (GO)[15] reveals that the consensus motif is over-represented (again with a very high statistical confidence) in development-related genes (see Table S9).

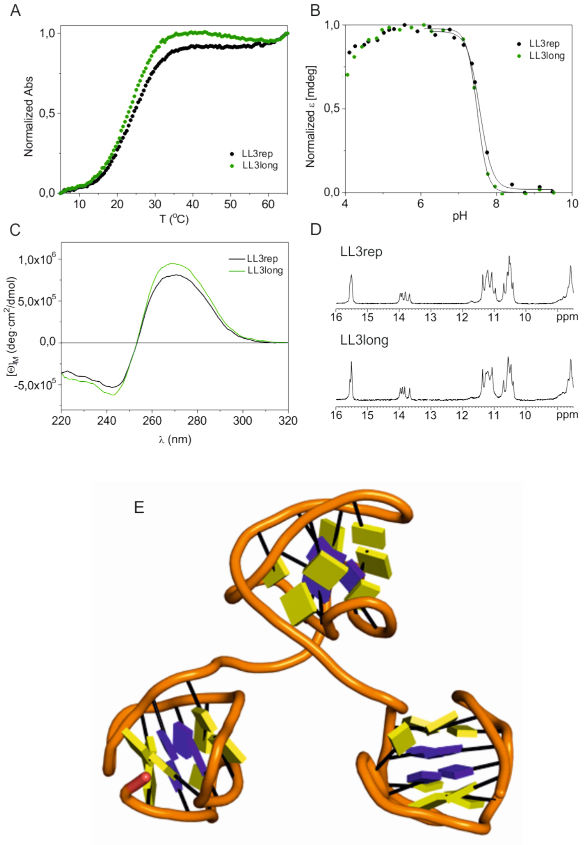


Figure 4. Studies on LL3rep and LL3long (A) UV melting experiments B) pH titration C) CD spectra and D) exchangeable protons region of the ^{1}H-NMR at T=5°C, 10 mM phosphate buffer, pH7. [oligonucleotide]= 3.5 μM for UV/CD and 1mM for NMR. E) Model of the formation of three mini i-motifs in tandem.

In summary, we have found that single stranded DNA can fold into tandem repeats of a novel i-motif like quadruplex structures at physiological pH. Bioinformatics analysis very strongly suggest that mini i-motif forming sequences are not only prevalent in the human genome, but are present in key regulatory elements associated to genes which need to be tightly controlled during development and differentiation. All these findings very strongly support an important functional role for the suggested structure. It is tempting to believe, that as found for telomeres, the equilibrium between B-type duplex and mini i-motif can help in the control of the expression of these development-related genes.

ASSOCIATED CONTENT

Supporting Information.

Detailed descriptions of the experimental procedures and NMR assignments; 10 tables with assignment, calculation statistics, and structural analysis; 16 figures showing UV melting, CD, electrophoretic experiments, NMR data, and details on the structural models and sequence analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

nescaja@ub.edu; cgonzalez@iqfr.csic.es.

Funding Sources

No competing financial interests have been declared. This work was supported by the MINECO grants CTQ2014-52658-R, BFU2014-61670-EXP and BFU2014-52864-R. The European Research Council (ERC SimDNA), the European Union's Horizon 2020 program (grant No 676556), Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030), and the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). B.M. was supported by an ADR University of Barcelona fellowship.

REFERENCES

[1] a) A. Bacolla, R. D. Wells, *Molecular carcinogenesis* **2009**, *48*, 273-285; b) R. D. Wells, *Trends Biochem Sci* **2007**, *32*, 271-278.

[2] a) K. Gehring, J-L. Leroy, M. Guéron, *Nature* **1993**, *363*, 561-565; b) S. Benabou, A. Aviñó, R. Eritja, C. González, R. Gargallo, *RSC Adv.* **2014**, *4*, 26956-26980; c) H. A. Day, P. Pavlou, Z. A. Waller, *Bioor. Med. Chem.* **2014**, 22, 4407-4418.

[3] a) A. T. Phan, M. Guéron, J. L. Leroy, *J Mol Biol* **2000**, *299*, 123-144; b) M. Garavís, N. Escaja, V. Gabelica, A. Villasante, C. González, *Chem-Eur J* **2015**, *21*, 9816-9824; c) Y. W. Chen, C. R. Jhan, S. Neidle, M. H. Hou, *Angew Chem Int Ed* **2014**, *53*, 10682-10686.

[4] a) S. Kendrick, H. J. Kang, M. P. Alam, M. M. Madathil, P. Agrawal, V. Gokhale, D. Z. Yang, S. M. Hecht, L. H. Hurley, *J. Am. Chem. Soc.* **2014**, *136*, 4161-4171; b) H. J. Kang, S. Kendrick, S. M. Hecht, L. H. Hurley, *J. Am. Chem. Soc.* **2014**, *136*, 4172-4185.

[5] a) J. A. Brazier, A. Shah, G. D. Brown, *Chem. Comm.* **2012**, *48*, 10739-10741; b) A. M. Fleming, Y. Ding, R. A. Rogers, J. Zhu, J. Zhu, A. D. Burton, C. B. Carlisle, C. J. Burrows, *J. Am. Chem. Soc.* **2017**, *139*, 4682-4689; c) E. P. Wright, J. L. Huppert, Z. A. E. Waller, *Nucleic Acids Res.* **2017**, *45*, 2951-2959; d) A. Kovanda, M. Zalar, P. Sket, J. Plavec, B. Rogelj, *Sci. Rep.* **2015**, *5*, 17944.

[6] J. Cui, P. Waltman, V. H. Le, E. A. Lewis, *Molecules* **2013**, *18*, 12751-12767.

[7] H. Abou Assi, R. W. Harkness, N. Martin-Pintado, C. J. Wilds, R. Campos-Olivas, A. K. Mittermaier, C. González, M. J. Damha, *Nucleic Acids Res.* **2016**, *44*, 4998-5009.

[8] N. Escaja, J. Viladoms, M. Garavís, A. Villasante, E. Pedroso, C. González, *Nucleic Acids Res.* **2012**, *40*, 11737-11747.

[9] J. Gallego, S. H. Chou, B. R. Reid, *J. Mol. Biol.* **1997**, *273*, 840-856.

[10] J. Viladoms, N. Escaja, E. Pedroso, C. González, *Bioor. Med. Chem.* **2010**, *18*, 4067-4073.

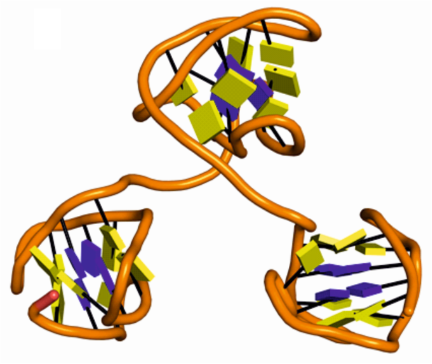
[11] a) N. Escaja, I. Gómez-Pinto, E. Pedroso, C. González, *J. Am. Chem. Soc.* **2007**, *129*, 2004-2014; b) J. Viladoms, N. Escaja, M. Frieden, I. Gómez-Pinto, E. Pedroso, C. González, *Nucleic Acids Res.* **2009**, *37*, 3264-3275; V. Kocman, J. Plavec, *Nat. Commun.* **2017**, *8*, 15355.

[12] N. Escaja, E. Pedroso, M. Rico, C. González, *J. Am. Chem. Soc.* **2000**, *122*, 12732-12742.

[13] N. Escaja, J. L. Gelpi, M. Orozco, M. Rico, E. Pedroso, C. González, *J. Am. Chem. Soc.* **2003**, *125*, 5654-5662.

[14] N. Escaja, I. Gómez-Pinto, J. Viladoms, E. Pedroso, C. González, *Org. Biomol. Chem.* **2013**, *11*, 4804-4810.

[15] McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. *Nat. Biotech.*, **2010**, *28*, 495–501.



Insert Table of Contents artwork here

ELECTRONIC SUPPORTING INFORMATION

Prevalent sequences in the human genome can form mini i-motif structures at physiological pH.

Bartomeu Mir[‡,§], Israel Serrano[†], Diana Buitrago[&], Modesto Orozco[&,%], Núria Escaja[‡,§]* and Carlos González[†,§]*.

[†]Instituto de Química Física 'Rocasolano', CSIC, Serrano 119, 28006 Madrid, Spain

[‡] Inorganic and Organic Chemistry Department, Organic Chemistry Section, and IBUB, University of Barcelona, Martí i Franquès 1-11, 08028 Barcelona, Spain

[&] Institute for Research in Biomedicine (IRB), 08028 Barcelona, Spain

[%]Departament de Bioquímica I Biomedicina. Facultat de Biología. Universitat de Barcelona, 08028, Barcelona, Spain

[§] BIOESTRAN associated unit UB-CSIC.

Supplementary Figures

Supplementary Tables

Methods

Oligonucleotides synthesis. Oligodeoxynucleotides **LL3** and **LL4** were synthesized on an ABI 3400 DNA synthesizer by using standard solid-phase phosphoramidite chemistry at 1 μmol scale. Cleavage from the solid support and nucleobases deprotection were carried out with concentrated aqueous ammonium hydroxide at room temperature for 12 h. Crude products were purified by reverse phase HPLC (250x10 mm Jupiter C18 column from Phenomenex, solvent A: 0.1M TEAA pH=7, solvent B: ACN). Oligonucleotides were further desalted by EtOH precipitation and characterized by MS-MALDI-TOF. Synthesis results are summarized in Supplementary Table S10. The other oligonucleotides were purchased (IDT).

