How to cite: *Angew. Chem. Int. Ed.* **2023**, *62*, e202309327
doi.org/10.1002/anie.202309327

DNA Structures

A Screening Protocol for Exploring Loop Length Requirements for the Formation of a Three Cytosine-Cytosine$^+$ Base-Paired i-Motif

Michele Ghezzo, Marko Trajkovski, Janez Plavec, and Claudia Sissi**

Abstract: DNA sequences containing at least four runs of repetitive cytosines can fold into tetra-helical structures called i-Motifs (iMs). The interest in these DNA secondary structures is increasing due to their therapeutical and technological applications. Still, limited knowledge of their folding requirements is currently available. We developed a novel step-by-step pipeline for the systematic screening of putative iM-forming model sequences. Focusing on structures comprising only three cytosine-cytosine$^+$ base pairs, we investigated what the minimal lengths of the loops required for formation of an intra-molecular iM are. Our data indicate that two and three nucleotides are required to connect the strands through the minor and majorgrooves of the iM, respectively. Additionally, they highlight an asymmetric behavior according to the distribution of the cytosines. Specifically, no sequence containing a single cytosine in the first and third run was able to fold into intra-molecular iMs with the same stability of those formed when the first and the third run comprise two cytosines. This knowledge represents a step forward toward the development of prediction tools for the proper identification of biologically functional iMs, as well as for the rational design of these secondary structures as technological devices.

Introduction

i-Motifs (iMs) are tetra-helical nucleic acid structures in which two parallel cytosine-cytosine$^+$ (CC$^+$) base-paired duplexes intercalate into each other with opposite polarity.[1] While initially explored as tetra-molecular assemblies,[2] bi-molecular and intra-molecular iMs have been reported as well. To fold into intra-molecular iMs, sequences must contain at least four runs of repetitive cytosines. These sequence requirements prompted the interest in iMs, since a non-random distribution of such C-rich domains through the human genome was envisaged, with a significant enrichment at telomeres and gene promoters.[3–5] Moreover, iM occurrence within the promoter of oncogenes was related to the regulation of oncogene transcription in vitro.[6] It was tough to extend the same function in vivo. Indeed, a slightly acidic pH, seemingly incompatible with the *in vivo* environment, is needed for iM folding. This is due to the requirement of cytosine hemi-protonation for the formation of the planar CC$^+$ pairing, a condition maximized at pH close to 4.5, corresponding to the cytosine pKa. Nevertheless, iMs were directly detected *in cell* with the use of different techniques, including in-cell NMR and imaging of cell nuclei with an antibody fragment able to selectively recognize iMs.[7,8] More recently, the same antibody has been used to map the genomic distribution of iMs in living human and rice cells through chromatin immunoprecipitation (ChIP)-sequencing studies.[9,10] This reinforced the current interest in iMs from a biological and pharmaceutical point of view.

In addition, iMs represent promising building blocks in nanotechnology. In particular, the reversible formation of iM as a function of pH is widely explored for bio-sensing, drug delivery and logic device development.[11–13]

Within an iM, the CC$^+$ intercalation frame may shift, thereby leading to structural polymorphism. This behavior is most conveniently explained in iMs with an even number of CC$^+$ pairs where the two topologies are classified as 3′E and 5′E when the outmost CC$^+$ pair is in the 3′ and the 5′ end, respectively.[14,15] Conversely, for iMs with an odd number of CC$^+$, only one topology grants the maximal number of base-pairings, thus representing the most favorable one.[16] Additionally, as a result of the geometry of the CC$^+$ pairing, the backbones of the four paired strands in the core of the structure are not equidistant thereby defining two major and two minor grooves. Consequently, starting from the 5′ end of an intra-molecular iM, the first, second and third loops connect the four strands through a major-minor-major or a minor-major-minor grooves combination pattern (Figure 1).

All cytosines in an iM core are confined to adopt *anti*-glycosidic conformation. This imposes the CC$^+$ pairing to form between parallel strands and limits the major-minor-major and the minor-major-minor topologies to be anti-clockwise and clockwise, respectively.

Importantly, the minor grooves in iMs are extremely narrow, and this is a unique feature among the currently

[*] Dr. M. Ghezzo, Prof. C. Sissi
Department of Pharmaceutical and Pharmacological Science,
University of Padua
Via Marzolo 5, 35131 Padua (Italy)
E-mail: claudia.sissi@unipd.it

Dr. M. Trajkovski, Prof. J. Plavec
Slovenian NMR Centre, National Institute of Chemistry
Hajdrihova 19, 1000 Ljubljana (Slovenia)
E-mail: janez.plavec@ki.si

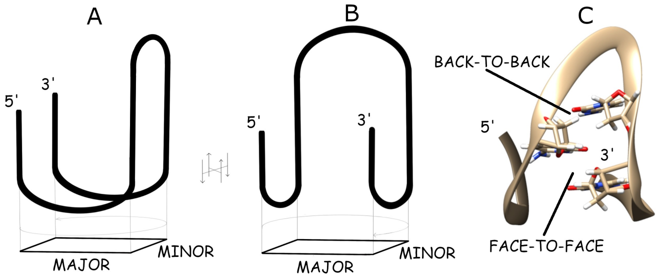

Figure 1. Two topologies for intra-molecular iM in terms of loop connection path: the major-minor-major (Panel A) and the minor-major-minor (Panel B). View of the face-to-face and back-to-back sugar orientation throughout the minor groove (PDB ID 8PMB) (Panel C).

characterized DNA secondary structures. Along the minor grooves, inter-strand sugar moieties are alternatively face-to-face (ff) and back-to-back (bb) oriented. In the ff orientation, the inter H4′–H4′ distances are about 3.5 Å, while in the bb orientation the H2′′–H2′′ distances can be shorter than 3.0 Å. Moreover, unique sugar-sugar interactions occur within the minor grooves; in particular, a systematic inter-strand H1′–O4′ hydrogen bonding network highlighted by computational studies, contributes to iM stability.[17]

The formation of iM structures relies on a balance between the unfavorable entropic change (ΔS) and the favorable reduction in enthalpy (ΔH) associated with the folding. Many factors contribute to the ΔH, but theoretical calculations indicate that the three hydrogen bonds of the CC^+ base pairs and the H1′–O4′ hydrogen bonding network through the minor grooves are the most relevant. Indeed, in contrast to B-DNA, π–π stacking interactions between consecutive CC^+ base pairs do not significantly contribute to iM stability.[17] Globally, each CC^+ base pair corresponds to about 10–12 $\mathrm{kcal\,mol^{-1}}$ to the standard enthalpy change of folding (ΔH^{Θ}), and this value can be applied to predict the iM core length.[18,19]

To date, few high-resolution structures of DNA sequences folded into intra-molecular iMs are available (PDB codes: 7O5E, 1A83, 8BQY, 8BV6, 1G22, 5OGA, 1ELN and 1EL2, see Figure S1 and Table S1).[14,20–24] All of them were solved by NMR. Recently, the iM from a sequence of the *HRAS* oncogene promoter has been resolved by X-ray. However, although in solution this sequence folds into monomeric iM, the crystallization conditions shifted the process toward a dimeric arrangement.[25] Moreover, there are no algorithms specifically designed to predict iM folding from the primary nucleotide sequence. As a result, the currently used tools to screen the genome for iM-forming sites are those designed for the search of G-quadruplexes (G4 s), the more extensively studied tetra-helical nucleic acid structures that form at the complementary G-rich strand. However, there is no evidence that the requirements for G4 and iM formation fully overlap. For instance, while G4 may accommodate loops with 1 or even no nucleotides,[26–29] the minimum loop length sustaining an intra-molecular iM folding is still a matter of discussion. Systematic studies of designed iM models have been reported to rationalize the energetic contributions of the different structural components (i.e. each CC^+ and each nucleotide in the loops) to the iMs stability.[30–33] These data can help to uncover unique folding features and to implement iM-specific folding prediction algorithms. However, multiple variables such as the C-runs length, the type and the number of nucleotides within the loops can be potentially interconnected thus giving rise to an unmanageable number of different combinations.[34]

There are two complementary approaches to solving this issue. One covers the screening of a huge number of model sequences, and it has the great advantage of providing a general overview of the folding requirements.[30–33] The alternative strategy relies on the study of fewer model sequences, with the advantage of providing a description at molecular level of the system under investigation.

Here we applied this second working model. Specifically, we developed an interactive protocol for the rational selection of a limited number of sequences to be studied by spectroscopic and calorimetric analysis. We applied it to identify the minimal length of the loops in sequences forming an iM core limited to three CC^+ base pairs. The model was finally validated by NMR spectroscopy and the structure of the selected iM model was resolved. These results allowed us to better address the yet poorly explored structural contribution of the loops in iM.

