PNAS
www.pnas.org

Main Manuscript for

i-DNA stability: confronting in vitro experiments with models and in-cell NMR data

Mingpan Cheng[†,‡], Dehui Qiu[†], Liezel Tamon[§], Eva Maturová[¶], Pavlína Víšková[¶], Samir Amrane[‡], Aurore Guédin[‡], Jielin Chen[†,‡], Laurent Lacroix[#], Huangxian Ju[†], Lukáš Trantírek[¶], Aleksandr B. Sahakyan[§], Jun Zhou[†,*] & Jean-Louis Mergny[†,‡,||]

[†] State Key Laboratory of Analytical Chemistry for Life Science, School of Chemistry & Chemical Engineering, Nanjing University, Nanjing 210023, China.

[‡] ARNA Laboratory, Université de Bordeaux, Inserm U 1212, CNRS UMR5320, IECB, Pessac 33607, France.

[§] MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, UK.

[¶] Central European Institute of Technology, Masaryk University, Brno 62500, Czech Republic.

[#] IBENS, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, Paris 75005, France.

[||] Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, Brno 61265, Czech Republic.

* Corresponding author Jun Zhou; Email: *jun.zhou@nju.edu.cn*

Classification

Biophysics and Computational Biology; Biochemistry

Keywords

i-motif; pH and sequence effects; in-cell NMR; thermal stability; stability prediction

Author Contributions

M.C., J.Z. and J.L.M. designed research project; M.C., D.Q., E.M., P.V. performed the experiments; Liezel T. and A.B.S. performed the calculation of predication model; M.C., Liezel T., E.M., P.V., S.A., A.G., J.C., L.L., H.J., Lukáš T., A.B.S., J.Z. and J.L.M. analyzed data; M.C., Lukáš T., A.B.S., J.Z. and J.L.M wrote the paper and all other authors revised the paper.

This PDF file includes: Main Text; Table 1; Figures 1 to 8

Abstract

i-DNA, also known as the i-motif, is a four stranded DNA structure stemming from the interlocking of two equivalent parallel-stranded right-handed duplexes. Since its discovery in 1993, efforts have been made to understand i-DNA sequence requirements. Recent studies indicate that the i-motif is actually present within human cells; however, a systematic study on the effect of spacer, loop and C-track length has been missing. In the present work, an unprecedented number of different sequences (236) bearing four runs of 3-6 cytosines with different spacer lengths were tested under identical conditions. The central spacer played a special role on i-motif stability, but i-DNA stability was found to be nearly independent on total spacer length while stability significantly increased with the length of the C-tracts at both acidic and neutral pH. This study provides a global picture on i-DNA stability thanks to the large size of the introduced data set; it revealed unexpected fatures and allowed to conclude that determinants of i-DNA stability do not mirror those of G-quadruplexes (G4s) that may be formed with the complementary G-rich strand. Our results illustrate the structural roles of loops and C-tracks on i-DNA stability, and allow establishing rules to predict the stability of i-motifs.

Significance Statement

i-DNA, also called i-motif, is a fascinating DNA structure with extreme pH-sensitivity. i-DNA plays an important role in the regulation of gene functions and serves as an environmentally-responsive element for DNA-based nanotechnology. We established models to predict the stability of i-DNAs based on the analysis of hundreds of different sequences. Stabilities determined *in vitro* were compared with results in cells. This work provides a global picture of i-DNA under near physiological conditions and provides tools to design i-DNA-based programmable nanodevices.

Introduction

i-DNA (also referred to as the i-motif) is a fascinating nucleic acid structure discovered in the 1990s by Maurice Guéron and colleagues in *Ecole Polytechnique,* France (1). It is a four-stranded structure stemming from the interlocking of two equivalent parallel-stranded right-handed duplexes. Such structure, which relies on the formation of hemi-protonated C•C^+ base pairs (see **Figure 1A**), can be formed with two or more independent strands, or be intramolecular, as depicted in **Figure 1B**, provided that four runs of cytosines are present (2-6). The systematic intercalation between each C·C^+ base pair forces each individual duplex to be severely underwound and extended. i-DNA is somewhat polymorphic, even if strand orientation is strictly imposed as the two diametrically distant strands must remain parallel to each other and adjacent strands are always running in opposite orientations. For example, the intramolecular structure formed by four repeats of the human telomeric motif (CCCTAA) can adopt four configurations, in which all cytosines are base-paired and with minimal energetic differences (2). In addition, bi- or tetra-molecular complexes may coexist with intramolecular structures. These observations may explain why spectroscopic measurements often reveal a multi-step denaturation/renaturation process (7-9).

Since its discovery, i-DNA remained in the shade of other unusual structures such as G-quadruplexes (abbreviated as “G4s”) formed by complementary G-rich sequences, given i-DNA’s limited stability under physiological conditions. Formation of each C·C^+ base pair requires the protonation of cytosine at its N3 position, given the relatively low pKa of this group (<5), and the stability of this motif is optimal under mildly acidic conditions but remains questionable at neutral pH (10). Increasing the solution pH by one unit typically leads to a decrease in T_m of 15°C or more (10). This extreme pH dependency can actually become an asset to design sensitive pH-responsive devices (reviewed in (11)) and i-DNA may therefore be relevant for analytical chemistry (12), nanotechnology (11, 13), and therapeutics (14). Regarding its biological relevance, two recent independent studies indicate that i-DNA is actually present within human cells; their conclusions are based on specific i-DNA antibodies (15) and in-cell NMR (16). Similar to the G4s, i-DNAs have

been found to modulate telomerase activity (17), transcription of genes (18, 19), and DNA biosynthesis (20).

Our understanding of i-DNA sequence requirements is still far from complete. Increasing cytosine tract lengths results in increased thermal stability; sequences with at least five cytosines per tract fold into i-motif at room temperature and neutral pH (7, 9, 10). Additional interactions involving hydrogen bonding also stabilize the structures (21, 22). Burrows and colleagues analyzed dC homo-oligonucleotides, and found that pure cytosine tracts may also adopt stable i-motif conformations (23, 24). Interestingly, the relation between C tract and stability was not linear. The authors found that dC_n strands of length 15, 19, 23, and 27 nucleotides (*i.e.*, 4n−1) have optimal stabilities, with high pHs of mid-transition and thermal stabilities above 37 °C at pH 7.0. These requirements somewhat mirror those for G4 formation (25); as a consequence, the complementary strand of a G4-forming sequence is generally prone to i-DNA formation. Besides C-tracts, loop regions including loop length and base composition also play roles in i-DNA formation (26-31). However, contradictory conclusions have been drawn upon how loop length influences i-DNA stability (23, 27, 29, 32). These results came from the investigations of a limited number of sequences; systematic studies based on large numbers of examples are needed to achieve an objective conclusion.

In addition, i-DNA stability depends on a number of solution parameters. First and foremost, pH plays a critical role as discussed above. In contrast with B-DNA, i-DNA is also favored by crowding (33-35). On the other hand, ionic strength plays a limited role, as the electrostatics of the i-motif are unusual: while i-DNA involves four negatively charged strands, its net linear charge is decreased by the positive charge on each $C{\cdot}C^+$ base pair and the extended backbone resulting from intercalation. As a consequence, increasing the ionic strength or adding dication has limited, if any, effect on i-motif stability: when working at pH 5.5 or higher, stability is actually *higher* with no salt added (2, 10, 36).

In this study, we performed a systematic analysis of i-motif stability on an unprecedented large selection of model C-rich sequences (236 different sequences, listed in **Tables S1-S3**). This unique dataset unveiled important parameters governing the stability of this structure. Global trends were easily identified and more subtle effects were found using machine learning and other modeling approaches within our subset of i-DNA sequences, allowing us to predict i-DNA stability from primary sequence with reasonable accuracy. i-DNA formation in cells with motifs stable *in vitro* at near neutral pH was confirmed by in-cell NMR for the most stable motifs *in vitro*.

Results

Sequence design

Table S2 summarizes the results obtained for 60 groups of three sequences with different spacer arrangements. Each sequence has four C-tracts of equal length (C_3 to C_6) separated by three spacer regions, which should allow the formation of an intramolecular i-DNA structure (10). The C_3 to C_6 range was chosen as i-DNA becomes unstable for shorter (C_2) C-tracts, and is prone to form competing structures (inter- or intra-molecular) when C-tract length is longer than six (7, 9, 10). In order to reduce the number of spacer arrangements, most sequence groups contain two identical spacers. Each spacer involves one to six thymine nucleotides, and total spacer length was capped at twelve nucleotides. Note that the term "*spacer*" corresponds here to the non-C nucleotides connecting C-tracks: as some cytosines may also participate to loops rather than to the i-motif stem, the operational *loop* length may therefore be longer than the *spacer* composed of thymines only.

Evidence for i-DNA formation

First, i-DNA formation was checked for all sequences under acidic (pH 5.0) or neutral (pH 7.0) conditions. Thermal difference spectra (TDS) of an i-DNA structure exhibits two major peaks, one

positive and one negative, at 239 ±1 and 294 ±1 nm, respectively (37). TDS spectra of all sequences are provided in **Figure S1**. It is clear from these spectra that all sequences fold into an i-motif in pH 5.0. In addition, i-DNA formation was also demonstrated by the presence of imino proton peaks at 15 to 16 ppm in 1D ^{1}H NMR spectra recorded at 20 ^{o}C. These measurements were performed on a selection of twelve sequences with C_5 or C_6-tracts. These peaks are assigned to the imino protons of protonated cytosines (**Figure S2**) (2, 5).

