UNIVERSITY OF PADUA

MOLECULAR SCIENCES

Ph.D. THESIS

DNA i-Motif: from structure and thermodynamic to genome distribution.

Ph.D. student: Michele Ghezzo

Supervisor: Professor Claudia Sissi

Index

Abstract

Sequences containing at least four runs of repetitive cytosines or guanines can form tetra-helical nucleic acid structures called i-Motifs and G-quadruplexes, respectively. The folding into the tetra-helixes is based on the non-canonical cytosine-cytosine$^+$ and guanine-guanine Hoogsteen base pairs that form between the cytosines and guanines within the runs of the i-Motifs and G-quadruplexes.

At now, the general patterns for potential i-Motif and G-quadruplex forming sequences, commonly reported as C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5} and G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5}, are considered perfectly complementary, thus clustering them at the same genome locations. On this basis, bioinformatics tools were developed to search for these secondary structures, and, interestingly, a significant enrichment of potential quadruplex forming sequences was found at the telomers and the promoters of oncogenes. Moreover, it has been proved that they are folded *in-vivo* and play a role in the regulation of transcription *in-vitro*, making them promising targets for the development of new anti-cancer therapies.

Therefore, we started by screening the proximal promoter of the *EGFR* oncogene for potential quadruplex-forming sequences having at least 3 guanines or cytosines in all the runs. Two sequences were found: EGFR-272 and EGFR-37, 272 and 37 nucleotides upstream of the transcription starting site of the oncogene, respectively. Focussing on the cytosine-rich strand of EGFR-37, we observed that it folds into an i-Motif that forms fewer cytosine-cytosine$^+$ base pairings than the predicted ones. From the *in-silico* analysis, the central loop should have been one nucleotide long to properly form 6 cytosine-cytosine$^+$ base pairs but our results indicated that two cytosines, one from each of the second and the third runs, did not pair to make the central loop longer.

This result prompted us to consider that one nucleotide long loops may not fit with the i-Motifs as they do with G-quadruplexes. Therefore, we developed a novel step-by-step pipeline for the systematic screening of i-Motif models, and we applied it to determine the minimal length of the loops allowing the folding into an intra-molecular i-Motif with a focus on structures comprising only three cytosine-cytosine$^+$ base pairs. Our data indicate that two and three nucleotides are required to connect the strands through the major and the minor grooves of the i-Motif, respectively.

Moreover, as it was for the i-Motif of EGFR-37, we noticed that very often thymine-thymine base pairs are found to be in staking on the outermost cytosine-cytosine$^+$ base pairs of the i-Motifs. Therefore, we decided to verify if, beyond the outside of the i-Motif core, they can form within the cytosine-cytosine$^+$ base pairs as well. Thus, we studied CT dinucleotide repeats and we proved that they fold into i-Motif structures with alternating intercalation of cytosine-cytosine$^+$ and thymine-thymine base pairs.

These results prove that the oligonucleotide pattern of a potential i-Motif forming sequence is different from the G-quadruplex one and this potentially clusters these secondary structures at different genome locations, from which may derive different biological functions.

This knowledge represents a step forward to the development of prediction tools for the proper identification of bio-functional i-Motifs as well as for the rational design of these secondary structures for technological applications.

Introduction

DNA secondary structures

DNA is the genetic material present in all living organisms. It is a long polymer made up of nucleotides, each consisting of a sugar, a phosphate group, and a nitrogenous base. The four types of bases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these bases encodes the genetic information that is responsible for the inherited traits of an organism.

In addition to its well-known role as the carrier of genetic information, DNA can also adopt various types of secondary structures that play important roles in its biological functions.

The double helix is the most familiar form of DNA and is characterized by two complementary strands of DNA that are held together by hydrogen bonds between the bases. This structure, proposed by James Watson and Francis Crick in 1953, is the basis for the stable and faithful transmission of genetic information (Figure 1) [1].

Figure 1. B-DNA, the right-handed double helix.

The double helix is not the only secondary structure of DNA, there are other conformations including hairpin, cruciform, triplex and quadruplexes (Figure 2).

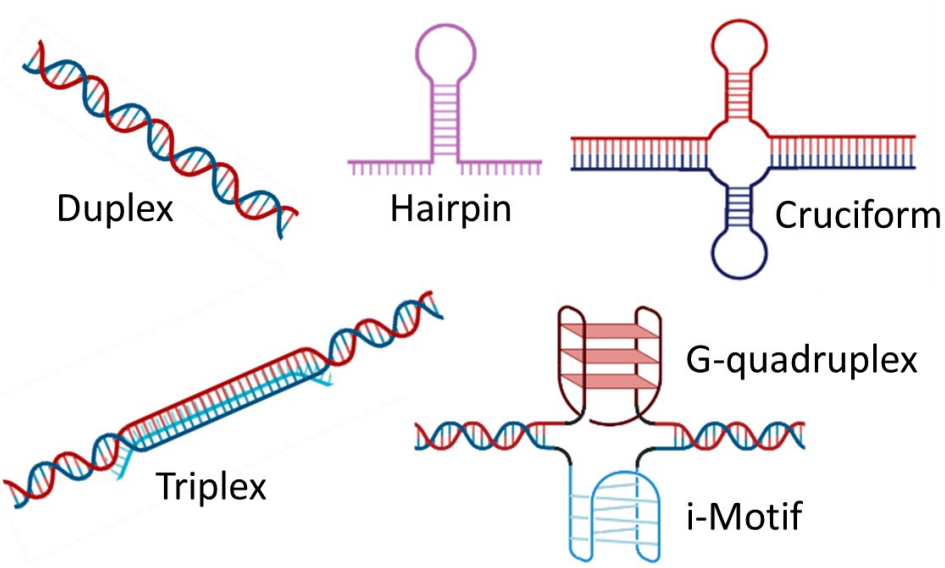

Figure 2. DNA secondary structures

Hairpin and cruciform DNA structures are important secondary structures that play important roles in various biological processes. Hairpin DNA is formed when a single strand of DNA folds back on itself to form a loop, stabilized by hydrogen bonding between complementary base pairs. Cruciform DNA is formed when two opposing DNA strands cross over each other to form a cruciform shape, with the arms of the cruciform held together by hydrogen bonding between complementary base pairs. Both hairpin and cruciform may form in palindromic sequences and they have been found to play important roles in DNA metabolism, including DNA replication, recombination, and repair [2,3].

Triplex DNA is a structure formed when a third strand of DNA or RNA binds to the major groove of a double-stranded DNA helix through non-canonical base pairing. This structure has been proposed to play a role in the regulation of gene expression and DNA repair [2,4].

Quadruplexes are structures formed by the arrangement of guanine or cytosine -rich DNA sequences to form G-quadruplex (G4) or i-Motif (iM) structures, respectively. To date, the general patterns associated with the formation of G4 and iM structures are considered to be G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5} and C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5}, respectively. Therefore, potential G4 and iM forming sequences are located at the same genome sites as the promoter of oncogenes and telomers where it has been proposed to play a role in the regulation of gene expression and genome architecture [2,5,6].

In summary, DNA secondary structures are important for various biological processes and have garnered significant attention in the scientific community. Understanding these structures and their functions will provide insights into the molecular mechanisms underlying these processes and may have potential applications in the development of therapies for diseases related to DNA metabolism.

G-quadruplex

G-quadruplexes (G4s) are formed by the alignment of four guanine-rich DNA or RNA strands in a tetra-helical arrangement. The formation of these structures is driven by the ability of guanine bases to form Hoogsteen hydrogen bonds with each other, leading to the formation of a planar arrangement called G tetrad [5].

G tetrads stack on each other to form the tetraplex structure with the coordination of a monovalent cation in the internal cavity (Figure 3).

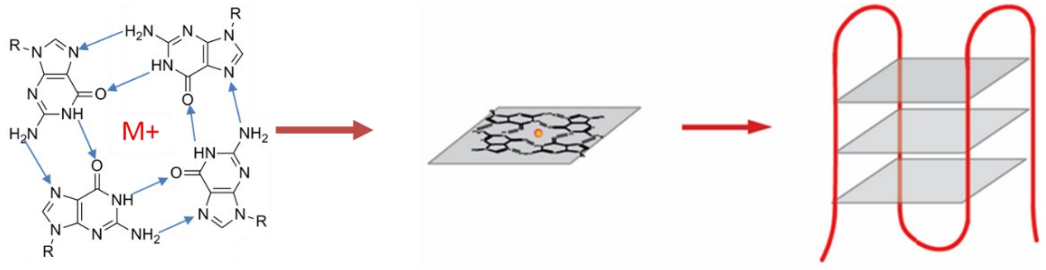

Figure 3. G tetrad and G4 structure.

On this basis, potential G4 forming sequences must contain 4 runs of at least two guanines separated by 1-7 nucleotides, and this pattern is commonly reported as G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5}-N_{1-7}-G_{2-5} [7].

G4 structures are highly polymorphic and, commonly, a classification of G4 polymorphism is made on these main features: molecularity, number of tetrads, and orientation of the strands. Concerning molecularity, G4s can be mono, bi, or tetra-molecular, based on the number of individual strands forming the structure. The number of G-tetrads, two at least, is another common feature that greatly affects the stability of G4s and on which G4s can be classified. Lastly, the mutual orientation of the strands, also, leads to three different classes of G4s: parallel, antiparallel and hybrid G4s.

G4s have been observed in a variety of contexts, including in the promoter regions of genes and at telomeres [8].

The stability of G4s is influenced by a variety of factors, including the number of G tetrads, the presence of potassium or sodium cations, and the binding of small molecules. G4s are highly ordered structures and their unfavorable negative change in the entropy (ΔS) is balanced by a favorable negative change in the enthalpy (ΔH).

G4s are thought to play a variety of roles in cells, including gene regulation and telomere maintenance. In gene regulation, G4s may function as transcriptional repressors, inhibiting or activating the expression of genes by blocking or activating the binding of transcriptional machinery to promoter regions. At telomeres, G4s may play a role in genome architecture, protect the ends of chromosomes from degradation and contribute to the maintenance of genomic stability.

Given their potential roles in a variety of biological processes, G4s have attracted interest as potential therapeutic targets for a range of diseases. In particular, G4s have been explored as potential targets for cancer therapy, due to their involvement in the regulation of oncogenes and the maintenance of telomeres in cancer cells. Small molecules that specifically target G4s, known as G4 ligands, are being developed as potential cancer therapies [5].

i-Motif

I-motifs (iMs) are tetra-helical nucleic acid secondary structures that may form at sequences containing four runs of repeated cytosines [9]. This nucleotide pattern is commonly reported as C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5}-N_{1-7}-C_{2-5} which is complementary to the G4 one, thus potentially locating iMs and G4s at the same genomic sites. As above mentioned, there is an interesting enrichment of potential iM/G4 forming sequences in the human genome at regions endowed with relevant biological functions, such as telomeres and oncogene promoters [10,11]. A detailed description of the molecular events that relate the tetra-helices formation to specific functional roles is not fully disclosed but, it has been proven that their conversion from the double-stranded DNA at gene promoters largely impacts the efficiency of gene transcription [12,13].
iMs are based on non-canonical hemi-protonated cytosine-cytosine base pairs (CC^+) [14,15] that form between two parallel-oriented strands (Figure 4).

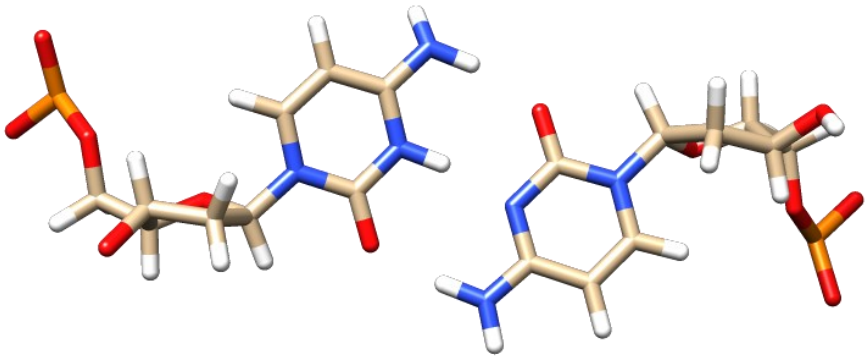

Figure 4. Cytosine-cytosine$^+$ base pair

Two of these parallel duplexes intercalate in an antiparallel orientation to form the iM structure. Therefore, iMs are highly ordered structures, and the adverse entropy change of folding (ΔS) is balanced by a favorable reduction of the enthalpy (ΔH) [16,17]. Many factors contribute to the ΔH, but, theoretical calculations indicate that the three hydrogen bonds of the CC^+ base pairs and the bonding networks in the minor grooves between the H1' and O4' of two consecutive sugar residues are the most relevant stabilizing components and, unlike what is observed in the canonical B-DNA double helix, π-π stacking interactions

between consecutive CC^+ do not significantly contribute to the enthalpy of the folded state [18]. Globally, each CC^+ base pair corresponds to about 10-12 kcal mol^{-1} to the ΔH, and this value can be applied to predict the iM core length [19,20].

To support the formation of three hydrogen bonds within the CC^+ base pair, the N3 of cytosines must be hemi-protonated. This makes iMs pH-dependent structures and the apparent Gibbs free energy change of folding ($\Delta G_{app}(T, pH)$) is generally minimized at pH 4.5 which corresponds to the pKa of cytosine N3 [21]. For a long time, the acidic condition required for their folding was considered incompatible with the cellular environment. However, by using iM fluorescent-labeled antibodies and *in-cell* NMR experiments, it has been proven that they exist in the cell nuclei as well [22,23]. Noteworthy, the intracellular environment of cancer cells is often slightly acidic and this further increases iM formation [24]. This evidence confirmed iMs as promising targets for the development of new anticancer therapies. On these bases, a detailed structural characterization of the genomic regions potentially prone to fold into iMs is the first step to achieving a better understanding of the finely tuned mechanisms that correlate their formation with the regulation of oncogene transcription. In addition to biological and pharmacological purposes, 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 [25–27].

Within an iM, the CC^+ intercalation frame may shift 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 or 5'E when the outmost CC^+ are in the 3' or the 5' ends, respectively [28,29]. Conversely, for iM with an odd number of CC^+, only one topology grants the maximal number of base-pairings, thus representing the most favorable one [30]. 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. As a consequence, 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 5).

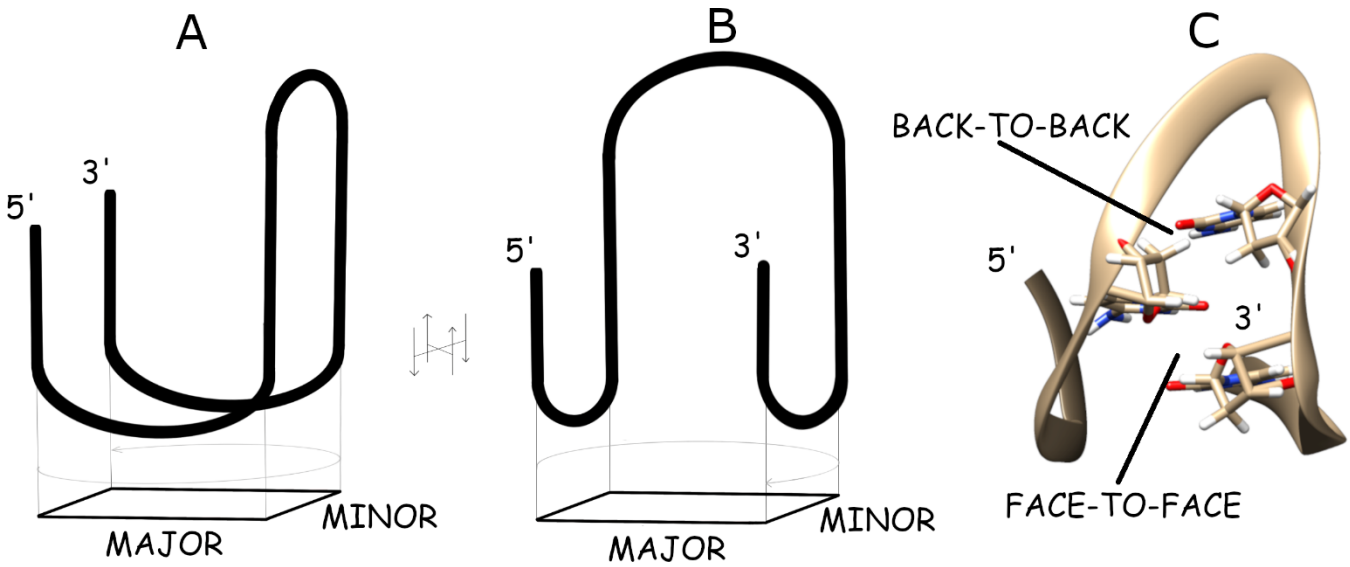

Figure 5: Two topologies for intra-molecular iM in terms of loop connection path: the major-minor-major (A) and the minor-major-minor (B). View of the face-to-face and back-to-back sugar orientation throughout the minor groove (PDB ID 7QDC) (C).

Noteworthy, all cytosines in the iM core are confined to adopt an *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 anticlockwise and clockwise, respectively (Figure 5).

Importantly, the minor grooves in iMs are extremely narrow, and this is a unique feature among the currently characterized DNA secondary structures. Throughout 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. This unique feature of the iMs, as above described, leads to the formation of the systematic inter-strand H1'-O4' hydrogen bonding network which contributes to iM stability [31]. Another unique feature of the iMs is the helical rise between two consecutive cytosines belonging to the same strand which is doubled if compared to B-DNA [31].

iM structures are thought to play a variety of roles in cells, including in gene regulation and DNA-protein interactions.

Given their potential roles in a variety of biological processes, iMs have attracted interest as potential therapeutic targets for a range of diseases. In particular, iMs have been explored as potential targets for cancer therapy, due to their involvement in the regulation of oncogenes. Small molecules that specifically target iM DNA structures, known as iM ligands, are being developed as potential cancer therapies.

Aim

My Ph.D. started with the structural and thermodynamic characterization of an iM forming sequence that was identified by screening the promoter of the *EGFR* oncogene (Chapter 1). I was searching for potential iM structures that could form 6 CC^+, thus matching with the C_{3-5}-N_{1-7}-C_{3-5}-N_{1-7}-C_{3-5}-N_{1-7}-C_{3-5} oligonucleotide pattern and I found one match 37 nucleotides upstream of *EGFR* the transcription starting site. The experimental data were not in agreement with the *in-silico* prediction, and I noticed that the central loop should have been 1 nucleotide long but the cytosines around it did not pair to make the loop longer. This evidence prompted me to reconsider the minimal length of the loops allowing the iM folding. Indeed, G4s can fold with one nucleotide long loops, and it was commonly considered possible for iMs as well. Nevertheless, there was no evidence of it and, in general, there was not enough data available on this topic.

Therefore, I developed a novel step-by-step pipeline for the systematic screening of iM models and we applied it to determine the minimal length of the loops allowing the folding into an intra-molecular iM (Chapter 2).

The results in Chapters 1 and 2 highlighted another interesting side evidence related to the non-canonical thymine-thymine base pair (TT). In both works, we found this interesting interaction in stacking on the outermost CC^+, as an extension of the iM core. Therefore, I decided to check whether it could form into the iM core as well, because this possibility would lead to the possibility of CT mixed iMs, greatly expanding the general pattern and thus the number of potential iM forming sequences (Chapter 3).

I aimed to check whether G4s and iMs have different patterns, in contrast with what is generally considered by the scientific community. Different patterns would lead to different genome locations of the iM and G4 structures, from which it would be possible to derive their biological functions with higher accuracy. To achieve my aim I followed an *in-vitro* approach based on several biophysical techniques commonly used to characterize the structure and thermodynamics of iM foldings. Spectroscopic techniques such as NMR, circular dichroism, and UV-Vis spectroscopy were used to obtain high and low-resolution data on the iM structures I worked on. Concerning the thermodynamics data I took advantage of differential scanning calorimetry which is the gold standard for obtaining the energetics driving the folding of macromolecules such as DNA. This approach proved to be useful in achieving the aim of my Ph.D. and the results represent a step forward to the general knowledge of the iMs which are interesting DNA secondary structures from a biological as well as technological point of view.

Chapter 1: Structural characterization of a cytosine-rich potential quadruplex forming sequence in the EGFR promoter

Many cancers, such as breast or lung cancers and glioblastoma, are related to an over-expression of the Epidermal Grow Factor Receptor (*EGFR*) oncogene [32,33]. At now, targeted treatments are represented by monoclonal antibodies and tyrosine-kinase inhibitors which specifically target the EGFR transmembrane receptor are used in therapies but, frequently these approaches result ineffective because of drug resistance development [34,35]. A new valid therapeutic strategy is to down-regulate the oncogene over-expression on his root, by targeting the gene itself.

The proximal promoter of EGFR has large guanine-cytosine-rich regions with several potential iMs forming sites. In the region comprising 500 nucleotides upstream of the TSS, we identified two potential iM forming sequences having four runs with at least three cytosines, EGFR-272 and EGFR-37, located at 272 and 37 nucleotides upstream of the TSS, respectively. In previous works, we characterized the conformational features of both the guanine and cytosine-rich strands of EGFR-272 [36,37]. We observed that the guanine-rich strand folds into two different G4s in thermodynamic equilibrium under our experimental condition [37]. Conversely, the cytosine-rich strand folds into a single iM structure that can be targeted by small molecules [36].

Here we report comprehensive calorimetric and spectroscopic analyses of the iM assumed *in vitro* by the cytosine-rich strand of EGFR-37.

Materials and methods

Materials

All tested oligonucleotides (sequences reported in Table 1) were synthesized and purified by Eurogentec (Liege, Belgium). They were dissolved in Milli-Q water to obtain 1 mM stock solutions and strand concentrations were determined from the absorbance at 260 nm. Before use, all DNA samples were melted for 5 minutes at 95 °C and then slowly cooled to room temperature.

Table 1. Sequences of oligonucleotides used in this study

EGFR-37	5’-CCCTCCTCCTCCCGCCCTGCCTCCCC-3’
EGFR-37-FAM	5’-CCCTCCTCCTCCCGCCCTGCCTCCCC-FAM-3’
EGFR-37MUT	5’-CCCTCCTCCTCCCACCCTGCCTCCCC-3’
EGFR-37MUT-FAM	5’-CCCTCCTCCTCCCACCCTGCCTCCCC-FAM-3’
22 bases Marker	5'-GGATGTGAGTGTGAGTGTGAGG-3'

Electrophoretic mobility shift assay

Samples were prepared at 4 µM or 200 µM DNA (strand concentration) in 10 mM Na-cacodylate pH 5.5. They were melted at 95 °C and slowly cooled to room temperature before loading them on the gel (250 ng DNA/lane). DNA electrophoresis was performed on a 15% native polyacrylamide gel (acrylamide/bis-acrylamide 19:1) in 1X TAE (40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 5.5). The run was performed at 25 °C for 2 hours by applying a voltage of 15 V/cm. Gels were stained by soaking them in a 1X Sybr Green II in 1X TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA) solution. DNA bands were visualized on a Geliance system (Roche).

S1 footprinting

Reaction mixtures were prepared using 3’- FAM-labelled DNA (100 ng/µL) in 10 mM Na-cacodylate, 4.5 mM $ZnSO_4$, pH 5.5. They were melted for 5 minutes at 95 °C and then slowly cooled to room temperature. After the annealing step, 1 U/µL of S1 endonuclease (Promega) was added. Samples were incubated at 25 °C at different times and the reactions were stopped by adding 2.5 µL of the reaction mixture to 9 µL of stop solution (80% formamide, 60 mM EDTA). Samples were melted for 5 min at 95 °C, chilled on ice, and then loaded on a 20% polyacrylamide (acrylamide/bis-acrylamide 19:1) denaturing gel, 7 M urea, 1x TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA). Reaction products were visualized on a Geliance system (Roche).