Results and Discussion

Design of cytosine-rich models and interactive workflow for the selection of i-Motif forming sequences

Distinctly from other screening approaches, we considered that the identification and characterization of a minimal iM building block comprising both base pairing and loops might provide a more comprehensive model to rationally define the sequence requirements for iM formation. In this view, here we focused on sequences that may adopt an intra-molecular iM comprising 3 CC^+ base pairs at the most. The stability of such models is expected to be low thus making easier the identification of the sequences with pronounced propensity to adopt iM structures. Moreover, these model sequences cannot allow the relative sliding of the two intercalated duplexes thus avoiding the coexistence of different folding topologies, a useful condition for a fine characterization of the iM.

Such sequences can be clustered according to two different subgroups: 5′-CC–N_x–C–N_y–CC–N_z–C-3′ (C21 subgroup) and 5′-C–N_x–CC–N_y–C–N_z–CC-3′ (C12 subgroup), where C is cytosine, N is any nucleobase, while x, y and z are the numbers of nucleotides within the first, second and third loop from the 5′ end, respectively.

With the aim of identifying the minimum length of the loops compatible with the iM formation, we considered $1 \leq x, y, z \leq 4$. These constraints were selected because it was reported that three nucleotides per loop are compatible with both minor and major grooves.[14,24] When considering N as any among G, C, T or A, a total number of 39304000 different combinations is obtained. To reduce this number, we restrained N to T. This choice was crucial: C was not considered in the loops to keep a fine control of the number of forming CC^+ base pairs and G was discarded to avoid any base pair competition with the cytosines potentially involved in the iM core. On the other hand, between T and A, we chose T that can form additional T-T base pairs that stack on the outmost CC^+ base pair. This interaction does not compete with CC^+ base pairing and it is quite common in iMs.[2,20,35] Consistently, several previous works on iM models accommodated T in loops as well.[16,18,32,33] By restraining N to T, the number of combinations decreases to 128, i.e. 64 for each of the C21 and the C12 subgroups.

Here, we screened the C21 and C12 subgroups separately. For both of them, we started by filling a list with all the 64 possible sequences (*P_list*). At each step, we experimentally checked the folding of one sequence from the *P_list* with its defined combination of x, y and z. Since we were looking for the minimal length of the loops, whenever the selected sequence was proved to fold into a 3 CC^+ base-paired intra-molecular iM, we removed from *P_list* all the sequences having simultaneously all the three loops longer (i.e. the first loop $\geq$ x, second loop $\geq$ y and third loop $\geq$ z). Conversely, when the tested sequence did not fold, we removed from *P_list* all those having simultaneously the three loops shorter (i.e. the first loop $\leq$ x, second loop $\leq$ y and third loop $\leq$ z). Therefore, the number of sequences in the *P_list* progressively decreased at each step. The screening ended when *P_list* was empty.

To reach this goal within the most limited number of steps (thus, by studying the minimal number of sequences), we did not change any single loop hand-by-hand. Conversely, we developed an algorithm that at each step selected from the *P_list* the sequence that would allow for the removal of the maximal number of components as a result of the input “yes” (it folds) or “no” (it does not fold) as derived from the experimental data (Supporting Information). This was the faster protocol to reach the emptiness of *P_list* and the identification of the minimal folding sequences. Moreover, we designed the algorithm to allow the user to define the first sequence to check at the beginning of the screening (step 0). We converted it into a runnable Python code to make it unambiguously interpretable. The code can be used for other C-run systems, and it is free to download from GitHub (https://github.com/micheleghezzo/I-motif_loop_minimizer).

Since reported iM structures showed that three nucleotides per loop are compatible with both minor and major grooves,[14,20] for both the C21 and the C12 subgroups, we decided to start with the sequences containing three T in all loops ($x=3$, $y=3$ and $z=3$).

Experimentally, at each step, the folding of the selected sequence into a 3 CC^+ base-paired intra-molecular iM was assessed by nuclear magnetic resonance spectroscopy (NMR), circular dichroism spectroscopy (CD) and differential scanning calorimetry (DSC) experiments.

The formation of CC^+ base pairs was monitored by recording ^{1}H NMR spectra, whereby assuming that their formation is coupled to a reduction of the exchange rate of the corresponding imino proton, i.e. H3, with the water solvent, thus resulting in a detectable signal in the range between δ 15 and 16 ppm. Whether the presence of such peaks safely supported iM formation, on the contrary, from the ^{1}H NMR spectra we could not derive the number of strands involved in the pairing. Thus, to complement NMR data, we assessed the formation of iM structures by acquiring the chiroptical fingerprint. Indeed, CD spectroscopy allowed us to record data on DNA samples in a wider concentration range, enabling us to distinguish the folding into inter-molecular (concentration-dependent) from an intra-molecular (concentration-independent) iM.

Finally, by DSC, we directly measured the folding enthalpy of intra-molecular iM models to derive the number of CC^+ in the iM core. Indeed, as above described, the two parameters are linearly related.[18,19] The experimentally studied sequences are reported in Table S2.

Screening of the C21 subgroup

As anticipated, to screen the C21 subgroup, we rationally selected to start from the C21T333 sequence since this length of all loops has been already reported to be compatible with iM formation.[14,24] Starting from this leading sequence, 10 steps (from 0 to 9) were sufficient to conclude the screening (see Output of the screening algorithm for the C21 run model in Supporting Information).

As reported in Figure 2, the ^{1}H NMR spectrum of C21T333 showed two well-resolved signals at δ 15.36 and 15.58 ppm both supporting the formation of CC^+ base pairs. The CD spectra of the same sequence acquired in the 1–400 μM DNA concentration range, were almost superimposable, consistent with the formation of an intra-molecular structure (Figure 3A). Additionally, the formation of an iM was corroborated by the positive dichroic band centered at 285 nm and the negative one at 260 nm which were lost at neutral pH, a condition that suppressed also the NMR fingerprint associated to the presence of CC^+ base pairs (Figure S2, S3 and S4)..[36]

As previously reported, for iMs the intensity of the CD signal at 285 nm is directly related to the number of CC^+.[32] Worth noting, at all tested C21T333 concentrations at pH 5.5, this value was in line with the one expected for the formation of an iM with a 3 CC^+ base pair core, thus corresponding to the recruitment of all the 6 available cytosines. To quantitatively validate this prediction, we analyzed the thermodynamic profiles of this structure. Both CD and DSC, performed at different DNA concentrations, showed a fully reversible single transition process, corroborating the intra-molecular features of the iM adopted by C21T333 (Figure 3B). By applying a global fitting analysis on the spectroscopic and calorimetric data, we derived a ΔH^{Θ} of (-37.8 ± 0.1) kcal mol^{-1} (Table 1). Since in an iM, the enthalpic contribution of a single CC^+ is about -12 kcal mol^{-1},[18] this evidence fully supported the folding of C21T333 into an intra-molecular iM characterized by a 3 CC^+ base-paired core.

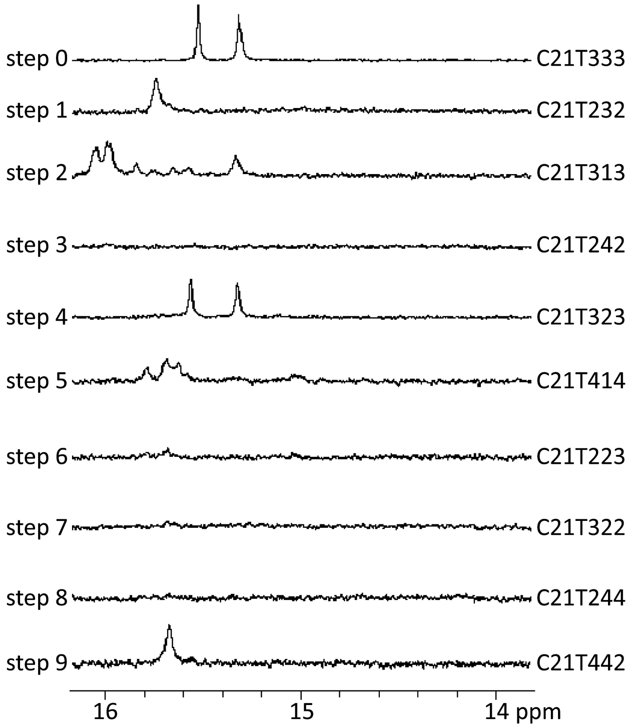

Figure 2. ^{1}H NMR imino region of potentially iM forming oligonucleotides from the C21 group. Spectra were recorded on a 600 MHz NMR spectrometer at 0.2 mM oligonucleotide concentration in 50 mM Na-cacodylate buffer with 10% ^2H_2O, pH 5.5, at 0 °C.