The molecularity (number of independent strands involved) of the 49 sequences with a C_5-tract in **Table S2** was checked at both pH 5.0 and 7.0 by non-denaturing PAGE (**Figures S3** and **S4**). All sequences tested fold into intramolecular species as expected, in agreement with previous studies (7, 9). The conclusions drawn from these work therefore apply to intramolecular i-DNAs, which are more likely to be physiologically relevant at the genome level. Once intramolecular i-DNA formation was established, we wished to examine their stability.

In contrast with TDS recorded at pH 5.0, the situation was more contrasted at neutral pH. We divided the 60 groups into four classes, based on the number of sequences in the same group that fold into an i-DNA at neutral pH (**Figure 2,** dashed lines):

I. None of the three sequences in a given group fold into an i-DNA structure at neutral pH. This category includes all groups with C_3-tract (**Figures 2A**, **2B** and **S1A**), the *T336-4* group (**Figure 2C**) and the *T336-5* group (**Figure 2D**);

II. Only one of three sequences within the same group folds into an i-DNA structure at neutral pH. This category includes T121-4 (**Figure 2E**), T161-4 (**Figure 2F**), T262-4 (**Figure 2G**) as well as T353-5 (**Figure 2H**);

III. Two of the three sequences in the same group fold into an i-DNA structure at neutral pH. In this category, sequences that do not fold within a group include T225-4 (**Figure 2J**), T335-4 (**Figure 2K**), and T225-5 (**Figure 2L**);

IV. All three sequences in the same group fold into an i-DNA structure (**Figures 2M-P** and **S1B-C**).

An interesting trend already emerges from this classification. In types *II* and *III* categories (for which some, but not all, group members form an i-DNA structure at pH 7.0), the sequence with a longer central spacer folds into an i-DNA while one or the two other group members do not form, or only partially form, an i-DNA structure. This suggests that sequences with a longer central spacer are relatively more stable at neutral pH.

i-DNAs with a long central spacer exhibits higher pH mid-transition point and melting temperature

To confirm the differences in stability inferred from TDS recorded at pH 7.0, we performed UV-melting and CD measurements. C-rich sequences switch from stable i-DNA to random coil when increasing pH from acidic to neutral or basic, and this transition can be followed by changes in ellipticity (**Figures S5-S9**) and absorbance (**Figures S10-S14**). pH transition mid-points (pH_T) of all sequences are depicted in **Figures S9** and **S14** for pH-dependent CD and UV absorbance spectra, respectively; the consistency between pH_T obtained by ellipticity and absorbance was checked (**Figure S15**). pH_T values were calculated from these curves and are provided in **Tables S1-S2**.

A consistent trend emerged from the comparison of pH_T values: in most groups, the sequence with a longer central spacer has a higher pH_T than other sequences in the same group. For example, in the T112-3 group, pH_T of T121-3 (6.30) is higher than the ones of T112-3 (6.11) and T211-3 (6.12) (**Figure S5B**). A precise count of groups obeying this *"long central spacer is better"* rule is presented in **Table 1**. Based on CD and UV absorbance spectra, 48 or 46 of the 60 groups (80% and 77%) follow this tendency, respectively.

The thermal stability of i-DNA can be followed by UV-absorbance at 260 or 295 nm; in the latter case an inverted transition is obtained as i-DNA denaturation leads to a *decrease* in absorbance at this wavelength (38). Herein, the denaturation of i-DNAs at pH 5.0 and pH 7.0 was tracked by monitoring UV-absorbance at 295 nm (**Figures S16-S18**). At neutral pH, only sequences with longer C-tracts such as C_5 and C_6 were considered, as sequences with shorter C-tracts do not fold or exhibit a low stability (T_m < 12 °C) preventing an accurate determination.

Folding and unfolding processes follow relatively fast kinetics under mildly acidic conditions, as expected for intramolecular folding. However, this is no longer the case at near-neutral pH, where a hysteresis phenomenon occurs, leading to large differences in apparent mid-transition point (T_m) upon heating and cooling at near neutral solution of an intramolecular process (7, 10). For some sequences, such as T444-6, T336-6, T363-6 and T633-6, this difference in melting/cooling temperatures can reach 19 °C (**Figure S17**). As a first approximation, T_m at pH 7.0 is assumed to be equal to the average of half-transition values for heating and cooling curves (39).

The analysis of T_m values confirmed the *"long central spacer is better"* rule inferred from TDS analysis (at both pH 5.0 and 7.0): for most groups, the sequence with a longer central spacer has a higher T_m than the other sequences in the same group (**Figure S18**). For example, in the *T114-5* group at pH 5.0, the T_m of the sequence T141-5 (74.2 °C) is higher than the one of T114-5 (69.5 °C) or T411-5 (70.6 °C). At pH 7.0, a similar result is found, although all T_m are much lower: the T_m of sequence T141-5 is 17.0 °C only, but still higher than the ones of T114-5 (13.6 °C) and T411-5 (14.8 °C) (**Table S2**). The counts of groups obeying this rule are summarized in **Table 1**: 24/30 and 60/60 follow this trend at pH 7.0 and 5.0, respectively. Statistical analyses are provided for all hypotheses related to this *"long central spacer is better"* rule.

Analyses of effects of spacer permutation are presented in **Figures 3A-F**. Sequences are divided into two categories: *i)* sequences with two long (L) and one short (S) spacers and *ii)* sequences with two short and one long spacers. Average and median values of pH_T and T_m of sequences with a relative longer central spacer, including *SLS* (**Figures 3A-C**), *LLS* and *SLL* (**Figures 3D-F**) are obviously higher than that of the corresponding sequences with a shorter central spacer (*SSL* and *LSS*, *LSL*). Considering that three sequences in the same group are generated by spacer permutations, any two sequences of them are treated as a paired sample. Then hypotheses of pair-sample *t*-test are performed between every two spacer combinations. Except for 3 comparisons (*LLS* versus *LSL* and *LSL* versus *SLL* shown in **Figure 3F**, and *LLS* versus *LSL* in **Figure 3D**), all 9 other *t*-tests support the conclusion that pH_T and T_m of the sequences with a longer central spacer are significantly higher (p < 0.05; *SLS* versus *SSL* or *LSS* in **Figures 3A-C**; *LLS* or *SLL* versus *LSL* in **Figures 3D-E**). In addition, except for one comparison (*SSL* versus *LSS* in **Figure 3B**), all 5 other *t*-tests show that the differences of pH_T and T_m values between two sequences from the same group that have the identical central spacer are not significant (p > 0.05). This *"stability-spacer length-symmetry"* may come from the linking pattern of three loops in intramolecular i-DNAs, by Leroy *et al.* in 1994 (2) and depicted in **Figures 1B** and **C**. Three loops stretch and pass through either major-minor-major grooves (conformation I) or minor-major-minor grooves (conformation II). Given the results obtained here, assuming spacer length would allow both possibilities, conformation I generally appears less stable than conformation II.

Thermal stabilities of 12 sequences in 4 groups (*T112-5*, *T225-5*, *T112-6* and *T225-6*) at pH 5.0 and 7.0 were also evaluated by DSC (**Figure S19**) and values of T_m and hysteresis are summarized in **Table S4**. These thermal data are consistent with those obtained by UV experiments. The *"long central spacer is better"* was also observed for 7 of 8 group datasets.

Stability depends on length of C-tract but not on spacer length

pH_T (from 196 sequences) or T_m (from 196 or 98 sequences at pH 5.0 and 7.0, respectively) are presented in **Table S2** (at pH 7.0, only the sequences with C_5 and C_6-tracts are used) and plotted

as a function of total spacer length in **Figures 3G-F**. Values of pH_T or T_m are widely distributed for each spacer length from 3 to 12, and little or no correlation was found between T_m and total spacer length. These results indicate that the total spacer length (when considered in the 3-12 range) has a limited effect on the stability of i-DNA at both acidic and neutral pH. This maybe the reason why previously reported studies about the effects of loop length on i-DNA are contradictory or negligible (7, 9, 23, 27, 29).

Longer C-tracts stabilize the i-DNA by providing more stacked $C{\cdot}C^+$ base pairs, as demonstrated by us and others (7, 9, 10). However, quantitative analyses of the relationships between pH_T or T_m *vs* C-tract length were missing. We provide here a quantitative assessment of C-tract role on i-DNA stability as depicted in **Figure 4**. Stability increases with C-tracts length: averages of pH_T with C_3, C_4, C_5 and C_6-tracts are 6.11, 6.39, 6.56 and 6.68, respectively (**Figure 4A**). A similar relationship was found between T_m and C-tract length under both acidic and neutral solutions (**Figures 4B** and **C**). The increase in pH_T is monotonous but not linear: the average difference between C_4 and C_3, C_5 and C_4 or C_6 and C_5 is 0.28, 0.17 or 0.12, respectively. Of note, sequences with C-tracts longer than six are prone to intermolecular i-DNA formation, and the corresponding pH_T increase with C-tract length becomes small (7, 9).

Unfolding and folding rates depend on both C-tract and loop lengths

As noted before, thermal transitions at near-neutral pH are no longer reversible, and a hysteresis phenomenon is observed: the apparent melting transition upon heating is shifted towards higher temperatures than the value deduced from cooling profiles (**Figures 5 and S17**). The analysis of melting (heating) profiles alone would only lead to an overestimation of i-DNA thermal stability at neutral pH. Previous observations allowed to conclude that the average of $T_{Heating}$ and $T_{Cooling}$ provides a reasonable estimate of the thermodynamic T_m (at equilibrium, using an infinitely slow temperature gradient). What was not reported before is the strong dependency of the hysteresis phenomenon on total loop length, found both for C_5 and C_6 sequences (**Figure 5**): in other words, sequences with long T-loops fold and unfold slower than motifs with shorter T-loops. The hysteresis, induced by longer sequence length and higher pH value, is also observed in the DSC-melting and annealing experiments (**Figure S19** and **Table S4**). For this reason, the analysis of melting (heating) curves only would provide a wrong picture of i-DNA stability and lead to the inaccurate conclusion that stability increases with loop length. Restricting the analysis to cooling profiles would actually lead to the opposite conclusion.