Mass spectrometry. MS-MALDI-TOF spectra of **LL3** and **LL4** were acquired in the negative ion mode on an ABSciex 4800 plus device. Samples were prepared by mixing 1μL of oligonucleotide solution (100-500 μM) with 1μL of ammonium citrate (50 mg/mL) and allowed to interact for few seconds. Next 1 μL of the mixture and 1μL of the matrix (2,4,6-trihidroxyacetophenone, THAP, 10 mg/mL in H2O/ACN 1:1) were mixed and deposited onto the plate.

CD and UV spectroscopy. Circular dichroism spectra were recorded on a Jasco J-815 spectropolarimeter. UV spectra were recorded on a Jasco V-730 spectrophotometer. Both fitted with a thermostated cell holder. Spectra were recorded in 25 mM sodium cacodylate buffer or 25 mM sodium acetate buffer at different pH values. Samples were initially heated at 90°C for 5 min, and slowly allowed to cool to room temperature and stored at 4°C until use. CD and UV melting curves were recorded at the wavelength of the larger positive band, ~ 265 nm, with a heating rate of 0.5 °C·min^{-1}. Associated error in T_m values determination has been estimated in 0.2 °C.

For pH titration experiments, the pH was adjusted by adding aliquots of concentrated solutions of HCl or NaOH. The effective pH-transition midpoint (pH_T) values were obtained from the plot of observed ellipticity (mdeg), at around 265 nm, versus pH in 25 mM NaPi buffer at 5°C. CD versus pH data were fit to a standard titration model involving a single protonation event using a dose-response equation (OriginPro 8):

$$\varepsilon = \varepsilon_{unfolded} + \frac{\varepsilon_{folded} - \varepsilon_{unfolded}}{1 + 10^{cooperativity\cdot(pH-pH_T)}}$$

This equation allows obtaining the pH_T and the cooperativity parameter (See Table S1). Associated error in pH_T determination has been estimated in 0.1 pH units.

Molar extinction coefficients for all sequences (see Supplementary Table S10) were calculated by applying the nearest-neighbor method.

Gel electrophoresis. Polyacrylamide gel electrophoresis under native conditions was performed to confirm the molecularity of the structures. 20% polyacrylamide gels (acrylamide/bisacrylamide 19:1) were prepared at pH 7 using phosphate buffer. 0.05 OD oligonucleotide samples were prepared in water/sucrose 2M 1:1 and equilibrated as described in the CD and UV section. Running of the gels was done in phosphate buffer 50 mM at pH 7 for 6-7 h at 180 V. Visualization of the gels was achieved by treatment with Stains-All dye from Sigma. Poly-dT references (dT_{20}, dT_{25}, dT_{30}, dT_{35}, dT_{40} and dT_{50}) were used as oligonucleotide length controls.

NMR. Samples for NMR experiments were dissolved (in Na^+ form) in either D_2O or 9:1 H_2O/D_2O, 10 mM sodium phosphate buffer. Experiments were carried out at different pH values, ranging from 4 to 7. The pH was adjusted by adding aliquots of concentrated solution of either DCl or NaOD. All NMR spectra were acquired on Bruker spectrometers operating at 600 and 800 MHz, equipped with cryoprobes and processed with the TOPSPIN software. For the experiments in D_2O, presaturation was used to suppress the residual H_2O signal. A jump-and-return pulse sequence[1] was employed to observe the rapidly exchanging protons in 1D H_2O experiments. NOESY[2] spectra in D_2O and 9:1 H_2O/D_2O were acquired with mixing times of 150, 250 and 300 ms. TOCSY[3] spectra were recorded with the standard MLEV-17 spin-lock sequence and a mixing time of 80 ms. In most of the experiments in H_2O, water suppression was achieved by including a WATERGATE[4] module in the pulse sequence prior to acquisition. The spectral analysis program SPARKY was used for semiautomatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities.

Assignment of the NMR spectra. The NMR spectra of all sequences exhibit very similar features, although we focused on **LL3** and **LL4**. To overcome the intrinsic difficulties of the specific assignment of repetitive sequences, we studied two additional oligonucleotides in which one or two cytosines were replaced by 5-methyl-cytosine residues (^{m}C). The NMR spectra of these two oligonucleotides **LL4-M1:** d(T^{m}CGTTCCGT-T4-TCGTTCCGT); and **LL4-M2**: d(T^{m}CGTTCCGT-T4-

T^{m}CGTTCCGT) were partially assigned. Unambiguous assignment of C2 and C15 residues in **LL4** could be carried out by observing the changes in the H5-H6 region of the TOCSY spectra (Figure S7). From these starting points, sequential assignment of all the residues could be carried out by analyzing TOCSY and NOESY spectra following standard methods. Assignment details for **LL4-M1** are given in Figures S8 and S9, for **LL4-M2** in Figure S10, **LL4** in Figure S11, and for **LL3** in Figure S12. In these figures, cross-peaks involving multiple overlapped resonances are labelled using a slash symbol (i.e. H1'C7/C20-H8G8/G21). Chemical shift lists are given in Tables S2-5.

NMR constraints. Qualitative distance constraints were obtained from NOE intensities. NOEs were classified as strong, medium or weak, and distances constraints were set accordingly to 3, 4 or 5 Å. In addition to these experimentally derived constraints, hydrogen bond and planarity constrains for the base pairs were used. Target values for distances and angles related to hydrogen bonds were set to values obtained from crystallographic data in related structures[5]. Due to the relatively broad line-widths of the sugar proton signals, J-coupling constants were not accurately measured, but only, roughly estimated from DQF-COSY cross-peaks. Loose values were set for the sugar dihedral angles δ, ν_1 and ν_2 to constrain deoxyribose conformation to North or South domain.

Structural calculations. Structures were calculated with the program DYANA[6] and further refined with the SANDER module of the molecular dynamics package AMBER 12.0[7]. Resulting DYANA structures were taken as starting points for the AMBER refinement, consisting of an annealing protocol in vacuo, followed by trajectories of 500 ps each in which explicit solvent molecules were included and using the Particle Mesh Ewald method to evaluate long-range electrostatic interactions. The specific protocols for these calculations have been described elsewhere[8]. The BSC1 force field[9] was used to describe the DNA, and the TIP3P model was used to simulate water molecules. Analysis of the representative structures was carried out with the program MOLMOL[10].

The structural models for the sequences with different poly-thymine loops shown in Figures S2 and S3 were built on the basis of their similar spectra to **LL3**. The experimental distances observed for **LL3** were renumbered according to the different sequences and the models were calculated with the program DYANA. No refinement with AMBER package was carried out in this case.

Supplementary Figures

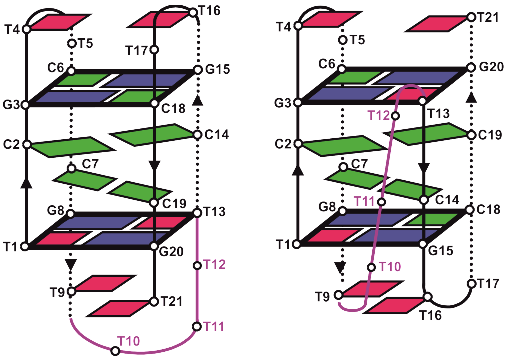


Figure S1. Possible unimolecular folding patterns for **LL3** are shown. In the head-to-head like structure one G:C:G:C and one G:C:G:T tetrads are formed, whereas in the head-to-tail like structure two mixed G:C:G:T tetrads are formed. Moreover, the pattern of C:C$^+$ base-pairs differs in each of the two arrangements. In the head-to-head like orientation (left), C2:C14 and C7:C19 base pairs are expected, whereas in the head-to-tail one (right), the base pairs should be C2:C19 and C7:C14. The assignment of the exchangeable protons spectra of **LL3** and **LL4** (see Figures 2 and S11) clearly shows the formation of the latter structure. Colour code: C in green, G in blue, and T in magenta