NMR spectroscopy

^{1}H NMR spectra were recorded on a Bruker (Billerica, MA, USA) DMX 600 MHz spectrometer, equipped with a 5 mm TXI probe XYZ-Gradient at 25 °C. Samples were prepared at 150 µM DNA in 10 mM Na-phosphate, pH 5.5 with 10% D_2O. Before data acquisition, they were melted for 5 minutes at 95 °C and then slowly cooled down to room temperature. Suppression of the water signal was achieved by applying WATERGATE pulse sequence. The NMR spectrum was processed and analyzed by using TOPSPIN software.

Thermal difference spectra

Thermal difference spectra were obtained by subtracting the DNA UV-Vis spectrum acquired at 1°C from the one recorded at 85°C, thus below and above the oligonucleotide melting temperature, respectively. The experiments were performed in 10 mM Na-cacodylate pH 5.5.

Circular dichroism

CD spectra were acquired using a Jasco J 810 spectropolarimeter equipped with a Peltier as a temperature controller device in a 1 cm length quartz cuvette. Signals were reported as Δε.
pH-titration experiments were performed by adding HCl to 4 µM DNA samples in 10 mM Na-cacodylate pH 8.5 at 298.15 K. The CD signal recorded at 287 nm was fitted with equation 1 according to the Hill formalism:

$$\Delta\varepsilon = \frac{a + b\ 10^{n(pH_T - pH)}}{1 + 10^{n(pH_T - pH)}} \quad (1)$$

where n is the Hill coefficient, pH_T is the pH of the middle transition, a and b are the Δε of the unfolded and folded species, respectively.
Melting and annealing experiments were performed by setting ± 20 K h^{-1} temperature slopes and recoding the spectra every 2 K in the 230-330 nm wavelength range. The CD signal recorded at 287 nm during the melting and annealing steps was fitted with equation 2 according to Van't Hoff's formalism:

$$\Delta\varepsilon = \frac{a + b\ e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right)}}{1 + e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right)}} \quad (2)$$

where T is the temperature (K), ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}), T_m is the melting temperature (K), R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}) and a, b are the molar ellipticities of the unfolded and folded species, respectively.

Differential scanning calorimetry

Differential scanning calorimetry experiments were performed on a Microcal VP-DSC with cells of 502.7 µL in the 1-80 °C temperature range at stated heating-cooling rates. Multiple water-water, buffer-water, and buffer-buffer scans were performed before the analysis to derive the baseline thermogram and to check there were no heat exchanges due to the buffer in the set experimental condition. The measurements were performed in 10 mM Na-cacodylate pH 5.5 whose ionization in water is characterized by a ΔG^{o} of 35.8 kJ

mol^{-1} and a ΔH^{o} of -3.0 kJ mol^{-1}. Therefore, a little contribution from the protonation of the buffer to the enthalpy of folding could influence the measurement of -3.0 kJ mol^{-1} for each proton involved in the formation of the CC^{+} base pairs. Samples were prepared at 200 µM DNA concentration in the required buffer. Data were reported as molar excess of heat capacity (ΔCp) as a function of the temperature. Thermograms were fitted according to equation 3:

$$\Delta Cp = \frac{\Delta H^{o2}\, e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}}{R\,T^{2}\left(1+e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}\right)^{2}} \quad (3)$$

where T is the temperature (K), ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}), T_m is the melting temperature (K), and R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}).

Global fitting analysis

Global fitting analyses were performed on all the datasets derived from CD and DSC experiments according to the simplest model mechanism that successfully described the folding process:

$$U + nH^{+} \rightleftarrows H_nF^{n+} \quad (4)$$

$$\Delta G^{o}(T) = -RT\left(\frac{[H_nF^{n+}]}{[U][H^{+}]^{n}}\right) \quad (5)$$

where U is the unfolded species, H_nF^{n+} is the folded species, n is the number of protons recruited for the folding, T is the temperature (K), $\Delta G^{o}(T)$ is the standard Gibbs free energy change referred to T and R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}).

Worth noting that in a DNA folding process, it is more appropriate to consider n as the Hill coefficient reflecting cooperativity rather than the number of binding sites. Moreover, it should be taken into account that this model could fit iM folding when the pH is ≥ 5.5, at lower pHs one should consider that cytosines are already protonated and there is no uptake of protons from the buffer.

Thermodynamic parameters such as $\Delta G^{o}(298.15\ K)$, ΔH^{o}, and n were kept as shared fitting parameters.

The dataset corresponding to CD dependence upon pH, temperature (melting and annealing), and DSC thermograms (melting and annealing) were simultaneously fitted according to equations 6, 7, and 8 respectively:

$$\Delta\varepsilon = \frac{a + b\, e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - n\,pH\ln(10)}}{1 + e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - n\,pH\ln(10)}} \quad (6)$$

$$\Delta\varepsilon = \frac{a + b\, e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - \frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_o}\right) - n\,pH\ln(10)}}{1 + e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - \frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_o}\right) - n\,pH\ln(10)}} \quad (7)$$

$$\Delta Cp = \frac{\Delta H^{o\,2}\, e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - \frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_o}\right) - n\,pH\ln(10)}}{R\,T^{2}\left(1 + e^{-\frac{\Delta G^{o}(T_o)}{R\,T_o} - \frac{\Delta H^{o}}{R}\left(\frac{1}{T} - \frac{1}{T_o}\right) - n\,pH\ln(10)}\right)^{2}} \quad (8)$$

where n is the Hill coefficient, T is the temperature (K), T_o is 298.15 K, $\Delta G^{o}(T_o)$ is the standard Gibbs free energy change of folding (kJ mol^{-1}) referred to T_o, ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}), R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}) and a and b are the Δε of the unfolded and folded species, respectively.

ΔH^{o} was considered a temperature-independent parameter (ΔCp=0).

Considering the unimolecular model mechanism, instead, the apparent standard Gibbs free energy change is pH-dependent:

$$U \rightleftarrows F \quad (9)$$

$$\Delta G^{o}_{app}(T, pH) = -RT\left(\frac{[F]}{[U]}\right) \quad (10)$$

where U is the unfolded species, F is the folded species, T is the temperature, $\Delta G_{app}{}^{o}(T, pH)$ is the apparent standard Gibbs free energy change referred to T and pH, and R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}).

$\Delta G_{app}{}^{o}(T, pH)$ was derived at different temperatures ad pHs according to equation 11:

$$\Delta G^{o}_{app}(T, pH) = \Delta G^{o}(T_o)\frac{T}{T_o} + \Delta H^{o}\left(1 - \frac{T}{T_o}\right) + nRTpH\ln(10) \quad (11)$$

where n is the Hill coefficient, T is the temperature (K), T_o is 298.15 K, $\Delta G^{o}(T_o)$ is the standard Gibbs free energy change of folding (kJ mol^{-1}) referred to T_o, ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}) and R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}).

Python scripts based on NumPy and SciPy libraries were developed to perform data analysis and global fitting.

Singular Value Decomposition

Multiple wavelength CD experiments were analyzed by Singular Value Decomposition (SVD). For each experiment, the dataset was built as a matrix A in which A[i, j] is the Δε at a given i wavelength and j temperature or pH. A matrix was decomposed, into a product of three matrixes: USV^T where U is an orthogonal matrix in which column vectors are the eigenvectors of AA^T, V is an orthogonal matrix in which column vectors are the eigenvectors of A^TA and S is a rectangular diagonal matrix which values are the eigenvalues of both AA^T and A^TA.
The A matrix can be well approximated according to equation 12:

$$A \simeq \sum_{k=1}^{n} U_k S_k V_k^T \quad (12)$$

where n is the number of significant optical components contributing to the signal.
Only the species (k) that simultaneously have autocorrelation coefficient of U_k and $V_k \geq 0.75$ and the maximum absolute value of $U_k S_k V_k^T \geq 10^{-3}$ cm^{-1} were considered.
Python scripts based on NumPy and SciPy libraries were developed to perform this analysis.

Results

Characterization of EGFR-37 folding

The chiroptical properties of DNA change upon its folding into secondary structures, and this can be followed by circular dichroism (CD). In particular, the antiparallel orientation of the strands in the iM structure correlates with a positive CD signal around 290 nm and a negative one around 260 nm. Therefore, the ability of EGFR-37 to assume iM structures was preliminarily evaluated following the CD signal in the 330-230 nm range while moving from basic (pH 8.5) to acidic (pH 4.0) conditions upon progressive additions of HCl to the DNA solution (Figure 1). The spectrum acquired at pH 5.5 was more intense than the ones recorded at basic conditions and it showed a positive signal at 287 nm and a negative one at 264 nm. This optical fingerprint corresponds to the one expected for an iM. The folding of EGFR-37 into an iM in acidic conditions was further supported by the UV-Vis thermal difference spectrum (TDS) (Figure S1). Indeed, also this spectroscopic profile varies according to the specific DNA secondary structures. The TDS spectrum of EGFR-37 at pH 5.5 showed a negative signal at 295 nm and a positive one at 240 nm, in line with those reported for iM structures [38].

Additionally, native polyacrylamide gel electrophoresis performed in TAE pH 5.5 indicated that EGFR-37 samples in the 4-200 µM concentration range, constantly run as a single band with electrophoretic mobility slightly greater than the 22 residues DNA marker one (Figure S2). This evidence supported that EGFR-37 folds into intra-molecular iM structures in acidic conditions.

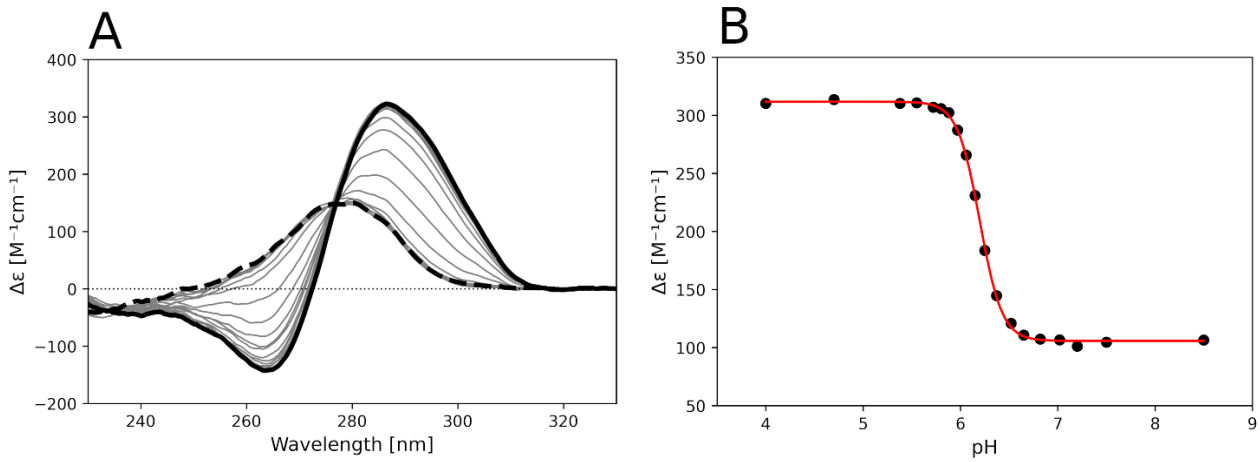

Figure 1: CD titration of EGFR-37 with HCl. Panel A: CD spectra of 4 µM EGFR-37 in 10 mM Na-cacodylate at 25°C from pH 8.5 (black solid line) to pH 4.0 (black dashed line) upon addition of HCl. Panel B: CD signals at 287 nm (black dots) plotted as a function of pH and fitted according to equation 1 (red solid line).

The CD spectra acquired at variable pH showed an iso-dichroic point at 277 nm, thus supporting the presence of only two optical species during the protonation-induced folding process. This was further confirmed by the analysis of the optical components performed on the U, S, and V matrices derived by SVD (Table S1).

The lack of intermediate species also indicated that under these experimental conditions, the DNA folding process was highly cooperative. On these bases, the CD data acquired at 287 nm (maximum signal intensity for an iM) were fitted according to equation 1, based on the Hill formalism. From this analysis, we derived the pH of the middle transition (pH_T) and the Hill coefficient (n) which resulted in 6.2 ± 0.1 and 3.8 ± 0.1, respectively. The pH_T at 298.15 K was significantly higher than the pKa of cytosine N3 which is ≃ 4.5 referred to 298.15 K [21]. This shift was expected since, beyond the protonation of cytosines, other energetic contributions (base pairing, hydrogen bonding, etc) play part in determining the ΔG^e(298.5 K) of the iM folding. As it concerns the n coefficient, its value indicated that the cooperativity of the folding process was markedly positive, a result that fits with the above-described absence of long-living intermediated species. In these conditions, the Hill coefficient can be considered to approach the number of binding sites, in this case corresponding to the recruited protons responsible for the structural rearrangement [39]. Nevertheless, it should be considered that by the Hill equation, non-specific interactions of H3O+ cations with other DNA domains cannot be ruled out. However, since our data were acquired in the 4.5 -8.0 pH range and the pKa of phosphate groups is ≃ 2, we can safely consider that n did

not comprise interactions of protons with the DNA backbone. Thus, n is a reliable estimation of the number of protons directly involved in the iM formation corresponding to the number of CC^+ base pairs.

To better define the EGFR-37 folding model, we analyzed the temperature dependency of the iM of EGFR-37 at pH 5.5 by following the melting and annealing processes by CD (Figure S3). The process resulted to be fully reversible, and we confirmed the presence of only two significant optical components in the solution by SVD (Table S1). Thus, it was possible to fit the data corresponding to the CD signal at 287 nm according to equation 2 based on Van't Hoff's formalism (Figure 2, S4). Through this analysis, we derived the melting temperature (43.9 ± 0.1) °C and the associated ΔH° (-238 ± 5) kJ mol^{-1} reported in Table 2.

Due to the monomeric feature of the iM assumed by EGFR-37, it was possible to integrate these data by DSC experiments applying the same heating-cooling rate (± 20 C h^{-1}) at 200 µM DNA samples. The process was confirmed to be fully reversible under these conditions as well. Additionally, melting and annealing temperatures as well as the enthalpic contributions derived by CD and DSC well overlapped (Table 2).

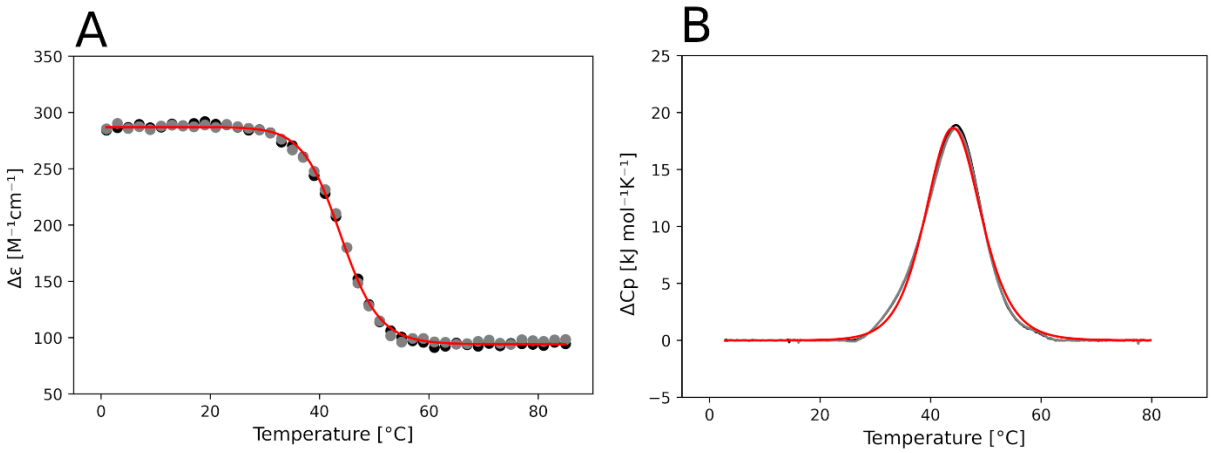

Figure 2: EGFR-37 melting and annealing in 10 mM Na-cacodylate pH 5.5 performed at ± 20 K h^{-1} heating-cooling rate. Panel A: CD signals at 287 nm of CD melting (black dot) and annealing (grey dots) of 4 µM EGFR-37 and fitting curve (red solid line) according to equation 2. Panel B: DSC curves acquired during the melting (black solid line) and annealing (grey solid line) scans of 200 µM EGFR-37 and fitting curve (red solid line) according to equation 3.

Table 2. Thermodynamic parameters of EGFR-37 folding in 10 mM Na-cacodylate pH 5.5 referred to 298.15 K as derived by different data set analyses.

	Hill coefficient	**ΔG° kJ mol^{-1}**	**ΔH° kJ mol^{-1}**	**$-T\Delta S^{\circ}$ kJ mol^{-1}**	**T_m (pH 5.5)**	**pH_T**
CD HCl-titration	3.8 ± 0.1					6.2 ± 0.1
CD melting			-238 ± 5		43.9 ± 0.1	
CD annealing			-238 ± 5		43.9 ± 0.1	
DSC melting			-250 ± 5		44.1 ± 0.1	
DSC annealing			-250 ± 5		44.1 ± 0.1	

Global analysis	3.8 ± 0.1	-134 ± 5	-250 ± 5	116 ± 5		

This result further supported that the EGFR-37 folding process was concentration-independent as expected for an intra-molecular structure and occurred with no formation of other metastable states. Therefore, we used the ensemble of CD (HCl-titration, thermal melting, and annealing) and DSC experiments to run a global fitting according to equations 6, 7, and 8, respectively (Figure S5). In the analysis n, $\Delta G^{o}(298.15\ K)$, ΔH^{o} were kept as shared fitting parameters. It is appropriate to specify that, in equations 7 and 8, ΔH^{o} was considered independent from the temperature ($\Delta Cp=0$) and pH. The assumption of $\Delta Cp=0$ is reasonable considering that the reported ΔCp values for i-Motif structures are very small [16]. Concerning the independence from pH, performing the heating-cooling experiments at pHs ≥ 5.5, the enthalpic contribution the derives from the protonation of the cytosines is already at its maximum and it does not change by increasing the pH because all the couples of cytosines that form CC^{+} base pairs in the iM folded state have to be protonated [16]. On the other hand, working at pHs in 4.5-5.5 range we should consider the reduction of this contribution. Furthermore, at pHs ≤ 4.5 we should consider the enthalpy contribution deriving from the deprotonation of the cytosines forming CC^{+} base pairs. On these bases, we performed the globally fitting analysis, and the derived parameters, which are consistent for pHs ≥ 5.5, are well compared to those obtained by the individual analyses of every single dataset. This approach allowed us to further test the model mechanism of folding and derive all the associated thermodynamic parameters. The distribution of the folded and unfolded species through pH and temperature are reported in Figure 3.

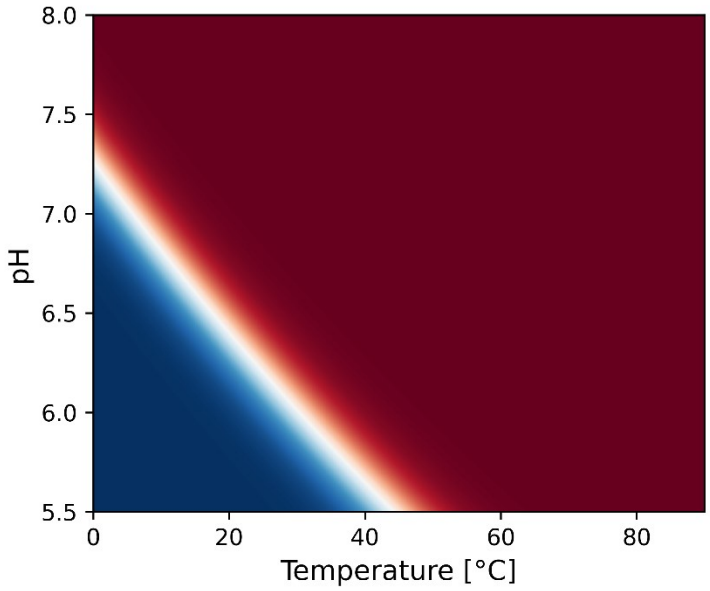

Figure 3: Distribution of the unfolded (red surface) and the folded (blue surface) species of EGFR-37 as a function of temperature and pH.

Noteworthy, the ΔH° resulted in (-250 ± 5) kJ mol^{-1} (Table 2), which was significantly lower than those previously reported for iM models with 6 CC^{+} base pairs [17]. Moreover, the Hill coefficient indicated that only 4 protons were involved in the iM folding whereas the *in-silico* prediction indicated that EGFR-37 was able to form 6 CC^{+} base pairs.

To clarify these discrepancies, we mapped the cytosines paired within the iM structure by performing S1 cleavage footprinting. This enzyme is a nuclease that cuts only single-stranded residues and well performs under acidic conditions. Thus, it can be efficiently exploited to identify unpaired nucleotides within a folded iM structure. The enzymatic cleavage pattern of EGFR-37 at pH 5.5 is reported in Figure 4. It showed strong cleavage sites at T4, T7-C8, and G19-C20 along with the neighboring residues T18 and C21 that were cleaved although with lower efficiency. This pattern allowed us to locate all these residues in the loops of the iM. Additionally, most of the cytosines at 3' and 5' terminals (C1, C2, C25, and C27) were significantly cleaved thus ruling out their involvement in stable CC^{+} base pairings. Therefore, it resulted that the four runs of cytosines involved in the iM core should be C5-C6, C11-C12-C13, C15-C16-C17, and C23-C24. However, it must be reminded that iM structures have a conserved folding topology corresponding to two parallel duplexes intercalated in antiparallel orientation, from which it derives that cytosines of the first and second runs must pair with those of the third and fourth runs, respectively. Consequently, C5-C6 paired with two cytosines among C15-C16-C17 and, similarly, C23-C24 with two among C11-C12-C13. As a result, within the iM core, only 4 CC^{+} base pairs can be present.

Notably, T10 and T22 were fully protected from S1 cleavage. This suggested their involvement in the formation of a non-canonical TT base-pair, a frequently observed capping module for iM cores [23]. In our sequence, this element would extend the CC+ pairing occurring between the second and fourth cytosine runs. This structural feature implied that C11 and C12 pair to C23 and C24, respectively. According to this folding model, the second loop should comprise C13, G14, and C15, although it was not efficiently cleaved by the enzyme.

However, C3 was protected from S1 whereas C2 and T4 were cleaved (Figure 4), and based on that, we attributed C3-G14 to the formation of an additional GC base pair.