Expanding the "long central spacer is better" rule

All oligonucleotides studied above belong to a relatively narrow sequence space, in which *(i)* loops are entirely composed of thymines, *(ii)* total loop length is 12 or lower, *(ii)* two loops are of identical size and *(iv)* no individual loop involves more than 6 nucleotides. In order to validate our conclusions for a wider variety of motifs, we analyzed i-DNA stability for sequences that escape one or more of the conditions listed above. These oligonucleotides are listed in the last part of **Table S1**. i-DNA formation was confirmed by TDS (**Figure S20**). pH_T and T_m were also evaluated (**Figure S21**) and are given in **Table S3**. For example, stability of sequences containing a longer central loop was analyzed, from 7 to 15 nucleotides, and results are summarized in **Figure S22**. These sequences allow us to conclude that i-DNA motif is still possible with a relatively long central loop (T_m is moderately affected while the drop in pH_T is more significant). In addition, this bell curve indicates that an optimal central loop length is 2-7 nucleotides for both T_m and pH_T.

Then *t*-tests show that the differences in pH_T and T_m values (**Table S3** and **Figures S23-S24**) between two sequences produced by swapping positions of two relatively short loops (*SLM* versus *MLS*, where *S, M* and *L* are short for relatively *short*, *middle*, *long* loop length, respectively) are not significant ($p > 0.05$) (**Figure S25**). This *"stability-loop length-symmetry"* is similar to the one disclosed above (**Figure 3**).

Replacement of one or two thymine residues in loops by adenine of three sequences from T115-5 groups produces 24 sequences in 7 groups (**Table S1**). pH_T and T_m were measured (**Figures S26-S27**) and given in **Table S3** and **Figure S28**. Sequences with longer central loops from 6 of 7 groups and all 7 groups have higher pH_T and T_m, respectively. It demonstrates that the *"long central spacer is better"* rule can be extended to i-DNAs with different loop compositions.

Relative i-DNA stabilities in the intracellular environment parallel those found *in vitro*

As noted before, formation of i-DNAs from natively occurring C-rich DNA motifs in cells was recently demonstrated by using specific antibodies (9) and in-cell NMR spectroscopy (16). Notably, the NMR study revealed that stabilities of i-DNAs in cells might be slightly higher compared to those observed under simplistic conditions *in vitro* and that natively occurring intracellular environmental factors modulate i-DNA stability. To assess whether "rules of thumb" for formation of i-DNA accounting for length of C-tract and size of the loop derived within this study under *in vitro* conditions are applicable for assessment of i-DNA stabilities in living cells, we performed in-cell NMR experiments for four selected constructs (T212-4, T121-5, T121-6, and T343-6) differing by the virtue of their T_m and pH_T (**Table S2**). In-cell NMR spectra were acquired on a suspension of living HeLa cells transfected separately with individual constructs at 20 ºC (**Figure 7A**). As evidenced from confocal microscopy images, all transfected constructs were localized in the nuclei (**Figure S29**). Observation of signals in region of the in-cell NMR spectra specific for imino protons involved in $C{\cdot}C^+$ base pairs (15-16 ppm) corroborated formation of i-DNAs for T121-5, T121-6, and T343-6, while absence of signals indicated no formation of i-DNA for T212-4 (**Figure 7A**). Notably, the order of relative intensities of the imino signals in the in-cell NMR spectra of individual constructs (T121-6 > T343-6 > T121-5 >> T212-4) essentially paralleled that obtained from corresponding NMR spectra acquired under conditions *in vitro*. Although imino signal intensities provide only rough estimate of i-DNA stability (16), it is noteworthy that the derived relative order of the stabilities estimated from (in-cell) NMR data agrees with that derived from in vitro measured T_m values by UV-melting experiments (see **Table S2**). Altogether these data suggest that the rules derived on the basis of *in vitro* data are reasonable approximation for behavior of i-DNAs in cells: i-DNAs forming at near physiological pH *in vitro* are also likely to form in the intracellular environment.

However, at the same time, it needs to be considered that the absolute stabilities of i-DNAs in cells might differ from those observed *in vitro* (16). As shown in **Figures 7B** and **C**, the intensities of imino signals in temperature resolved in-cell NMR spectra of T121-6 are perturbed by increasing temperature to lower extent than those in the corresponding in vitro NMR spectra: while the absence of imino signals in *in vitro* NMR spectrum acquired at 32ºC suggests complete unfolding of T121-6, the detectable imino signals in the corresponding in-cell NMR spectrum measured at 36°C (and even 40 ºC) indicate that a significant population of folded species (i-DNA) is present in intracellular space at physiological temperature. In agreement with the observations by Dzatko *et al.* (16), these data demonstrate that i-DNAs might be more stable in cells than under simplistic conditions *in vitro*.

Predicting i-DNA stability

Models for i-DNA stability were generated using three distinct approaches G4Hunter-based (40), machine learning based (41), and analytical equation (42), as detailed in the Materials and Methods section in Supporting Material. We used the C/T-only restricted space for the i-DNAs, for which this work contributes an extensive set of systematic experimental data, therefore our models for melting temperatures (T_m at pH 5.0 or 7.0) or the pH transition mid-points (pH_T) can be used only to draw conclusions for C/T-based i-DNA structures (for instance, we do not take into account effect that may rise from competing Watson-Crick base-pairing while having G nucleobases in the loops) with similar restricted relation of the three spacer lengths (mostly with the two having the same length).

The first approach of creating a G4Hunter analogue for i-DNAs, while accounting for C-tract-based (instead of G4Hunter G-tracts) scores and optimizing the scoring coefficients, resulted in models

that assign overall scores (iM_{score}) to i-DNAs while capturing the T_m (T_m *pred* = 55.15 + 0.6440 iM_{score}^{Tm}, Pearson's R = 0.958, **Figure S31A**) and pH_T (pH_T *pred* = 6.13 + 0.0188 iM_{score}^{pHt}, Pearson's R = 0.915, **Figure S31B**) dependencies.

In general, the above approach resulted in all the scoring coefficients for the C-tracts of length 3 and shorter to be optimized to 0, retaining only the coefficients for the tracts of length 4 (24.1 for iM_{score}^{Tm}, 20.2 for iM_{score}^{pHt}), 5 (37.0 for iM_{score}^{Tm}, 31.4 for iM_{score}^{pHt}) and at-or-above-6 (45.1 for iM_{score}^{Tm}, 38.3 for iM_{score}^{pHt}). The 0 coefficients reflected the fact that all our sequences had at least 3 Cs in their C-tracts, thus eliminating the need to have a differentiating contributor from C tracts of length 3 or below.

Since the above model arrived to a simple scoring of only the C-tracts that are longer than 3, we embarked on developing an independent model by using all four metrics of the i-DNAs used in this study (see Materials and Methods in Supporting Materials). This approach with gradient boosting machines (GBM) as machine learning framework, resulted in models that capture the T_m and pH_T measurements with great performance (data from the 20 % left-out validation dataset, see **Figure 8**). The optimal GBM architecture for T_m was found to have 0.01 learning rate, interaction depth of 4, subsampling ratio of 0.6, minimum child weight of 5, and contained 1000 trees as individual learners. This resulted in a model with 1.210 RMSE (root mean squared error) and 0.990 Pearson's R while predicting the T_m values from the validation dataset (**Figure 8**). In contrast, the model developed for pH_T measurements had 0.01 learning rate, interaction depth of 6, subsampling ratio of 0.6, minimum child weight of 10, and contained 1500 trees as individual learners. The pH_T model had 0.053 RMSE and 0.973 Person's R, as applied on the validation dataset (**Figure 8**).

In our third, Eureqa-based, approach, we searched for a simple and interpretable non-linear analytical equation for both T_m and pH_T, expressing those as a function of C-tract and T-spacer lengths. With some compromises in the model performance, we arrived to the following mathematical expressions:

$$T_m = 102 - T_3 - (137 - T_2T_3 + T_1)/C \qquad \textit{(equation 1)}$$

$$pH_T = 7.38 - 3.70/C - (0.00565\ L)/T_2 \qquad \textit{(equation 2)}$$

in which L is the total sequence length:

$$L = 4C + T_1 + T_2 + T_3 \qquad \textit{(equation 3)}$$

C is the C-tract length (common for all four C-tracts), T_1, T_2 and T_3 are the lengths of the first, second and third spacers respectively (in 5'-to-3' direction), and the equations result in Pearson R values of 0.979 and 0.960 for T_m and pH_T values respectively, based on Eureqa's internal validation. As for all the other models above, these mathematical models are applicable for only the C/T-based sequence space with equally sized C-tracts used in this study for most experimental measurements. Furthermore, the found other Eureqa solutions show comparable performance, due to the internal restrictions on the spacer lengths in the used experimental dataset (in most cases, two spacers being equal in length, hence some candidate solutions eliminating some of the spacers). The equations are consistent with our observations in the explored i-DNA subspace, and capture the stabilizing role of the lengthy middle spacer length (T_2) within a given overall length of i-motifs.