Based on this positive output, in the next step (step 1), the algorithm indicated to test C21T232. The ^{1}H NMR spectrum of this sequence showed one imino signal at δ 15.45 ppm (Figure 2). However, for this sequence, the intensity and position of the maximum of the positive CD band were strongly DNA concentration-dependent (Figure S2 and S3) which pointed to C21T232 forming intermolecular iM structure(s).

The low CD intensity detected at 4 μM oligonucleotide concentration, at 0 °C can rely on a modest fraction of intermolecular species or on an intra-molecular iM, favored by the diluted conditions, endowed by a low thermal stability. To address this issue, we acquired the melting and annealing profiles at 4 μM in solutions containing 40 % ethylene glycol, to extend our analyses down to −20 °C (Figure 4). The advantage of using this crowding agent is that it poorly affects the thermal stability of iMs with no relevant effects on the associated ΔH.[37,38] Under this condition, the annealing was slow and hysteresis was always observed. The melting showed a smooth transition but again, at the lowest tested temperature, the chiroptical signal did not reach the intensity observed for the C21T333 under the same experimental conditions (Figure S5). On the contrary, by increasing the oligonucleotide concentration to 40 μM, CD intensity (and hysteresis) incremented. Moreover, the melting profile showed a clear sigmoidal transition, in line with the assembly of inter-molecular structures occurring at increased C21T232 concentrations. Overall, although these results do not fully exclude its potential folding into an intra-molecular arrangement in the low concentration range, they indicated it did not correspond to an iM with a 3 CC$^+$ core.

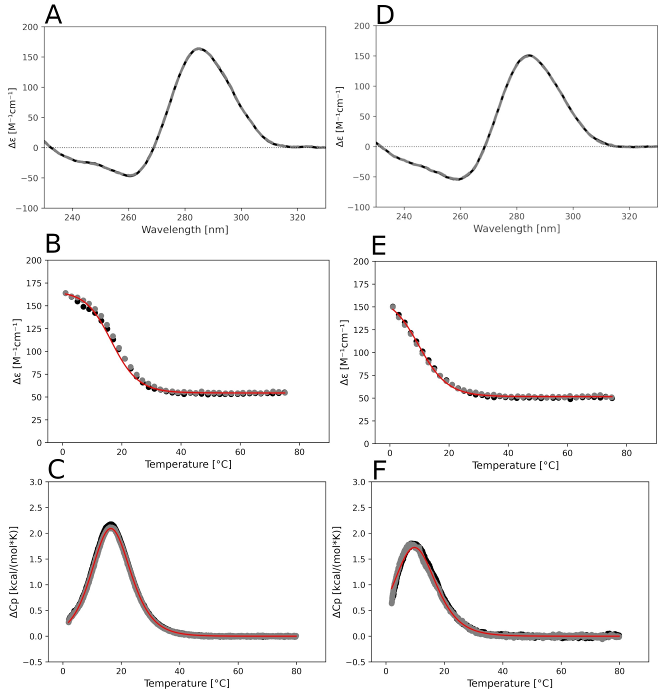

Figure 3. CD spectra of 4 μM (grey dashed lines) and 400 μM (black solid lines) C21T333 (Panel A) and C21T323 (Panel D) at 0 °C. CD melting (black dots) and annealing (grey dots) profiles of 4 μM C21T333 (Panel B) and C21T323 (Panel E). DSC melting (black dots) and annealing (grey dots) scans of 400 μM C21T333 (Panel C) and C21T323 (Panel F). All data were acquired in 50 mM Na-cacodylate pH 5.5. Red lines correspond to global fitting of CD and DSC melting/annealing curves according to a single transition model (equation 1 and 2 in Supporting Information).

Table 1: Thermodynamic parameters for the folding process of C21T333 and C21T323 in 50 mM Na-cacodylate, pH 5.5 derived from global fitting of CD and DSC data. Reported parameters refer to 273.15 K.

	$\Delta G^{\ominus}$	$\Delta H^{\ominus}$	$-T\Delta S^{\ominus}$	Tm [°C]
	[kcal/mol]	[kcal/mol]	[kcal/mol]	
C21T333	−2.1 ± 0.1	−37.8 ± 0.1	35.7 ± 0.1	16.3 ± 0.1
C21T323	−1.2 ± 0.1	−33.7 ± 0.1	32.5 ± 0.1	9.3 ± 0.1

Having removed C21T232, the sequence to be tested in step 2 was C21T313. The presence of multiple ^{1}H NMR signals in the δ 15.0–15.8 ppm range indicated that C21T313 formed a structure exhibiting CC$^+$ base pairs (Figure 2). However, again, the CD spectra were DNA concentration-dependent and matched the characteristic features for iM structures only at higher tested DNA concentrations (Figure S2, S3 and S5). As above discussed for C21T232, this pointed towards the preferential formation of multimeric assemblies.

Subsequently, the folding of C21T242 was analyzed (step 3). However, the ^{1}H NMR spectrum, showed no signals in the range between δ 15 and 16 ppm (Figure 2) and, at any

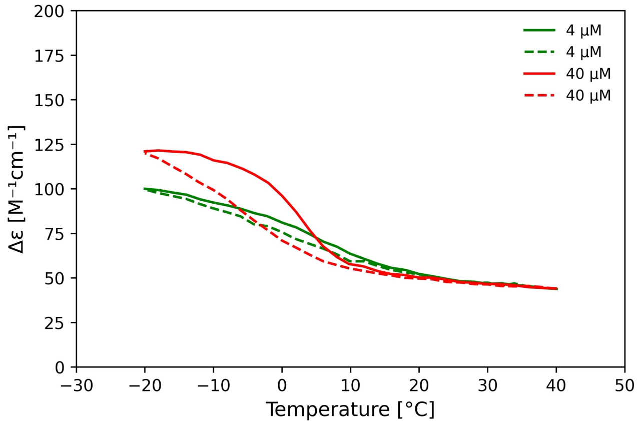

Figure 4. Melting (solid lines) and annealing (dashed lines) profiles of CD spectra of 4 µM (green lines) and 40 µM (red lines) C21T232 acquired by CD in 50 mM Na-cacodylate, 40% ethylene glycol, pH 5.5 at a 20 C h^{-1} rate.

tested oligonucleotide concentrations, the CD features did not match those expected for an iM (Figure S6). These results indicated that C21T242 does not fold into iM.

In the following step we tested the sequence C21T323 (step 4). The ^{1}H NMR spectrum showed well-defined signals at δ 15.36 and 15.61 ppm only under acidic conditions (Figure 2 and S4). The CD spectra recorded at 4 and 400 µM DNA concentration, were almost superimposable with a positive signal at 285 nm and a negative one at 260 nm, a fingerprint that was lost by increasing the pH or the temperature (Figure 3, S2 and S3). As above described for C21T333, the signal intensity at 285 nm in the spectrum of C21T323 was in line with the one expected for the formation of 3 CC^+ base pairs. CD and DSC melting/annealing profiles acquired at pH 5.5 showed a single reversible transition process (Figure 3). From a global fitting analysis of these spectroscopic and calorimetric data, we derived a ΔH^{Θ} of (-33.7 ± 0.1) kcal mol^{-1} (Table 1) that supported the folding of C21T323 into an intra-molecular iM comprising 3 CC^+ base pairs.

Moving to C21T414 (step 5), we observed the presence of multiple signals in the 15.2–15.6 ppm range of the ^{1}H NMR spectrum (Figure 2) but again, the CD signal was DNA concentration-dependent, thus supporting the preferential folding of C21T232 into inter-molecular iM structures (Figure S2, S3 and S5 A).

The screening went progressively through C21T223, C21T322 and C21T244 (step 6, step 7 and step 8), but no signals in the 15–16 ppm range were detected in the ^{1}H NMR spectrum for any of them (Figure 2). Consistently, also CD failed to show the presence of iM (Figure S6).

At step 9, we tested C21T442. The corresponding ^{1}H NMR spectrum showed a signal at 15.45 ppm (Figure 2). As far it concerned its chiroptical features, we detected only a modest dependence of the optical signal upon the DNA concentration. However, neither its intensity nor shape well fitted with the formation of an iM with similar stability of C21T333 or C21T323 (Figure S2 and S3). This conclusion was reinforced by the lack of the expected chiroptical signature also in the low temperature range obtained in the presence of ethylene glycol (Fig S5 A).

Thanks to this result, the P_list list turned out to be completely emptied.

Overall, the end of the screening for the C21 subgroup allowed us to indicate C21T323 as the minimal sequence that folded into an intra-molecular 3 CC^+ base-paired iM under acidic conditions.