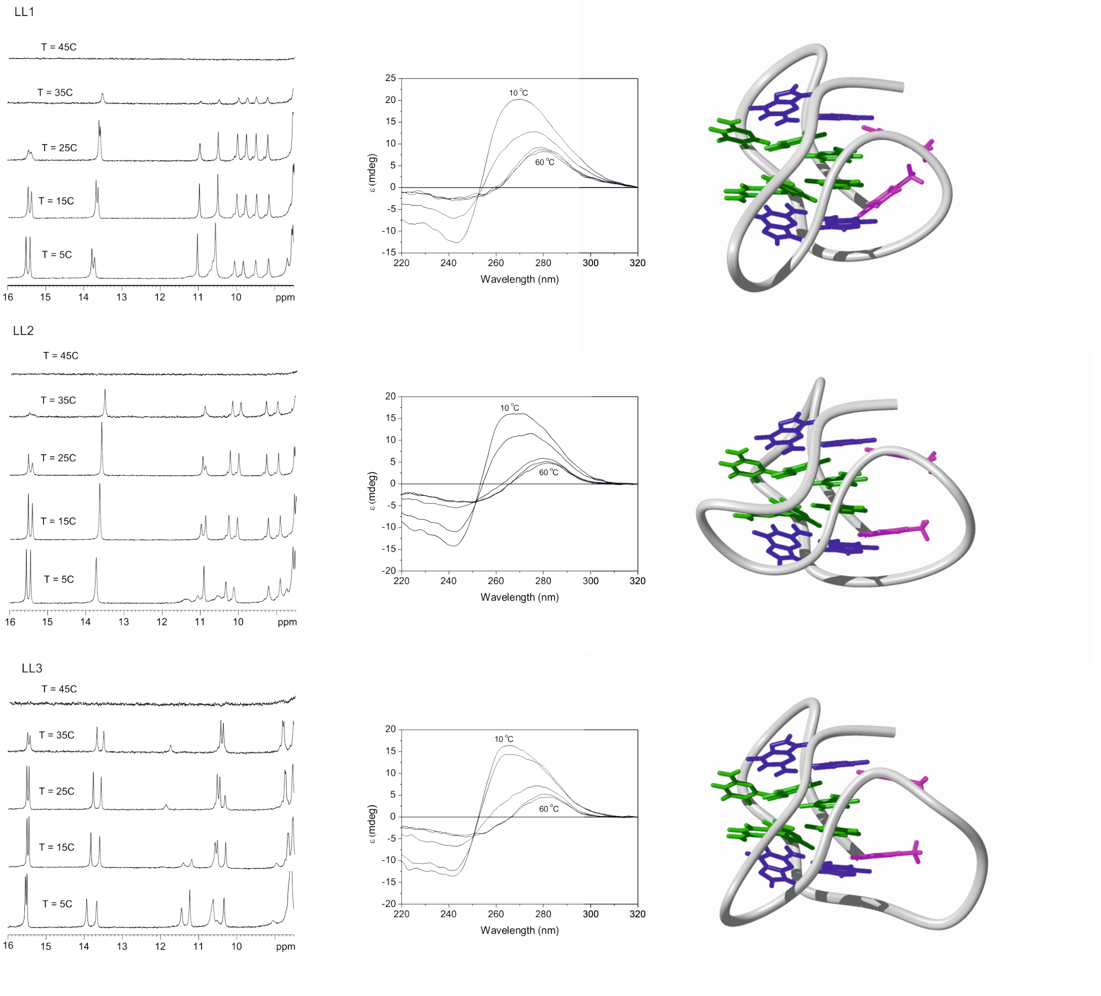


Figure S2. NMR and CD melting experiments at pH 7 (phosphate buffer, [oligonucleotide]=1 mM for NMR and 3.5 μM for CD) and structural models of **LL1**, **LL2** and **LL3**. Only the nucleotides forming the C:C$^+$ and G:C:G:T tetrads are shown. Colour code: cytosine in green, guanines in blue, and thymine in magenta.

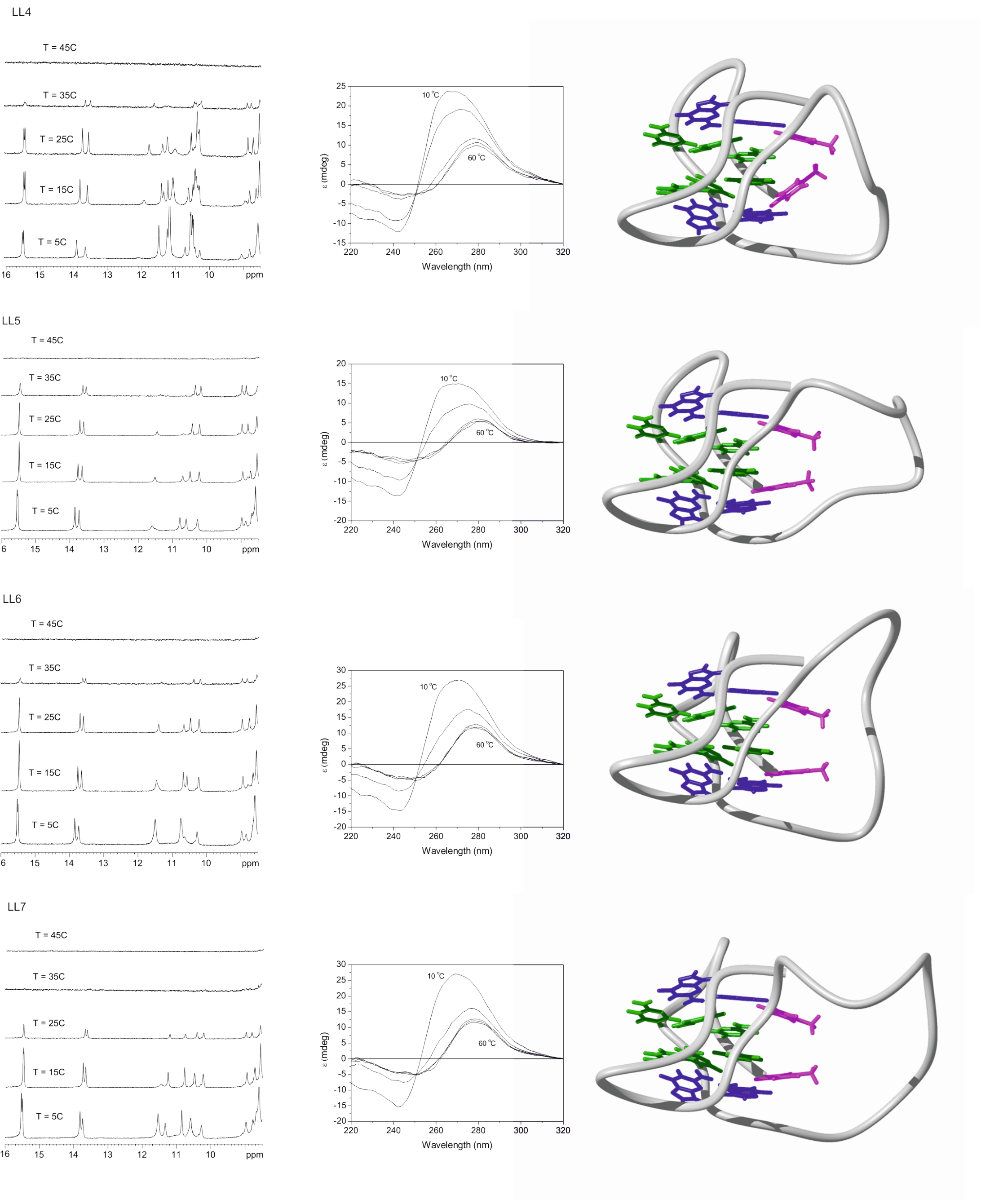


Figure S3. NMR and CD melting experiments at pH 7 (phosphate buffer, [oligonucleotide]=1 mM for NMR and 3.5 μM for CD) and and structural models of **LL4**, **LL5, LLA6** and **LL7**. Only the nucleotides forming the C:C$^+$ and G:C:G:T tetrads are shown. Colour code: cytosine in green, guanines in blue, and thymine in magenta.

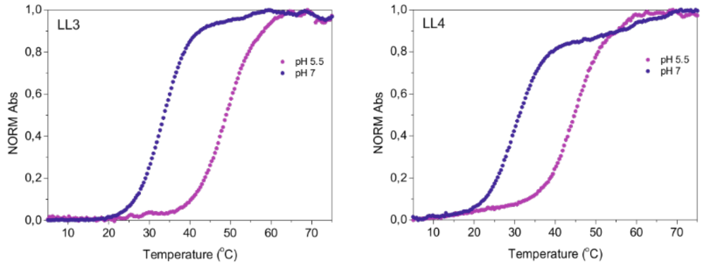


Figure S4. UV melting curves of **LL3** and **LL4** at different pH. [oligonucleotide]=3,5 μM, 25 mM cacodylate buffer.

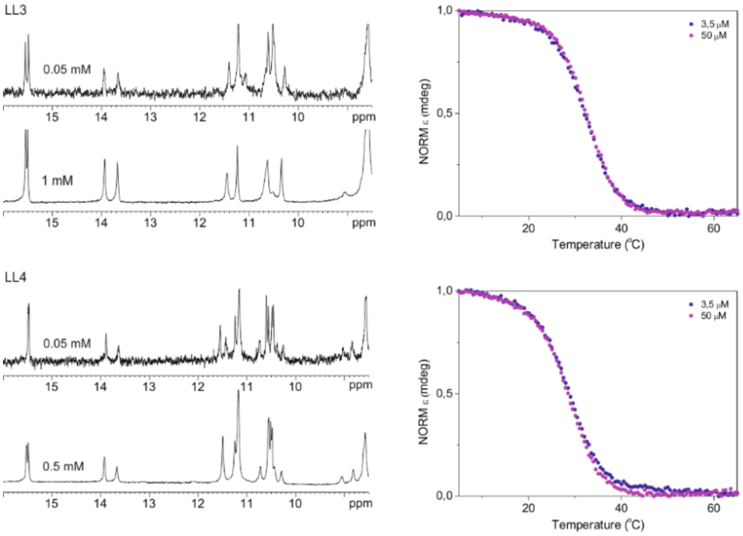


Figure S5. Left: NMR spectra of **LL3** and **LL4** at different oligonucleotide concentration, phosphate buffer pH 7, T=5°C. Right: CD melting curves of **LL3** and **LL4** at different oligonucleotide concentration, phosphate buffer, pH 7.

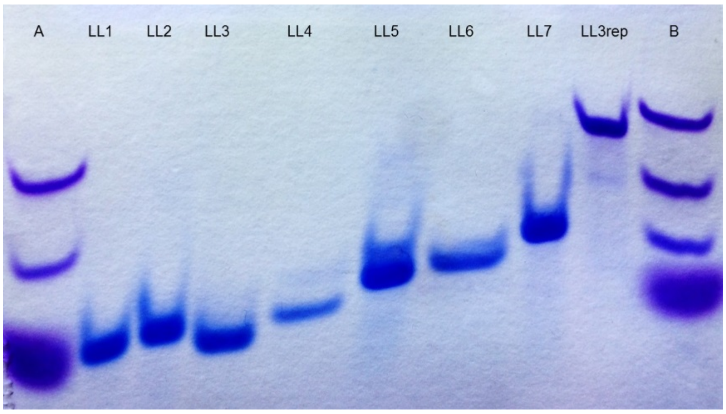


Figure S6. Native PAGE of **LL** sequences. 20% polyacrylamide gels (acrylamide/bisacrylamide 19:1), pH 7, phosphate buffer 0.05 M. Experimental conditions: 180 V for 6-7 hours, T=5°C. A = dT_{15}, dT_{25}, dT_{40} + bromophenol. B = dT_{20}, dT_{30}, dT_{50} + bromophenol.