Discussion

Our work on i-DNA sequence requirements is of unprecedented magnitude, with 236 different sequences tested under a variety of conditions. Even if impressive, this dataset does not allow to explore the full sequence space of i-DNA-prone sequences: most oligonucleotides studied above have *(i)* loops that are entirely composed of thymines with *(ii)* total loop length of 12 nucleotides or lower, *(ii)* two loops of identical size and *(iv)* no individual loop longer than 6 nucleotides. Even with these restrictions, and because we tested a few sequences escaping this sequence space, our data already provides key information on i-DNA stability.

Two parameters are useful to monitor i-DNA stability: pH_T (at a given temperature) or T_m (at a given pH). As we found difficult to discard one, both were used in this manuscript, and it is difficult to conclude that one is superior to the other. If biological applications are contemplated, T_m at physiological pH and pH_T at physiological temperature would be recommended, although the accurate determination of intracellular (intranuclear) pH may prove harder than expected (see below). Of note:

- pH_T was determined by two independent measurements (absorbance and ellipticity) which give very consistent results (**Figure S15**) while our experimental settings did not allow us to monitor CD melting profiles for all samples: T_m values are based on UV-absorbance profiles only.
- For both pH_T and T_m, one should remember that these transitions may not be at thermodynamic equilibrium and exhibit a hysteresis: the profiles obtained by varying a parameter (temperature or pH) in one direction are not superimposable when doing the reverse experiment (for example: cooling the sample instead of heating) (8). Hysteresis is determined for each melting experiment described in this paper, and T_m average between cooling and heating was taken as a proxy for thermodynamic stability, as previously found for other i-DNA structures (38). For pH_T determination, each sample was allowed to anneal at a given pH for a long period of time, allowing thermodynamic equilibrium.

We determined how well correlated these values are. The analyses of pH_T versus T_m (**Figures 6A-B**), and T_m at pH 7.0 versus 5.0 (**Figures 6C**) revealed good but not perfect positive correlations between these figures (Pearson'r values between 0.79 and 0.95). This indicates that a higher pH_T generally translates into a higher T_m, both at pH 5.0 and 7.0, and that a higher T_m at pH 5.0 means a higher thermal stability at pH 7.0.

i-DNA sequence constraints do not mirror G4 requirements. A quick glance at our experimental results reveal several trends for i-DNA sequence requirements:

- Stability increases with the length of the cytosine tract. This increase is monotonous but not linear: it tends to plateau for longer runs as shown in **Figure 4**. Averages of pH_T with C_3, C_4, C_5 and C_6-tracts are 6.11, 6.39, 6.56 and 6.68, respectively. A similar relationship was found between T_m and C-tract length under both acidic and neutral solutions: increasing C-tract from 3 to 6 translates into T_m of 54.7, 66.0, 72.5 and 77.3 ^oC at pH 5.0 for 3, 4, 5 and 6 cytosines, respectively.
- The nature of the spacer regions does not play a critical role on stability. Correlation coefficients of pH_T and T_m at pH 5.0 and 7.0 versus total spacer length are close to zero. These results indicate that total *spacer* length, assumed to reflect total *loop* length, does not affect the stability of i-DNA at both acidic and neutral pH.
- The "long central spacer is better" rule seems to hold for both G4s (described in our recent report (43)) and i-DNA. For G4s, sequences with long loop in the central position not only exhibit a relative high thermal stability, but are also more prone to form non-parallel conformations. As a consequence of this property shared by i-DNA and G4s, a double-stranded nucleic acid bearing a C-rich and a G-rich strand may be more prone to dismutation into G4 + i-DNA if a relatively long central spacer is present, keeping in mind that a longer loop also means more base pairs to disrupt, and therefore a higher energetic cost.
- In light of the fact that it is nearly impossible to tell apart the different i-DNA conformations attributed to their minimal energetic differences, it is hard to analyze the effects of loops on i-DNA folding patterns (**Figure 1B**). i-DNA stability is basically independent from the total loop length in the 3-12 nucleotides range (**Figures 3G-J**), however, stability of G4 is significantly relied on the total loop length (43-47). When the loop length goes longer, such as, above 15 nucleotides (**Figure S22**), stabilities of both i-DNA and G4 will decrease with elongation of loops.

Overall, these observations confirm that i-DNA requirements do not perfectly match those of G4s. Increasing the number of quartets does lead to an increase in quadruplex stability, but this effect is hard to monitor when dealing with sequences with runs of 5 or more guanines. In addition, loop

effects were more pronounced for G4 forming sequences, with large differences in T_m (and topology) found when changing spacer length (43). These observations argue that the sequence requirements for i-DNA formation do not perfectly mirror those for G4 formation; in other words, the complementary strand of a very stable G4-forming sequence is not necessarily forming a very stable i-DNA. This indicates that the prediction tool we designed for G4 prediction, G4-Hunter (40, 48) is not optimized for i-DNA formation and should be recalibrated for this motif. For this reason, we decided to tackle i-motif prediction by using the three different approaches discussed below.

Modifying the G4-Hunter algorithm to make it applicable for i-DNAs. In G4Hunter, guanines are given a positive score (between 1 for isolated guanine and 4 for each guanine in a run of four or more consecutive guanines) while A or T are considered neutral and C detrimental (40). We modified the G4Hunter algorithm as described in Materials and Methods in Supporting Materials, to make it applicable for the C/T-based i-DNAs. *Optimus* (49) was used to find the optimum positive scoring for each cytosine (counterpart of guanine in the case of G4s) in a given C-tract length while Ts were considered neutral. The model for T_m and pH_T reached a good performance, however, due to all the training sequences having C-tracts with at least three Cs within, the scoring coefficients for the C-tracts of length 3 and shorter were optimized into 0.

Gradient boosting machines (GBM). We built a *de novo* machine learning model to predict the experimental T_m and pH_T stability measures, for the limited sub-universe of C/T-based i-DNAs. The models used only four features - equally sized C-tract and three spacer lengths. Feature importance analysis from the GBM machine learning approach revealed that the most important feature in defining the stability of the i-motifs (in the given sub-universe) both in terms of T_m and pH_T is the length of the C tracts. For T_m prediction, the length of the 3rd spacer (T_3) is slightly more important than that of the other two. For pH_T prediction, this is unclear because the importance ranking of the 3 spacers differs whether total sequence length is included or not as a feature (data not shown).

Defining a simple analytical equation expressing T_m / pH_T as a function of the primary sequence. The mathematical models from Eureqa analytical equation fitting process shows that, with some compromise in prediction quality, we can have a simple, transparent analytical equation that expresses T_m and pH_T as a function of the chosen C-tract and spacer lengths. Both equations capture the interplay between the C-tract length and the spacer lengths 1-3 in modulating the T_m of i-DNAs in the given sub-universe. For pH_T, the chosen equation outlines the observed stabilizing role of the length of the central spacer. Overall, equations would perform better as we expand the investigated space of i-DNA sequences in future, by including sequences with varying C-tract lengths and spacer length relations.

Comparing these predictions. Within a restricted sub-universe of the C/T-based i-DNAs, the three approaches described above perform reasonably well, allowing fairly accurate T_m and pH_T predictions. Unsurprisingly, Eureqa results agree with GBM's in that C-tract length is far more important in predicting the stability in terms of T_m and pH_T of this sub-universe of i-DNAs, as already visible from the plots shown in **Figure 4**. The distilled simple mathematical equation for pH_T also outlined the preferential role of the second spacer in stabilizing the i-DNA, in agreement with the experimental observations. In general, the models will become more extrapolatable as we increase the experimentally studied sequence space by allowing a wider spread of the relations/ratios of the C-tract and spacer lengths. The sequences tested here only cover a limited sequence space, and more data should be collected to apply these prediction tools to mixed motifs containing spacers of any sequence or C-runs of unequal length. A "theory of every i" has yet to emerge!

Stability *in vitro* at near-neutral pH is still marginal, even for the most stable sequences. Extrapolating maximum pH_T and T_m values from the data provided in **Figure 4** would indicate plateaus of 7.28 and 27 ^{o}C at neutral pH, implying that i-DNA stabilized only by $C{\cdot}C^{+}$ base pairs such as the sequences studied in this work, is barely stable at physiological temperature. These results impact the predictions made: for example, with the formula obtained by Eureqa, to obtain a

pH of mid-transition above 7.0, one would need a sequence containing 11 or more consecutive cytosines according to *equation **2***.

***In cellulo* i-DNA formation was confirmed for i-DNAs stable *in vitro:* implications for biology**. Although the in-cell NMR data do not allow quantitative determination of i-DNA stability in the intracellular space, they demonstrate that sequences forming i-DNA *in vitro* at neutral pH are also to form i-DNA in living cells. We observed that the stabilities for selected i-DNA constructs *in cellulo* as estimated from in-cell NMR data followed the same relative order as those observed in vitro at neutral pH. This observation suggests that the rules derived on the basis of *in vitro* data are reasonable approximation for i-DNA behavior in cells. i-DNA relative instability may be an asset for regulation of pH homeostasis, as modest and transient changes in intracellular pH should lead to important variations in i-DNA stability. For example, the physiological normal intracellular pH (intracellular means *cytoplasmic* for most studies) has been reported to vary between 7.0 and 7.4, depending on tissues and phase of the cell cycle (50). Changes in metabolism and respiratory chain activity modify the acid balance in cells not only affecting mitochondria but also altering intracellular pH. Invasive tumor cells tend to acidify their extracellular environment while keeping their pH_i more alkaline (51, 52). All these observations point out the role of pH regulation in normal and pathologic processes, and i-DNA formation may be affected by these changes. i-DNA formation in the promoter of pH-sensitive genes may therefore represent a pH-responsive transcriptional regulator. It is therefore important to correlate *in vitro* and *in cellulo* observations. In-cell NMR measurements suggest that i-DNA stability may be slightly higher than what is found *in vitro*. Differences in crowding (33-35), water activity, dielectric constant, local concentration of free ions, pH, may affect the stability of the structure of interest, as well as the presence of cellular competitors or natural ligands. This is a problem of general importance for biochemists, to make sure that the conclusions reached in the test tube reflect what is happening in the cell. We hope that further *in cellulo* - *in vitro* comparisons will provide decisive answers.