Screening of the C12 subgroup

The above described protocol was applied to the C12 subgroup as well. By starting from the C12T333 sequence, we concluded the screening in 7 steps (see Output of the screening algorithm for the C21 run model in Supporting Information).

The ^{1}H NMR spectrum of C12T333 showed two broad signals at δ15.60 and 15.73 ppm (Figure 5). The CD signal was DNA concentration-dependent (Figure S7). Moreover, also at the highest tested concentrations, the positive band was of low intensity and centered at lower wavelengths than expected for a fully folded iM. The CD spectra did not change considerably even working at lower temperatures (Figure S5B). These data pointed to C12T333 as marginally involved in inter-molecular interactions.

In the next step (step 1), the algorithm indicated to test C12T242, for which, however, no imino ^{1}H NMR signal was detected, which together with the lack of iM chiroptical fingerprints in the CD spectra indicated that it did not fold into this secondary structure (Figure 5 and S6).

At step 2, we moved to C12T343. While the ^{1}H NMR spectrum showed a well-resolved signal at δ 15.61 ppm, the CD showed that iM folding was DNA concentration-dependent although the signal was always of low intensity

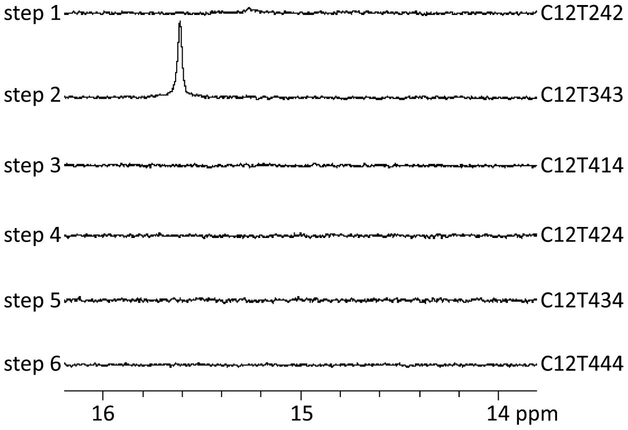

Figure 5. ^{1}H NMR imino region of potentially iM forming oligonucleotides from the C12 group. Spectra were recorded on a 600 MHz NMR spectrometer at 0.2 mM oligonucleotide concentration in 50 mM Na-cacodylate buffer with 10% ^2H_2O, pH 5.5, at 0 °C.

(Figure 5, S5B and S7). This suggested that C12T343 formed only inter-molecular structures.

Then the screening indicated to test C12T414, C12T424, C12T434 and C12T444 (step 3, step 4, step 5 and step 6). None of these sequences, however, exhibited imino ^{1}H NMR signals and their CD spectra did not match the expected profile of an iM (Figure 5 and S6). Thus, we ended the screening for the C12 subgroup without finding any sequence able to fold into an intra-molecular iM with 3 CC$^+$ base pairs.

In-depth structural characterization of C21T333 and C21T323

2D NMR experiments were performed in sodium-phosphate buffer to avoid the signal from the methylic protons of cacodylate. The equivalence of these two different buffers in DNA folding was tested (Figure S8). ^{1}H NMR spectrum of C21T333 exhibits two imino signals at δ 15.36 and 15.58 ppm, suggesting the formation of iM comprising two CC$^+$ base pairs. Additionally, five major and several minor ^{1}H NMR signals are observed in the region between δ 10.58 and 11.37 ppm, consistent with several thymine imino protons protected from the exchange with the solvent, possibly *via* non-canonical base pair formation. Most of the imino ^{1}H NMR resonances were unequivocally assigned by employing ^{15}N-edited HSQC spectra on partially ^{13}C- and ^{15}N-isotope residue-specifically labelled C21T333 at key positions (Figure 6A–B). This approach furthermore revealed that the twelve well-resolved signals observed in the ^{1}H NMR spectrum of C21T333 between δ 8.05 and 9.98 ppm correspond to hydrogen-bonded amino protons of the six cytosine residues (Figure 6C). In parallel, heteronuclear NMR experiments on C21T323 with partially ^{15}N-isotopically residue-specific residues were used to unambiguously assign imino signals at ^{1}H δ 15.37 and 15.61 ppm, corresponding to CC$^+$ base pairing in the iM (Figure 6D). The NMR data also enabled the assignment of thymine imino and cytosine amino ^{1}H NMR signals corresponding to iM adopted by C21T323, which are observed in the range between δ 10.58 and 11.32 ppm, and between δ 8.09 and 9.92 ppm, respectively (Figure 6E–F).

Pairwise comparison of the ^{1}H and ^{15}N NMR chemical shifts corresponding to amino groups in iM adopted by

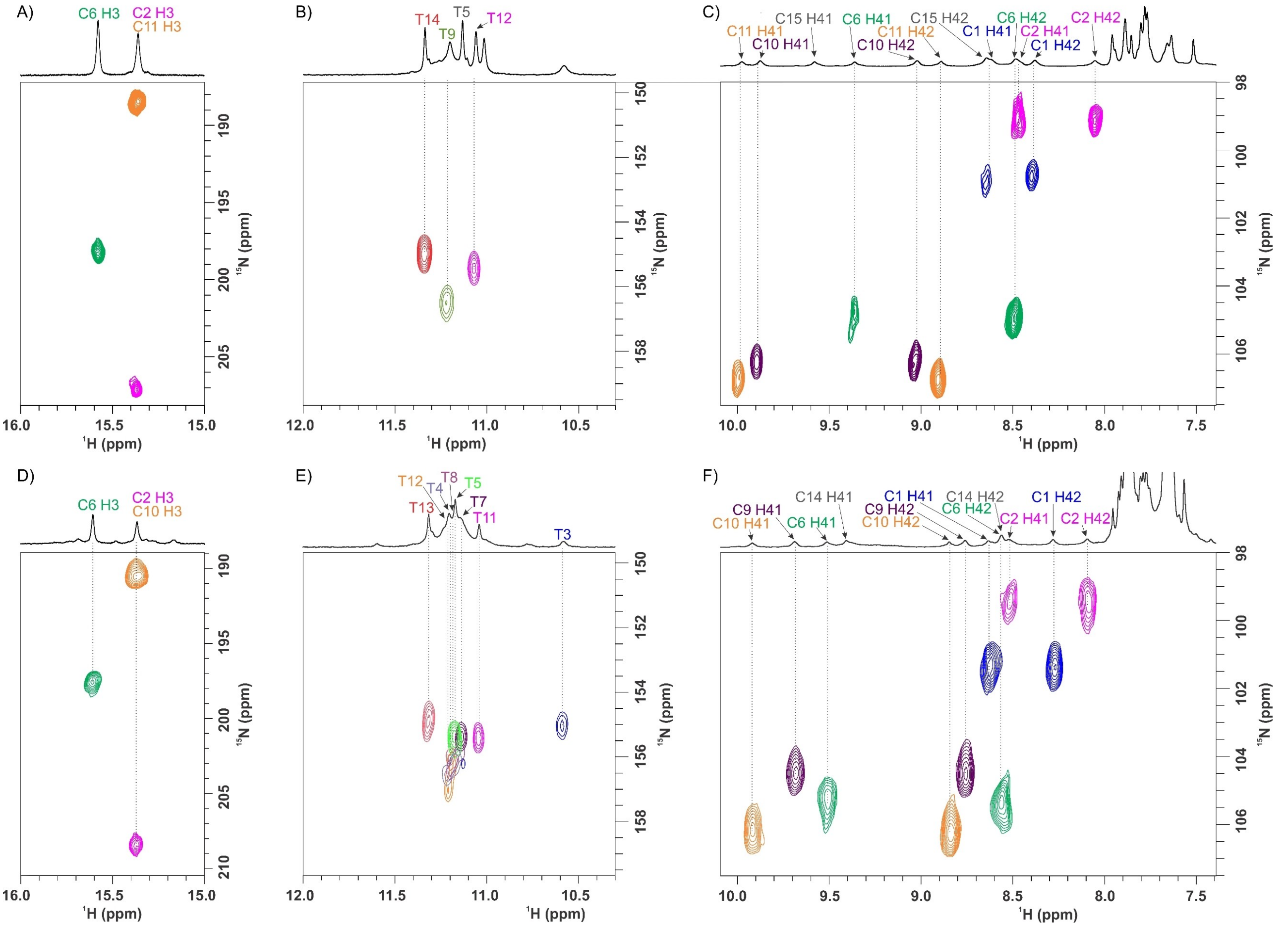

Figure 6. Stacked ^{15}N-edited HSQC spectral regions of residue-specifically partially ^{15}N-isotopically labelled (A–C) C21T333 and (D–F) C21T323. The cross-peaks corresponding to H3–N3 correlations are shown in A and D for cytosine and in B and E for thymine residues, while H41/H42–N4 cross-peaks corresponding to cytosine residues are shown in panels C and F. The cross-peaks in the plots of 2D spectra are colored-matched with the ^{1}H NMR assignments shown in the corresponding ^{1}H NMR spectral regions shown on top.