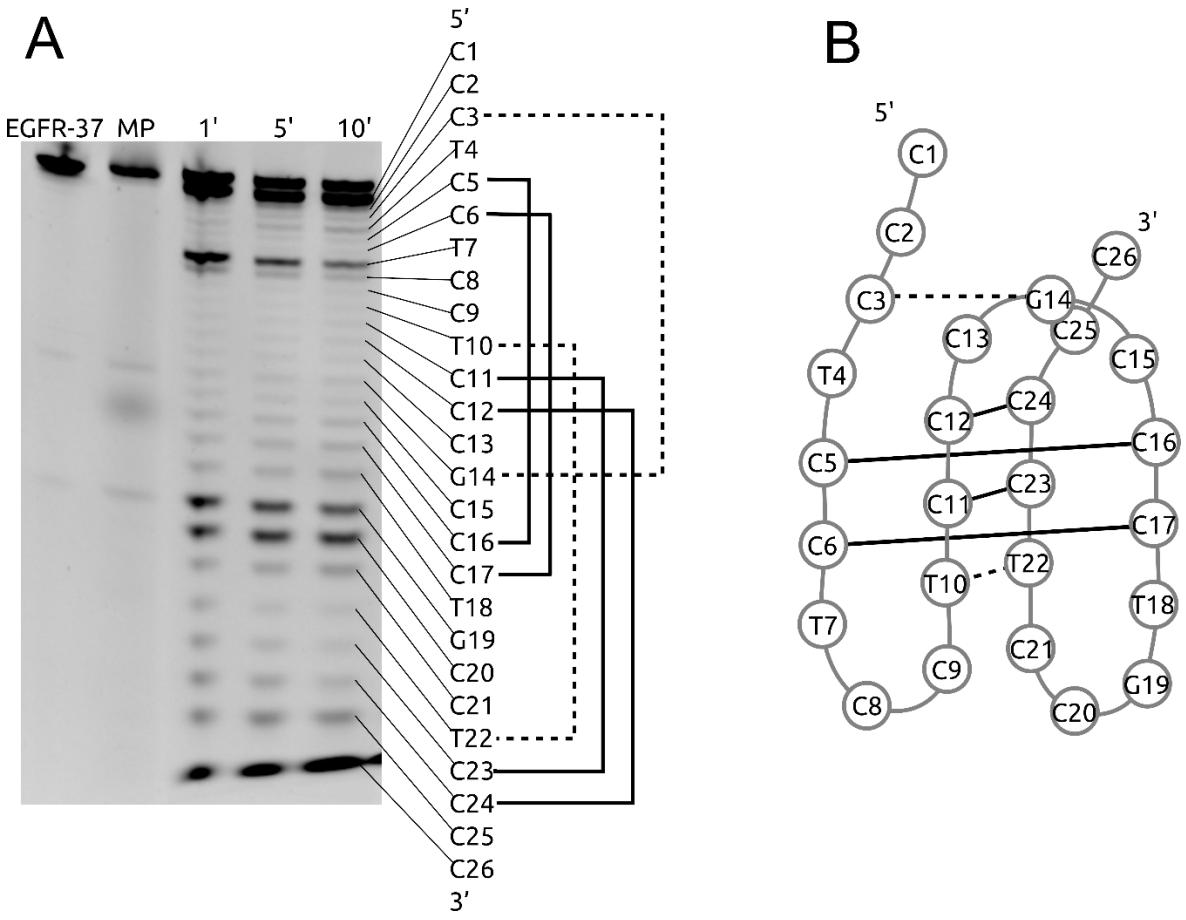

Figure 4: Folding topology of EGFR-37. Panel A: S1 footprinting of 100 ng/µL EGFR-37-FAM in 10 mM Na-cacodylate, 4.5 mM $ZnSO_4$, pH 5.5. In line 1, EGFR-37-FAM reference sample. In line 2, EGFR-37-FAM purine marker. In lines 3, 4, and 5 EGFR-37-FAM after 1, 5, and 10 min of cleavage reaction, respectively. Panel B: schematic representation of the iM folding topology.

To verify the presence of the TT and GC additional interactions we performed 1D ^{1}H NMR. The imino region of the spectrum showed two well-solved signals at 11.18 and 11.10 ppm deriving from the H3 of thymines involved in the formation of the TT base pair (Figure 5). The signals at 15.54 ppm and 14.84 ppm belonged to H3 of cytosines in CC^+ pairings, but, the overlap of contributions prevented us from clearly deriving the effective number of CC^+.

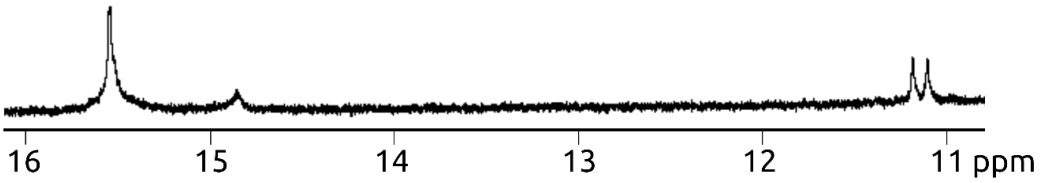

Figure 5: ^{1}H NMR imino region of 150 µM EGFR-37 acquired at 25 °C in 10 mM Na-phosphate, pH 5.5

Worth of mention that we did not detect any signal belonging to H1 of the guanine involved in GC interaction. However, distinctly from the TT cupping element, the G14-C3 base pair does not directly form on the top of a CC^+, thus it is highly exposed to the solvent. This condition favors a fast exchange rate of the G14 imino proton with the water thus preventing its detection by NMR.

Mutated sequences of EGFR-37

To validate the presence of the GC base pair in the iM of EGFR-37, we designed a mutated sequence, EGFR-37MUT, in which G14 was substituted with adenine, and we performed the S1 footprinting on the mutated sequence (Figure 6).

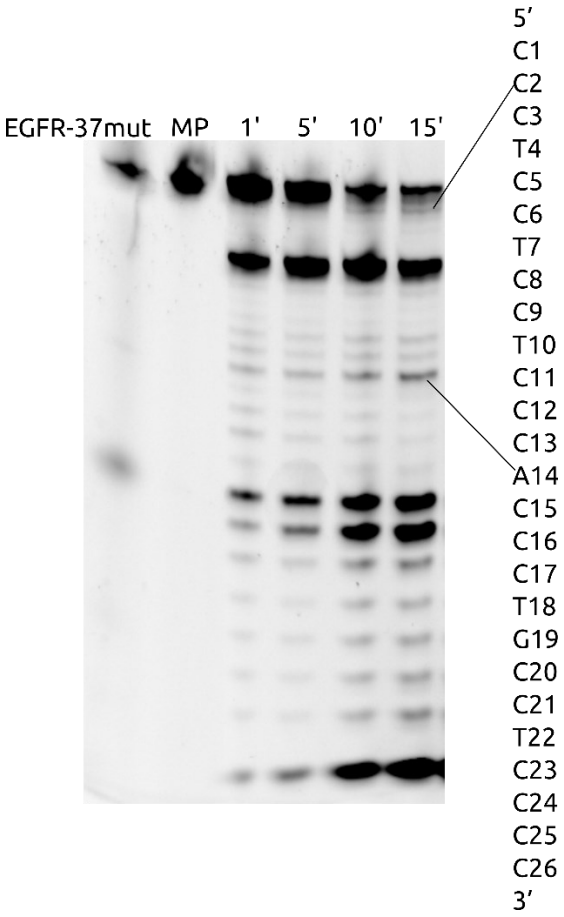

Figure 6: S1 footprinting of 100 ng/µL EGFR-37MUT-FAM in 10 mM Na-cacodylate, 4.5 mM $ZnSO_4$, pH 5.5. In line 1, EGFR-37MUT-FAM reference sample. In line 2, EGFR-37MUT-FAM purine marker. In lines 3, 4, 5, and 6 EGFR-37MUT-FAM after 1, 5, 10, and 15 min of cleavage reaction, respectively.

As can be observed in Figure 6, cleavage sites were detected at C3 and A14 of EGFR-37MUT, thus confirming they were more accessible to S1 enzymatic cleavage.

Worth noting that HCl-titration, melting, and annealing experiments followed by CD confirmed that the EGFR-37MUT folds into an iM comparably to EGFR-37. Indeed, the recorded CD signals of the two folded sequences almost overlapped (Figure S6) as well as the T_m e pH_T. Also, the TDS of EGFR-37MUT showed the

same shape and intensity as the one acquired for EGFR-37 with a positive pick at 240 nm and a negative one at 290 nm (Figure S1). The only difference rested in the shoulder at 260 nm which slightly decreased in intensity for the mutated sequence and notably, this is the wavelength range in which a GC base pair formation mainly contributes to the hyperchromic effect [38].

Again, SVD statistical analysis sustained the presence of only two significant optical components in all the CD experiments (Table S2). However, DSC experiments performed at 200 µM DNA concentration, showed that the heating and cooling curves did not overlap. In particular, multiple transitions in the melting scan were recorded (Figure 7). As supported by gel electrophoresis, they were associated with the formation of inter-molecular species (Figure S2) promoted at high EGFR-37MUT concentrations.

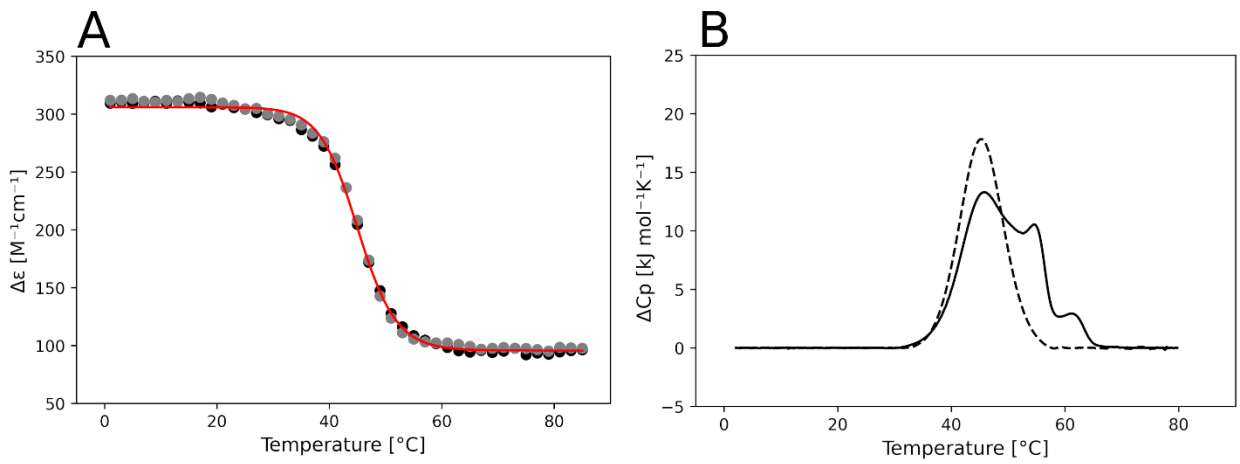

Figure 7: EGFR-37MUT melting and annealing in 10 mM Na-cacodylate pH 5.5 performed at ± 20 K h^{-1} heating-cooling rate. Panel A: CD signals at 287 nm of CD melting (black dot) and annealing (grey dots) of 4 µM EGFR-37MUT and fitting curve (red solid line) according to equation 2. Panel B: DSC curves acquired during the melting (black solid line) and annealing (black dashed line) scans of 200 µM EGFR-37MUT.

This was an interesting difference between EGFR-37 and EGFR-37MUT, which can only be attributed to the substitution of G14 with an adenine.

Ionic strength further stabilizes the multimeric species, indeed, the DSC profiles recorded in 50 mM Na-cacodylate showed an increase in the enthalpy associated with the multimeric species in the melting scans, and the hysteresis remained even lowering the scanning rate to 5 K h^{-1} (Figure S7).

This behavior prevented us to derive the thermodynamic parameters from DSC data, thus, we limited the global fitting to the CD experiments, which were performed at 4 µM DNA concentration (Figure S8, S9).

In this condition, the molar fraction distribution of the folded and unfolded species (Figure S10) and, consistently, the thermodynamic parameters (Table 3) of EGFR-37 and EGFR-37MUT were well compared.

Worth to note that, if paired with the common thermodynamic fingerprint of the two sequences, this result confirmed that the G14-C3 pairing in EGFR-37 is a weak interaction.

Table 3. Thermodynamic parameters for the iM folding of 4 µM EGFR-37 and EGFR-37MUT in 10 mM Na-cacodylate pH 5.5 referred at 298.15 K as derived from global fitting analyses of spectroscopic data.

	Hill coefficient	ΔG° kJ mol^{-1}	ΔH° kJ mol^{-1}	$-T\Delta S^{\circ}$ kJ mol^{-1}	Tm (pH 5.5)	pH_T
EGFR-37	3.8 ± 0.1	-134 ± 5	-250 ± 5	116 ± 5	44.1 ± 0.1	6.2 ± 0.1
EGFR-37MUT	3.8 ± 0.1	-134 ± 5	-247 ± 5	113 ± 5	43.9 ± 0.1	6.2 ± 0.1

Ionic strength induces the formation of intermolecular species.

It was surprising to find that the two tested sequences folded into analog intra-molecular iM structures from a thermodynamic point of view, but with a divergent attitude towards multimerization. Indeed, whether the GC interaction seemed to play a negligible effect on the iM features, still it appeared that the substitution of G14 with an adenine largely impacted the formation of multimeric species.

Since the pairing of multiple strands can be promoted by increasing the ionic strength, we decided to test the behavior of EGFR-37 in the presence of increasing salt concentrations by performing melting and annealing experiments at 50 mM Na-cacodylate with CD and DSC. At this ionic strength CD and DSC experiments, recorded at different DNA concentrations, provided different pictures (Figure 8).

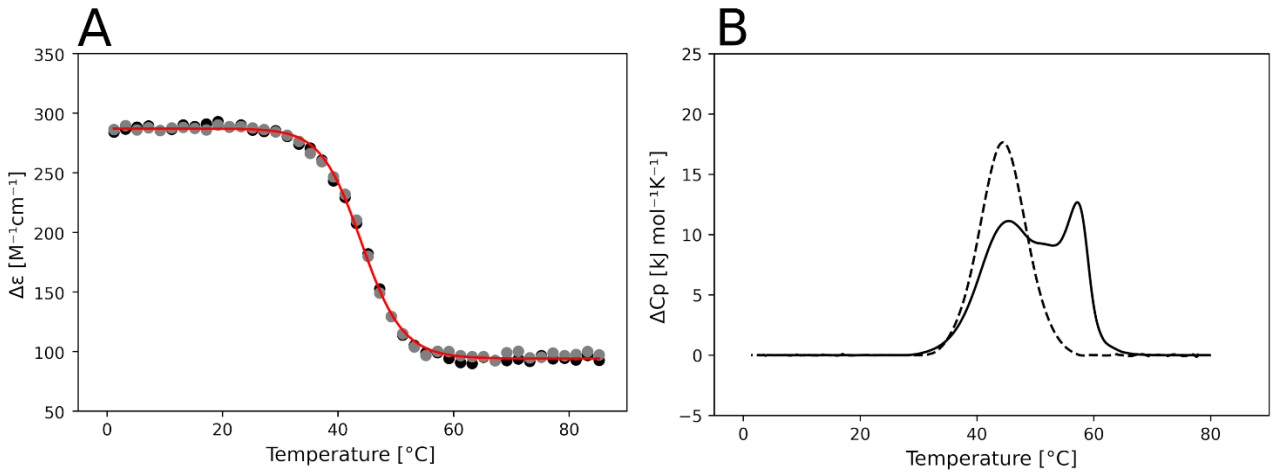

Figure 8: EGFR-37 melting and annealing in 50 mM Na-cacodylate pH 5.5 performed at 20 K h^{-1} heating-cooling rate. Panel A: CD signals at 287 nm of CD melting (black dot) and annealing (grey dots) of 4 µM EGFR-37 and fitting curve (red solid line) according to equation 2. Panel B: DSC curves acquired during the melting (black solid line) and annealing (black solid line) scans of 200 µM EGFR-37.

Working at 4 µM DNA concentration, CD data showed a single, fully reversible transition with a melting temperature of (44.1 ± 0.1) °C, a profile that was well compared to the one acquired at lower ionic strength. Conversely, when we increased the DNA concertation up to 200 µM, the reversibility of the

process was lost and additional transitions occurring at higher temperatures appeared along the melting scan. Nevertheless, during the annealing scan, a single peak was still recorded and it had the maximum at 44.5 °C, as well as the peak corresponding to the first transition in the melting scan. Thus, we attributed it to the intra-molecular iM structure of EGFR-37. Worth noting that, the melting temperature of the intra-molecular folding did not significantly increase from the one recorded in 10 mM Na-cacodylate. On the other hand, the inter-molecular foldings associated with the higher temperature transitions were significantly stabilized by the higher ionic strength.

Discussion

Here we performed an integrated thermodynamic characterization of an iM structure that was expected to form at the site located 37 nucleotides upstream of the transcription starting site of the *EGFR* oncogene. The sequence was selected to potentially form 6 CC^+ base pairs which should grant significant stability even under conditions approaching the physiological ones. However, our results are interestingly in contrast with the *in-silico* prediction. The thermodynamic parameters we derived through a global analysis of calorimetric and spectroscopic data indicated that the iM core is held by 4 CC^+, a conclusion supported also by electrophoretic experiments. This surprising discrepancy prompts us to reconsider that having 4 runs of 3 cytosines is not the only requirement to fold into an iM with 6 CC^+ base pairs.

Our data indicated that the maintenance of a significantly stable iM is supported by the presence of additional structural elements, in detail a TT and a GC base pair. The presence of a TT base pair that works as a cupping element was not unexpected since several reported high-resolution structures included it at the 3' or 5' terminal of the iM core [14,40,41]. The same role could be played by the GC which, in principle, should provide a more prominent effect. However, in our sequence, the occurrence of this base pairing is spaced out from the iM core thus preventing efficient stacking. This was in line with the fast exchange in the NMR and confirmed by the overlapping thermodynamic profiles of the EGFR-37 and EGFR-37MUT as well. This addresses the C3-G14 as a weak interaction. Remarkably, despite the lack of any iM stabilization roles by this novel motif, it turned out that the single residue mutation of the guanine at position 14 with an adenine greatly stabilizes inter-molecular species. It would be possible to argue that this is a direct consequence of the presence of adenine. However, we proved that an increment of the ionic strength can induce the same process also on the wild-type sequence while it does not significantly affect the stability of the intra-molecular iM.

These data indicate that a lot of information concerning the structural properties of iM is lacking. Here, we showed how the global analysis of calorimetric and spectroscopic data is a highly valuable strategy to obtain thermodynamic and structural information on iM foldings. This "low-resolution" approach is flexible

and reliable and can thus be easily extended to a wide panel of C-rich sequences. The output would help in better addressing the sequence requirements behind iM formation paving the way towards a more precise prediction of iM forming sites and, consequently, in their screening as a potential pharmacological target or in the set up of solid pH-sensible nanodevices.

Supplementary information

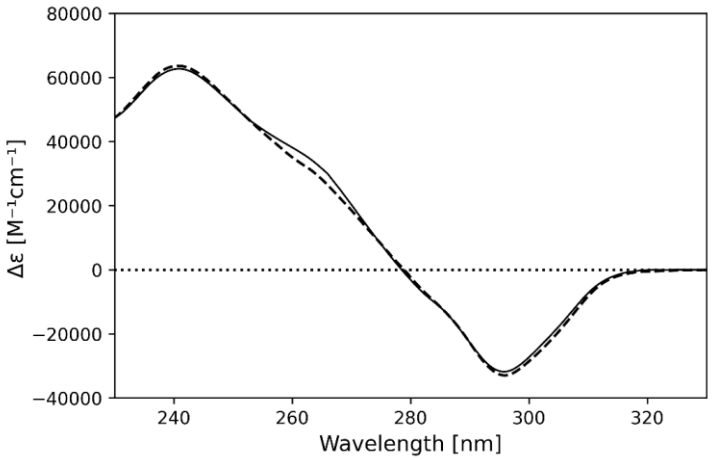

Figure S1: UV-Vis thermal difference spectra (TDS) of EGFR-37 and EGFR-37MUT. 4 µM EGFR-37 (black solid line) and EGFR-37MUT (black dashed line) in 10 mM Na-cacodylate pH 5.5

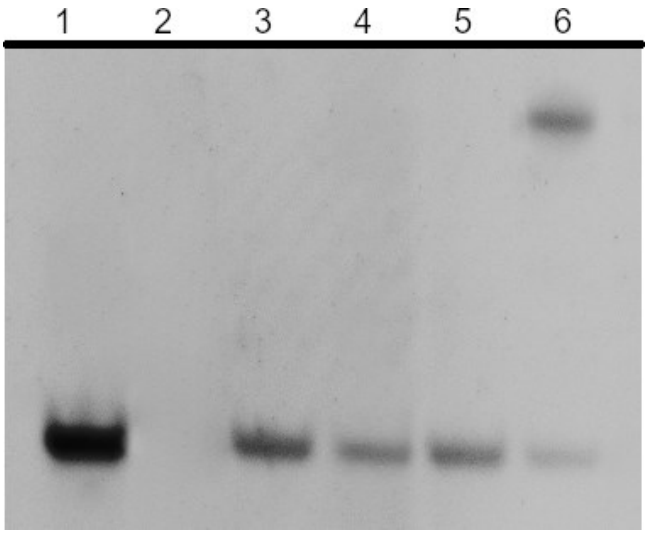

Figure S2: Native polyacrylamide gel electrophoresis of EGFR-37 and EGFR-37MUT in 10 mM Na-cacodylate pH 5.5 at different DNA concentrations. In line 1, 250 ng of 22 bases marker oligonucleotide. In line 3, 250 ng of 4 µM EGFR-37.

In line 4, 250 ng of 200 µM EGFR-37. In line 5, 250 ng of 4 µM EGFR-37MUT. In line 6, 250 ng of 200 µM EGFR-37MUT. Samples were melted for 5 min at 95 °C and then slowly cooled to room temperature before loading. The gel was run in TAE 1X pH 5.5 buffer.

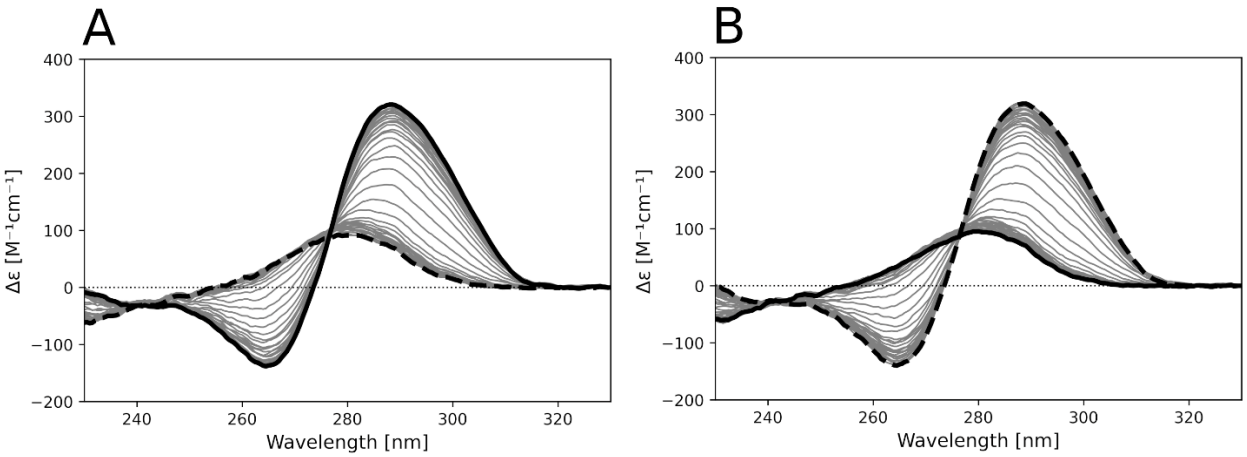

Figure S3: CD melting and annealing of EGFR-37 at ± 20 K h-1. Panel A: CD spectra of 4 µM EGFR-37 in 10 mM Na-cacodylate pH 5.5 from 1 °C (solid black line) to 85 °C (dashed black line). Panel B: CD spectra of 4 µM EGFR-37 in 10 mM Na-cacodylate pH 5.5 from 85 °C (solid black line) to 1 °C (dashed black line).