Conclusion

By performing an exhaustive experimental analysis of i-DNA formation on a dataset of unprecedent magnitude, we were able to provide a global picture of i-DNA formation *in vitro*, and propose tools to predict its stability as a function of primary sequence. The most stable candidates were confirmed to adopt an i-DNA conformation in cells. This work will be invaluable not only for those interested in the biological functions of this structure, but also when considering nano- or biotech applications with these pH-sensitive devices.

Materials and Methods

Experimental details are provided in the following Supplementary Materials.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (21977045 to J.Z.), Fundamental Research Funds for the Central Universities (02051430210 to J.Z.), and Excellent Research Program of Nanjing University (ZYJH004). J.L.M. acknowledges funding from Nanjing University (020514912216) and ERDF (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000477). M.C. acknowledges the China Postdoctoral Science Foundation (2019M661793). E.M., P.V., and Lukáš T. were supported by grant from the Czech Science Foundation (19-26041X). Liezel T. is grateful to the Jardine Foundation for supporting her DPhil studies. A.B.S thanks Medical Research Council (UK) for the core funding of his laboratory. The Ministry of Education, Youth and Sports of the Czech Republic is gratefully acknowledged for the support of the access to a research infrastructure (CIISB research infrastructure project LM2015043; CEITEC 2020, LQ1601).

References

1. K. Gehring, J.-L. Leroy, M. Gueron, A tetrameric DNA structure with protonated cytosine·cytosine base pairs. *Nature* **363**, 561-565 (1993).
2. J.-L. Leroy, M. Guéron, J.-L. Mergny, C. Hélène, Intramolecular folding of a fragment of the cytosine-rich strand of telomeric DNA into an i-motif. *Nucleic Acids Res.* **22**, 1600-1606 (1994).
3. S. Nonin-Lecomte, J.-L. Leroy, Structure of a C-rich strand fragment of the human centromeric satellite III: a pH-dependent intercalation topology. *J. Mol. Biol.* **309**, 491-506 (2001).
4. A. T. Phan, M. Gueron, J.-L. Leroy, The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *J. Mol. Biol.* **299**, 123-144 (2000).
5. X. Han, J.-L. Leroy, M. Gueron, An intramolecular i-motif: the solution structure and base-pair opening kinetics of d(5mCCT3CCT3ACCT3CC). *J. Mol. Biol.* **278**, 949-965 (1998).
6. K. Snoussi, S. Nonin-Lecomte, J.-L. Leroy, The RNA i-motif. *J. Mol. Biol.* **309**, 139-153 (2001).
7. E. P. Wright, J. L. Huppert, Z. A. E. Waller, Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH. *Nucleic Acids Res.* **45**, 2951-2959 (2017).
8. R. A. Rogers, A. M. Fleming, C. J. Burrows, Unusual isothermal hysteresis in DNA i-motif pH transitions: a study of the RAD17 promoter sequence. *Biophys. J.* **114**, 1804-1815 (2018).
9. P. Skolakova *et al.*, Systematic investigation of sequence requirements for DNA i-motif formation. *Nucleic Acids Res.* **47**, 2177-2189 (2019).
10. J.-L. Mergny, L. Lacroix, X. Han, J.-L. Leroy, C. Helene, Intramolecular folding of pyrimidine oligodeoxynucleotides into an i-DNA motif. *J. Am. Chem. Soc.* **117**, 8887-8898 (1995).
11. J.-L. Mergny, D. Sen, DNA quadruple helices in nanotechnology. *Chem. Rev.* **119**, 6290-6325 (2019).
12. J. J. Alba, A. Sadurni, R. Gargallo, Nucleic acid i-motif structures in analytical chemistry. *Crit. Rev. Anal. Chem.* **46**, 443-454 (2016).
13. K. Leung, K. Chakraborty, A. Saminathan, Y. Krishnan, A DNA nanomachine chemically resolves lysosomes in live cells. *Nat. Nanotechnol.* **4**,176-183 (2018).
14. M. Debnath, K. Fatma, J. Dash, Chemical gegulation of DNA i-motifs for nanobiotechnology and therapeutics. *Angew. Chem., Int. Ed. Engl.* **58**, 2942-2957 (2019).
15. M. Zeraati *et al.*, I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.* **10**, 631-637 (2018).
16. S. Dzatko *et al.*, Evaluation of the stability of DNA i-motifs in the nuclei of living mammalian cells. *Angew. Chem., Int. Ed. Engl.* **57**, 2165-2169 (2018).
17. X. Li, Y. Peng, J. Ren, X. Qu, Carboxyl-modified single-walled carbon nanotubes selectively induce human telomeric i-motif formation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 19658-19663 (2006).
18. K. Niu *et al.*, BmILF and i-motif structure are involved in transcriptional regulation of BmPOUM2 in Bombyx mori. *Nucleic Acids Res.* **46**, 1710-1723 (2018).
19. H. J. Kang, S. Kendrick, S. M. Hecht, L. H. Hurley, The transcriptional complex between the BCL2 i-motif and hnRNP LL is a molecular switch for control of gene expression that can be modulated by small molecules. *J. Am. Chem. Soc*. **136**, 4172-4185 (2014).
20. S. Takahashi, J. A. Brazier, N. Sugimoto, Topological impact of noncanonical DNA structures on Klenow fragment of DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9605-9610 (2017).
21. B. Mir *et al.*, Prevalent Sequences in the Human genome can form mini i-motif structures at physiological pH. *J. Am. Chem. Soc.* **139**, 13985-13988 (2017).
22. I. V. Nesterova, E. E. Nesterov, Rational design of highly responsive pH sensors based on DNA i-motif. *J. Am. Chem. Soc.* **136**, 8843-8846 (2014).
23. A. M. Fleming, K. M. Stewart, G. M. Eyring, T. E. Ball, C. J. Burrows, Unraveling the 4n - 1 rule for DNA i-motif stability: base pairs *vs.* loop lengths. *Org. Biomol. Chem.* **16**, 4537-4546 (2018).
24. A. M. Fleming *et al.*, 4n-1 is a "sweet spot" in DNA i-motif folding of 2'-deoxycytidine homopolymers. *J. Am. Chem. Soc.* **139**, 4682-4689 (2017).
25. A. Sengar, B. Heddi, A. T. Phan, Formation of G-quadruplexes in poly-G sequences: structure of a propeller-type parallel-stranded G-quadruplex formed by a G_{15} stretch. *Biochemistry* **53**,

7718-7723 (2014).
26. S. Benabou *et al.*, Understanding the effect of the nature of the nucleobase in the loops on the stability of the i-motif structure. *Phys. Chem. Chem. Phys.* **18**, 7997-8004 (2016).
27. S. M. Reilly, R. K. Morgan, T. A. Brooks, R. M. Wadkins, Effect of interior loop length on the thermal stability and pK_a of i-motif DNA. *Biochemistry* **54**, 1364-1370 (2015).
28. I. V. Nesterova, J. R. Briscoe, E. E. Nesterov, Rational control of folding cooperativity in DNA quadruplexes. *J. Am. Chem. Soc.* **137**, 11234-11237 (2015).
29. S. P. Gurung, C. Schwarz, J. P. Hall, C. J. Cardin, J. A. Brazier, The importance of loop length on the stability of i-motif structures. *Chem. Commun. (Camb)* **51**, 5630-5632 (2015).
30. T. Fujii, N. Sugimoto, Loop nucleotides impact the stability of intrastrand i-motif structures at neutral pH. *Phys. Chem. Chem. Phys.* **17**, 16719-16722 (2015).
31. M. McKim, A. Buxton, C. Johnson, A. Metz, R. D. Sheardy, Loop sequence context influences the formation and stability of the i-motif for DNA oligomers of sequence (CCCXXX)4, where X = A and/or T, under slightly acidic conditions. *J. Phys. Chem. B* **120**, 7652-7661 (2016).
32. S. Kendrick, Y. Akiyama, S. M. Hecht, L. H. Hurley, The i-motif in the bcl-2 P1 promoter forms an unexpectedly stable structure with a unique 8:5:7 loop folding pattern. *J. Am. Chem. Soc.* **131**, 17667-17676 (2009).
33. J. Zhou, G. Jia, Z. Feng, C. Li, Properties of i-motif under molecular crowding conditions. *Chin. J. Chem. U.* **31**, 309–311 (2010).
34. A. Rajendran, S. Nakano, N. Sugimoto, Molecular crowding of the cosolutes induces an intramolecular i-motif structure of triplet repeat DNA oligomers at neutral pH. *Chem. Commun. (Camb)* **46**, 1299-1301 (2010).
35. Y. P. Bhavsar-Jog, E. Van Dornshuld, T. A. Brooks, G. S. Tschumper, R. M. Wadkins, Epigenetic modification, dehydration, and molecular crowding effects on the thermodynamics of i-motif structure formation from C-rich DNA. *Biochemistry* **53**, 1586-1594 (2014).
36. T. Nguyen, C. Fraire, R. D. Sheardy, Linking pH, temperature, and K^+ concentration for DNA i-motif formation. *J. Phys. Chem. B* **121**, 7872-7877 (2017).
37. J.-L. Mergny, J. Li, L. Lacroix, S. Amrane, J. B. Chaires, Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.* **33**, e138-143 (2005).
38. J.-L. Mergny, L. Lacroix, Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.* **26**, 4797-4803 (1998).
39. J.-L. Mergny, L. Lacroix, Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).
40. A. Bedrat, L. Lacroix, J.-L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* **44**, 1746-1759 (2016).
41. A. B. Sahakyan *et al.*, Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535-14545 (2017).
42. M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data. *Science* **324**, 81-85 (2009).
43. M. Cheng *et al.*, Loop permutation affects the topology and stability of G-quadruplexes. *Nucleic Acids Res.* **46**, 9264-9275 (2018).
44. A. Piazza *et al.*, Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J.* **34**, 1718-1734 (2015).
45. A. Guedin, J. Gros, P. Alberti, J.-L. Mergny, How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.* **38**, 7858-7868 (2010).
46. N. Kumar, B. Sahoo, K. A. Varun, S. Maiti, S. Maiti, Effect of loop length variation on quadruplex-Watson Crick duplex competition. *Nucleic Acids Res.* **36**, 4433-4442 (2008).
47. P. Hazel, J. Huppert, S. Balasubramanian, S. Neidle, Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.* **126**, 16405-16415 (2004).
48. V. Brazda *et al.*, G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* **35**, 3493-3495 (2019).
49. N. A. G. Johnson, L. Tamon, A. B. Sahakyan (Optimus: a general purpose adaptive optimisation engine, GitHub link to the code: http://github.com/SahakyanLab/Optimus, accessed in November 2019.