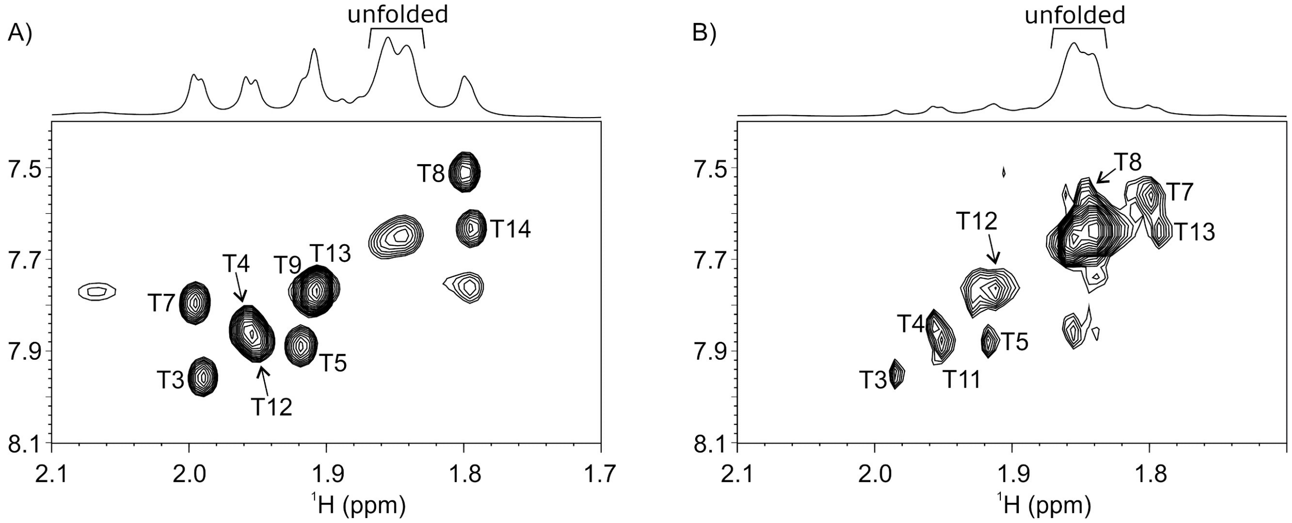

Figure 7. Regions of NOESY spectra of A) C21T333 and B) C21T323 with designated intra-residual aromatic H6-methyl cross-peaks, above which the ^{1}H NMR spectral regions are shown together with indicated methyl groups signals corresponding to unfolded species. The spectra were recorded in 90% H_2O/10% ^2H_2O, at 273 K, 0.4 mM oligonucleotide concentration per strand, 25 mM sodium-phosphate buffer (pH 5.5) and A) (τ_m 200 ms) and B) (τ_m 100 ms).

C21T333 and C21T323 (Table S3 and Figure S9) shows that the variance of a single thymine residue in the second T-run is coupled to the largest difference for C10 vs. C9. These results, however, may reflect variations in T7–T8–T9 versus T7–T8 arrangements, rather than differences in the overall iM topologies of C21T333 and C21T323, especially when considering that hydrogen-bonding patterns are in-line with both oligonucleotides adopting similar (iM) structures. Notably, analysis of integral values of ^{1}H NMR signals for methyl protons showed that the ratio between the iM and unfolded species is ca. 6:4 for C21T333, while it is considerably lower, i.e. 3:7, for C21T323 (Figure 7). This is supported also by analysis of the series of ^{13}C-HSQC spectra of C21T323 carrying partially ^{13}C-isotopically labelled thymine residue at individual positions, in which the most intensive signals at ^{1}H δ 1.84–1.86 ppm corresponded to the methyl groups of unfolded species (Figure 8).In addition to exchangeable and methyl protons, also aromatic and sugar ^{1}H NMR signals corresponding to C21T333 (Figures S10 and S12–S15) as well as to C21T323 (Figure S11) were assigned by analyzing NOESY together with ^{13}C- and ^{15}N-HSQC spectra recorded with the use of residue-specifically partially ^{13}C- and ^{15}N-isotopically labelled samples. The combination of the different 2D NMR experiments was particularly insightful for resolving individual cytosines' aromatic H6 and anomeric H1′ ^{1}H NMR resonances, which are observed in extremely narrow ranges, i.e. δ 7.75–7.94 ppm and δ 6.27–6.57 ppm for C21333 and δ 7.76–7.89 ppm and δ 6.25–6.57 ppm for C21T323 (Figure S16). Weak-to-medium intensities of intra-residual H1′–H6 NOESY cross-peaks corresponding to the predominant (iM) species are consistent with *anti*-glycosidic bond angle disposition for all residues in C21T333. Detailed structural analysis of C21T323 was precluded due to the signal overlap in NOESY spectra.

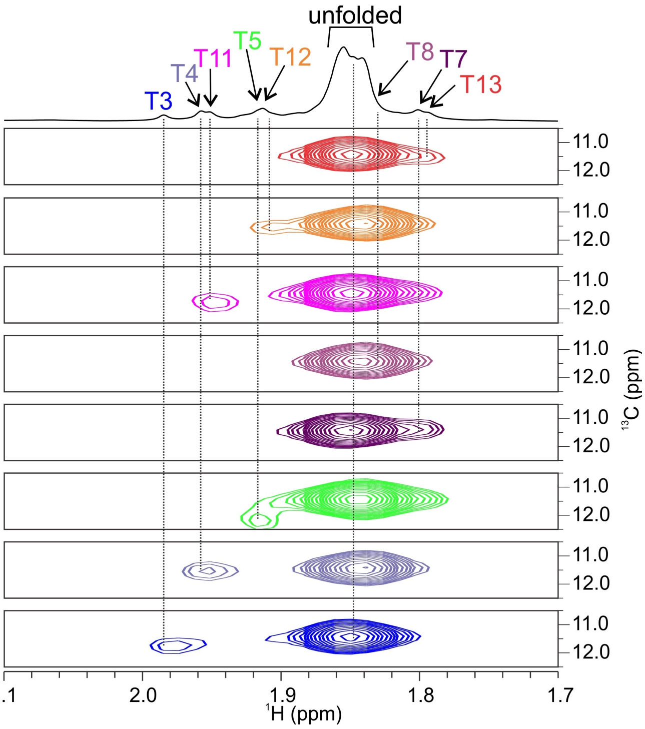

Figure 8. Methyl region of ^{13}C-HSQC spectra of C21T323 recorded on partially residue-specifically ^{13}C-isotopically labelled samples. In the corresponding ^{1}H NMR region above the 2D plots, the assignments of thymine methyl signals are indicated in black for the unfolded oligonucleotide, while in other colors for the iM. The spectra were recorded in 90% H_2O/10% ^2H_2O, at 273 K, 0.4 mM oligonucleotide concentration per strand and 25 mM sodium-phosphate buffer (pH 5.5).

Inter-residual imino-imino and imino-amino correlations observed in the NOESY spectra of C21T333 (Figures S10 and S12–15) are consistent with the formation of closely positioned hemi-protonated C2–C11 and C6–C15 base pairs. Furthermore, NOE interaction is observed between H3 of T5 and T14, with both imino protons exhibiting also NOESY cross-peaks with imino and amino protons of C2/C11. These correlations together with the NOESY cross-peaks between amino (H41 and H42) and H2′/H2′′ observed for the pairs C1–C6, C2–T5, C10–C15 and C11–T14 (Figure S15), are consistent with C21T333 adopting iM comprising two CC$^+$ base pairs in the core, further stabilized by stacking of base pairs C2–C11 to T5–T14 on one side and capping of C6–C15 by C1–C15 on the other. Noteworthy, the absence of ^{1}H NMR signal for hemi-protonation in the C1–C15 base pair could be attributed to the fast exchange of the imino proton with the solvent, as previously noted for external CC$^+$ base pairs in tetra-molecular iMs.[39] The perusal of NOESY spectra of C21T333 furthermore indicates antiparallel directions of the neighboring segments comprising cytosines, which are linked by three T-tracts (T3–T4–T5, T7–T8–T9 and T12–T13–T14) arranged into lateral loops. The intra-molecular iM folding topology is corroborated by the NOESY cross-peaks corresponding to correlations of C1 and C10 with T7, T8 and T9, as well as of C2 and C11 with T5 and T14. Additionally, interactions of T3 with T5–T14 non-canonical base pair are evident from several inter-nucleotide NOESY cross-peaks between sugar and nucleobase moieties of T3 and T14, as well as the interactions of the T5 methyl group with T3 H1′ and H2′′. The NOE interactions between sugar moieties, especially of C1 H1′, H2′ and H2′′ with C15 H1′; of C2 H1′ with all sugar protons of C15; and C2 H1′ with T14 H1′ and H2′′ are consistent with a narrow groove between C1–C2 and T14–C15 strands. The NOE correlations of C10 H1′ with C6 H2′ and H2′′; of C10 H2′ with C6 H1′; of C6H1′ with C11 H1′ and of C11 H1′ with T5 H1′, H2′ and H2′′ indicate that the iM exhibits another narrow groove between T5–C6 and C10–C11 strands.

spectra. These and above noted correlations between amino-H2′/H2′′ protons are consistent with C21T333 adopting iM with the T3–T4–T5, T7–T8–T9 and T12–T13–T14 loops bridging minor, major and minor grooves, respectively.