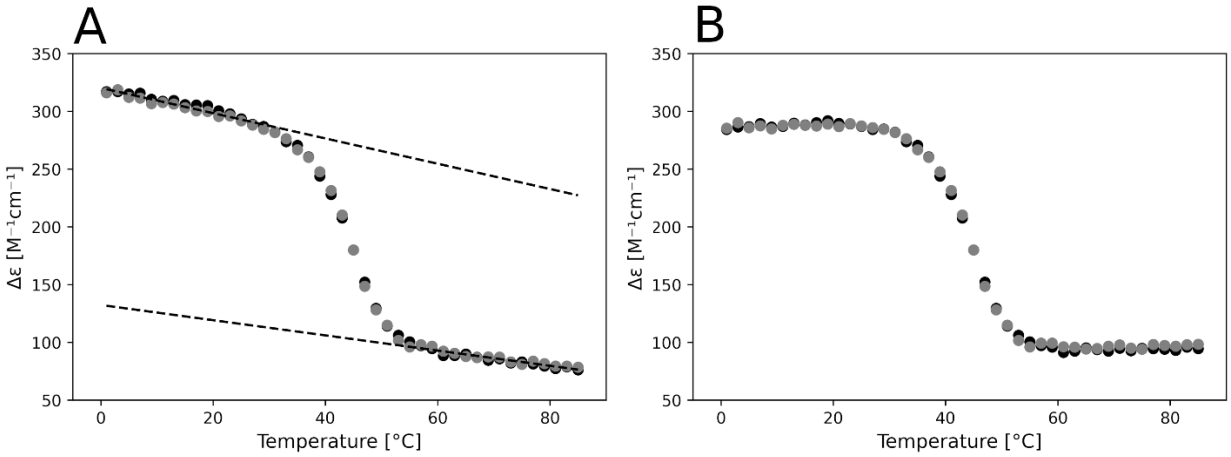

Figure S4: Drift correction of the CD signal at 287 nm of melting and annealing of EGFR-37 at ± 20 K h-1. Panel A: CD melting (black dots) and annealing (grey dots) and signal drifts (dashed lines). Panel B: CD melting (black dots) and annealing (grey dots) corrected from the drift.

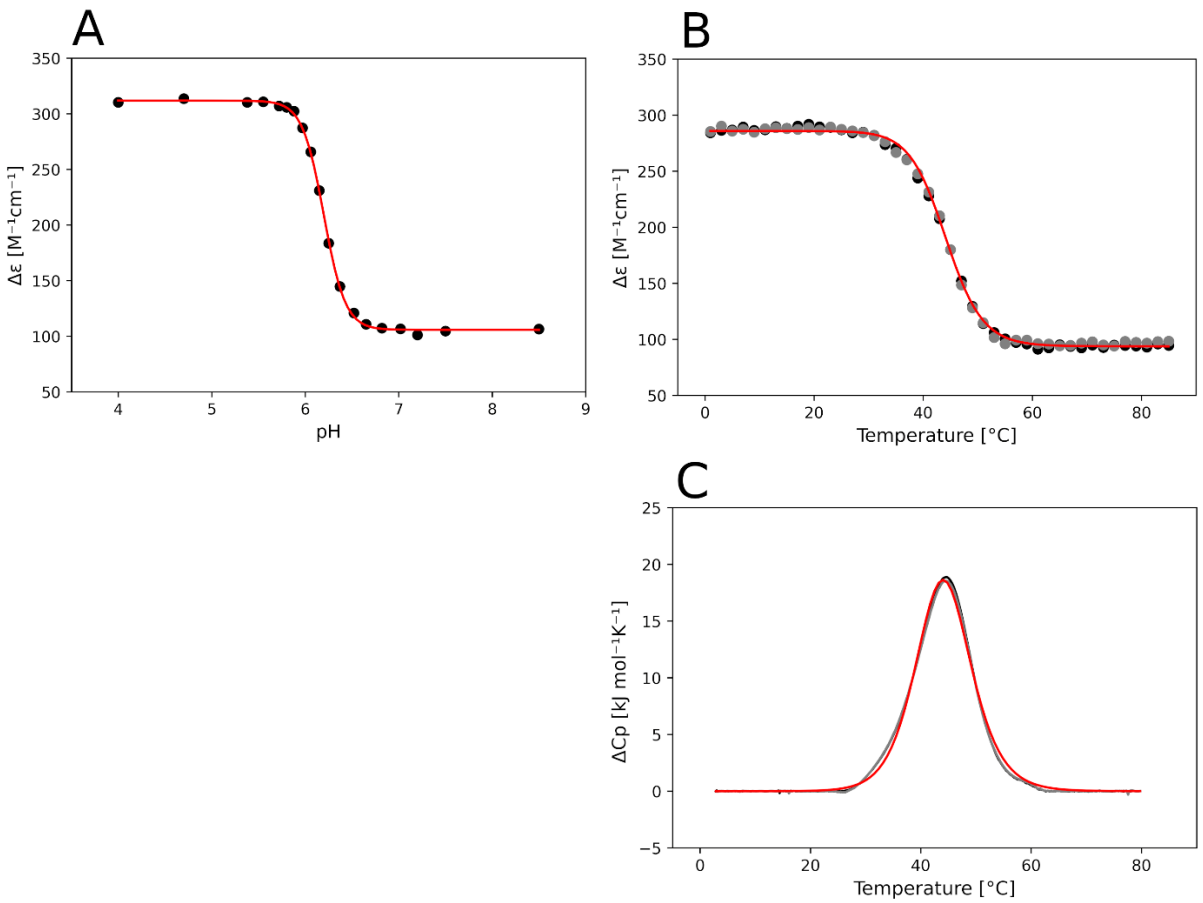

Figure S5: Global analysis of CD HCl-titration, CD melting-annealing, and DSC melting-annealing experiments performed on EGFR-37. CD signal at 287 nm of HCl-titration (black dots) in panel A, CD signal at 287 of melting (black dots) and annealing (grey dots) in panel B, and DSC melting (black solid line) and annealing (grey solid line) in panel C were globally fitted according to equation 6, 7 and 8, respectively. Fitting curves are reported as red solid lines. Thermodynamic parameters were kept as shared parameters.

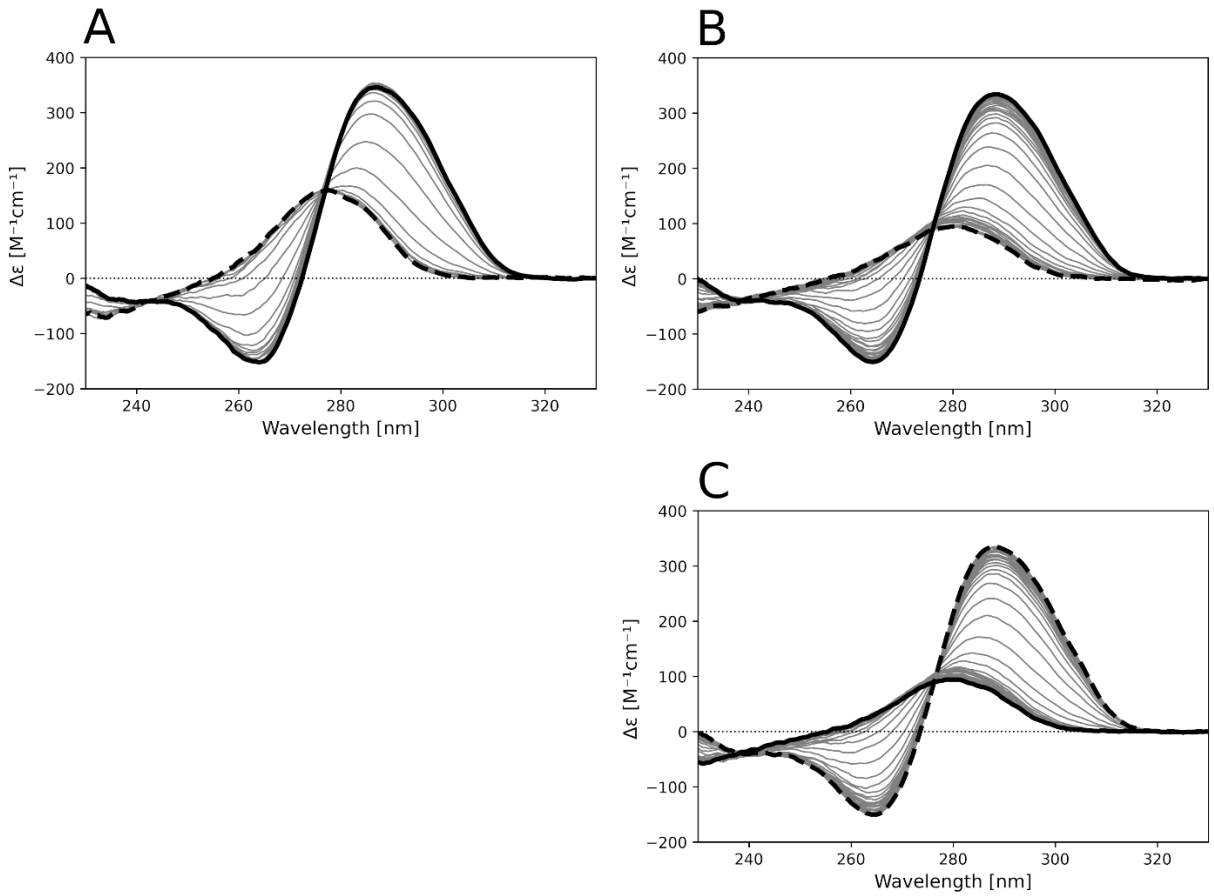

Figure S6: CD of EGFR-37MUT. CD spectra of 4 µM EGFR-37MUT in 10 mM Na-cacodylate at 25°C from pH 8.5 (black solid line) to pH 4 (dashed line) upon addition of HCl in panel A, CD spectra of 4 µM EGFR-37MUT in 10 mM Na-cacodylate pH 5.5 from 1 °C (solid black line) to 85 °C (dashed black line) in panel B. CD spectra of 4 µM EGFR-37MUT in 10 mM Na-cacodylate pH 5.5 from 85 °C (solid black line) to 1 °C (dashed black line) in panel C.

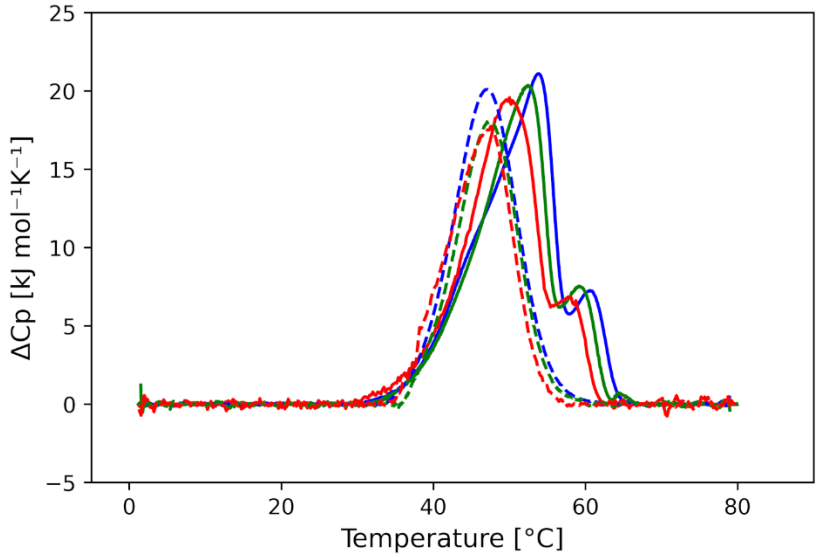

Figure S7: EGFR-37MUT melting and annealing acquired in 50 mM Na-cacodylate, pH 5.5. DSC curves acquired at different heating-cooling rates: + 20 K h-1 (blue solid line), - 20 K h-1 (blue dashed line), + 10 K h-1 (green solid line), - 10 K h-1 (green dashed line), + 5 K h-1 (red solid line) and - 5 K h-1 (red dashed line) scans of 200 µM EGFR-37MUT.

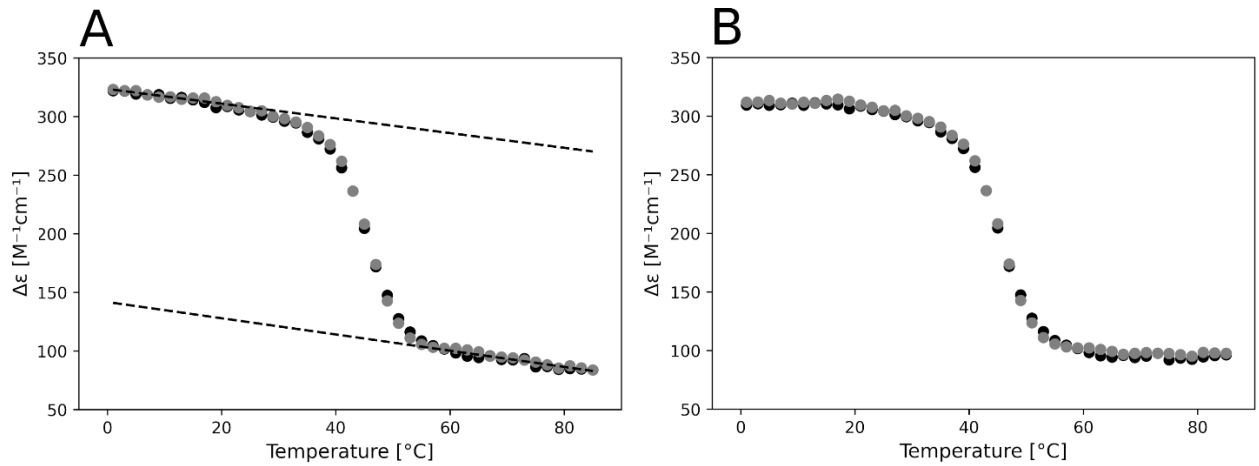

Figure S8: Drift correction of the CD signal at 287 nm of melting and annealing of EGFR-37MUT at ± 20 K h-1. Panel A: CD melting (black dots) and annealing (grey dots) and signal drifts (dashed lines). Panel B: CD melting (black dots) and annealing (grey dots) corrected from the drift.

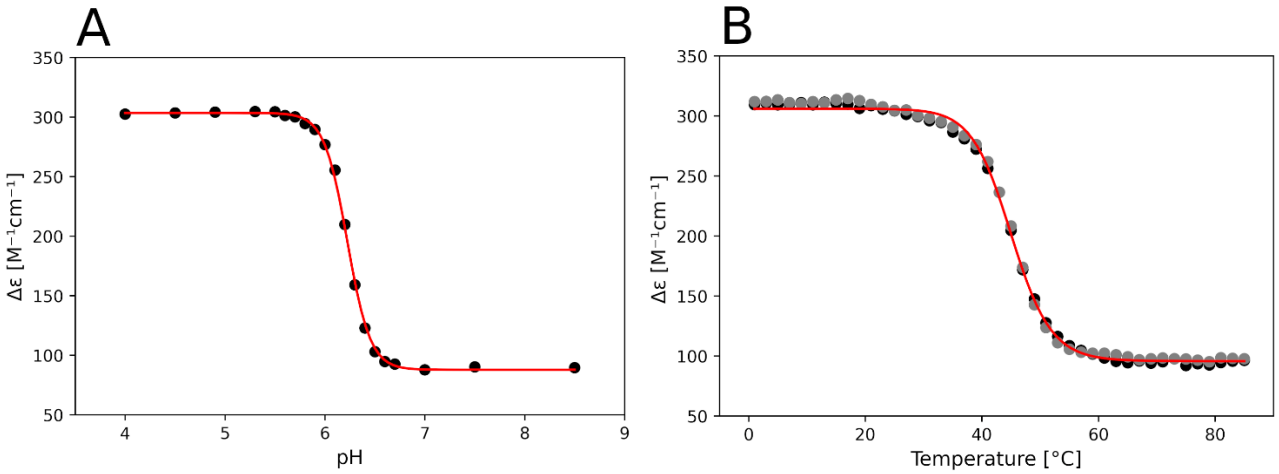

Figure S9: Global analysis of CD HCl-titration and CD melting-annealing experiments performed on EGFR-37MUT. CD signal at 287 nm (black dots) of HCl-titration in panel A and CD signal at 287 of melting (black dots) and annealing (grey dots) in panel B were globally fitted according to equations 6 and 7, respectively. Fitting curves are reported as red solid lines. Thermodynamic parameters were kept as shared parameters.

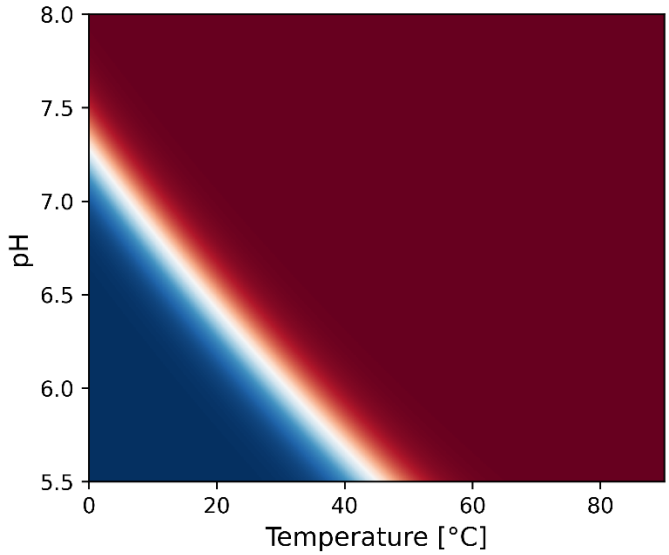

Figure S10: Distribution of the unfolded (red surface) and the folded (blue surface) species of EGFR-37MUT as a function of temperature and pH.

Table S1. Significant components analysis of the CD experiments performed on EGFR-37

	HCl-titration			Melting			Annealing		
Component	**U[a]**	**M[b]**	**V[c]**	**U[a]**	**M[b]**	**V[c]**	**U[a]**	**M[b]**	**V[c]**
1	0.9962	5.6431	0.9428	0.9990	5.9063	0.9726	0.9990	5.8898	0.9725
2	0.9948	1.9560	0.9099	0.9939	1.4637	0.9696	0.9937	1.4811	0.9690
3	0.9608	0.1854	0.4064	0.9595	0.2025	0.5895	0.9732	0.2042	0.5571
4	0.8276	0.3399	0.5995	0.9927	0.0954	0.4963	0.9930	0.0855	0.3395
5	0.7952	0.0884	0.3867	0.9662	0.0477	0.3037	0.9737	0.0731	0.2446

[a]Autocorrelations of U vectors

[b]Max signal [10^{-3} cm^{-1}]

[c]Autocorrelations of V vectors

Table S2. Significant components analysis of the CD experiments performed on EGFR-37MUT

	HCl-titration			Melting			Annealing		
Component	**U[a]**	**M[b]**	**V[c]**	**U[a]**	**M[b]**	**V[c]**	**U[a]**	**M[b]**	**V[c]**
1	0.9970	5.6762	0.9515	0.9990	5.6552	0.9736	0.9990	5.6391	0.9738
2	0.8327	1.9902	0.9069	0.9945	2.0551	0.9700	0.9947	2.1174	0.9702
3	0.9069	0.1615	0.7098	0.9685	0.1816	0.7923	0.9714	0.1915	0.6967
4	0.8916	0.0430	0.0766	0.9923	0.3406	0.5215	0.9905	0.3786	0.4146
5	0.8864	0.1216	-0.4786	0.9804	0.1992	0.0631	0.9867	0.2923	0.4908

[a]Autocorrelations of U vectors

[b]Max signal [10^{-3} cm^{-1}]

[c]Autocorrelations of V vectors

Chapter 2: Exploring loops length requirements for the formation of a 3 cytosine-cytosine+ base-paired i-Motif.

To date, few high-resolution structures of intra-molecular iMs are available [28,40–43] and 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 G4s, 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 [44–47], 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 [20,30,48–51]. These data could be the way to uncover unique folding features and implement iM-specific folding prediction algorithms. Nevertheless, the variation of the C-runs length, the type, and the number of nucleotides within the loops would give rise to an unmanageable number of different combinations.

There are two complementary approaches to solving this issue. One consists of the screening of a huge number of models, and it has the great advantage of providing a general perspective [20,30,48–51]. The alternative strategy relies on the study of fewer models, with the advantage of providing a more detailed description of the system under observation.

Here we applied this second working model. In particular, we set up an interactive protocol for the rational selection of a limited number of sequences to be studied by spectroscopic and calorimetric analyses to efficiently dissect sequence and structural requirements for the formation of these non-canonical DNA structures. In particular, we applied it to derive 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 which allowed us to resolve the structure of the so-selected iM to better address the yet poorly explored structural contribution of the loops in iM.

Material and methods

Material

The non-labeled and the residue-specific partially (8 %) ^{15}N- and ^{13}C-isotopically labeled oligonucleotides (Table 1) were synthesized on K&A Laborgeraete GbR DNA/RNA Synthesizer H-8 by standard

phosphoramidite chemistry. Deprotection was achieved with overnight incubation in aqueous ammonia at 55 °C. Purification and desalting were achieved with the use of Amicon Ultra-15 Centrifugal Filter Units.

Table 1. Synthesized oligonucleotides

Name	Sequence from 5' to 3' end
C21T333	CCTTTCTTTCCTTTC
C21T232	CCTTCTTTCCTTC
C21T313	CCTTTCTCCTTTC
C21T242	CCTTCTTTTCCTTC
C21T323	CCTTTCTTCCTTTC
C21T414	CCTTTTCTCCTTTTC
C21T223	CCTTCTTCCTTTC
C21T322	CCTTTCTTCCTTC
C21T244	CCTTCTTTTCCTTTTC
C21T442	CCTTTTCTTTTCCTTC
C12T333	CTTTCCTTTCTTTCC
C12T242	CTTCCTTTTCTTCC
C12T343	CTTTCCTTTTCTTTCC
C12T414	CTTTTCCTCTTTTCC
C12T424	CTTTTCCTTCTTTTCC
C12T434	CTTTTCCTTTCTTTTCC
C12T444	CTTTTCCTTTTCTTTTCC

Circular dichroism

CD spectra were acquired using a Jasco J 1500 spectropolarimeter equipped with a Peltier as a temperature controller device. Samples were prepared in 50 mM Na-cacodylate pH 5.5 at DNA concentrations ranging between 4 µM and 400 µM. CD signals were reported as Δε.

Differential scanning calorimetry

Differential scanning calorimetry experiments were performed on a Microcal VP-DSC with cells of 502.7 µL between 1 and 80 °C at ± 20 °C/h heating-cooling rates. Multiple water-water, buffer-water, and buffer-buffer scans were performed before the analysis to derive the baseline thermogram and to check there were no heat exchanges due to the buffer in the set experimental condition. The measurements were performed in 50 mM Na-cacodylate pH 5.5 whose ionization in water is characterized by a ΔG° of 35.8 kJ mol^{-1} and a ΔH° of -0.7 kcal mol^{-1}. Therefore, a little contribution from the protonation of the buffer to the enthalpy of folding could influence the measurement of -0.7 kcal mol^{-1} for each proton involved in the

formation of the CC^+ base pairs. The sample cell was filled with a solution of 400 µM DNA per strand in 50 mM Na-cacodylate pH 5.5. Data were reported as molar excess of heat capacity (ΔCp).
The acquired thermograms were analyzed according to equation 1:

$$\Delta Cp = \frac{\Delta H^{o\,2}\, e^{-\frac{\Delta G^o(T^\circ)}{R\,T^\circ} - \frac{\Delta H^o}{R}\left(\frac{1}{T} - \frac{1}{T^\circ}\right)}}{R\,T^2\left(1 + e^{-\frac{\Delta G^o(T^\circ)}{R\,T^\circ} - \frac{\Delta H^o}{R}\left(\frac{1}{T} - \frac{1}{T^\circ}\right)}\right)^2} \qquad (1)$$

where T is the temperature (°K), T° is 298.15 °K (25 °C), $\Delta G^o(T^\circ)$ is the standard Gibbs free energy change of folding (kcal mol^{-1}) referred to T°, ΔH^o is the standard enthalpy change of folding (kcal mol^{-1}) and R is the ideal gas constant (1.987 10^{-3} kcal mol^{-1} K^{-1}). In this analysis, ΔH^o was considered a temperature-independent parameter (ΔCp = 0).
Data analysis and fitting were performed by using NumPy and SciPy Python libraries.