50. J. R. Casey, S. Grinstein, J. Orlowski, Sensors and regulators of intracellular pH. *Nat. Rev. Mol. Cell Biol.* **11**, 50-61 (2010).
51. E. Persi *et al.*, Systems analysis of intracellular pH vulnerabilities for cancer therapy. *Nat. Commun.* **9**, 2997-3007 (2018).
52. B. A. Webb, M. Chimenti, M. P. Jacobson, D. L. Barber, Dysregulated pH: a perfect storm for cancer progression. *Nat. Rev. Cancer* **11**, 671-677 (2011).

Figures and Tables

Table 1. Counts of the groups obeying the *"long central spacer is better"* rule.[a]

Counts	i-DNAs in the same group				Total (percentage)
	C_3 tract	C_4 tract	C_5 tract	C_6 tract	
pH_T^{CD}	11/15	13/15	12/15	12/15	48/60 (80%)
pH_T^{UV}	10/15	13/15	12/15	11/15	46/60 (77%)
$T_{1/2}^{pH\ 5.0}$	15/15	15/15	15/15	15/15	60/60 (100%)
$T_{1/2}^{pH\ 7.0}$	--	--	12/15	12/15	24/30 (80%)

[a] Counts based on results presented in **Tables S1**-**S2**, **Figures S9** and **S14**. The thermal stability of sequences with C_3- and C_4-tracts at pH 7.0 was not evaluated.

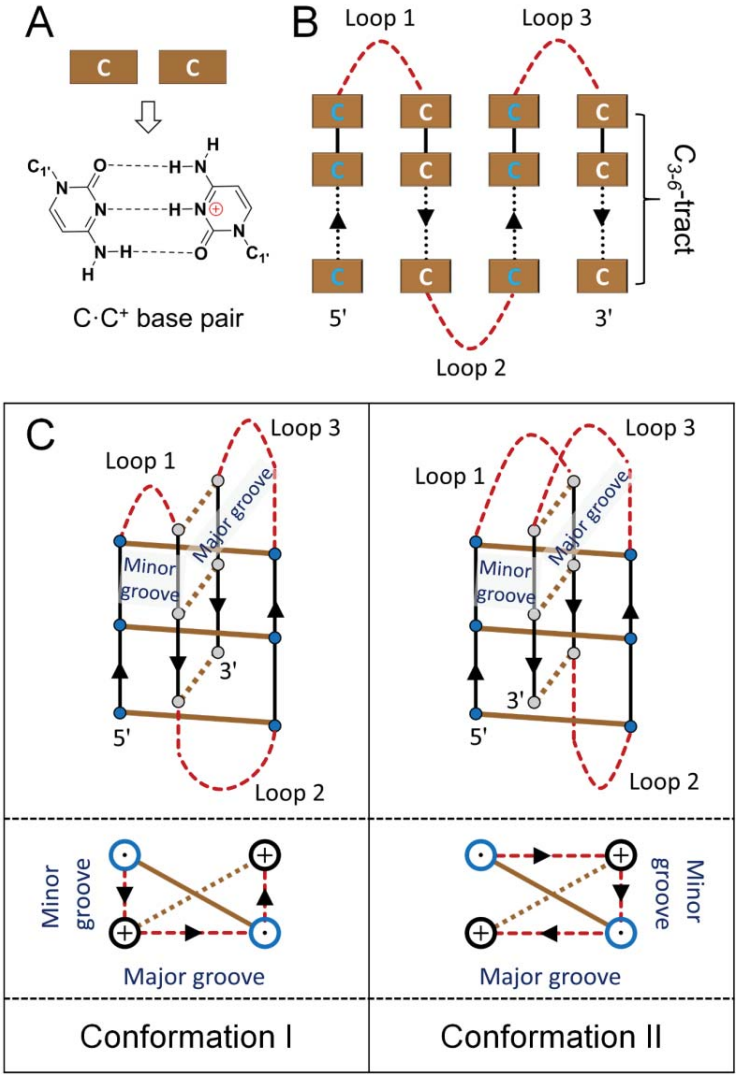


Figure 1. (**A**) Hemi-protonated C·C⁺ base pair; one of the two cytosine being protonated at the N3 position (**B**) example of a sequence prone to intramolecular i-DNA formation, with four C-tracts connected by three loops (red dotted lines). (**C**) Possible loop arrangements in an intramolecular i-DNA structure (only the so-called 5'E conformation is shown here, in which the external, solvent-exposed C·C⁺ base pairs are at the 5' ends of both parallel duplexes); Simplified diagram of two linking directions between strands: Central loop can across either major (*left*, conformation I) or minor (*right*, conformation II) groove (2). The width ranges of major and minor grooves of i-DNA are 10.8-12.0 and 4.8-6.5 Å, respectively (4-6). Note that the loops include the nucleotides founds in the spacer (non C) regions, and possibly adjacent cytosines that contribute to the loop rather than to the i-DNA stem: the terms *spacer* and *loop* are therefore not interchangeable.

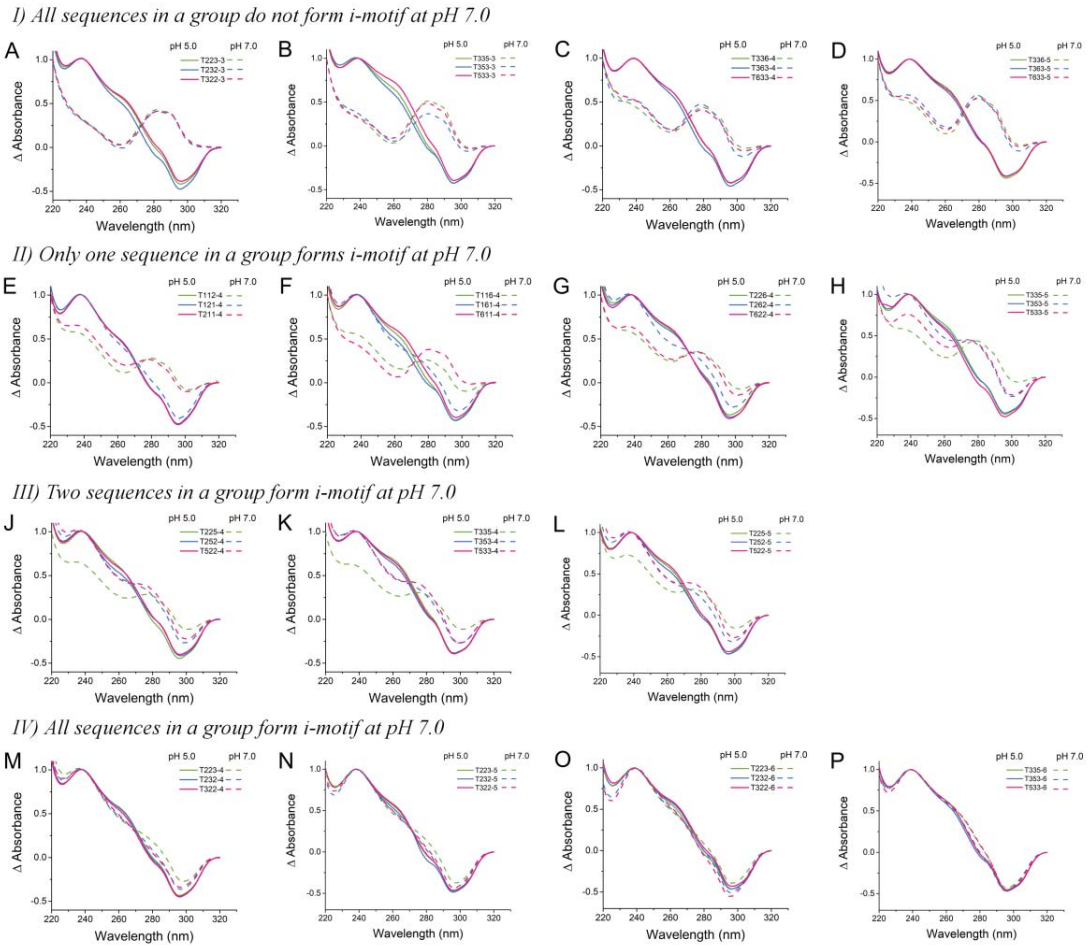


Figure 2. Thermal difference spectra (TDS) of selected groups. All sequences fold into an i-DNA structure at pH 5.0 (solid line). TDS spectra of different groups are divided into four types according to i-DNA formation at pH 7.0 (dash line). (**A**-**D**) Type I: No sequences in a group fold into i-DNA completely at pH 7.0; (**E**-**G**) Type II: Only one of three sequences within the same group folds into an i-DNA at neutral pH; (**J**-**L**) Type III: Two of the three sequences in the same group fold into an i-DNA at neutral pH; (**M**-**P**) Type IV: All sequences in a group folds into i-DNA.