High-resolution structure of C21T333 i-Motif

The high-resolution structure of C21T333 was calculated with the use of simulated annealing protocol that relied on 277 NOE-derived distance restraints along with 8 hydrogen-bond and 15 torsion-angle restraints (Table 2; PDB ID 8PMB).

The structure is characterized by antiparallel arrangement of the neighboring strands at the core of the structure (Figure 9).

Moreover, C2–C11 intercalates between C1–C10 and C6–C15, featuring virtually no overlap between nucleobases of the consecutive base pairs. T8 caps C1–C10, with its' base moiety positioned almost directly over the base pairs cross-section (Figure 9C above). Interestingly, also T7 exhibits a well-defined orientation, with its' base moiety almost co-planar to the one of C1 (Figure S17). Inspection of the ensemble of C21T333 iM structures shows the distance between T7 O2 and C1 H42 is in the range of 1.8–1.9 Å, suggesting potential hydrogen bonding. Orientation of T9, on the other hand, is more flexible and orientated away from the core of the structure (Figure S17). The vicinity of T5 and T14, especially the mutual closeness of their H3 and O4 atoms indicates the formation of T–T base pair, albeit the corresponding nucleobases are not perfectly coplanar (Figure S17). Moreover, T5–T14 is positioned beneath C2–C11 (Figure 9C, bottom panel), seemingly extending the feature of consecutive intercalated base pairs from 3 to 4. This continuous run of (three CC$^+$ and a single T–T) base pairs may even be extended by considering closely positioned inter-residual H3 and O4 atoms of T3 and T12, which, however, are tilted with respect to each other and protrude towards surroundings. These extensive hydrogen bonding, including the T–T base pairs suggested by structure analysis, is corroborated by the RMSD of 0.719 Å considering all residues, with the overall high convergence related particularly to the well-defined CC$^+$ base pairs in the center and the lateral T3–T4–T5 and T12–T13–T14 loops (Figure S17).

Table 2: NMR restraints and structural statistics for iM adopted by C21T333.

	non-exchan-geable	Exchan-geable
Intra-nucleotide NOE-derived distance restraints	168	2
Sequential (i, i+1) NOE-derived distance restraints	53	0
Long-range (i, >i+1) NOE-derived distance restraints	48	6
Torsion angle restraints	15	
Hydrogen-bond restraints [a]	8	
NOE violations >0.3 Å	0	
Pairwise atom RMSD [Å]	0.72 (0.65) [b]	
Pairwise atom RMSD [Å] without T3, T4, T5	0.74 (0.69) [b]	
Pairwise atom RMSD [Å] without T7, T8, T9	0.53 (0.44) [b]	
Pairwise atom RMSD [Å] without T12, T13, T14	0.77 (0.72) [b]	
Pairwise atom RMSD [Å] only cytosine residues	0.52 (0.48) [b]	

[a] H-bonding restraints corresponding to hydrogen-bonding in CC+ base pairs: six accounting for reciprocal H41–O1 in each of the C1–C10; C2–C11 and C6–C15 base pairs; two accounting to H3–N3 in C6–C15 and C11–C2 base-pairs. [b] Number in the bracket corresponds to RMSD value when considering heavy atoms (C, O, N, P) only.

Validation of i-Motif loop requirements on sequences with longer C-tracts

The data so far collected applied to very constrained iM models comprising only 3 CC$^+$ base pairs. A search for C21T333 and C21T323 on T2T human genome showed that these sequences globally appear 2360 times (Table S4). Nevertheless, as herein shown, the relative iMs are poorly stable. Thus, we decided to validate the herein identified loop requirements on model sequences containing longer C-tracts. For this comparative analysis, we selected already

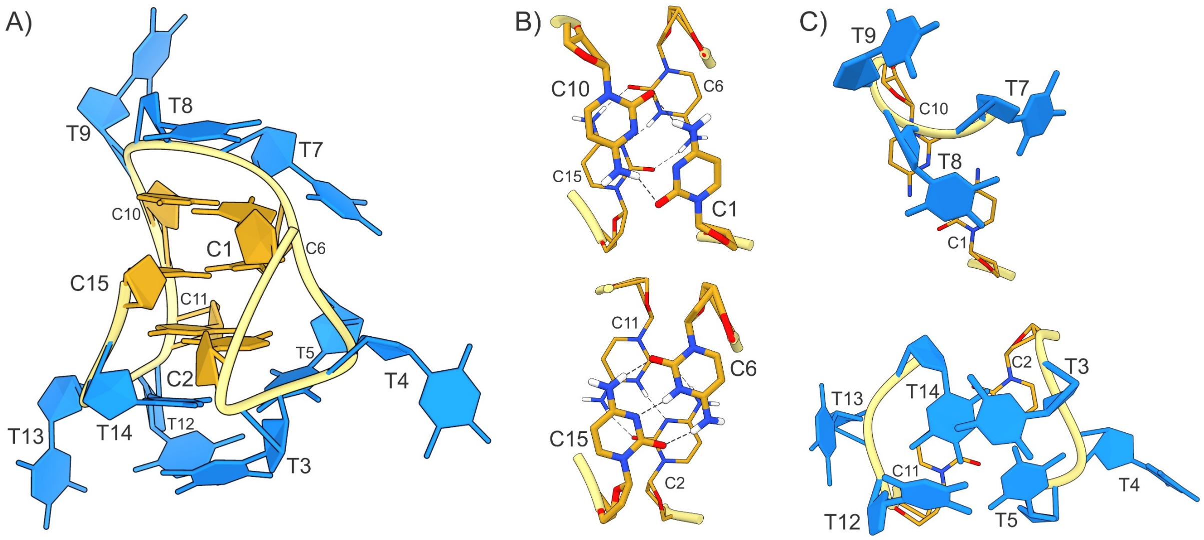

Figure 9. Solution-state structure of iM adopted by C21T333. The lowest energy structure is shown in A), along with the details focused on B) top views on consecutive CC$^+$ base pairs and C) TTT loops arrangements and the nearby CC$^+$ base pairs. Cytosine and thymine residues are depicted in gold and blue, respectively. The dashed lines shown in B) correspond to hydrogen-bonding between C1–C10, C2–C15 and C6–C11 base pairs.

reported sequences that can accommodate up to 6 CC$^+$ with loops comprising 2 or 3 thymines (Table S2).[16,32]

For these longer sequences ^{1}H NMR, CD and DSC have been acquired under the same experimental conditions applied to the shorter ones. (Figure 10, Figures S18–S23).

All of them showed multiple ^{1}H NMR signals between δ 15 and 16 ppm, consistent with the formation of CC$^+$ base pairs. However, under the same experimental conditions, DSC melting and annealing thermograms highlighted a single fully reversible process only for C33T333 and C33T232. For all the other sequences the DSC melting profiles showed multiple peaks with variable intensities and distributions. Conversely, the melting-annealing curves acquired by CD in the low micromolar oligonucleotide concentration range (4 μM) were perfectly overlapping for all the tested sequences. Overall, these data inferred the occurrence of multimeric species, a behavior further supported by EMSA where slow migrating bands were detected when samples were prepared at 100 μM, while they were absent at 4 μM (Figure S24).