NMR

NMR samples were prepared by dissolving DNA oligonucleotides at final 0.2-0.7 mM concentrations in 90%/10% $H_2O/^2H_2O$ and 50 mM Na-cacodylate or 25 mM Na_2HPO_4 pH of 5.5. The equivalence of these two different buffers in DNA folding was tested (Figure S9). NMR spectra were recorded on 600 and 800 MHz spectrometers equipped with HCN cryogenic probes at 0°C, if not stated differently. NOESY spectra were acquired at mixing times between 80 and 200 ms. ^{15}N- and ^{13}C-edited HSQC experiments were recorded on 8 % residue-specifically ^{15}N- and ^{13}C-isotopically labelled oligonucleotides. NMR spectra were processed and analyzed by using TopSpin (Bruker) and Sparky (UCSF) software.[52]

Restraints and structure calculation

The interproton distance restraints were assessed from the corresponding integral volumes of cross-peaks in NOESY spectra recorded in the range between 80 and 200 ms mixing times and classified as strong (1.8−3.6 Å), medium (2.6−5.0 Å), and weak (3.5−6.5 Å) according to the averaged values obtained for cytosine(s) H5-H6 cross-peaks (2.5 Å). Additionally, six restraints corresponding to hydrogen-bonding in CC^+ base pairs were included, accounting for H41-O1 in each of the C1-C10; C2-C11, and C6-C15 base pairs; and between C6H3-C15N3 and C2N3-C11H3 (all 1.6–1.9 Å). The glycosidic bond torsion angles (χ) for all fifteen residues were restrained to adopt *anti*-conformation (240 ± 70°) based on the weak-to-medium intra-residual H6-H1' NOE correlations. The structure calculations were performed by using Amber 20 software and parmbsc1 force field, including parm χOL4 and parm ε/zOL1 modifications. Born implicit solvent model and random starting velocities were used. The partial charges for protonated cytosines at positions 6 and 11 were derived with the use of RESP ESP charge Derive (R.E.D) Server. Calculations relied on two rounds of

restrained 1000 ps long simulated annealing (SA). The first SA relied on using force constants at 2 kcal mol^{-1} $Å^{-2}$ for NOE-based distance and glycosidic torsion angles (χ) restraints and at 200 kcal mol^{-1} $Å^{-2}$ for chirality restraints. In the second and final SA, the NOE-based distances, glycosidic torsion angles (χ) as well as hydrogen-bonds were restrained by setting the corresponding force constants at 20 kcal mol^{-1} $Å^{-2}$, while restraints for chirality were set at 200 kcal mol^{-1} $Å^{-2}$. Both SA rounds were run according to the same protocol, in which the force constants for restraints were scaled from the initial value of 0.1 to the final value of 1.0 in the first 500 ps and held constant until the end of the calculation. The SA protocol included heating over the first 100 ps from 0 to 1000 K, followed by 400 ps equilibration at 1000 K, cooling over 400 ps from 1000 to 0 K, and equilibration for 100 ps at 0 K. Initial starting structures for the first SA calculations were generated with the use of 3DNA software (48) and leap model of AMBER 20. The lowest energy structure among the twenty structures calculated in the first step was used as the initial structure in the second and the final SA step, in which 100 structures were calculated, and among them, ten with the lowest energy were subjected to energy minimization with a maximum of 10000 steps to yield representative ensemble.

Results

Design of cytosine-rich models and interactive workflow for the selection of iM 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 took into account sequences that can accommodate only 3 CC^+ base pairs when folded intramolecularly. The stability of such models is expected to be low thus making easier the definition of those supporting iM formation. Moreover, they 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. Intending to identify 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 [28]. When considering N as any among G, C, T, or A, a total 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, G and A were discarded to avoid any base pair competition with the cytosines and thymines

potentially involved in the iM core and loops, respectively [53]. On the other hand, T can form additional TT 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 [54–57]. Consistently, several previous works on iM models accommodated T in loops as well [20,30,48–51]. As a result of 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 checked the folding of one sequence from the *P_list* with its defined combination of x, y, and z. Whenever the selected sequence was proved to fold into an intramolecular iM, we removed from *P_list* all the sequences having simultaneously the first loop ≥ x, the second loop ≥ y, and the third loop ≥ z. Conversely, when it did not fold, we removed from *P_list* all those having simultaneously the first loop ≤ x, second loop ≤ y, and third loop ≤ z. Therefore, the number of sequences in the *P_list* progressively decreased at each step. The screening ended when *P_list* was empty. We developed an algorithm that indicated the sequence to check with the intent to maximize the number of sequences to remove from the *P_list* at each step (Supplementary Information). This was the faster protocol to reach the emptiness of the *P_list* and the identification of the minimal folding sequences. 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).

Finally, we designed the algorithm to allow the user to define the first sequence to check at the beginning of the screening (step 0). Since reported iM structures showed that three nucleotides per loop are compatible with both minor and major grooves [28], 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 NMR, circular dichroism (CD), and differential scanning calorimetry (DSC) experiments. The formation of CC^+ base pairs was monitored by recording ^{1}H NMR spectra, whereby assuming that the formation of the CC^+ base pair is coupled to a lower exchange rate of the corresponding H3 proton with the water solvent, thus resulting in a detectable signal in the range between δ 15 and 16 ppm. To complement NMR data, the formation of iM structures was assessed through the chiroptical fingerprint. Indeed, CD spectroscopy allowed recording data on DNA samples in a wider concentration range thus allowing us to distinguish the inter-molecular iM formation (a process concentration-dependent) from an intra-molecular iM, whose folding is concentration-independent.

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 [20].

Screening of the C21 subgroup

To screen the C21 subgroup, we rationally selected to start from the C21T333 sequence since it contains all the loops compatible with iM formation. Starting from this leading sequence, 10 steps (from 0 to 9) were sufficient to conclude the screening (Figure 2).

As reported in Figure 2, the NMR spectrum of C21T333 showed two well-resolved signals at δ15.36 and 15.58 ppm both suggesting the formation of a CC^+ base pair. The CD spectra of the same sequence acquired at 4 and 400 µM DNA concentration, were almost superimposable with a positive signal at 285 nm and a negative one at 260 nm, consistent with an iM structure (Figure 3) [58]. Moreover, the CD signal intensity at 285 nm was in line with the one expected for the formation of 3 CC^+ base pairs [50]. DSC scans showed a reversible single transition process, corroborating an inter-molecular characteristic of iM adopted by C21T333 (Figure 3). By fitting the calorimetric data to equation 1, we derived a $\Delta H^{\ominus}$ of (-37.8 ± 0.1) kcal mol^{-1} (Table 2) thus supporting the folding of C21T333 into an intra-molecular 3 CC^+ base-paired iM.

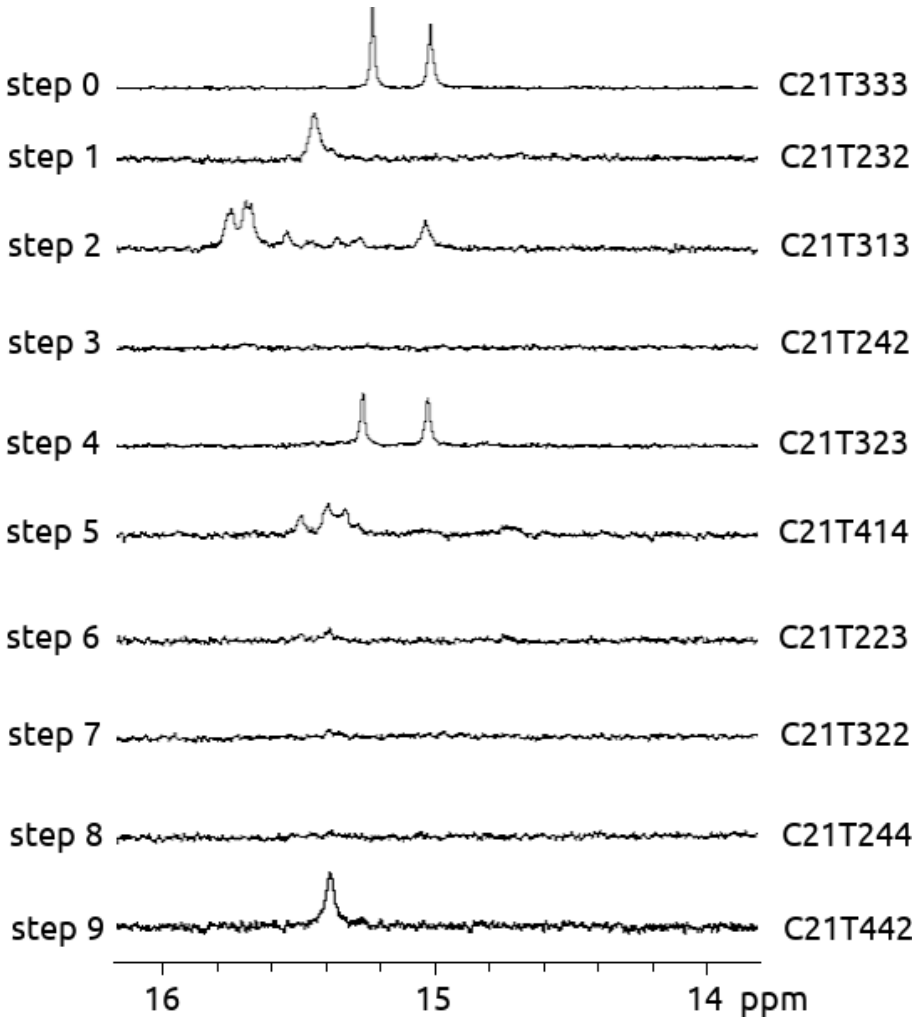

Figure 2: ^1H NMR imino region of potentially iM forming oligonucleotides in the C21 group. The spectra were recorded on a 600 MHz NMR spectrometer in 90% H_2O/10% ^2H_2O, at 0 °C, 0.2 mM oligonucleotide concentrations, and 50 mM Na-cacodylate buffer, pH 5.5.

Based on this positive output, in the next step (step 1), the algorithm indicated to test C21T232. The NMR spectrum of this sequence showed a single signal at δ 15.45 ppm (Figure 2). However, for this sequence, the CD spectra recorded at 4 and 400 µM were remarkably different. In particular, the intensity of the CD signal at 285 nm was strongly DNA concentration-dependent (Figure S1). These results were consistent with C21T232 forming inter-molecular iM structures.

As a sequence to be tested in the next step (step 2), the algorithm indicated C21T313. Also here, the presence of multiple signals in the δ 15.0-15.8 ppm range of the NMR spectrum indicated that C21T313 formed a structure exhibiting CC^+ base pairs (Figure 2). However, the CD signal of C21T313 was DNA concentration-dependent indicating that the oligonucleotide tends to form multimeric assemblies, and, additionally, the spectral features did not match the ones characteristic of iM structures (Figure S1).

Subsequently, the folding of C21T242 was analyzed (step 3). However, the ^{1}H NMR spectrum, showed no signals around δ 15 ppm (Figure 2), indicating the absence of an iM structure.

Then, the algorithm indicated testing the sequence C21T323 (step 4). The NMR spectrum showed well-defined signals at δ15.36 and 15.61 ppm (Figure 2), and 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 (Figure 3). Likewise, 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. DSC scans showed a reversible single transition process (Figure 3). By fitting the calorimetric data according to equation 1, we derived a ΔH° of (-33.7 ± 0.1) kcal mol^{-1} (Table 2) thus supporting the folding of C21T323 into an intra-molecular iM comprising 3 CC^+ base pairs.

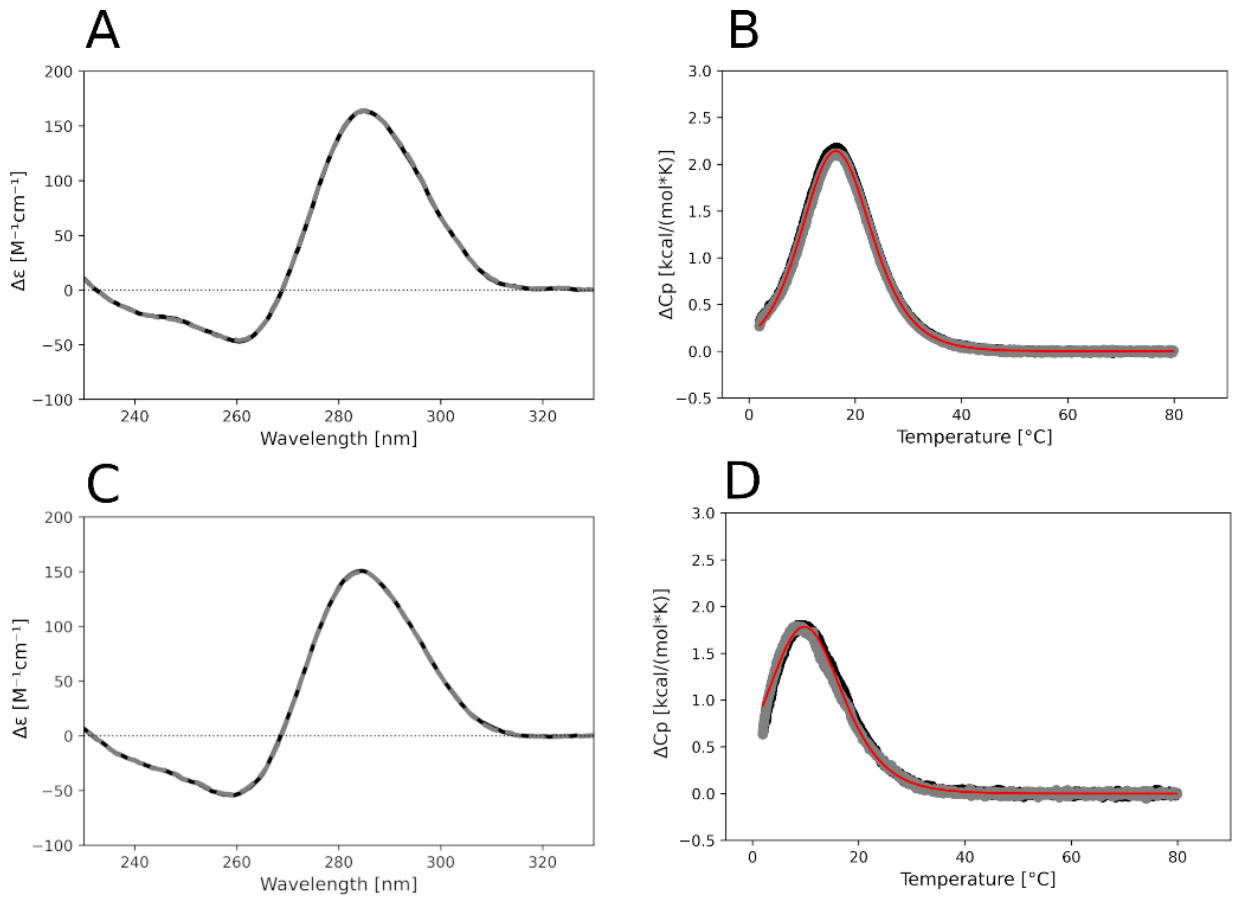

Figure 3: A) CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T333 in 50 mM Na-cacodylate pH 5.5 at 0°C. B) DSC melting (black dots) and annealing (grey dots) scans of 400 µM C21T333 in 50 mM Na-cacodylate pH 5.5 and corresponding fitting curve (red solid line). C) CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T232 in 50 mM Na-cacodylate pH 5.5 at 0°C. D) DSC melting (black dots) and annealing (grey dots) scans of 400 µM C21T323 in 50 mM Na-cacodylate pH 5.5 and corresponding fitting curve (red solid line).

Moving to C21T414 (step 5), we observed the presence of multiple signals in the 15.2-15.6 ppm range of the NMR spectrum (Figure 2), but the CD signal was DNA concentration-dependent, thus supporting that C21T232 folded into inter-molecular iM structures (Figure S1).

Then the screening went progressively through C21T223, C21T322, and C21T244 (step 6, step 7, and step 8), but no signals in the 16-15 ppm range were detected in the NMR spectrum for any of them (Figure 2).

Table 2. Thermodynamic parameters of the C21T333 and C21T323 models derived in 50 mM Na-cacodylate pH 5.5 and referred to 273.15 K (0 °C).

	ΔG° **kcal/mol**	**ΔH°** **kcal/mol**	**-TΔS°** **kcal/mol**	**Tm °C**
C21T333	-2 ± 1	-38 ± 1	36 ± 1	16.3 ± 0.1
C21T323	-1 ± 1	-34 ± 1	33 ± 1	9.3 ± 0.1

In the last step (step 9), the algorithm indicated to test C21T442. The NMR spectrum showed a single signal at 15.45 ppm (Figure 2). However, the CD signal was DNA concentration-dependent, supporting that C21T442 folded into inter-molecular iM structures (Figure S1).

At this step, the P_list was empty. 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.

Screening of the C12 subgroup

By applying the same protocol to the C12 subgroup, by starting from the C12T333 sequence we concluded the screening in 7 steps (Figure 4). The ^{1}H NMR spectrum of C12T333 showed two broad signals at δ15.73 and 15.60 ppm. However, the CD signal was DNA concentration-dependent suggesting that C12T333 adopts inter-molecular structures (Figure S1).

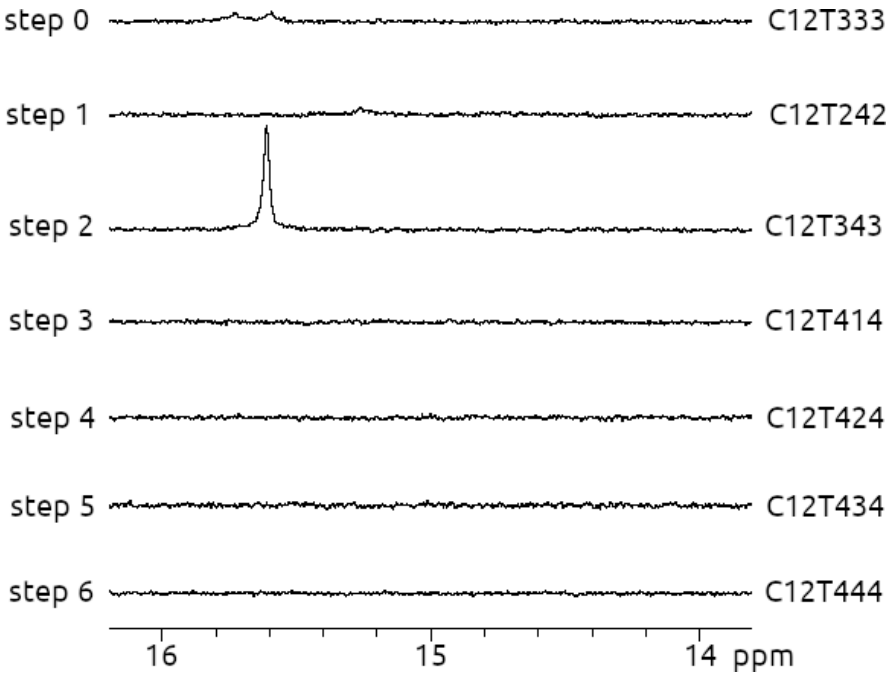

Figure 4: ^{1}H NMR imino region of potentially iM forming oligonucleotides in the C12 group. The spectra were recorded on a 600 MHz NMR spectrometer in 90% H_2O/10% ^2H_2O, at 0 °C, 0.2 mM oligonucleotide concentrations, and 50 mM Na-cacodylate buffer, pH 5.5.

In the next step (step 1), the algorithm indicated to test C12T242, for which, however, no ^{1}H NMR signal was detected around δ 15 ppm, indicating the absence of an iM structure (Figure 4).

The algorithm suggested moving to C12T343 (step 2). Although the ^{1}H NMR spectrum showed a well-defined signal at δ 15.61 ppm (Figure 4), the iM folding was DNA concentration-dependent. Noteworthy, the CD spectrum recorded at 4 μM oligonucleotide concentration was consistent with an unfolded sequence, thus supporting that C12T343 formed only inter-molecular structures (Figure S1).

Then the screening focused on C12T414, C12T424, C12T434, and C12T444 (step 3, step 4, step 5, and step 6), for which, however, ^{1}H NMR signals in the δ 15-16 ppm range were not detected (Figure 4). This led to the emptiness of the *P_list*, thus we ended the screening for the C12 subgroup without finding any sequence that folded into a 3 CC^{+} base-paired iM monomer.

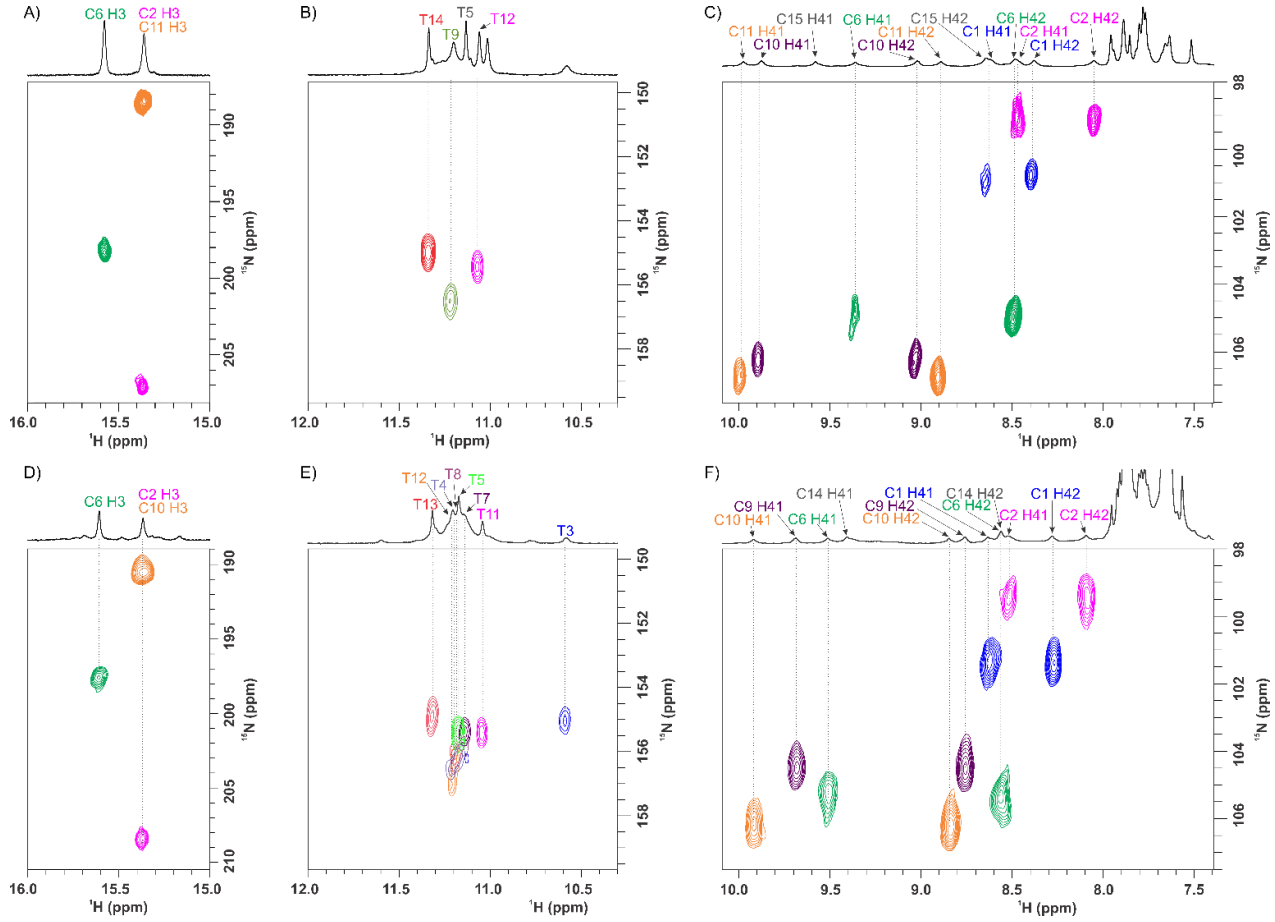

Figure 5: Stacked ^{15}N-edited HSQC spectral regions of residue-specifically partially ^{15}N-isotopically labeled (A-C) C21T333 and (D-F) C21T323. The cross-peaks corresponding to H3-N3 correlations are shown in A and D for cytosine and in B and E for thymine residues, while H41/H42-N4 cross-peaks 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. In-depth structural characterization of C21T333 and C21T323

^{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 a few 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 labeled C21T333 at key positions (Figure 5A-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 5C). 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 15.61 and 15.37 ppm, corresponding to CC^{+} base pairing in the iM (Figure 5D). 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 5E-F).