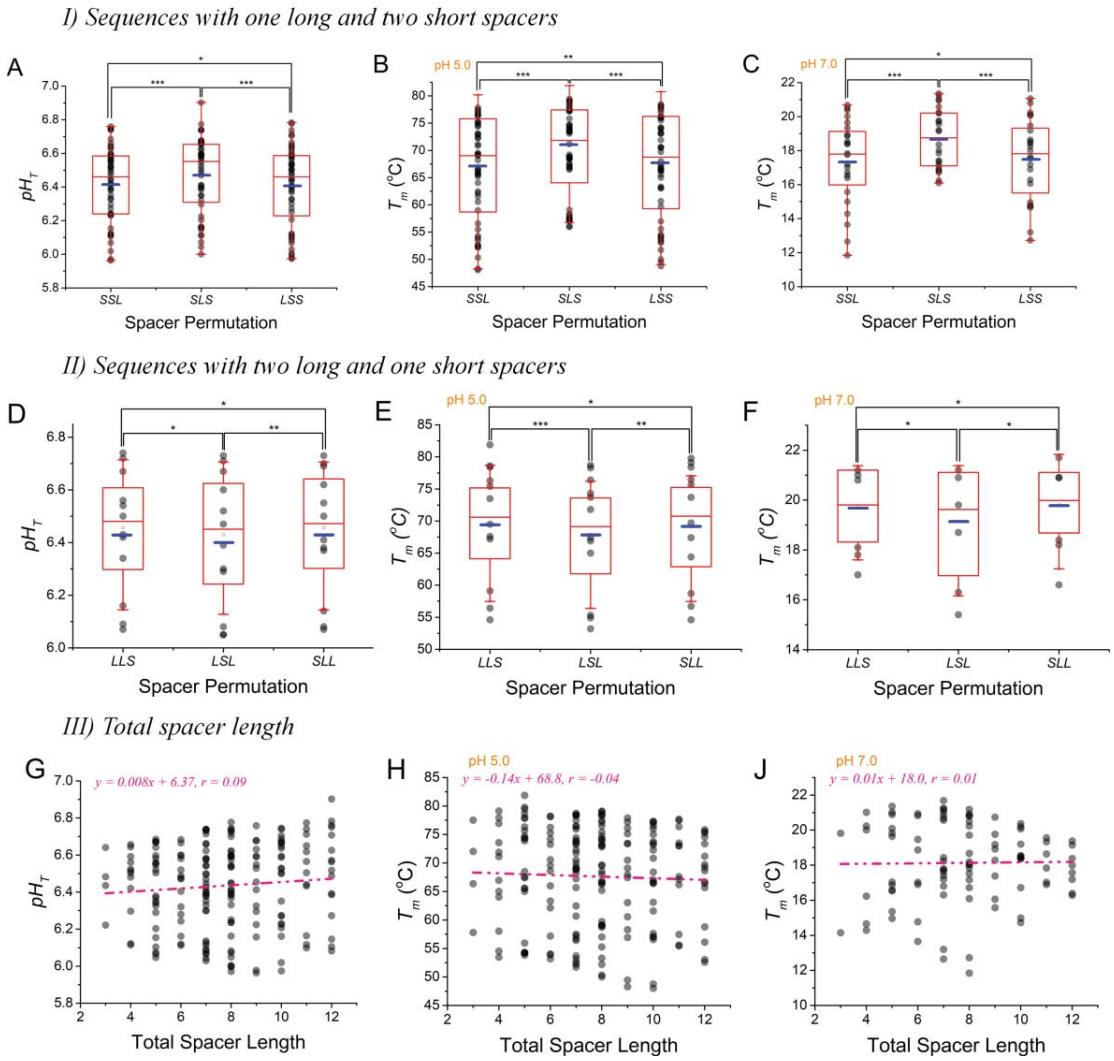


Figure 3. Effects of total spacer length and individual spacer permutation on pH_T and T_m. (**A** & **D**) pH_T versus spacer permutation; (**B** & **E**) Melting temperatures at pH 5.0 versus spacer permutation; (**C** & **F**) Melting temperatures at pH 7.0 versus spacer permutation. (**A**-**B**) 144 sequences with two short and one long spacers, and C_3, C_4, C_5 and C_6-tracts from 48 groups in **Table S2**. (**C**) 72 sequences with two short and one long spacers, and C_5 and C_6-tracts from 24 groups in right column of **Table S2**. (**D**-**E**) 36 sequences with two short and one long spacers from 12 groups in **Table S2**. (**F**) 18 sequences from with two short and one long spacers, and C_5 and C_6-tracts from 6 groups in right column of **Table S2**. Blue line and red line in the red box are the average and median values for each spacer combination, respectively. Two sequences from the same group are treated as a paired sample, hypotheses of pair-sample *t*-test are performed between every two spacer combinations: *, $p > 0.05$ (not obvious); **, $0.05 > p > 0.0005$ (significantly obvious); ***, $p < 0.0005$ (greatly obvious). (**G**) pH_T as function of total spacer length. (**H**) Melting temperatures at pH 5.0 as a function of total spacer length. All 196 sequences in **Table S2** were used in (**G** and **H**). (**J**) Melting temperatures at pH 7.0 as a function of total spacer length. The red dashed line corresponds to a linear fit. All 98 sequences with C_5 and C_6-tracts in the right column of **Table S2** were used in **J**.

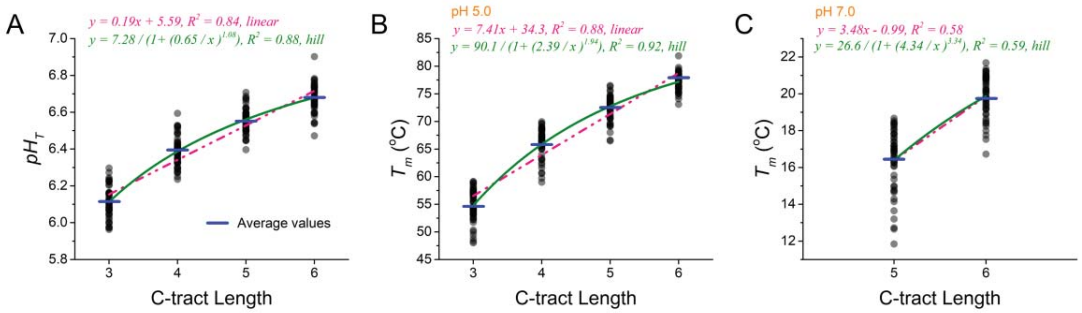


Figure 4. Effect of C-tract length on i-DNA stability. (**A**) pH_T as function of C-tract length. All 196 sequences in **Table S2** were used. Averages of pH_T with C_3, C_4, C_5 and C_6-tracts are 6.11, 6.39, 6.56 and 6.68, respectively. (**B**) Melting temperatures at pH 5.0 as a function of C-tract length. All 196 sequences in **Table S2** were used. Averages of T_m with C_3, C_4, C_5 and C_6-tracts are 54.7, 66.0, 72.5 and 77.3 °C, respectively. (**C**) Melting temperatures at pH 7.0 as a function of C-tract length. All 98 sequences with C_5 and C_6-tracts in the right column of **Table S2** were used in **C**. Averages of T_m with C_5 and C_6-tracts are 16.4 and 20.0 °C, respectively. Blue short lines provide the corresponding averages values of 49 sequences with eight C_3, C_4, C_5 or C_6-tracts. Data were fitted by linear (red) and non-linear (hill, green) functions, which depicted by red dashed and green solid lines, respectively.

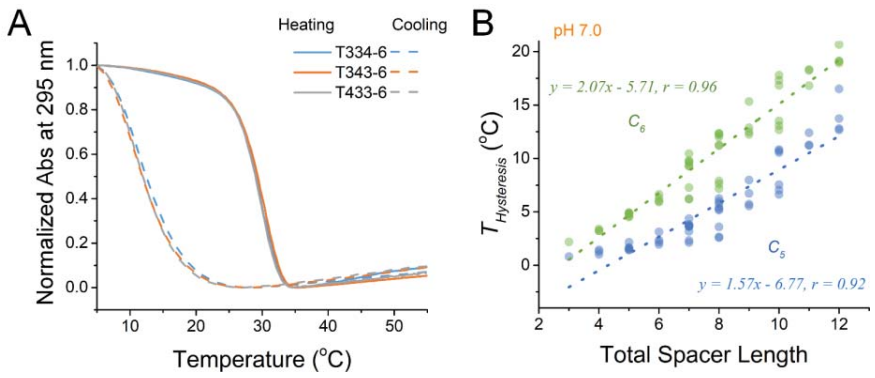


Figure 5. Hysteresis in the melting/annealing processes at neutral pH. UV-melting (solid line) and annealing (dashed line) curves at pH 7.0 of (**A**) *T334-6* group as an example. Curves of other sequences are provided in **Figure S17**. (**B**) Hysteresis as a function to total spacer length ($T_{hysteresis} = T_{heating} - T_{cooling}$). All 98 sequences with C_5 and C_6-tracts in the right column of **Table S2** were used. Linear fits between hysteresis and total spacer length of sequences with C_5 and C_6-tracts are presented in blue and green color, respectively. Hysteresis depends on temperature gradient (not shown).