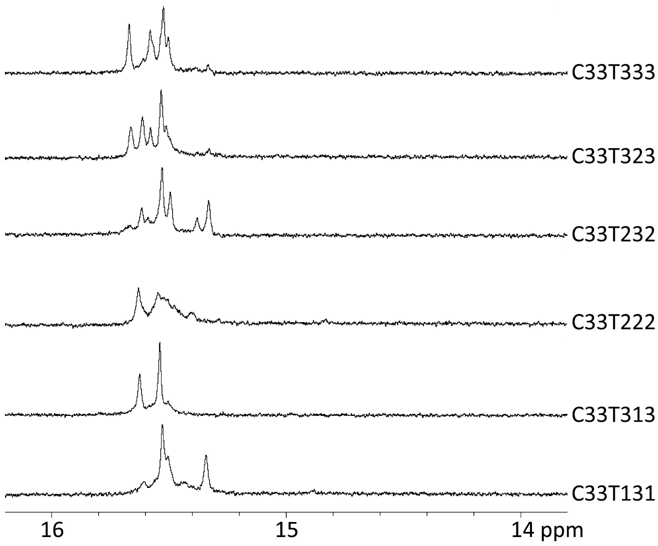

Figure 10. ^{1}H NMR imino region of iM forming oligonucleotides from the C33 group. Spectra were recorded on a 600 MHz NMR spectrometer at 0.2 mM oligonucleotide concentration in 50 mM Na-cacodylate buffer with 10% ^2H_2O, pH 5.5, at 0 °C.

Based on these results, to properly compare all these longer sequences, we derived the thermodynamic parameters by fitting according to equation 1 only the CD datasets acquired at the low oligonucleotide concentrations, at which intra-molecular structures were not occurring (Table 3). These results allowed us to cluster these iMs into two main groups, with C33T333, C33T323 and C33T232 showing a more favorable enthalpic contribution (≈ −10 kcal mol^{-1}) with reference to C33T222, C33T313 and C33T131. Such a drop is compatible with the loss of one CC$^+$ base pair. This is consistent with the accessibility to bromination of C3 and C11 in the C33T222.[32]

Table 3: Thermodynamic parameters for the folding process of 4 μM C33 model sequences in 50 mM Na-cacodylate, pH 5.5 as derived by fitting of the CD melting-annealing profiles according to a two-state model (equation 1). Reported parameters refer to 298.15 K.

	$\Delta G^\ominus$ [kcal/mol]	$\Delta H^\ominus$ [kcal/mol]	$-T\Delta S^\ominus$ [kcal/mol]	Tm [°C]
C33T333	−4.7± ±0.1	−67.6±1.4	62.9± ±1.4	47.2±0.1
C33T323	−3.9± ±0.1	−63.1±1.3	59.2± ±1.3	44.6±0.1
C33T232	−4.5± ±0.1	−64.9±1.4	60.4± ±1.4	46.8±0.1
C33T222	−3.0± ±0.1	−55.0±1.1	52.0± ±1.1	42.4±0.1
C33T313	−3.1± ±0.1	−56.7±1.0	53.6± ±1.0	41.7±0.1
C33T131	−3.6± ±0.1	−55.6±1.3	52.0± ±1.3	45.7±0.1

This output is in line with our finding about loop requirements for the C21 models that must contain at least 1 loop of 3 nucleotides long and are severely penalized by the presence of 1 nt long loops.

It is worth underlining that within this series both C33T323 and C33T232 folded into intra-molecular iMs. Apparently, this represented a discrepancy if compared to C21T323 vs C21T232 models, among which only the first one folded into an intra-molecular iM with 3 CC^+. This suggests that only the C33 sequences can change the topology from a major-minor-major to the minor-major-minor one when they need to accommodate two shorter loops. In this regard, similar iM topology of C33T333 and C33T323 is suggested by the similar number and chemical shifts of imino signals in their ^1H NMR spectra. In contrast, the spectral profile of the imino ^1H NMR region of C33T232 appeared different, indicating that iM topology did not match to the one of C33T333 and C33T323.

We can infer that iMs with an even number of CC^+ might be more prone to undergo such a structural rearrangement since they can accommodate the slipping from the 5‘E to 3‘E topology with no penalty related to a reduction of the number of intercalated CC^+.

Conclusion

Here, we applied an in-house developed algorithm based on a systematic step-by-step workflow that allowed us to identify the minimal loop lengths suitable for the formation of an intra-molecular iM containing only 3 CC^+ base pairs.

The experimental training of the supervised selection of the sequences to be tested was successful. Indeed, out of the 128 possible sequence combinations, we converged toward the minimal loop lengths by testing only 17 sequences. This protocol ensures the exploration of all the possible minimal folding sequences and, even if optimized for iMs, it can be easily applied to identify the minimal length of the loops for any other three-loops intra-molecular DNA structure. Still, it is worth emphasizing that it does not provide a folding prediction and that the supported selection of the sequence to be tested is based on experimentally derived inputs.

To train our workflow, we limited the base composition of all the loops to thymine. Interestingly, only C21T323 was identified as the minimum among our 3 CC^+ base-paired iM datasets. This sequence folds into a single iM topology that corresponds to the major-minor-major grooves combination pattern. This topology is conserved also in the C21T333 model, which contains one extra thymine in the central loop. The higher stability of C21T333 allowed us to obtain the high-resolution structure of this iM (PDB ID 8PMB). To date, it is the simplest deposited structure of an intra-molecular iM that contains an odd number of CC^+, i.e. only 3.

As an additional result, our screening procedure uncovered an intriguing asymmetry between the C21 and C12 series. Specifically, none of the assessed sequences from the C12 subgroup adopted an intra-molecular iM. This result fits with the reported higher stability of iMs with 5′E topology.[16] Still, it is worth mentioning that our data are limited to a 3 CC^+ system, hence such a difference cannot be generalized. Indeed, the stability of an iM may greatly depend on interactions among loop residues, such as the recently reported Watson–Crick hydrogen-bonding[40] and non-canonical base-pairing like the T-T pairs, present also in the C21T333 and C21T323 models studied here.

The constrained feature of the C21 core forces the loops toward a unique minimal length distribution, namely 323. Indeed, C21T232 did not accommodate all the 3 CC^+ base pairs. This result indicates that the minor-major-minor topology is not feasible in the C21 models for reasons that remain unidentified.

On the contrary, from our data it appears that such topological transition is permitted in the C33 models. To better rationalize this divergent feature of the C21 and C12 models compared to the C33 one, a comprehensive comparison of our solved structure with other iMs containing a higher odd number of CC^+ could be envisaged. However, currently, out of the 9 solved models deposited in the RCSB, only PDB 7O5E corresponds to an iM with an odd number of CC^+. Its core is folded according to the major-minor-major pattern but it is embedded between a G:T:G:T tetrad and a T–T base pair which continues in a Watson–Crick paired stem. As a result, the geometry of the iM grooves is altered.

As a future perspective, the screening of sequences corresponding to the C32 and C23 subgroups could provide deeper insights on this intriguing structural issue. These subgroups are expected to form up to 5 CC+ base pairs and, consequently, to demonstrate enhanced thermodynamic stabilities in comparison to the sequences analyzed in this study.

Overall, our data indicate that the folding of iMs is highly dependent on loop lengths with 3 and 2 nucleotides, representing the minimal number required to link the major and the minor grooves, respectively. This condition is remarkably different to that of G4 s, for which loops composed of 1 nucleotide are sufficient for the formation of thermodynamically stable intra-molecular structures.

To illustrate the consequences of these findings, we performed a comparison of the coverage of the genome by sequences potentially able to fold into G4 s or iMs, by applying the currently used string for G4 formation versus the one for iM formation refined herein (Figure 11, see regular expression in Supporting Information).

Since for G4 s the minimal length of all G-runs is fixed to 2, the number of iMs dominates. This approach likely overestimates the iM frequency, since most sequences potentially able to accommodate only 3 CC^+ iMs are expected to be unfolded in the cell nuclei. Still, it is interesting to remark that this search evidenced a fraction of sites where only G4 s can form. These are associated to G4 s with loops shorter than those required for iMs (i.e loops with 1 nucleotide, or sequences with all three loops covering only 2 nucleotides).

Overall, the evidence acquired in this work further addresses iMs and G4 s as distinct and independent sensing systems. Indeed, the folding of G- and C-rich structures

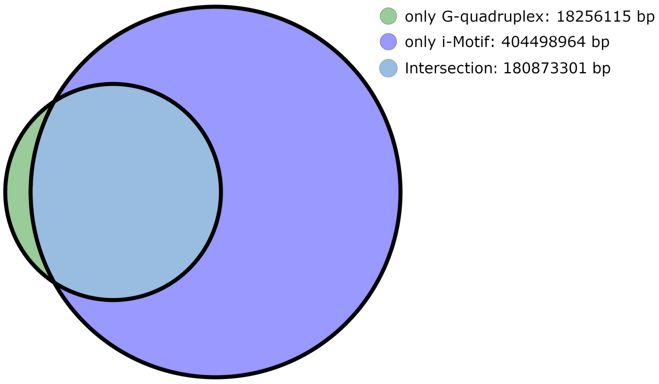

Figure 11. Venn diagram representing the intersection of the potential G4 and iM forming sequences found in the T2T reference human genome. The lengths of potential G4 and iM forming sequences are reported as number of base pairs (bp).

requires different environmental conditions to occur and, as we showed here, it further relies on different sequence patterns.