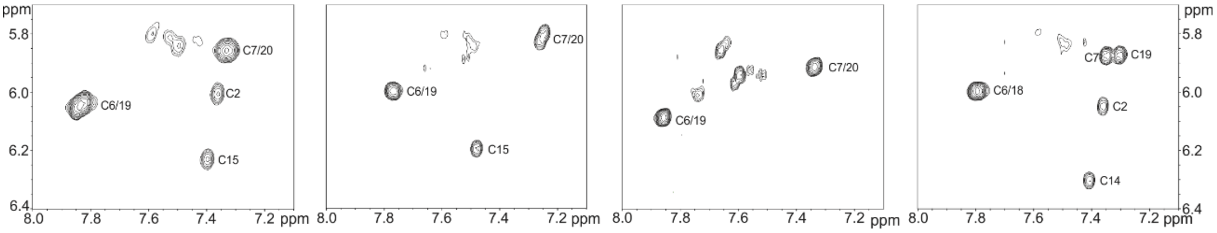


Figure S7. From left to right, H5-H6 cross-peaks region of TOCSY spectra of **LL4**, **LL4-M1**, **LL4-M2** and **LL3**. Comparison of H5-H6 cross-peaks region of the TOCSY spectra of **LL4** with two analogue sequences, **LL4-M1** and **LL4-M2,** in which a single cytosine in position 2, or two cytosines in positions 2 and 15, have been replaced by 5-methyl-cytosines. Unambiguous assignment of C2 and C15 residues was carried out by observing the H5-H6 cross-peaks that disappear in **LL4-M1** (C2) and in **LL4-M2** (C2 and C15). Chemical shifts of H5/H6 protons of C6/C19 and C7/C20 do not change significantly in any case upon 5-methyl-cytosine substitution.

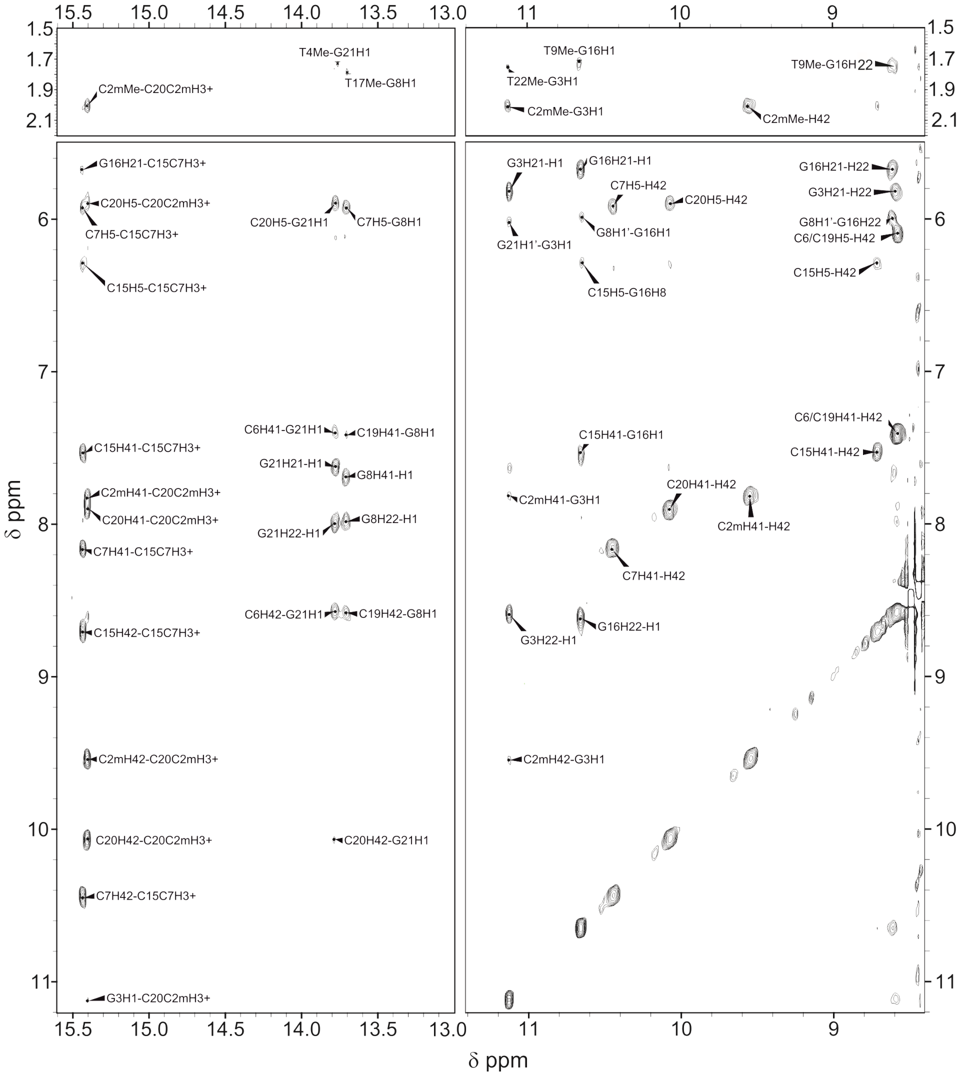


Figure S8. Exchangeable protons regions of NOESY spectrum (150 ms) of **LL4-M1**, phosphate buffer pH 7, T=5°C, [oligonucleotide]=1 mM.

Assignment details of exchangeable protons region of LL4-M1: As shown in Figure S8, **LL4-M1** exhibits two signals in the characteristic region of hemiprotonated C:C$^+$ base pairs (15.44 and 15.41 ppm) that show cross-peaks with two pairs of amino protons, consistent with the formation of C:C$^+$ pairs between non-equivalent cytosines. Two additional imino signals in the G:C WC base pairs region are also observed at 13.77 and 13.71 ppm. Starting from methylated ^{m}C2 residue (labelled as C2m in the spectrum), the other cytosine residues involved in C:C$^+$ base pairs could be unambiguously assigned: ^{m}C2 is hydrogen bonded to C15 and the other C:C$^+$ base pair is formed between C7 and C20. C6 and C19, that show degenerated signals, are involved in G:C WC base pairs formation (C6:G21 and C19:G8) and expected H1G21-H41/H42C6 and H1G8-H41/H42C19 cross-peaks are observed. Formation of mixed tetrads is also supported by the significant differences found in the chemical shift of the amino protons of paired hemiprotonated cytosines. Amino protons of cytosines capped by the G:C base pair of the tetrad are significantly downshifted than those capped by the G:T base pair. Some cross-peaks also support the minor groove interaction between guanine residues: H1'G8-H1/H22G16 and H1'G21/H1G3. Stacking contacts between guanine residues and hemiprotonated C:C$^+$ pairs are also observed: H5C7-H1/H21/H22G8 and H5/H42C21-H1G21.

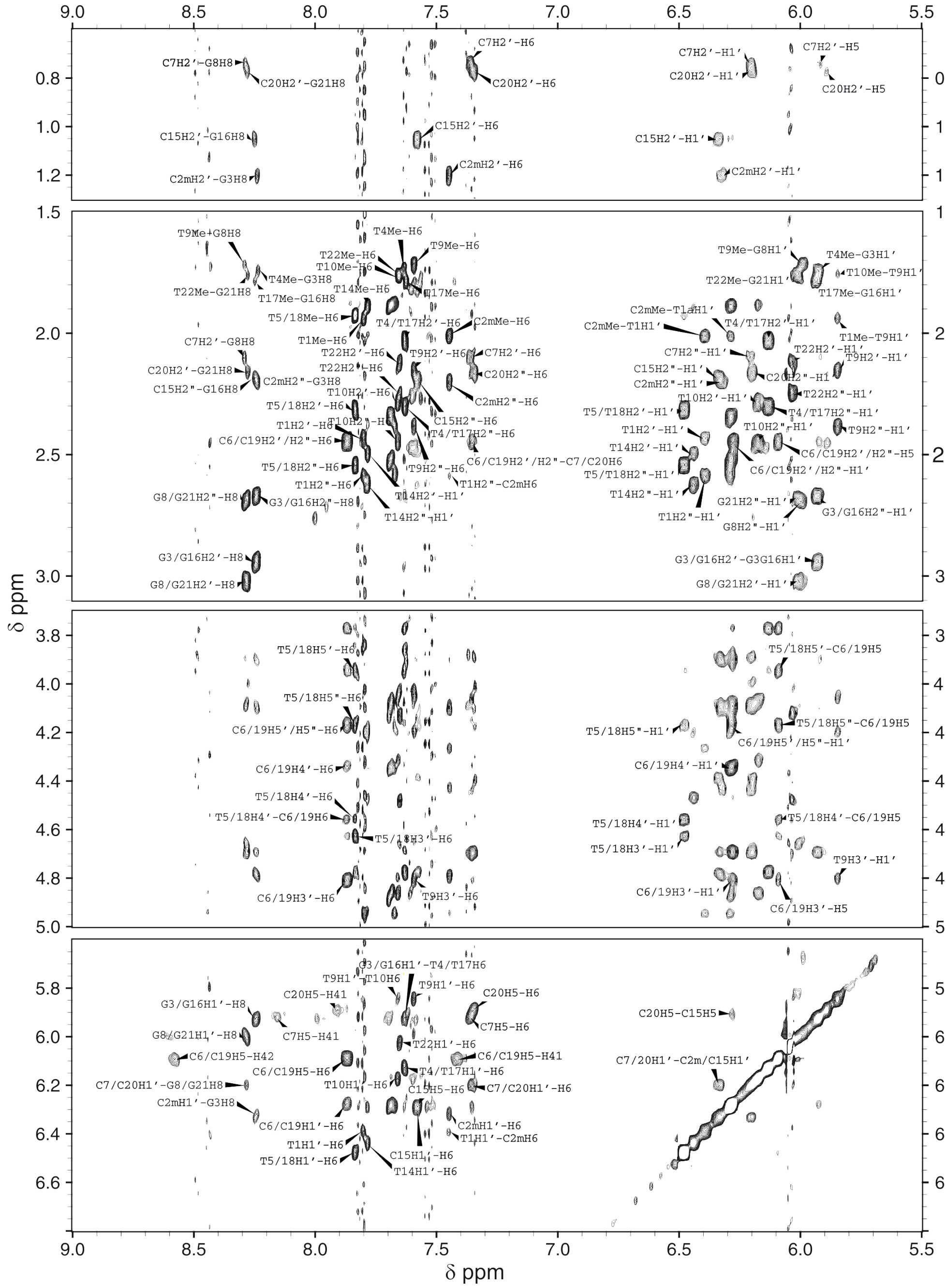


Figure S9. Non-exchangeable protons regions (Ar-H1'/H5 and Ar-H2'/H2"/Me) of NOESY spectrum (150 ms) of **LL4-M1**, phosphate buffer pH 7, T=5°C, [oligonucleotide]=1 mM. Cross-peaks involving multiple overlapped resonances are labelled with a slash symbol.