Pairwise comparison of the ^{1}H and ^{15}N NMR chemical shifts corresponding to amino groups in iM adopted by C21T333 and C21T323 (Table S1 and Figure 6) 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 labeled thymine residue at individual positions, in which the most intensive signal is observed at ^{1}H δ 1.84-1.86 ppm, corresponding to the methyl groups of unfolded species (Figure 8).

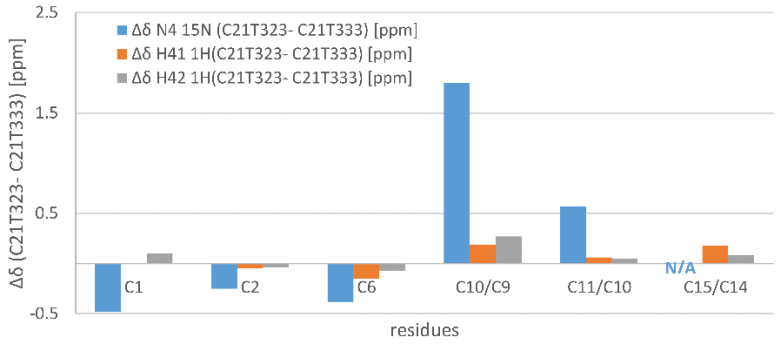

Figure 6: Differences in ^{15}N and ^{1}H NMR chemical shift of cytosines' amino group atoms (N4, H41, and H42) in C21T333 with respect to C21T323. The plotted differences correspond to δ(C21T323)-δ(C21T323) relying on the NMR chemical shift assignment derived from the spectra recorded in 90% H_2O/10% $^{2}H_2O$, at 273 K, 0.4 mM oligonucleotide concentration per strand and 25 mM sodium phosphate buffer (pH 5.5). Note that N4 of residues C15 and C14 in C21T333 and C21T323, respectively, were not assigned.

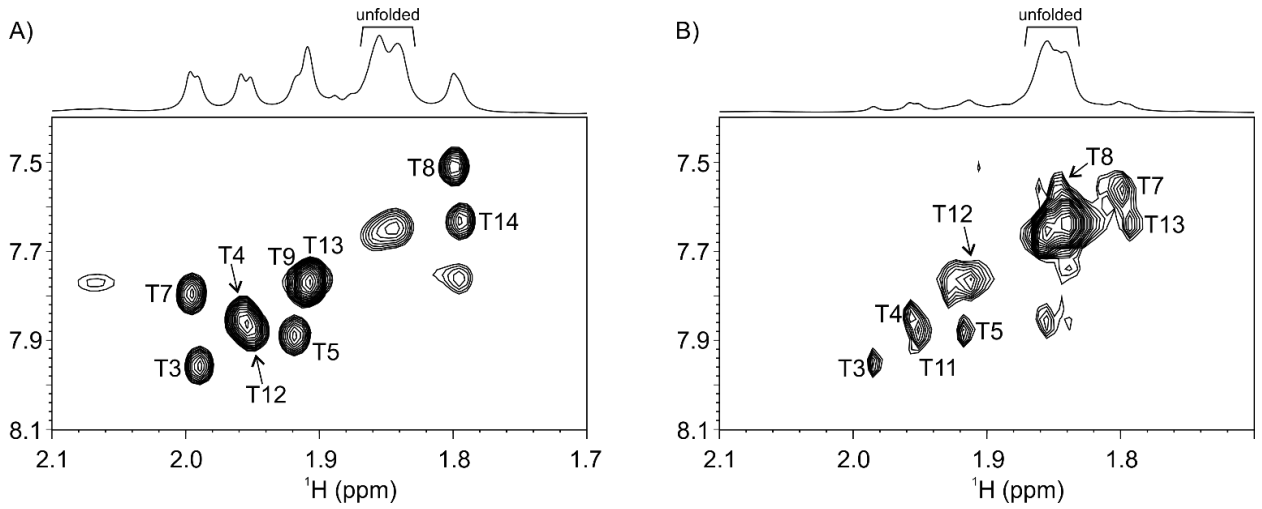

Figure 7: Regions of NOESY spectra of A) C21T333 and B) C21T323 with designated intra-residual aromatic H6-methyl cross-peaks, above which the ^{1}H NMR spectral regions are shown together with indicated methyl groups signals corresponding to unfolded species. The spectra were recorded in 90% H_2O/10% ^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).

In addition to exchangeable and methyl protons, also aromatic and sugar ^{1}H NMR signals corresponding to C21T333 as well as to C21T323 were assigned by analyzing NOESY (Figures S2-7) together with ^{13}C- and ^{15}N-HSQC spectra recorded with the use of residue-specifically partially ^{13}C- and ^{15}N-isotopically labeled 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 δ6.27-6.57 ppm for C21333 and δ 7.76-7.89 ppm and δ 6.25-6.57 ppm for C21T323 (Figure S8). 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.

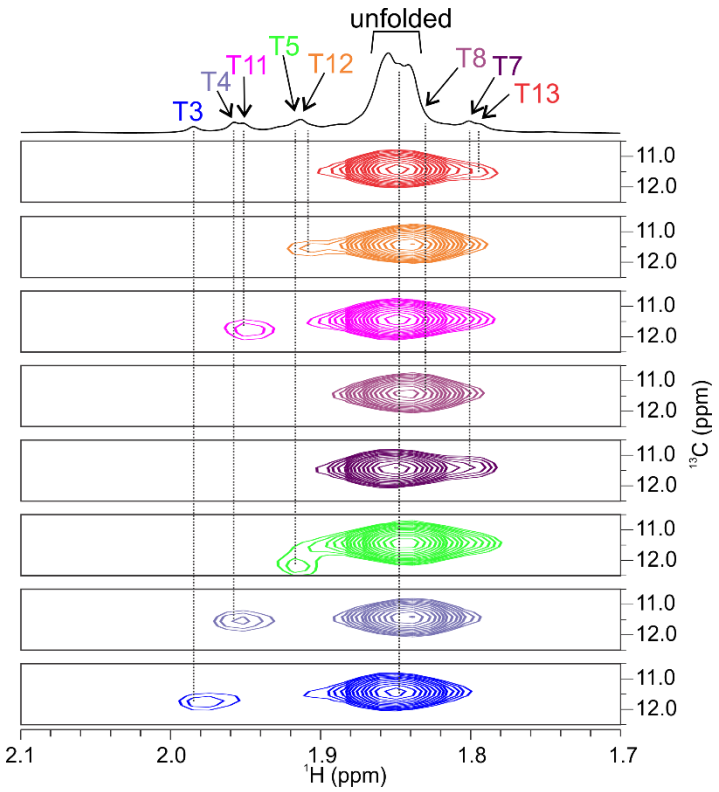

Figure 8: Methyl region of ^{13}C-HSQC spectra of C21T323 recorded on partially residue-specifically ^{13}C-isotopically labeled samples. In the corresponding ^{1}H NMR region above the 2D plots, the assignments of thymine methyl signals are indicated in black for the unfolded oligonucleotide, while in other colors for the iM. The spectra were recorded in 90% H_2O/10% $^{2}H_2O$, 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 S2 and S4-7) 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 S7), 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.[57] 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 the 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. 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

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 3; PDB ID 7QDC). 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. 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. The vicinity of T5 and T14, especially the mutual closeness of H3 and O4 atoms indicates the formation of T-T base pair, albeit the corresponding nucleobases are not perfectly coplanar. Moreover, T5-T14 is positioned beneath C2-C11 (Figure 9C below), seemingly extending the feature of consecutive intercalated base pairs from 3 to 4. This continuous run of (three CC^+ and 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.

Table 3. NMR restraints and structural statistics for iM adopted by C21T333

	distance and torsion angle restraints		
NOE-derived distance restraints		non – exchangeable	exchangeable
	Intra-nucleotide	168	2
	Sequential (i, i+1)	53	0
	Long-range (i, >i+1)	50	2
Torsion angle restraints	15		
Hydrogen-bond restraints	8		
Structural statistics (mean and s.d.)	NOE violations >0.3Å	0	
	[a]Pairwise atom RMSD (Å)	Overall	0.719 (0.651)
		Without T3, T4, T5	0.741 (0.694)
		Without T7, T8, T9	0.528 (0.439)
		Without T12, T13, T14	0.767 (0.716)
		Only cytosine residues	0.521 (0.478)

[a]Number in the bracket corresponds to RMSD value when considering heavy atoms (C,0, N, P) only

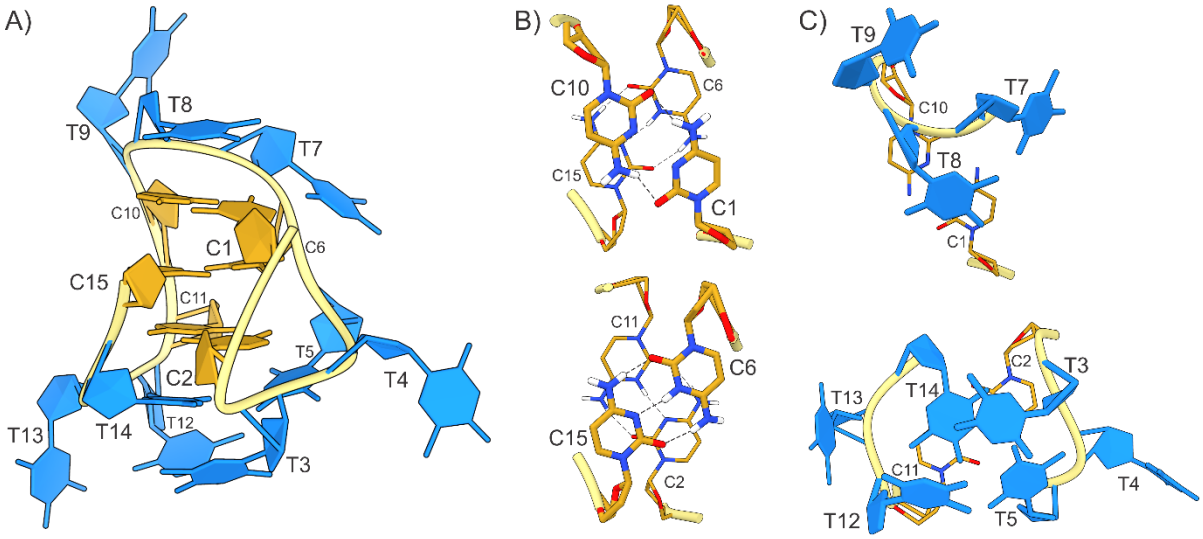

Figure 9: Solution-state structure of iM adopted by C21T333. The lowest energy structure is shown in A), along with the details focused on B) top views on consecutive CC^+ base pairs and C) TTT loops arrangements and the nearby CC^+ base pairs. Cytosine and thymine residues are depicted in gold and blue,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

Discussion

We developed an algorithm that allowed us to identify the minimal length of the loops amenable to the formation of an intra-molecular iM containing only 3 CC^+ base pairs through a systematic step-by-step workflow. 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 length of the loops 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.

To train such a protocol, we limited to thymine the base composition of all the loops. Interestingly, only C21T323 was identified as the minimum among our 3 CC^+ base-paired I-motif dataset. This sequence folds into a single iM topology which corresponds to the major-minor-major. This behavior is conserved even in the C21T333 model, derived by an increment of the central loop. The higher stability of C21T333 allowed us to obtain the high-resolution structure of this I-motif (PDB ID 7QDC). It is the first deposited structure of an I-motif with an odd number of CC^+ base pairs and, to date, it is the simplest model enclosing all the features of an intramolecular iM.

As an additional result, our screening procedure uncovered an intriguing asymmetry between the C21 and C12 series. In particular, none of the assessed sequences from the C12 subgroup adopted an intra-molecular iM. Notably, our data are limited to the 3 CC^+ system hence the different folding of sequences from the C21 and C12 series cannot be generalized, especially when considering that the stabilities of individual iMs may greatly depend on interactions among loop residues, such as recently reported Watson-Crick hydrogen-bonding[59] and non-canonical base-pairing observed herein for C21T333 and C21T323. Still, this result fits with the reported higher stability for 5’ iM.[30] As a future perspective, the screening of the sequences from C32 and C23 subgroups, which expectedly may form up to 5 CC^+ base pairs with enhanced thermodynamic stabilities compared to the ones analyzed in the present work, could provide interesting insights on this topic.

Overall, our data indicate that the folding of iM highly depends on loop lengths with 3 and 2 nucleotides representing the minimal requirements for connecting the C-rich strands in the core of the structure through the major and the minor grooves, respectively. This condition is remarkably different if compared to G-quadruplexes for which 1 nucleotide long loops are sufficient for the formation of thermodynamically stable intra-molecular structures, demonstrating that folding of G- and C-rich structures rely on different sequence requirements. Thus, iMs and G-quadruplexes might preferentially form at different genome locations, possibly related to their different biological function.

Supplementary information

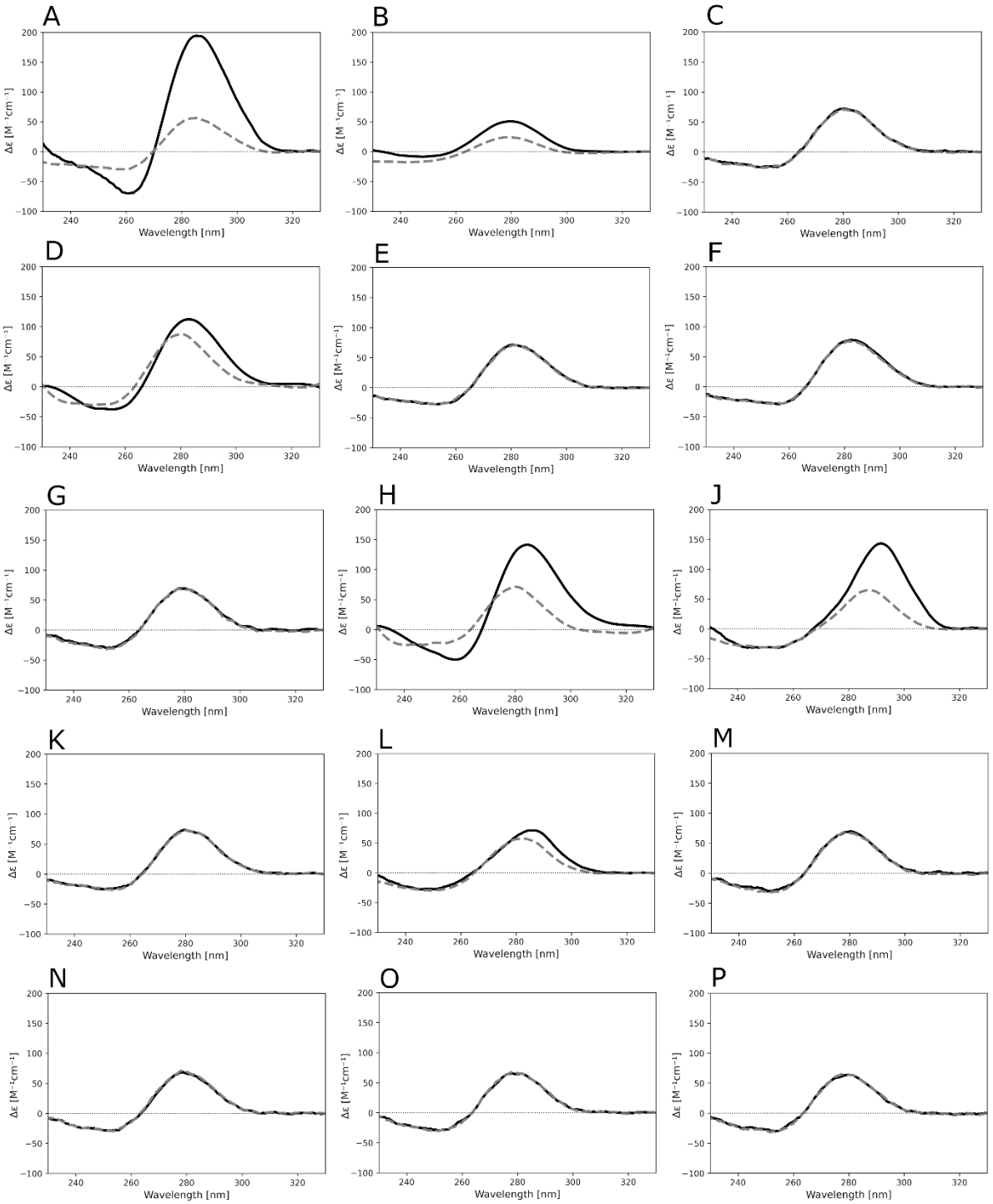

Figure S1: CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T232 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel A). CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T313 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel B). CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T424 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel C). CD spectra of 4 µM (grey dashed line) and 400 µM (black solid line) C21T414 in 50

mM Na-cacodylate pH 5.5 at 0°C (panel D). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C21T223 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel E). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C21T322 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel F). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C21T244 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel G). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C21T442 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel H). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T333 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel J). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T242 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel K). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T343 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel L). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T414 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel M). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T424 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel N). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T434 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel O). CD spectra of 4 μM (grey dashed line) and 400 μM (black solid line) C12T444 in 50 mM Na-cacodylate pH 5.5 at 0°C (panel P).

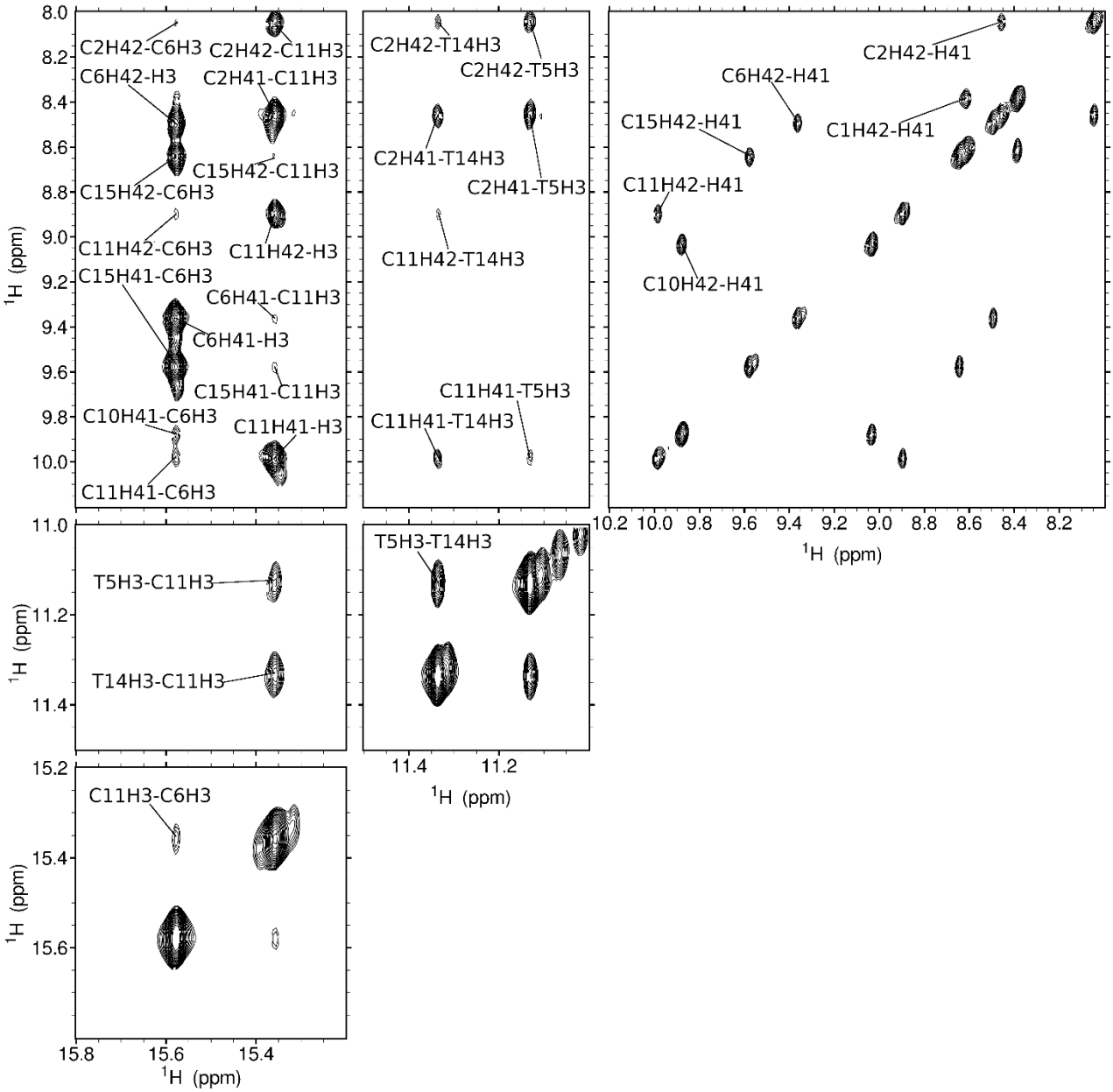

Figure S2: NOESY (τm 150 ms) spectrum of 500 µM C21T333 in 25 mM Na_2HPO_4 pH 5.5 at 0°C 10% ^2H_2O on an 800 MHz NMR spectrometer. Exchangeable protons.