Supporting Information

The authors have cited additional references within the Supporting Information.[41,42]

Acknowledgements

This work was supported by CARIPARO (PhD fellowship to MG), AIRC (grant IG 2021 - ID26474, PI CS), CERIC ERIC (grants No. 20207052 and 20202174) and European Union-Next GenerationEU (PNRR M4 C2-Investimento 1.4- CN00000041). MT and JP acknowledge financial support from the Slovenian Research Agency grants (P1-0242 and J1-1704).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: Calorimetry · DNA · i-Motif · Screening Algorithm · Spectroscopy

[1] H. Abou Assi, M. Garavís, C. González, M. J. Damha, *Nucleic Acids Res.* **2018**, *46*, 8038–8056.
[2] K. Gehring, J. L. Leroy, M. Guéron, *Nature* **1993**, *363*, 561–565.
[3] J. L. Huppert, S. Balasubramanian, *Nucleic Acids Res.* **2005**, *33*, 2908–2916.
[4] E. Belmonte-Reche, J. C. Morales, *NAR Genomics Bioinf.* **2020**, *2*, lqz005.
[5] R. A. Rogers, A. M. Fleming, C. J. Burrows, *ACS Omega* **2018**, *3*, 9630–9635.
[6] S. Kendrick, H.-J. Kang, M. P. Alam, M. M. Madathil, P. Agrawal, V. Gokhale, D. Yang, S. M. Hecht, L. H. Hurley, *J. Am. Chem. Soc.* **2014**, *136*, 4161–4171.
[7] S. Dzatko, M. Krafcikova, R. Hänsel-Hertsch, T. Fessl, R. Fiala, T. Loja, D. Krafcik, J.-L. Mergny, S. Foldynova-Trantirkova, L. Trantirek, *Angew. Chem. Int. Ed.* **2018**, *57*, 2165–2169.
[8] M. Zeraati, D. B. Langley, P. Schofield, A. L. Moye, R. Rouet, W. E. Hughes, T. M. Bryan, M. E. Dinger, D. Christ, *Nat. Chem.* **2018**, *10*, 631–637.
[9] X. Ma, Y. Feng, Y. Yang, X. Li, Y. Shi, S. Tao, X. Cheng, J. Huang, X. Wang, C. Chen, D. Monchaud, W. Zhang, *Nucleic Acids Res.* **2022**, *50*, 3226–3238.
[10] C. D. P. Martinez, M. Zeraati, R. Rouet, O. Mazigi, B. Gloss, C.-L. Chan, T. M. Bryan, N. M. Smith, M. E. Dinger, S. Kummerfeld, D. Christ, *bioRxiv preprint* **2022**, https://doi.org/10.1101/2022.04.14.488274.
[11] Y. Dong, Z. Yang, D. Liu, *Acc. Chem. Res.* **2014**, *47*, 1853–1860.
[12] S. Modi, S. M. G., D. Goswami, G. D. Gupta, S. Mayor, Y. Krishnan, *Nat. Nanotechnol.* **2009**, *4*, 325–330.
[13] D. Karna, M. Stilgenbauer, S. Jonchhe, K. Ankai, I. Kawamata, Y. Cui, Y.-R. Zheng, Y. Suzuki, H. Mao, *Bioconjugate Chem.* **2021**, *32*, 311–317.
[14] S. Nonin-Lecomte, J. L. Leroy, *J. Mol. Biol.* **2001**, *309*, 491–506.
[15] M. Guéron, J.-L. Leroy, *Curr. Opin. Struct. Biol.* **2000**, *10*, 326–331.
[16] M. Cheng, D. Qiu, L. Tamon, E. Ištvánková, P. Víšková, S. Amrane, A. Guédin, J. Chen, L. Lacroix, H. Ju, L. Trantírek, A. B. Sahakyan, J. Zhou, J.-L. Mergny, *Angew. Chem. Int. Ed.* **2021**, *60*, 10286–10294.
[17] I. Berger, M. Egli, A. Rich, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 12116–12121.
[18] J.-L. Mergny, L. Lacroix, X. Han, J.-L. Leroy, C. Helene, *J. Am. Chem. Soc.* **1995**, *117*, 8887–8898.
[19] J. Amato, F. D'Aria, S. Marzano, N. Iaccarino, A. Randazzo, C. Giancola, B. Pagano, *Phys. Chem. Chem. Phys.* **2021**, *23*, 15030–15037.
[20] X. Han, J.-L. Leroy, M. Guéron, *J. Mol. Biol.* **1998**, *278*, 949–965.
[21] I. Serrano-Chacón, B. Mir, N. Escaja, C. González, *J. Am. Chem. Soc.* **2021**, *143*, 12919–12923.
[22] I. Serrano-Chacón, B. Mir, L. Cupellini, F. Colizzi, M. Orozco, N. Escaja, C. González, *J. Am. Chem. Soc.* **2023**, *145*, 3696–3705.
[23] B. Mir, I. Serrano, D. Buitrago, M. Orozco, N. Escaja, C. González, *J. Am. Chem. Soc.* **2017**, *139*, 13985–13988.
[24] A. T. Phan, M. Guéron, J.-L. Leroy, *J. Mol. Biol.* **2000**, *299*, 123–144.
[25] K. S. Li, D. Jordan, L. Y. Lin, S. E. McCarthy, J. S. Schneekloth Jr, L. A. Yatsunyk, *Angew. Chem. Int. Ed.* **2023**, *62*, e202301666.
[26] M. Trajkovski, T. Endoh, H. Tateishi-Karimata, T. Ohyama, S. Tanaka, J. Plavec, N. Sugimoto, *Nucleic Acids Res.* **2018**, *46*, 4301–4315.
[27] M. Trajkovski, M. Webba da Silva, J. Plavec, *J. Am. Chem. Soc.* **2012**, *134*, 4132–4141.
[28] K. W. Lim, L. Lacroix, D. J. E. Yue, J. K. C. Lim, J. M. W. Lim, A. T. Phan, *J. Am. Chem. Soc.* **2010**, *132*, 12331–12342.

[29] M. Marušič, R. N. Veedu, J. Wengel, J. Plavec, *Nucleic Acids Res.* **2013**, *41*, 9524–9536.
[30] E. P. Wright, J. L. Huppert, Z. A. E. Waller, *Nucleic Acids Res.* **2017**, *45*, 2951–2959.
[31] 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.
[32] P. Školáková, D. Renčiuk, J. Palacký, D. Krafčík, Z. Dvořáková, I. Kejnovská, K. Bednářová, M. Vorlíčková, *Nucleic Acids Res.* **2019**, *47*, 2177–2189.
[33] N. Iaccarino, M. Cheng, D. Qiu, B. Pagano, J. Amato, A. D. Porzio, J. Zhou, A. Randazzo, J.-L. Mergny, *Angew. Chem. Int. Ed.* **2021**, *60*, 10295–10303.
[34] T. Fujii, N. Sugimoto, *Phys. Chem. Chem. Phys.* **2015**, *17*, 16719–16722.
[35] J. L. Leroy, *J. Mol. Biol.* **2003**, *333*, 125–139.
[36] J. Kypr, I. Kejnovská, D. Renčiuk, M. Vorlíčková, *Nucleic Acids Res.* **2009**, *37*, 1713–1725.
[37] S. Takahashi, N. Sugimoto, *Phys. Chem. Chem. Phys.* **2015**, *17*, 31004–31010.
[38] Y. P. Bhavsar-Jog, E. Van Dornshuld, T. A. Brooks, G. S. Tschumper, R. M. Wadkins, *Biochemistry* **2014**, *53*, 1586–1594.
[39] N. Esmaili, J. L. Leroy, *Nucleic Acids Res.* **2005**, *33*, 213–224.
[40] E. Ruggiero, S. Lago, P. Šket, M. Nadai, I. Frasson, J. Plavec, S. N. Richter, *Nucleic Acids Res.* **2019**, *47*, 11057–11068.
[41] J.-L. Mergny, L. Lacroix, *Curr. Protoc. Nucleic Acid Chem.* **2009**, *37*, 17.1.1–17.1.15.
[42] W. Lee, M. Tonelli, J. L. Markley, *Bioinformatics* **2015**, *31*, 1325–1327.

Manuscript received: July 3, 2023
Accepted manuscript online: August 23, 2023
Version of record online: September 4, 2023