Assignment details of non-exchangeable protons of LL4-M1: As shown in Figure S9, **LL4-M1** is completely structured at pH 7 and low temperature (T=5°C). A large number of sequential cross-peaks are observed: Me^{m}C2-H1'T1, H2"/H1'T1-H6^{m}C2, H1'/H2'/H2"^{m}C2-H8G3, MeT4-H1'G3, MeT9-H8G8, H2'/H2"C6/C19-H6C7/C20, H2'/H2"C7-H8G8, H1'C7/C20-H8G8/G21, H2'/H2"C15-H8G16, H2'/H2"C20-H8G21, MeT17-H1'G16, MeT22-H1'G21. Residues located in the core of the structure showed characteristic H1'C7/20-H1'^{m}C2/C15 cross-peaks across the minor groove and stacking contacts H5C20-H5C15. T5 and T18 residues are degenerated. These residues are exposed to the solvent, but they could be assigned on the basis of sequential sugar contacts: H4'/H5'/H5"T5/T18-H5/H6C6/C19. Some contacts involving thymine loop residues are observed: H1'T9-H6T10, MeT10-H1'T9 and Me^{m}C2-H1'Tla. (Tla indicates thymine loop residue that could not be sequentially assigned)

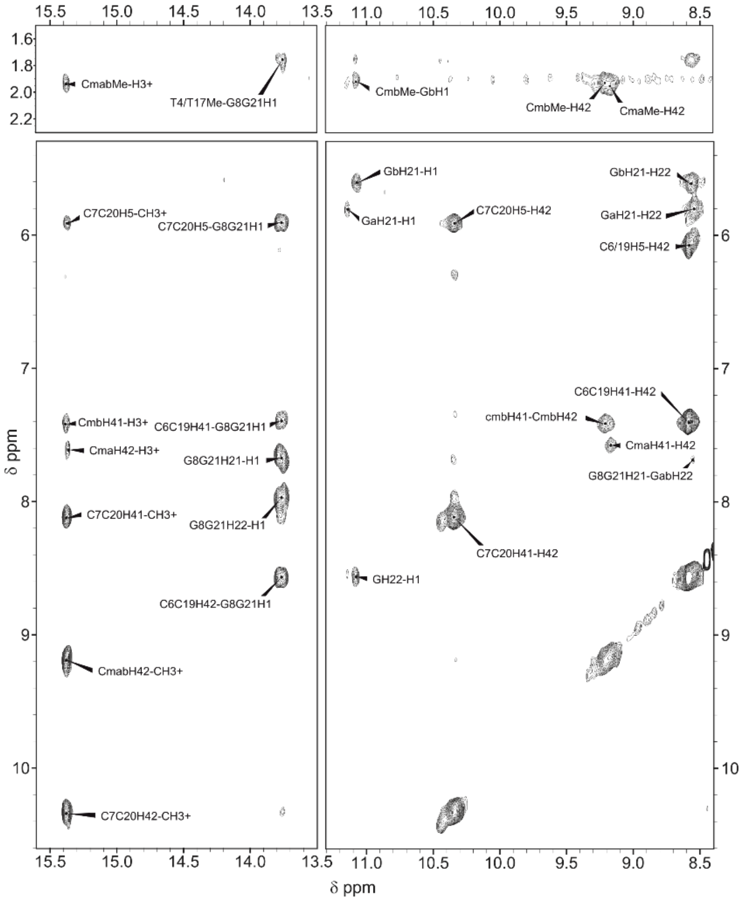


Figure S10. Exchangeable protons regions of NOESY spectrum (150 ms) of **LL4-M2**, phosphate buffer pH 7, T=5°C, [oligonucleotide]=1 mM. Cross-peaks involving multiple overlapped resonances are labelled with a slash symbol.

Assignment details of exchangeable protons region of LL4-M2: As can be observed in this Figure S10, only two H5/H6 cross-peaks are observed. One of them corresponds to the cytosines involved in Watson-Crick base pairs (C6/C19). The other corresponds to C7/C20, involved in C:C$^+$ base pairs. The observed degeneration of signals corresponding to C7/C20 and C6/C19 (exchangeable and non-exchangeable signals) points to a symmetric structure. The two methylated cytidines (labelled as Cm in the spectrum) show degenerated signals for exchangeable protons H41/H42, but not for the non-exchangeable ones. One C:C$^+$ imino signal is observed (15.37 ppm), showing cross-peaks with C7/C20/^{m}C2/^{m}C15 residues. It cannot be unambiguously deduced from this signal the composition of the base pairs (one C:C$^+$ and one ^{m}C:^{m}C$^+$ or two ^{m}C:C$^+$ base pairs), but it seems like ^{m}C2 and ^{m}C15 are not perfectly aligned in one H3$^+$ chemical shift (15.38 and 15.36 ppm). This observation would be only consistent with the formation of ^{m}C:C$^+$ base pairs and, hence a head to tail orientation. At 13.76 ppm, it is observed an imino proton signal that corresponds to two guanine residues (with degenerated exchangeable signals) involved in G:C WC base pairs with C6/C15. This imino signal also shows cross-peaks with amino protons of C7/C20 residues, indicating stacking interacions. At 11.15 and 11.08 ppm, two additional guanine imino signals are observed. These signals exhibit cross-peaks with their own amino protons and with the Me group of methylated cytidines ^{m}C2/^{m}C15.

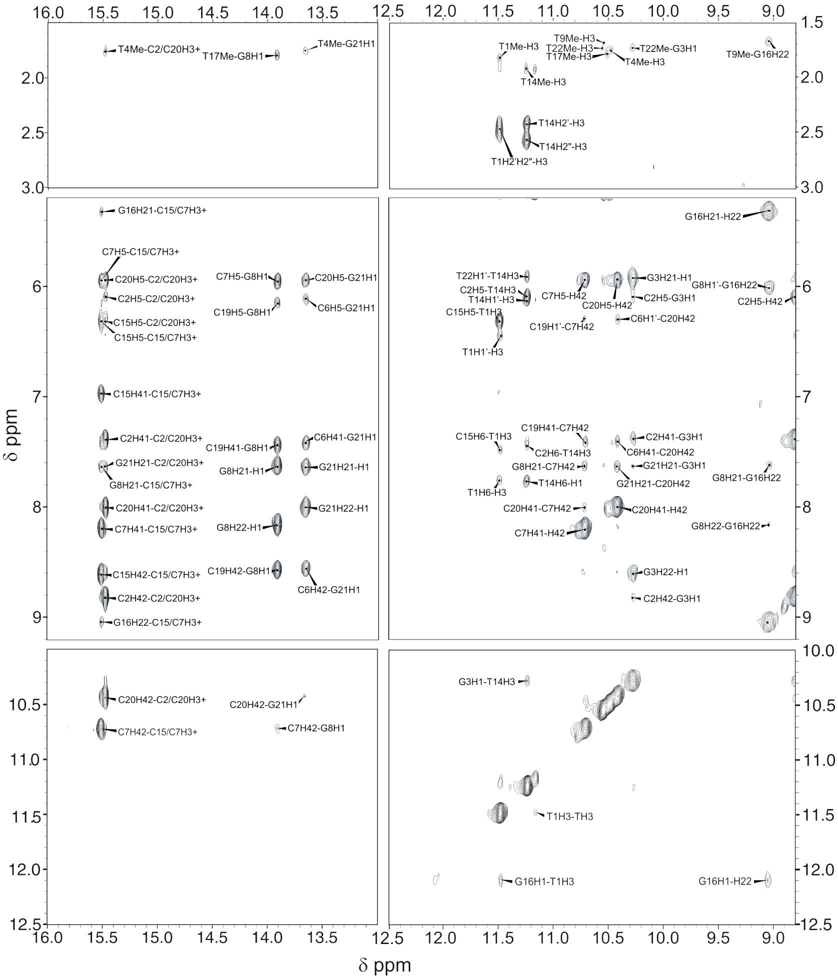


Figure S11. Exchangeable protons regions of NOESY spectrum (150 ms) of **LL4**, phosphate buffer pH 7, T=5°C, [oligonucleotide]=1 mM. Cross-peaks involving multiple overlapped resonances are labelled with a slash symbol.