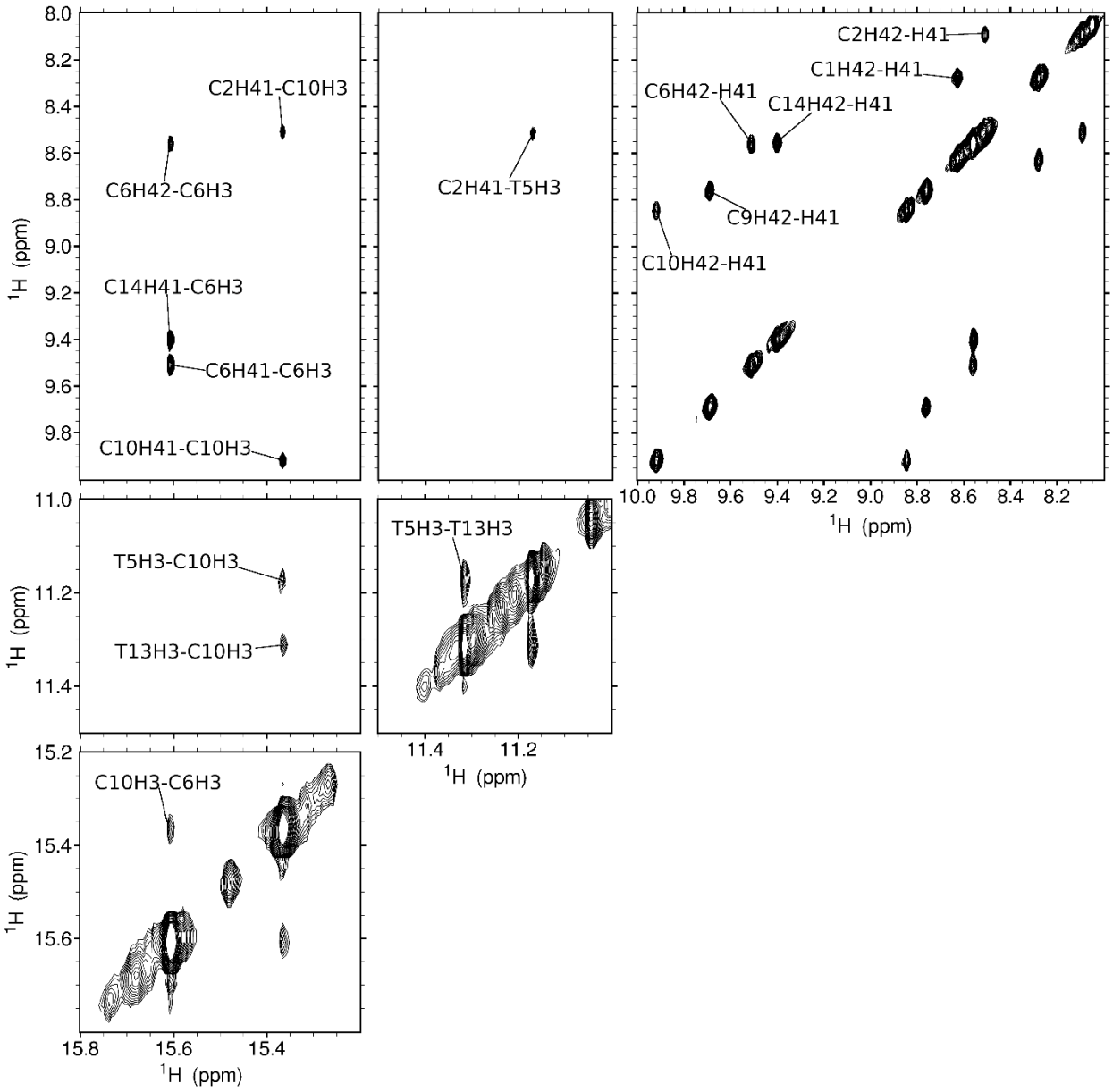

Figure S3: NOESY (τm 150 ms) spectrum of 500 µM C21T323 in 25 mM Na_2HPO_4 pH 5.5 at 0°C 10% ^2H_2O on an 800 MHz NMR spectrometer. Exchangeable protons.

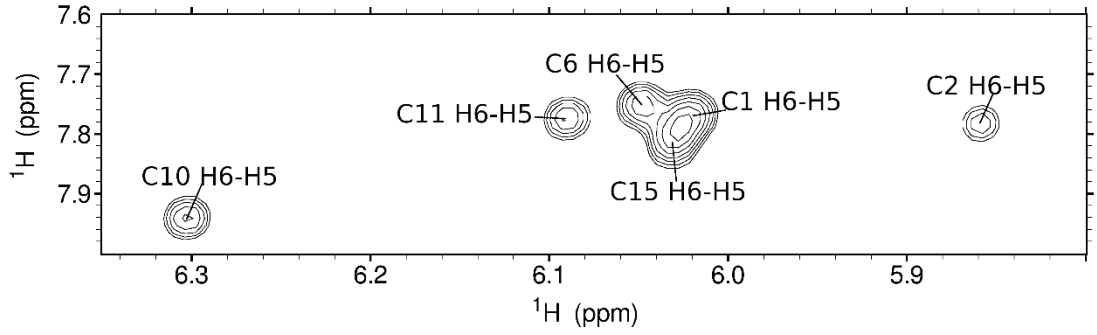

Figure S4: NOESY (τm 150 ms) spectrum of 500 µM C21T333 in 25 mM Na_2HPO_4 pH 5.5 at 0°C 10% ^2H_2O on an 800 MHz NMR spectrometer. H6-H5 intra-residue cross-picks of cytosines.

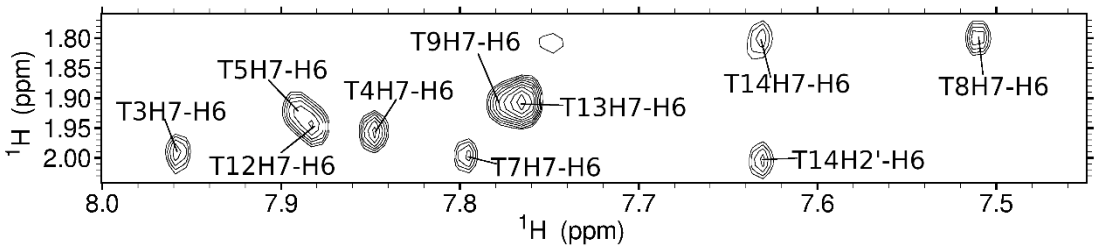

Figure S5: NOESY (τm 150 ms) spectrum of 500 µM C21T333 in 25 mM Na_2HPO_4 pH 5.5 at 0°C 10% ^2H_2O on an 800 MHz NMR spectrometer. H7-H6 intra-residue cross-picks of cytosines.

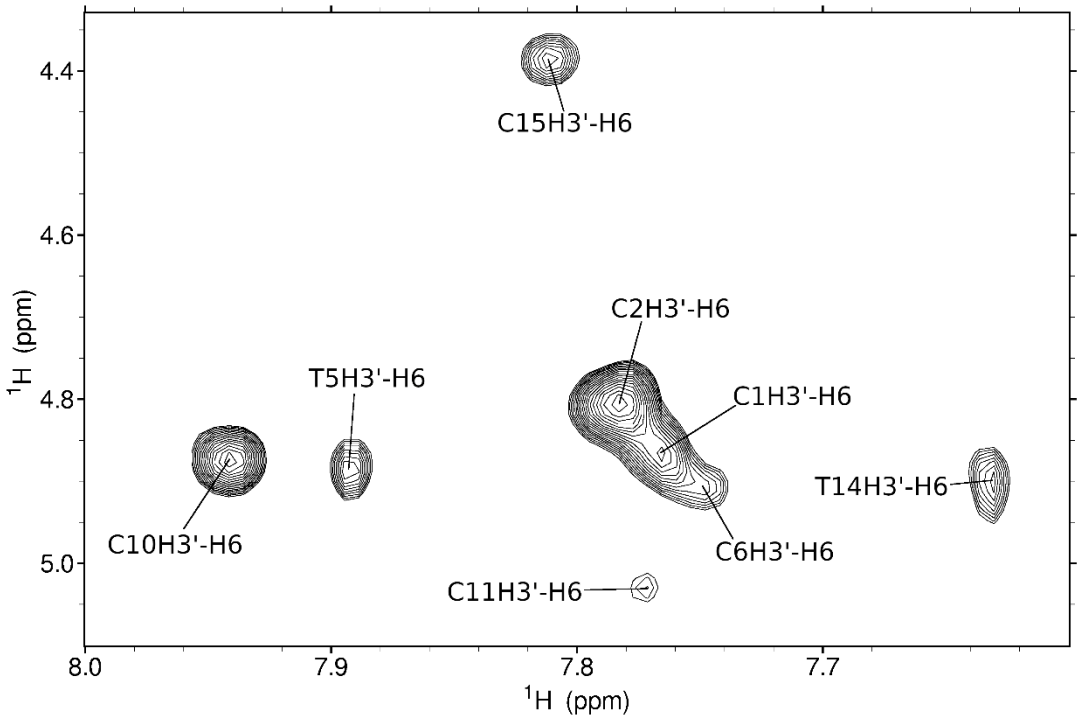

Figure S6: NOESY (τm 150 ms) spectrum of 500 µM C21T333 in 25 mM Na_2HPO_4 pH 5.5 at 0°C 10% ^2H_2O on an 800 MHz NMR spectrometer. H7-H6 intra-residue cross-picks of cytosines.

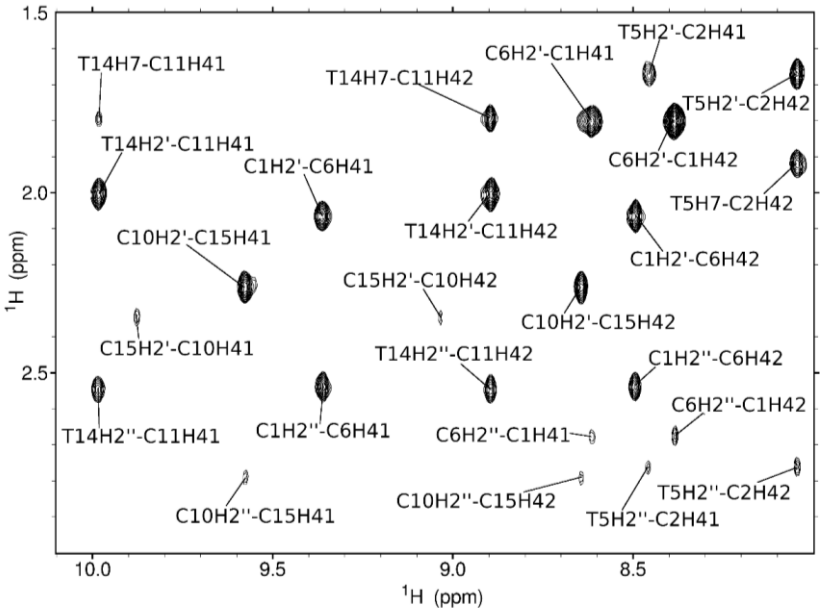

Figure S7: NOESY spectrum of C21T333 showing amino-H2'/H2'' cross-peaks. The spectrum was recorded at 150 ms mixing time, in 90% H_2O/10% ^2H_2O, 273 K, 0.5 mM oligonucleotide concentration per strand, and 25 mM sodium phosphate buffer (pH 5.5).

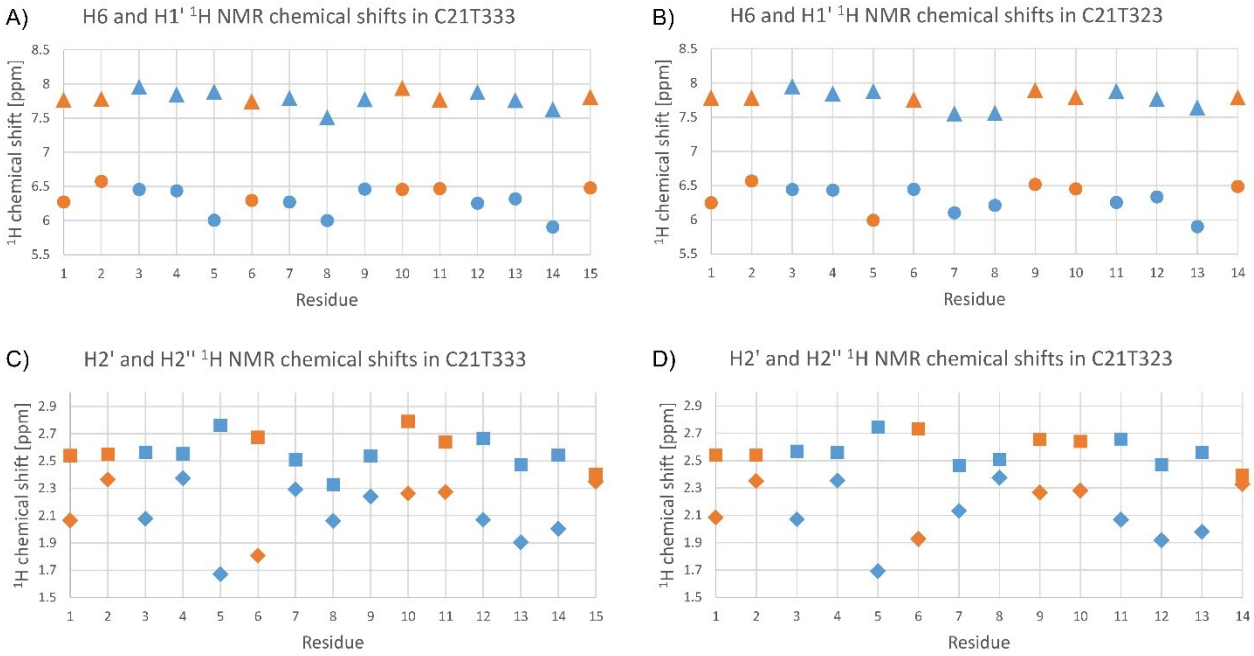

Figure S8: Chemical shifts of H6 (triangles), H1' (circles), H2' (squares), and H2'' (diamonds) corresponding to iM adopted by A and C) C21T333 and B and D) C21T323. For clarity, the data corresponding to cytosine and thymine residues are colored in orange and blue, respectively.

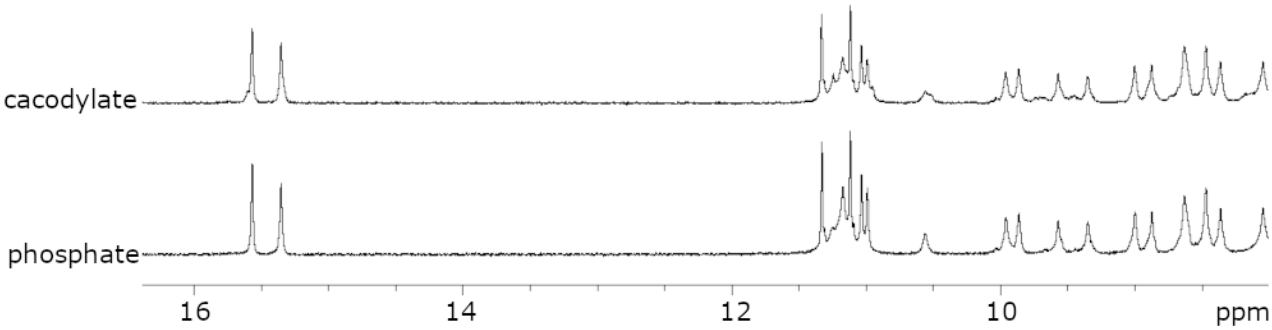

Figure S9: ^{1}H NMR exchangeable region of 0.2 mM C21T333 in 50 mM Na-cacodylate and 25 mM Na-phosphate (50 mM Na^{+}) buffer pH 5.5. The spectra were recorded on a 600 MHz NMR spectrometer in 90% H_2O/10% $^{2}H_2O$, at 0 °C.

Table S1. The differences in ^{15}N and ^{1}H NMR chemical shifts of N4, H41, and H42 in C21T333 and C21T323.

	N4 ^{15}N chemical shift [a]		Δδ N4 ^{15}N (C21T323-C21T333)[b]	H41 ^{1}H chemical shift [a]		Δδ H41 ^{1}H (C21T323-C21T333)[b]	H42 ^{1}H chemical shift [a]		Δδ H42 ^{1}H (C21T323-C21T333)[b]
residue	C21T333	C21323		C21T333	C21323		C21T333	C21323	
C1	100.88	101.36	-0.48	8.62	8.62	0.00	8.38	8.28	0.10
C2	99.17	99.42	-0.25	8.46	8.51	-0.05	8.05	8.09	-0.04
C6	104.91	105.29	-0.38	9.36	9.51	-0.15	8.49	8.56	-0.07
C10/C9	106.25	104.45	**1.80**	9.88	9.69	**0.19**	9.03	8.76	**0.27**
C11/C10	106.74	106.17	0.57	9.98	9.92	0.06	8.90	8.85	0.05
C15/C14	N/A	N/A		9.58	9.40	0.18	8.64	8.56	0.08

[a]The ^{15}N and ^{1}H NMR chemical shifts are given in [ppm] and were assigned with the use of NMR spectra recorded in 90% H_2O/10% $^{2}H_2O$, at 273 K, 0.4 mM oligonucleotide concentration per strand and 25 mM sodium phosphate buffer (pH 5.5).

[b] Chemical shift differences (Δδ) are given in [ppm] and correspond to δ(C21T323)-δ(C21T323). The largest Δδ values are designated by numbers in bold.

Screening algorithm

The output of the screening for the C21 run model:

The minimum pattern for this system is:

5'-CCTTTCTTCCTTTC-3'

Summary of the screening:

step0: ["5'-CCTTTCTTTCCTTTC-3'", 'yes', 'sequences in P_list from 64 to 56']

step1: ["5'-CCTTCTTTCCTTC-3'", 'no', 'sequences in P_list from 56 to 44']

step2: ["5'-CCTTTCTCCTTTC-3'", 'no', 'sequences in P_list from 44 to 39']

step3: ["5'-CCTTCTTTTCCTTC-3'", 'no', 'sequences in P_list from 39 to 35']

step4: ["5'-CCTTTCTTCCTTTC-3'", 'yes', 'sequences in P_list from 35 to 31']

step5: ["5'-CCTTTTCTCCTTTTC-3'", 'no', 'sequences in P_list from 31 to 24']

step6: ["5'-CCTTCTTCCTTTC-3'", 'no', 'sequences in P_list from 24 to 22']

step7: ["5'-CCTTTCTTCCTTC-3'", 'no', 'sequences in P_list from 22 to 20']

step8: ["5'-CCTTCTTTTCCTTTTC-3'", 'no', 'sequences in P_list from 20 to 10']

step9: ["5'-CCTTTTCTTTTCCTTC-3'", 'no', 'sequences in P_list from 10 to 0']

The output of the screening for the C21 run model:

There is no folding for this system

Summary of the screening:

step0: ["5'-CTTTCCTTTCTTTCC-3'", 'no', 'sequences in P_list from 64 to 37']

step1: ["5'-CTTCCTTTTCTTCC-3'", 'no', 'sequences in P_list from 37 to 33']

step2: ["5'-CTTTCCTTTTCTTTCC-3'", 'no', 'sequences in P_list from 33 to 28']

step3: ["5'-CTTTTCCTCTTTTCC-3'", 'no', 'sequences in P_list from 28 to 21']

step4: ["5'-CTTTTCCTTCTTTTCC-3'", 'no', 'sequences in P_list from 21 to 14']

step5: ["5'-CTTTTCCTTTCTTTTCC-3'", 'no', 'sequences in P_list from 14 to 7']

step6: ["5'-CTTTTCCTTTTCTTTTCC-3'", 'no', 'sequences in P_list from 7 to 0']

The algorithm is written in Python programing language.

Assigned variables:

sequence: potential iM forming sequence of nucleotides as a string.

first_run: first run of the iM sequence as a string.

first_loop: length of the first loop of the iM sequence as an integer.

second_run: second run of the iM sequence as a string.

second_loop: length of the second loop of the iM sequence as an integer.

third_run: third run of the iM sequence as a string.

third_loop: length of the third loop of the iM sequence as an integer.

foruth_run: fourth run of the iM sequence as a string.

upper_limit: integer value, it is the upper limit of the length of the loops.

loop_nucleotide_type: it is a string corresponding to the nucleotide type.

loop_string: string variable derived from the conversion of the first_loop, second_loop, or third_loop variable into its corresponding nucleotide sequence.

typing_error_list: list of the errors made by the user in typing the starting sequence.

step: integer value corresponding to any step of the screening.

user_answer: 'yes' or 'no' string, it is assigned by the user for each sequence after checking if it folds into a monomeric I-motif using all the expected cytosines or not.

screening_summary: dictionary that keeps each step and the associated sequence variable of the screening.

P_list: list with all the possible sequences that could potentially be a minimum.

P_list_reducted: list derived from the elimination of all the sequences that can be removed from P_list after each step.

P_list_equal_first_third_loop: list with all the sequences in P_list that have the first and third loops equally long.

P_list_different_first_third_loop: list with all the sequences in P_list that have the first and third loops differently long.

p_yes: integer value associated with a sequence, and it is the number of sequences that would be removed from P_list if it folds.

p_no: integer value associated with a sequence, and it is the number of sequences that would be removed from P_list if it does not fold.

p: integer value associated with a sequence, and it is defined differently if P_list_equal_first_third_loop is an empty list or not. It is defined as p = p_yes if p_yes ≤ p_no or p = p_no if p_no < p_yes when P_list_equal_first_third_loop is not an empty list, and p = p_yes if p_yes ≥ p_no or p = p_no if p_no > p_yes when P_list_equal_first_third_loop is an empty list.

next_sequence: list containing first_loop, second_loop, and third_loop of the next sequence to check.

minimum: list containing first_loop, second_loop, and third_loop of the minimum iM folding sequences.

We developed six functions:

sequence_decomposition(sequence , upper_limit, loop_nucleotide_type): it returns first_run, first_loop, second_run, second_loop, third_run, third_loop, fourth_run, typing_error_list of the input sequence.

loop_as_string(loop_integer , loop_nucleotide_type): it returns loop_as_string of the input loop_integer.

check(first_run, first_loop, second_run, second_loop, third_run, third_loop, fourth_run, loop_nucleotide_type, P_list): it asks the user to check the folding of the sequence corresponding to the input and it returns user_answer, P_list, sequence updated to this new step.

eliminate(P_list, user_answer, first_loop, second_loop, third_loop): this function eliminates from P_list all the sequences that have their loops simultaneously ≥ than the corresponding first_loop, second_loop, third_loop when user_answer = 'yes', or ≤ than the corresponding first_loop, second_loop, third_loop when user_answer = 'no'.

find_next(P_list, minimum): this function returns the next_sequence. It computes this task by selecting the sequence of P_list with the greater value of p or lowering the loops length of the sequences to minimum.

screening(): it is the main function, it asks the user to input the upper_limit, loop_nucleotide_type, and the starting sequence and it returns the screening_summary and it prints to console the minimum list at the end of the screening.

The program is free to download from GitHub (https://github.com/micheleghezzo/I-motif_loop_minimizer).

Chapter 3: CT dinucleotide repeats fold into intra-molecular i-Motif structures.

It is quite common to find thymine-thymine (TT) base pairs in stacking in the outermost CC^+ of the iM core [54–57]. This interaction was also found in a tetra-molecular iM structure within the iM core [54]. Thus, we hypothesized whether TT base pairs could form within the core of intra-molecular iMs. The possibility of having thymines within the runs of cytosines would be a remarkable difference between the iMs in comparison to the G4s folding requirements and it would greatly expand the potential iM forming sequences. To solve the issue, we decided to study CT dinucleotide repeats. Indeed, in such oligonucleotides, there are no runs of cytosines, and their folding into iMs would prove our hypothesis.

Materials and methods

Materials

All tested oligonucleotides (sequences reported in Table 1) were synthesized and purified by Eurogentec (Liege, Belgium). They were dissolved in Milli-Q water to obtain 1 mM stock solutions and strand concentrations were determined from the absorbance at 260 nm. Before use, all DNA samples were melted for 5 minutes at 95 °C and then slowly cooled to room temperature.

Table 1. Sequences of oligonucleotides used in this study

CT38	5’-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-3’
CT37	5’-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC-3’
CT36	5’-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-3’
CT35	5’-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC-3’
TC38	5’-TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC-3’
TC37	5’-TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-3’
TC36	5’-TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC-3’
TC35	5’-TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-3’

Circular dichroism

CD spectra were acquired using a Jasco J 810 spectropolarimeter equipped with a Peltier as a temperature controller device in a 1 cm length quartz cuvette. Signals were reported as Δε.

Melting and annealing experiments were performed by setting ± 20 K h^{-1} temperature slopes and recoding the spectra every 2 K in the 230-330 nm wavelength range. The CD signal recorded at 287 nm during the melting and annealing steps was fitted with equation 2 according to Van't Hoff's formalism:

$$\Delta\varepsilon = \frac{a + b\, e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}}{1 + e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}} \quad (1)$$

where T is the temperature (K), ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}), T_m is the melting temperature (K), R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}) and a, b are the molar ellipticities of the unfolded and folded species, respectively.