Assignment details of exchangeable protons region of LL4: As can be observed in Figure S11, **LL4** exhibits two hemiprotonated imino signals at 15.51 and 15.48 ppm. Each of these signals show cross-peaks with two pairs of amino protons. Unequivocal assignment of C2 and C15 (see Figure S7) allowed to determine cytosine residues involved in C:C$^+$ base pairs (C2, C7, C15 and C20). According to amino/imino cross-peaks, C2 and C15 are involved in different C:C$^+$ base pairs, thus confirming that the two protonated imino signals correspond to C2:C20$^+$ and C7:C15$^+$ base pairs. C7 and C20 amino protons are particularly unshielded (10.72 and 10.42 ppm, respectively). This base-pairing pattern is only compatible with a head-to-tail like strand folding (see Figure S1). Two different types of guanine imino protons are observed. Two guanine imino signals are found in the characteristic region of G:C WC base pairs, at 13.91 and 13.66 ppm, and show cross-peaks with amino protons of C6 and C19 (C6: G21 and C19:G8 base pairs). The other two guanine imino signals are found at 12.08 ppm, G16 (exhibiting a broader signal), and at 10.28 ppm, G3. Both signals show cross-peaks with imino protons of two thymine residues: T1 (11.49 ppm) and T14 (11.24), indicating the formation of G:T base pairs (G3:T14 and G16:T1 base pairs). The broad signal observed for G16 suggests that this guanine might be more exposed to the solvent than G3. Formation of G:C:G:T tetrads is supported by a number of cross-peaks: H1'G8-H22G16, H21/H22G8-H22G16 and the significant differences found in the chemical shift of amino protons of stacked hemiprotonated cytosines, depending on whether they are capped by the G:C or the G:T base pair of the tetrad. Thymine residues T4, T9, T17 and T22 are stacked over the tetrads and methyl groups of these thymines show cross-peaks with exchangeable and non-exchangeable protons of guanine residues: MeT22-H1/H22G3, MeT9-H22G16, MeT17-H1G8, MeT4-H1G21, MeT4-H8/H1'G3, MeT9-H8/H1'G8, MeT17-H8/H1'G16 and MeT22-H8/H1'G21. Moreover, imino signals corresponding to stacked thymine residues are also observed in the 10-12 ppm region: T4 (10.48 ppm), T17 (10.51 ppm), T9 (10.54 ppm) and T22 (10.55 ppm), indicating that these residues are solvent protected. Characteristic i-motif cross-peaks between cytidine residues across the minor groove are also observed: H1'C20-H1'C15 and H1'C6/C19-H5C20/C7.

Sequential assignment based on H6/H8-H2’/H2”/H1’ cross-peaks could be completed for C2-G3-T4, C6-C7-G8-T9, C15-G16-T17 and C19-C20-G21-T22 fragments. Thymine residues T1 and T14 can be identified by their own intraresidual Me/H2’/H2”-H3 cross-peaks. These residues show some non-sequential cross-peaks across the major groove (H2”T1-H6C15 and H2”T13-H6C2). Stacked thymine residues (T4, T9, T17 and T22) also show very weak intraresidual Me-H3 cross-peaks.

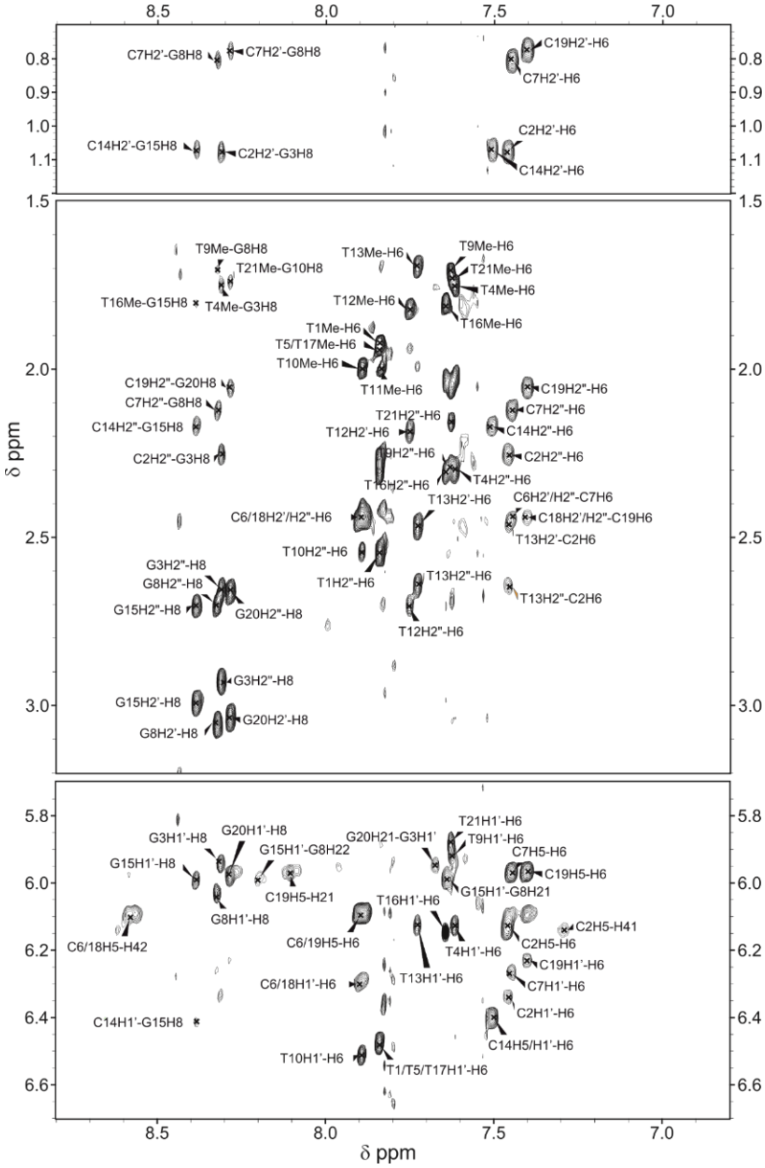


Figure S12. Non-exchangeable protons regions (Ar-H1’/H5 and Ar-H2’/H2”/Me) of NOESY spectrum (150 ms) of **LL3**, phosphate buffer pH 7, T=5°C, [oligonucleotide]=1 mM. Cross-peaks involving multiple overlapped resonances are labelled with a slash symbol.

Assignment details of non-exchangeable protons of LL3: Sequential assignment based on H6/H8-H2’/H2”/H1’ cross-peaks could be completed for C2-G3-T4, C6-C7-G8-T9, C14-G15-T16 and C18-C19-G20-T21 fragments. As shown in Figure S12, thymine residues T1 and T13 can be identified by own Me/H2’/H2”-H3 cross-peaks. These residues show some non-sequential cross-peaks across the major groove (H2”T1-H6C14 and H2”T13-H6C2). Characteristic i-motif cross-peaks between cytidine residues through the minor groove are also observed: H1’C19-H1’C14 and H1’C6-H5C19. In addition, some thymine residues in the loop could be assigned on the basis of Me-H1’ and Me-H6 cross-peaks. All NMR data is consistent with the formation of an unimolecular head-to-tail i-motif structure containing G:C:G:T minor groove tetrads.

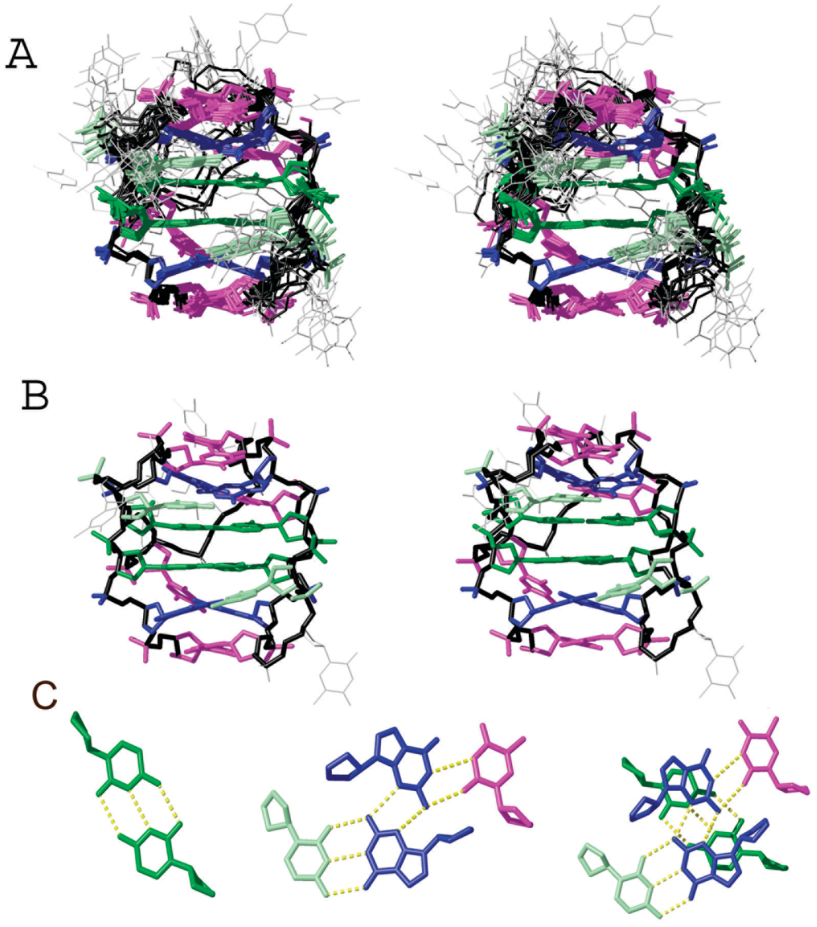


Figure S13. A) Stereoviews of the ensemble of ten structures of **LL3**. B) Stereoviews of the average structure. C) C:C$^+$ base pair, G:C:G:T minor groove tetrad and stacking of thymine residues on the tetrad. Colour code: Cytosines in green, guanines in blue, and well-defined thymines in magenta. Cytosines involved in CG base pairs are shown in light green. Thymines with non well-defined structures are shown in grey. Backbone is shown in black.

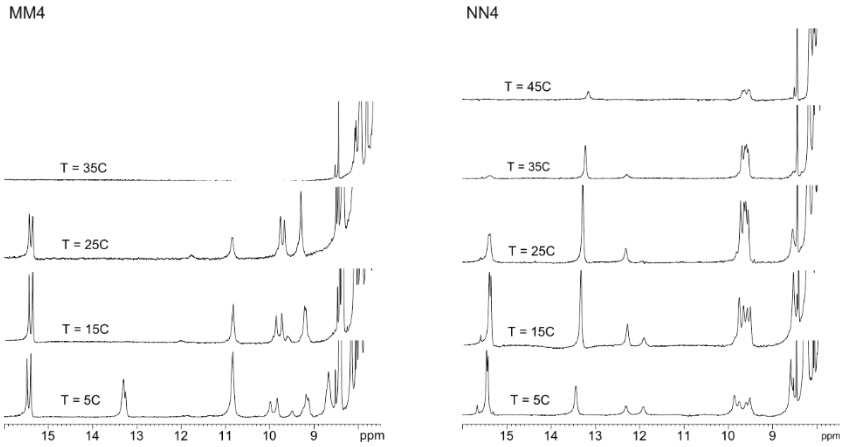


Figure S14. NMR melting experiments at pH 7 (phosphate buffer, [oligonucleotide]=1 mM) of **MM4** and **NN4**.

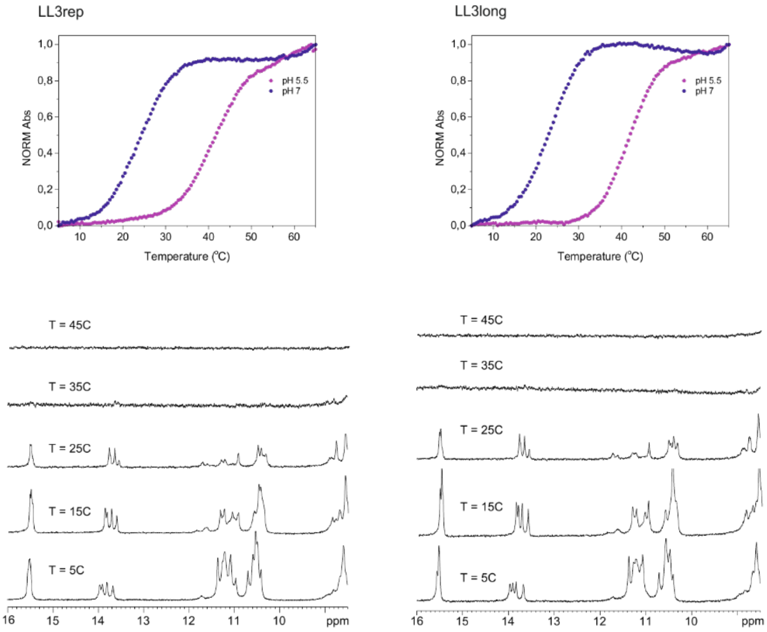


Figure S15. UV melting curves [oligonucleotide]= 3,5 μM, 25 mM cacodylate buffer at different pH and NMR melting experiments at pH 7 (phosphate buffer, [oligonucleotide]=1 mM) of **LL3rep** (left) and **LL3long** (right).

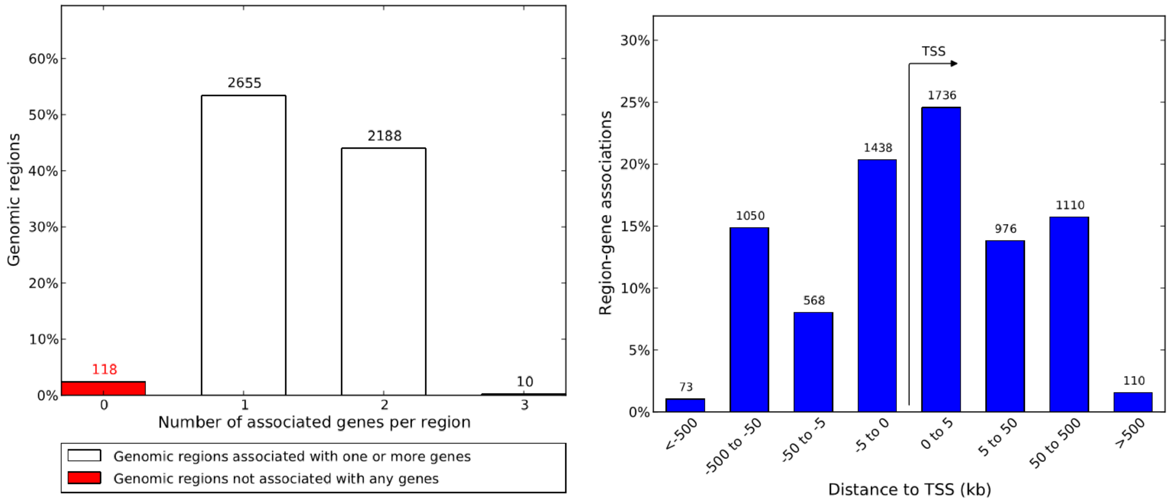


Figure S16. Annotation of regions with the consensus motif using Genomic Regions Enrichment Annotations (GREAT) Tool. **Left:** Number of hits associated to 0, 1, 2 or 3 genes, taking minimum distance upstream (5kb) and downstream (1kb) of the TSS. The number of hits not associated to any gene (118) is more than five times lower than the expected number indicating the concentration of i-motifs close to genes. **Right:** The bar chart shows the distribution of distances between hits and TSS of their associated genes. The hyper-concentration of i-motif-forming regions close to the origin of genes is very clear.

Supplementary Tables

Table S1. Dose-response fitting values of CD-monitored pH titration for **LL3**, **LL4**, **LL3long** and **LL3rep**.

Sequence	ε_{folded}		$\varepsilon_{unfolded}$		pH$_T$		Cooperativity parameter (Hill slope)	
	Value	Stand. Error	Value	Stand. Error	Value	Stand. Error	Value	Stand. Error
LL3	0.96387	0.00525	0.00622	0.00576	7.76202	0.00623	-2.84909	0.11266
LL4	0.92324	0.00442	0.00449	0.00495	7.94836	0.00697	-1.93829	0.05096
LL3long	0,97752	0,008	4,243E-4	0,00716	7,48845	0,0093	-3,17191	0,1794
LL3rep	0,96105	0,01154	0,02011	0,0112	7,55787	0,01436	-2,62543	0,19116

Table S2. Chemical shifts of **LL4-M1**, pH 7, T=5°C.

Residue	H1/H3^{+}	H42/H22	H41/H21	H6/H8	H5/Me	H1'	H2'	H2"	H3'	H4'	H5'/H5"
T1	n. o.	-	-	7.80	1.94	6.39	2.43	2.59			
^{m}C2	15.40	9.55	7.82	7.45	2.01	6.32	1.20	2.20			
G3	11.13	8.59	5.82	8.24	-	5.92	2.94	2.67			
T4	n. o.	-	-	7.63	1.74	6.13	2.03	2.30			
T5/T18	n. o.	-	-	7.84	1.93	6.48	2.32	2.54	4.63	4.56	4.17, 3.95
C6	-	8.57	7.40	7.87	6.09	6.28	2.45		4.81	4.34	4.17
C7	15.44	10.45	8.16	7.36	5.92	6.20	0.73	2.09			
G8	13.71	7.99	7.70	8.29	-	5.99	3.02	2.69			
T9	n. o.	-	-	7.59	1.72	5.85	2.15	2.38			
T10	n. o.	-	-	7.66	1.76	6.17	2.29	2.45			
T14	n. o.	-	-	7.79	1.89	6.85	2.49	2.63			
C15	15.44	8.71	7.53	7.58	6.29	6.33	1.05	2.19			
G16	10.66	8.62	5.67	8.51	-	5.94	2.94	2.66			
T17	n. o.	-	-	7.63	1.79	6.13	2.03	2.30			
C19	-	8.58	7.41	7.87	6.09	6.28	2.45		4.81	4.34	4.17
C20	15.40	10.07	7.91	7.35	5.89	6.20	1.37	2.17			
G21	13.78	8.00	6.62	8.28	-	6.01	3.02	2.68			
T22	n. o.	-	-	7.65	1.76	6.03	2.12	2.25			

no: not observed, na: not assigned.

T11, T12 and T13 cannot be unambiguously assigned.

Table S3. Chemical shifts of **LL4-M2**, pH 7, T=5°C.

Residue	H1/H3^{+}	H42/H22	H41/H21	H6/H8	H5/Me	H1'	H2'	H2"
^{m}C2/^{m}C15	15.39/15.37	9.20	7.62/7.42	7.45/7.41	1.91/1.96	6.31	1.10/1.17	2.16/2.22
G3/G16	11.15/11.08	8.56	5.81/5.59	8.19/8.23	-	5.91	2.94	2.67
T4/T9/T17/T22	no	-	-	7.64	1.76	na	na	na
C6/C19	-	8.58	7.39	7.86	6.08	6.28	2.44	
C7/C20	15.39/15.37	10.35	8.11	7.35	5.91	6.20	0.73	2.15
G8/G21	13.76	7.98	7.66	8.29	-	6.02	3.04	2.71

no: not observed, na: not assigned.

Table S4. Chemical shifts of **LL3**, pH 7, T=5°C.