Differential scanning calorimetry

Differential scanning calorimetry experiments were performed on a Microcal VP-DSC with cells of 502.7 µL in the 1-80 °C temperature range at stated heating-cooling rates. Multiple water-water, buffer-water, and buffer-buffer scans were performed before the analysis to derive the baseline thermogram and to check there were no heat exchanges due to the buffer in the set experimental condition. Samples were prepared at 200 µM DNA concentration in the required buffer. Data were reported as molar excess of heat capacity (ΔCp) as a function of the temperature.
Thermograms were fitted according to equation 3:

$$\Delta Cp = \frac{\Delta H^{o\,2}\, e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}}{R\, T^{2} \left(1 + e^{-\frac{\Delta H^{o}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)}\right)^{2}} \quad (2)$$

where T is the temperature (K), ΔH^{o} is the standard enthalpy change of folding (kJ mol^{-1}), T_m is the melting temperature (K), and R is the ideal gas constant (8.314 10^{-3} kJ mol^{-1}).

Results

To test whether TT can form within the iM core we decided to verify the folding into iM structure of CT dinucleotide repeats *in vitro*. CT dinucleotide repats, indeed, are sequences without any run of cytosines and with the only possibility to form CC^{+} and TT base pairs in an alternating frame.
Primarily, we decided to verify the presence and the lengths range of CT dinucleotide repeats in the human genome. To achieve this information, we developed a Python algorithm based on Re, FastaParser, NumPy,

and Matplotlib modules to search for dinucleotide repeats. We run the algorithm on the T2T-CHM13 new reference human genome, and interestingly we detected thousands of CT dinucleotide repeats longer than 20 nucleotides (Figure 1).

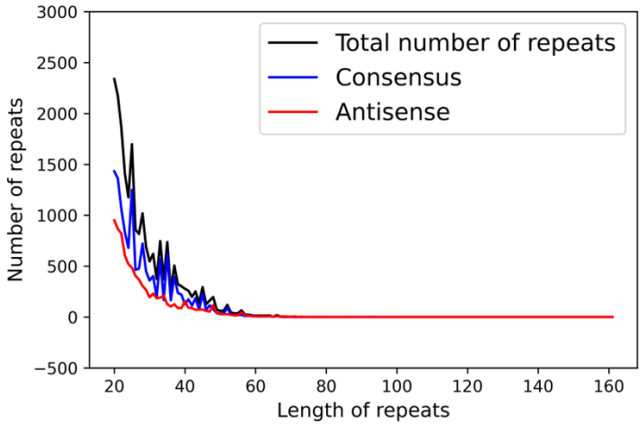

Figure 1: CT dinucleotide repeat search. Number of CT repeats per length of repeats in the consensus (blue line) and the antisense (red line) strand. The total number of repeats is reported as black line.

This very huge amount of long CT dinucleotide repeats suggests that they may have biological relevance. Therefore, we decided to test them in the length range of 35-38 nucleotides to verify if they may fold into iM structures *in vitro*.

The melting and annealing experiments performed on the CT repeats revealed that they fold into intra-molecular iM structures in our experimental conditions (Figure 9). The CD signals, indeed, recorded during the melting and annealing assays showed a perfectly reversible two-state transition with an iso-dichroic point at 275 nm. The processes were consistent with a thermodynamic equilibrium that evolves along temperature between an unfolded species, with a weak positive pick at 280 nm, and an iM one, with a strong positive pick at 285 nm and a negative one at 260 nm.

Worth noting that the folding process was DNA concentration independent, indeed DSC melting-annealing scans were consistent with the CD experiments which were performed at two orders of magnitude different DNA concentrations, and this confirmed that the iM species was an intra-molecular structure. Here only the experiments performed on CT38 are reported but the thermodynamic parameters of the studied CT repeats are reported in table 1.

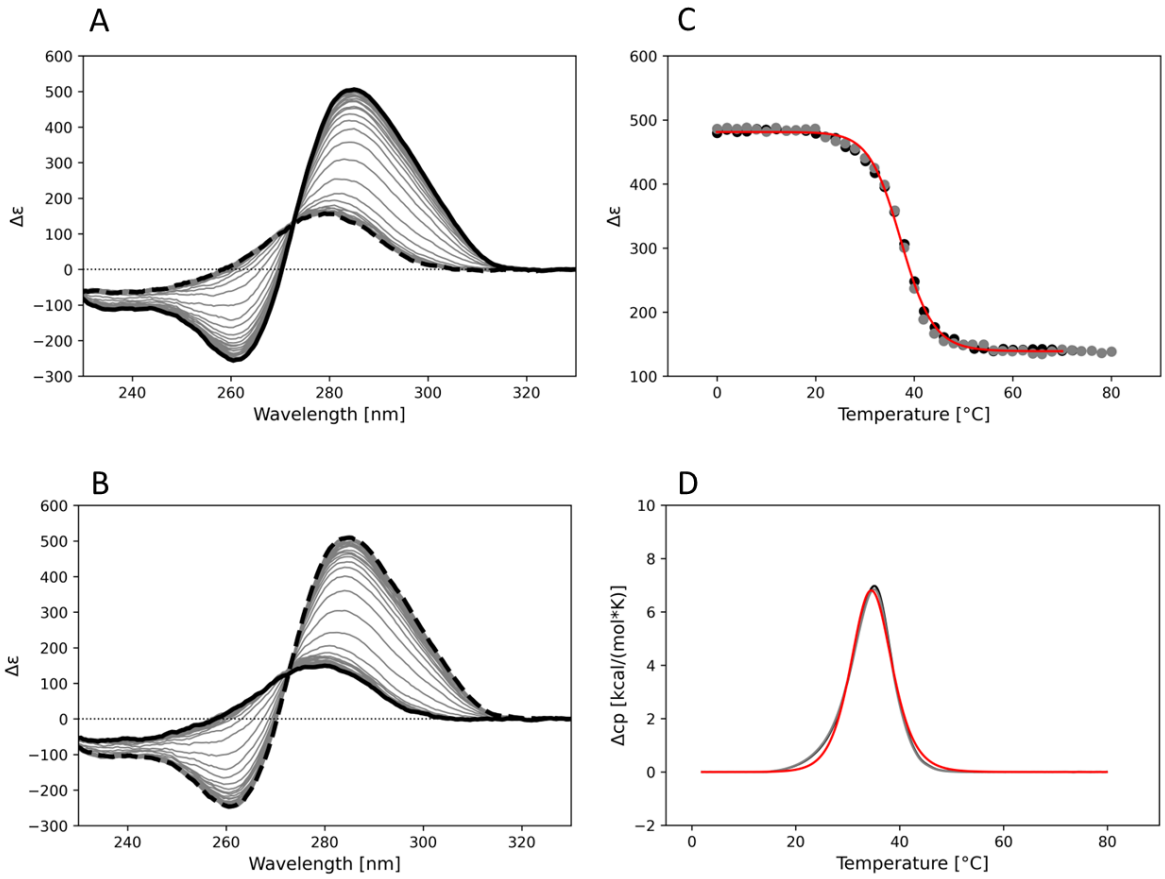

Figure 2: Melting-annealing experiments of CT38 oligonucleotide. A) CD spectra of 2 µM CT38 in 50 mM Na-cacodylate pH 5 from 0°C (solid black line) to 80°C (dashed black line). B) CD spectra of 2 µM CT38 in 50 mM Na-cacodylate pH 5 from 80°C (solid black line) to 0°C (dashed black line). C) CD signal at 285 nm of the melting (black dots) and annealing (grey dots) experiments and corresponding fitting curve (red solid line). D) DSC melting (black solid line) and annealing (grey solid line) scans of 200 µM CT38 in 50 mM Na-cacodylate pH 5 and corresponding fitting curve (red solid line).

Table 1. Thermodynamic parameters of the CT repeats in 50 mM Na-cacodylate pH 5.0 and referred to 298.15 K (25 °C).

oligonucleotide	ΔG kcal/mol	ΔH kcal/mol	-TΔS kcal/mol
CT38	-2 ± 1	-72 ± 1	70 ± 1
CT37	-2 ± 1	-70 ± 1	68 ± 1
CT36	-2 ± 1	-58 ± 1	56 ± 1
CT35	-2 ± 1	-63 ± 1	61 ± 1
TC38	-2 ± 1	-61 ± 1	59 ± 1
TC37	-2 ± 1	-61 ± 1	59 ± 1
TC36	-2 ± 1	-60 ± 1	58 ± 1
TC35	-2 ± 1	-62 ± 1	60 ± 1

These results prove that CT dinucleotide repeats can fold into iM structures, and this greatly expands the number of potential iM forming sequences in the human genome thus leading to reconsider the biological functions of this non-canonical DNA secondary structure.

Discussion

CT repeats can fold into intra-molecular iM structures in vitro. The thermodynamics of the iM structure in CT repeats are a bit different from iMs with C runs, in particular, ΔH/ΔS is lower for CT repeats and we suggest that this is due to the presence in the iM core of TT base pairs which, forming two hydrogen bonds, contribute less on the enthalpy than CC^+. These results highlight a remarkable difference between the G4s and iMs folding requirements. Indeed, while G4s need to have G runs, iMs can form with CT mixed runs. These findings have important implications for understanding the folding requirements of iM structures and for identifying potential iM forming sequences in the human genome. It suggests that the ensemble of potential iM forming sequences is much larger than previously thought and that the commonly used bioinformatic tools for screening potential iM forming sequences may not be sufficient to capture the full range of possibilities.

To explore this further, we conducted a search for CT repeats in the T2T-CHM13 new human reference genome and found thousands of sequences longer than 20 nucleotides that have the potential to form iM structures. These new potential iM forming sequences should be integrated with the ones that have already been identified for iMs. Moreover, accurately locating iM forming sequences in the human genome is crucial to better understanding the biological functions of these non-canonical DNA secondary structures. In conclusion, this study sheds light on the folding properties of iM structures with CT repeats, expands the range of potential iM forming sequences, and highlights the need for improved bioinformatic tools to accurately identify these sequences in the human genome.

Conclusion

My Ph.D. started with the structural and thermodynamic characterization of an iM that was expected to form 37 nucleotides upstream of the transcription starting site of the *EGFR* oncogene. The *in-silico* analysis matched the sequence as a potential 6 CC^+ base paired iM. We proved that it folds into an intra-molecular iM structure *in vitro*, nevertheless, the derived thermodynamic parameters as well as the S1 endonuclease cleavage assay indicated that the iM core is held by 4 CC^+ and additional TT and GC base pairs. We noticed that the central loop was predicted to be 1 nucleotide long, but the experimental data revealed that the 2 cytosines from the second and third runs did not pair to make this loop longer. This evidence prompted us to consider that 1 nucleotide long loops may not fit with the iMs as they do with G4s.

Therefore, we decided to investigate this topic, and we developed a novel step-by-step pipeline for the systematic screening of iM models. We applied it to determine the minimal length of the loops allowing the folding into an intra-molecular iM focussing on structures comprising only 3 CC^+ base pairs. We found 5'-CCTTTCTTCCTTTC-3' as the minimal iM folding model assuming the major-minor-major topology. This folding topology was conserved even in the 5'-CCTTTCTTTCCTTTC-3', derived by an increment of the central loop, and its higher thermal stability allowed us to obtain the high-resolution structure of this iM model (PDB ID 7QDC). It was the first deposited structure of an I-motif with an odd number of CC^+ base pairs and, to date, the simplest model enclosing all the features of an intramolecular iM. Interestingly, once again we detected a TT base pair in stacking on the 3 CC^+ base pairs of this iM structure. This interaction is quite common to find in iM structures and we noticed that it well fits with the geometry of the CC^+, as an extension of the iM core. Therefore, we speculated whether it could form within the iM core as well. To solve the issue, we decided to study CT dinucleotide repeats in which there are no runs of cytosines, and their folding would form an alternate intercalation of CC^+ and TT base pairs. Our results proved that CT mixed iMs can fold as well, greatly expanding the ensemble of potential iM forming sequences.

The results of my Ph.D. prove that the oligonucleotide pattern of a potential i-Motif forming sequence is different from the G-quadruplex one and this potentially clusters these secondary structures at different genome locations, from which may derive different biological functions.

This knowledge represents a step forward to the development of prediction tools for the proper identification of bio-functional i-Motifs as well as for the rational design of these secondary structures for technological applications.

Bibliography

1. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. Nature Publishing Group; 1953;171:737–8.

2. Bansal A, Kaushik S, Kukreti S. Non-canonical DNA structures: Diversity and disease association. Frontiers in Genetics [Internet]. 2022 [cited 2023 Jan 3];13. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2022.959258

3. Lobachev KS, Rattray A, Narayanan V. Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. Frontiers in Bioscience-Landmark. IMR Press; 2007;12:4208–20.

4. Jain A, Wang G, Vasquez KM. DNA triple helices: Biological consequences and therapeutic potential. Biochimie. 2008;90:1117–30.

5. Rigo R, Palumbo M, Sissi C. G-quadruplexes in human promoters: A challenge for therapeutic applications. Biochimica et Biophysica Acta (BBA) - General Subjects. 2017;1861:1399–413.

6. Abou Assi H, Garavís M, González C, Damha MJ. i-Motif DNA: structural features and significance to cell biology. Nucleic Acids Research. 2018;46:8038–56.

7. Monsen RC, Trent JO, Chaires JB. G-quadruplex DNA: A Longer Story. Acc Chem Res. American Chemical Society; 2022;55:3242–52.

8. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33:2908–16.

9. Abou Assi H, Garavís M, González C, Damha MJ. i-Motif DNA: structural features and significance to cell biology. Nucleic Acids Research. 2018;46:8038–56.

10. Huppert JL. Prevalence of quadruplexes in the human genome. Nucleic Acids Research. 2005;33:2908–16.

11. Belmonte-Reche E, Morales JC. G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. NAR Genomics and Bioinformatics. 2020;2:lqz005.

12. Kendrick S, Kang H-J, Alam MP, Madathil MM, Agrawal P, Gokhale V, et al. The Dynamic Character of the BCL2 Promoter i-Motif Provides a Mechanism for Modulation of Gene Expression by Compounds That Bind Selectively to the Alternative DNA Hairpin Structure. J Am Chem Soc. 2014;136:4161–71.

13. Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proceedings of the National Academy of Sciences. 2002;99:11593–8.

14. Gehring K, Leroy J-L, Guéron M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. Nature. 1993;363:561–5.

15. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature. 1988;334:364–6.

16. Amato J, D'Aria F, Marzano S, Iaccarino N, Randazzo A, Giancola C, et al. On the thermodynamics of folding of an i-motif DNA in solution under favorable conditions. Phys Chem Chem Phys. 2021;23:15030–7.

17. Mergny J-L, Lacroix L, Han X, Leroy J-L, Helene C. Intramolecular Folding of Pyrimidine Oligodeoxynucleotides into an i-DNA Motif. J Am Chem Soc. 1995;117:8887–98.

18. Berger I, Egli M, Rich A. Inter-strand C-H...O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4' in cytosine-rich DNA. Proceedings of the National Academy of Sciences. 1996;93:12116–21.

19. Amato J, D'Aria F, Marzano S, Iaccarino N, Randazzo A, Giancola C, et al. On the thermodynamics of folding of an i-motif DNA in solution under favorable conditions. Phys Chem Chem Phys. The Royal Society of Chemistry; 2021;23:15030–7.

20. Mergny J-L, Lacroix L, Han X, Leroy J-L, Helene C. Intramolecular Folding of Pyrimidine Oligodeoxynucleotides into an i-DNA Motif. J Am Chem Soc. American Chemical Society; 1995;117:8887–98.

21. Ts'o POP. Bases, nucleosides, and nucleotides. In: Ts'o POP, editor. Basic Principles in Nucleic Acid Chemistry. 1974. p. 453–584.

22. Dzatko S, Krafcikova M, Hänsel-Hertsch R, Fessl T, Fiala R, Loja T, et al. Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells. Angewandte Chemie International Edition. 2018;57:2165–9.

23. Zeraati M, Langley DB, Schofield P, Moye AL, Rouet R, Hughes WE, et al. I-motif DNA structures are formed in the nuclei of human cells. Nature Chem. 2018;10:631–7.

24. Shirmanova MV, Druzhkova IN, Lukina MM, Matlashov ME, Belousov VV, Snopova LB, et al. Intracellular pH imaging in cancer cells in vitro and tumors in vivo using the new genetically encoded sensor SypHer2. Biochimica et Biophysica Acta (BBA) - General Subjects. 2015;1850:1905–11.

25. Karna D, Stilgenbauer M, Jonchhe S, Ankai K, Kawamata I, Cui Y, et al. Chemo-Mechanical Modulation of Cell Motions Using DNA Nanosprings. Bioconjugate Chem. American Chemical Society; 2021;32:311–7.

26. Dong Y, Yang Z, Liu D. DNA Nanotechnology Based on i-Motif Structures. Acc Chem Res. American Chemical Society; 2014;47:1853–60.

27. Modi S, M. G. S, Goswami D, Gupta GD, Mayor S, Krishnan Y. A DNA nanomachine that maps spatial and temporal pH changes inside living cells. Nature Nanotech. Nature Publishing Group; 2009;4:325–30.

28. Nonin-Lecomte S, Leroy JL. Structure of a C-rich strand fragment of the human centromeric satellite III: a pH-dependent intercalation topology11Edited by K. Nagai. Journal of Molecular Biology. 2001;309:491–506.

29. Guéron M, Leroy J-L. The i-motif in nucleic acids. Current Opinion in Structural Biology. 2000;10:326–31.

30. Cheng M, Qiu D, Tamon L, Ištvánková E, Víšková P, Amrane S, et al. Thermal and pH Stabilities of i-DNA: Confronting in vitro Experiments with Models and In-Cell NMR Data. Angewandte Chemie International Edition. 2021;60:10286–94.

31. Berger I, Egli M, Rich A. Inter-strand C-H...O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4' in cytosine-rich DNA. Proc Natl Acad Sci U S A. 1996;93:12116–21.

32. Park HS, Jang MH, Kim EJ, Kim HJ, Lee HJ, Kim YJ, et al. High EGFR gene copy number predicts poor outcome in triple-negative breast cancer. Mod Pathol. 2014;27:1212–22.

33. Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer. 2007;7:169–81.

34. Cooper AJ, Sequist LV, Lin JJ. Third-generation EGFR and ALK inhibitors: mechanisms of resistance and management. Nat Rev Clin Oncol. 2022;1–16.

35. Wykosky J, Fenton T, Furnari F, Cavenee WK. Therapeutic targeting of epidermal growth factor receptor in human cancer: successes and limitations. Chin J Cancer. 2011;30:5–12.

36. Cristofari C, Rigo R, Greco ML, Ghezzo M, Sissi C. pH-driven conformational switch between non-canonical DNA structures in a C-rich domain of EGFR promoter. Sci Rep. 2019;9:1210.

37. Greco ML, Kotar A, Rigo R, Cristofari C, Plavec J, Sissi C. Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter. Nucleic Acids Res. 2017;45:10132–42.

38. Mergny J-L, Li J, Lacroix L, Amrane S, Chaires JB. Thermal difference spectra: a specific signature for nucleic acid structures. Nucleic Acids Res. 2005;33:e138.

39. Weiss JN. The Hill equation revisited: uses and misuses. Faseb J. 1997;11–835.

40. Han X, Leroy J-L, Guéron M. An intramolecular i-motif: the solution structure and base-pair opening kinetics of d(5mCCT3CCT3ACCT3CC)11Edited by I. Tinoco. Journal of Molecular Biology. 1998;278:949–65.

41. Phan AT, Guéron M, Leroy J-L. The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. Journal of Molecular Biology. 2000;299:123–44.

42. Mir B, Serrano I, Buitrago D, Orozco M, Escaja N, González C. Prevalent Sequences in the Human Genome Can Form Mini i-Motif Structures at Physiological pH. J Am Chem Soc. American Chemical Society; 2017;139:13985–8.

43. Serrano-Chacón I, Mir B, Escaja N, González C. Structure of i-Motif/Duplex Junctions at Neutral pH. J Am Chem Soc. American Chemical Society; 2021;143:12919–23.

44. Lim KW, Lacroix L, Yue DJE, Lim JKC, Lim JMW, Phan AT. Coexistence of Two Distinct G-Quadruplex Conformations in the hTERT Promoter. J Am Chem Soc. American Chemical Society; 2010;132:12331–42.

45. Trajkovski M, Webba da Silva M, Plavec J. Unique Structural Features of Interconverting Monomeric and Dimeric G-Quadruplexes Adopted by a Sequence from the Intron of the N-myc Gene. J Am Chem Soc. American Chemical Society; 2012;134:4132–41.

46. Trajkovski M, Endoh T, Tateishi-Karimata H, Ohyama T, Tanaka S, Plavec J, et al. Pursuing origins of (poly)ethylene glycol-induced G-quadruplex structural modulations. Nucleic Acids Research. 2018;46:4301–15.

47. Marušič M, Veedu RN, Wengel J, Plavec J. G-rich VEGF aptamer with locked and unlocked nucleic acid modifications exhibits a unique G-quadruplex fold. Nucleic Acids Research. 2013;41:9524–36.

48. Wright EP, Huppert JL, Waller ZAE. Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. Nucleic Acids Research. 2017;45:2951–9.

49. Fleming AM, Ding Y, Rogers RA, Zhu J, Zhu J, Burton AD, et al. 4n–1 Is a “Sweet Spot” in DNA i-Motif Folding of 2′-Deoxycytidine Homopolymers. J Am Chem Soc. American Chemical Society; 2017;139:4682–9.

50. Školáková P, Renčiuk D, Palacký J, Krafčík D, Dvořáková Z, Kejnovská I, et al. Systematic investigation of sequence requirements for DNA i-motif formation. Nucleic Acids Research. 2019;47:2177–89.

51. Iaccarino N, Cheng M, Qiu D, Pagano B, Amato J, Di Porzio A, et al. Effects of Sequence and Base Composition on the CD and TDS Profiles of i-DNA. Angewandte Chemie International Edition. 2021;60:10295–303.

52. Lee W, Tonelli M, Markley JL. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics. 2015;31:1325–7.

53. Hur JH, Kang CY, Lee S, Parveen N, Yu J, Shamim A, et al. AC-motif: a DNA motif containing adenine and cytosine repeat plays a role in gene regulation. Nucleic Acids Res. 2021;49:10150–65.

54. Leroy JL. T·T Pair Intercalation and Duplex Interconversion Within i-Motif Tetramers. Journal of Molecular Biology. 2003;333:125–39.

55. Gehring K, Leroy JL, Guéron M. A tetrameric DNA structure with protonated cytosine.cytosine base pairs. Nature. 1993;363:561–5.

56. Han X, Leroy J-L, Guéron M. An intramolecular i-motif: the solution structure and base-pair opening kinetics of d(5mCCT3CCT3ACCT3CC)11Edited by I. Tinoco. Journal of Molecular Biology. 1998;278:949–65.

57. Esmaili N, Leroy JL. i-motif solution structure and dynamics of the d(AACCCC) and d(CCCCAA) tetrahymena telomeric repeats. Nucleic Acids Res. 2005;33:213–24.

58. Kypr J, Kejnovská I, Renčiuk D, Vorlíčková M. Circular dichroism and conformational polymorphism of DNA. Nucleic Acids Research. 2009;37:1713–25.

59. Ruggiero E, Lago S, Šket P, Nadai M, Frasson I, Plavec J, et al. A dynamic i-motif with a duplex stem-loop in the long terminal repeat promoter of the HIV-1 proviral genome modulates viral transcription. Nucleic Acids Res. 2019;47:11057–68.