Residue	H1/H3$^+$	H42/H22	H41/H21	H6/H8	H5/Me	H1’	H2’	H2”	H3’	H4’
T1	11.40	-	-	7.84	1.93	6.47	2.35	2.53		
C2	15.48	8.62	7.28	7.44	6.11	6.32	1.06	2.26		
G3	10.26	8.60	5.84	8.31	-	5.94	2.93	2.66		
T4	no	-	-	7.62	1.75	6.13	na	na		
T5/T17	no	-	-	7.84	1.94	6.13	na	na		
C6	-	8.59	7.46	7.93	6.13	6.30	2.44			
C7	15.55	10.66	8.19	7.45	5.95	6.27	0.78	2.09		
G8	13.94	8.20	7.63	8.32	-	6.03	3.06	2.69		
T9	no	-	-	7.63	1.69	5.93	na	na		
T10	no	-	-	7.76	na	5.54	na	na		
T11	no	-	-	na	1.99	na	na	na		
T12	no	-	-	7.84	1.99	na	na	na		
T13	11.20	-	-	7.74	1.71	6.12	2.46	2.63		
C14	15.55	8.66	6.97	7.49	6.39	6.41	1.07	2.17		
G15	no	9.07	5.30	8.39	-	5.99	2.99	2.70		
T16	no	-	-	7.64	1.81	6.08	na	na		
C18	-	8.55	7.43	7.93	6.13	6.30	2.44			
C19	15.48	10.62	8.05	7.41	5.95	6.23	0.76	2.02	4.70	4.50
G20	13.65	7.96	7.67	8.27	-	5.98	3.04	2.66		
T21	no	-	-	7.63	1.73	5.96	na	na		

no: not observed, na: not assigned.

Table S5. Chemical shifts of **LL4**, pH 7, T=5°C.

Residue	H1/H3$^+$	H42/H22	H41/H21	H6/H8	H5/Me	H1’	H2’	H2”	H3’	H4’
T1	11.49	-	-	7.76	1.83	6.45	na	na		
C2	15.47	8.82	7.38	7.46	6.11	6.35	1.07	2.27	4.80	
G3	10.28	8.60	5.93	8.32	-	5.95	2.94	2.66	5.08	4.69
T4	10.48	-	-	7.63	1.77	na	na	na		
T5/T18	no	-	-	7.85	1.93	6.48	2.29	2.55		
C6	-	8.57	7.41	7.92	6.13	6.30	2.44		4.67	4.21
C7	15.51	10.72	8.20	7.43	5.95	6.28	0.79	2.12	4.71	
G8	13.91	8.16	7.63	8.33	-	6.02	3.06	2.70	5.15	4.61
T9	10.54	-	-	7.64	1.69	5.94	2.43	2.52		
T14	11.24	-	-	7.78	1.92	6.13	2.44	2.58		
C15	15.51	8.60	6.96	7.49	6.33	6.41	1.10	2.20	4.81	
G16	12.09	9.04	5.32	8.30	-	5.99	2.98	2.70	5.08	4.69
T17	10.51	-	-	7.64	1.80	na	na	na		
C19	-	8.57	7.43	7.92	6.15	6.29	2.44		4.67	4.21
C20	15.47	10.42	8.00	7.41	5.94	6.23	0.79	2.05	4.71	
G21	13.66	8.00	7.64	8.28	-	5.99	3.03	2.66	5.14	4.67
T22	10.55	-	-	7.63	1.75	5.92	2.03	2.15	4.29	

no: not observed, na: not assigned.

T10, T11, T12 and T13 cannot be unambiguously assigned.

Table S6. Experimental constraints and calculation statistics of LL3.

Experimental distance constraints		
Total number	125	
intra-residue	47	
sequential	41	
range > 1	37	
RMSD (Å)		
all well-defined* bases	0.5±0.1	
all well-defined* heavy atoms	0.8±0.2	
backbone	1.7±0.4	
all heavy atoms	2.8±0.7	
Residual violations	Average	Range
Sum of violation (Å)	1.23	0.84 … 1.69
Max. violation (Å)	0.22	0.28 … 0.16
NOE energy# (kcal/mol)	5.59	4.65 … 6.60
Total energy (kcal/mol)	-1517	-1717…-1334

* All except thymines 5,10,11,12,17

K_{NOE} = 20 kcal/(mol·Å^2)

Table S7. Average dihedral angles and order parameters of the dimeric structure of LL3.

Residue	Pseudorot.		α		β		γ		δ		ε		ζ		X	
	Phase	Ampli.	Averag.	OP	Averag.	OP	Averag.	OP	Averag.	OP	Averag.	OP	Averag.	OP	Averag.	OP
T1	64	31	-	-	-	-	4	1.0	6	1.0	2	1.0	-	-	5	1.0
C2	137	40	48	0.8	4	1.0	44	0.8	9	1.0	6	1.0	2	1.0	6	1.0
G3	166	36	6	1.0	4	1.0	3	1.0	2	1.0	17	1.0	5	1.0	6	1.0
T4	166	36	11	1.0	9	1.0	5	1.0	1	1.0	27	0.9	13	1.0	5	1.0
T5	153	39	84	0.3	23	0.9	55	0.7	9	1.0	28	0.9	16	1.0	12	1.0
C6	73	29	88	0.3	9	1.0	54	0.7	20	0.9	47	0.7	85	0.4	10	1.0
C7	153	41	44	0.8	21	0.9	47	0.8	4	1.0	3	1.0	48	0.7	5	1.0
G8	150	35	2	1.0	1	1.0	1	1.0	3	1.0	7	1.0	3	1.0	3	1.0
T9	166	33	6	1.0	7	1.0	3	1.0	4	1.0	40	0.8	7	1.0	5	1.0
T10	142	41	72	0.5	49	0.7	65	0.6	12	1.0	40	0.8	66	0.6	82	0.6
T11	147	41	95	0.0	14	1.0	93	0.2	10	1.0	49	0.7	83	0.5	92	0.2
T12	133	39	68	0.6	21	0.9	38	0.9	14	1.0	34	0.9	75	0.6	8	1.0
T13	156	30	73	0.6	12	1.0	33	0.9	8	1.0	5	1.0	47	0.7	3	1.0
C14	169	31	8	1.0	2	1.0	4	1.0	4	1.0	4	1.0	17	1.0	5	1.0
G15	164	36	4	1.0	1	1.0	4	1.0	3	1.0	14	1.0	2	1.0	1	1.0
T16	165	36	6	1.0	7	1.0	3	1.0	2	1.0	28	0.9	10	1.0	1	1.0
T17	157	40	65	0.6	6	1.0	97	0.1	7	1.0	40	0.8	7	1.0	21	0.9
C18	117	33	66	0.6	20	1.0	71	0.6	11	1.0	33	0.9	11	1.0	11	1.0
C19	133	41	53	0.7	17	1.0	52	0.7	20	0.9	44	0.8	34	0.9	10	1.0
G20	165	31	24	0.9	4	1.0	4	1.0	5	1.0	30	0.9	50	0.8	5	1.0
T21	164	34	15	1.0	8	1.0	2	1.0	2	1.0	-	-	23	0.9	3	1.0

Table S8. Genomic annotation of hits. Enrichment of hits for each category is computed as the ratio between the normalized number of hits (number of hits in a category divided by the total number of regions) and the normalized number of base-pairs per region type (percentage of base-pairs of the genome per category). Genomic sequences shorter than 1 Mbp were excluded to reduce noise. Note again the over-representation of i-motif-forming sequences in regulatory regions (promoter and 5UTR) and the under-representation in intron or intergenic regions.

Annotation	Number of hits	Total size (bp)	Log2 Enrichment
3UTR	28	23005177	-0.384
ncRNA	17	6409543	0.74
pseudo	11	2010139	1.785
Exon	358	36768443	2.616
Intron	1239	1240886382	-0.669
Intergenic	841	1749354789	-1.724
Promoter	2129	35414190	5.243
5UTR	264	2760547	5.912

Table S9. GO Enrichment analysis. Functional annotation terms significantly enriched at hits of the consensus motif as discovered by GREAT. For each term we report the p-value corrected for multiple testing (FDR p-value), the ratio of observed/expected regions with the annotation (Fold Enrichment) and the number of regions mapped to genes associated to the GO term.

Term Name	FDR p-value	Fold Enrichment	Observed Region Hits
regulation of metanephric glomerulus development	7.94876e-23	10.4771	38
positive regulation of kidney development	2.36364e-6	2.8127	39
enteric nervous system development	2.77212e-6	2.5710	45
Rac protein signal transduction	8.82346e-5	2.8911	28
phospholipid translocation	1.16460e-4	2.6258	32
intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress	2.04422e-3	2.7949	21
paraxial mesoderm morphogenesis	2.65032e-3	2.8223	20
corpus callosum development	4.54575e-3	2.8760	18
paraxial mesoderm development	2.48624e-2	2.1246	24
negative regulation of protein kinase activity by regulation of protein phosphorylation	3.17901e-2	4.7062	7

Table S10. Synthesis results for LL3 and LL4.

Sequence	ε ($mL\cdot\mu mol^{-1}\cdot cm^{-1}$)	Crude Yield %	HPLC Purity %	Purified Yield %	Theoretical Mass	MS-MALDI-TOF m/z
LL3	178.1	86	78	37	6336.1	6333.9
LL4	186.2	84	69	26	6640.1	6639.4

References

(1) Plateau, P.; Guéron, M. *J. Am. Chem. Soc.* **1982**, *104* (25), 7310.

(2) Kumar, A.; Ernst, R. R.; Wüthrich, K. *Biochem. Biophys. Res. Commun.* **1980**, *95* (1), 1.

(3) Bax, A.; Davis, D. G. *J. Magn. Reson.* **1985**, *65* (2), 355.

(4) Piotto, M.; Saudek, V.; Sklenář, V. *J. Biomol. NMR* **1992**, *2* (6), 661.

(5) Cai, L.; Chen, L.; Raghavan, S.; Rich, A.; Ratliff, R.; Moyzis, R. *Nucleic Acids Res.* **1998**, *26* (20), 4696.

(6) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273* (1), 283.

(7) D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Götz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman **2012**, AMBER 12, University of California, San Francisco.

(8) Soliva, R.; Monaco, V.; Gómez-Pinto, I.; Meeuwenoord, N. J.; Marel, G. a; Boom, J. H.; González, C.; Orozco, M. *Nucleic Acids Res.* **2001**, *29* (14), 2973.

(9) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. *Nat. Methods* **2015**, *13* (1), 55.

(10) Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graph.* **1996**, *14* (1), 51.