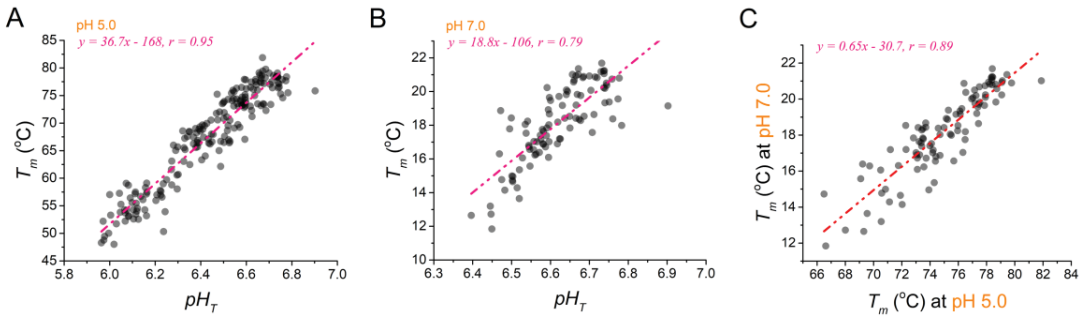


Figure 6. Relationship between T_m and pH_T. (**A**) pH_T as function of T_m at pH 5.0. All 196 sequences in **Table S2** were used. (**B**) pH_T as function of T_m at pH 7.0. Linear fit presented as a red dashed line. All 98 sequences with C_5 and C_6-tracts in the right column of **Table S2** were used in **B** and **C**. (**C**) T_m at pH 7.0 as a function of T_m at pH 5.0. Linear fit presented as a red dashed line.

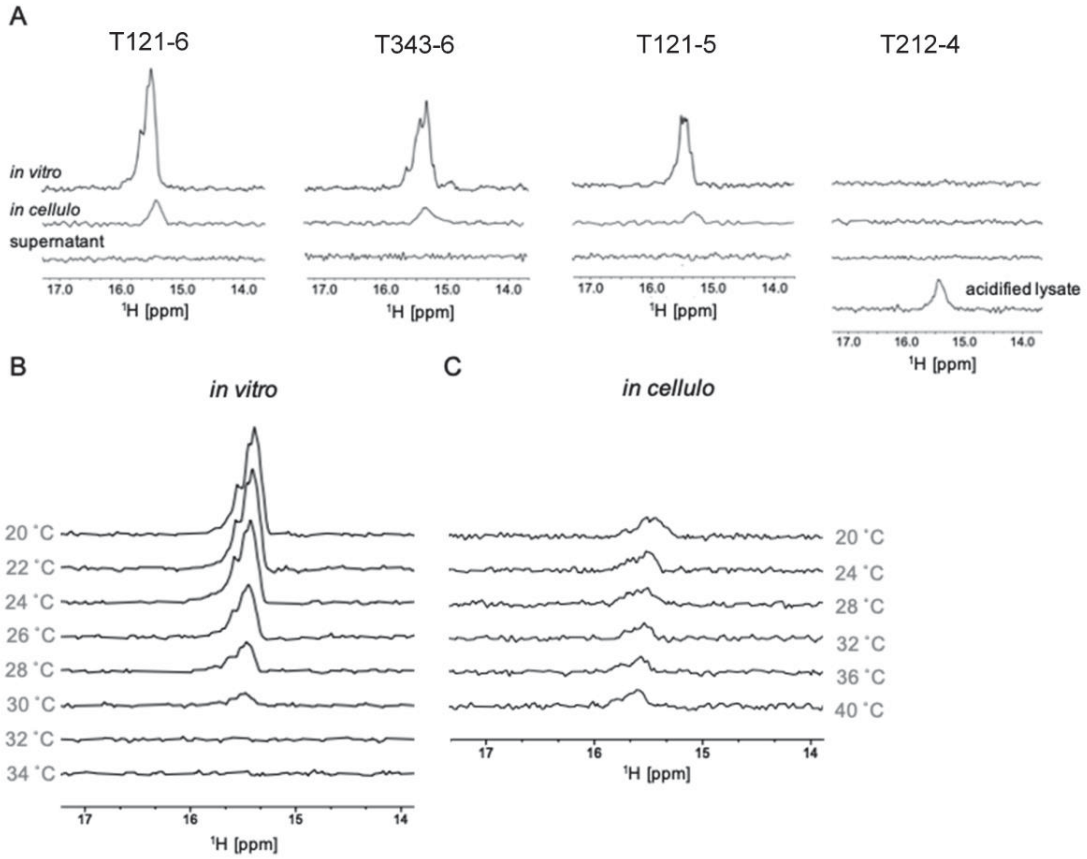


Figure 7. *in vitro* and in-cell NMR. (**A**) Imino regions of1D ^{1}H NMR spectra for T121-6, T343-6, T121-5, and T212-4 acquired at 20 ºC *in vitro* in potassium buffer (20 mM KPi, 120 mM KCl, pH=7.0), *in cellulo* (in living HeLa cells), and in supernatant (medium) collected from in-cell NMR sample post in-cell NMR spectra acquisition, respectively. Absence of signals in the "supernatant" spectra evidences that the signals observed in in-cell NMR spectra originates from DNA localized in cells. Presence of i-motif specific signals in the NMR spectrum of T212-4 acquired in acidified (pH < 6.5) lysate prepared from respective in-cell NMR sample (lysation control) confirmed that T212-4 was present, yet unfolded, in the intracellular space of living cells. For flow cytometry plots and confocal images of transfected cells used to acquire the in-cell NMR spectra reporting on cells viability, level of DNA transfection, and intracellular localization of transfected DNA - see **Figure S29**. **(B**, **C)** Imino region of 1D ^{1}H NMR spectra acquired at various temperatures under (**B**) *in vitro* conditions (20 mM KPi, 120 mM KCl, pH=7.0) and (**C**) in living cells for T121-6. For flow cytometry plots and confocal images of transfected cells used to acquire the in-cell NMR spectra reporting on cells viability (after in-cell NMR spectra acquisition), level of T121-6 transfection, and intracellular localization of transfected T121-6 - see **Figure S30**.

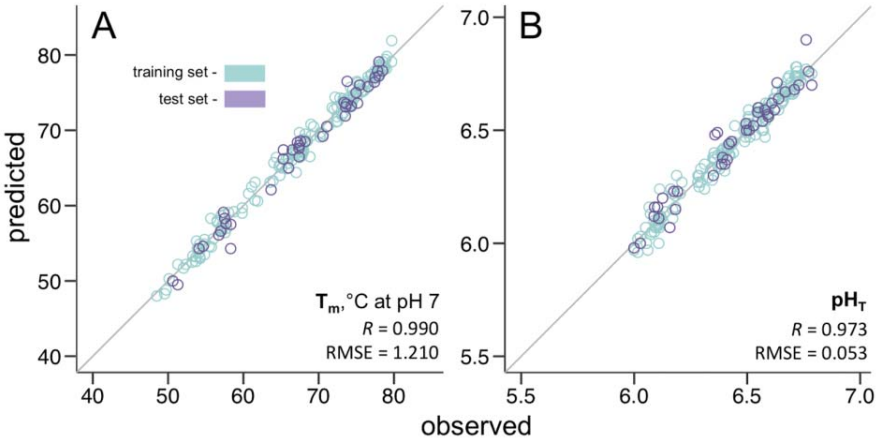


Figure 8. Correlation plots between the experimental stability measures (T_m at pH 7.0 and pH_T) and the i-DNA stability predicted via machine learning models obtained using gradient boosting machines. Plots are brought for both T_m (**A**) and pH_T (**B**) dependencies. The Pearson's correlation coefficients (R) and root mean squared errors (RMSE) are brought on the individual plots.

PNAS
www.pnas.org

Supplementary Materials for

i-DNA stability: confronting in vitro experiments with models and in-cell NMR data

Mingpan Cheng[†,‡], Dehui Qiu[†], Liezel Tamon[§], Eva Maturová[¶], Pavlína Víšková[¶], Samir Amrane[‡], Aurore Guédin[‡], Jielin Chen[†,‡], Laurent Lacroix[#], Huangxian Ju[†], Lukáš Trantírek[¶], Aleksandr B. Sahakyan[§], Jun Zhou[†,*] & Jean-Louis Mergny[†,‡,||]

[†] State Key Laboratory of Analytical Chemistry for Life Science, School of Chemistry & Chemical Engineering, Nanjing University, Nanjing 210023, China.

[‡] ARNA Laboratory, Université de Bordeaux, Inserm U 1212, CNRS UMR5320, IECB, Pessac 33607, France.

[§] MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, UK.

[¶] Central European Institute of Technology, Masaryk University, Brno 62500, Czech Republic.

[#] IBENS, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, Paris 75005, France.

[||] Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, Brno 61265, Czech Republic.

* Corresponding author Jun Zhou; Email: *jun.zhou@nju.edu.cn*

This PDF file includes:

Materials and Methods;
Table S1 to S4;
Figures S1 to S31;
Supplementary references.

Contents

Materials and Methods

Nomenclature of sequences, preparation of oligonucleotides and reagents

236 i-DNA forming sequences were investigated, divided into four types based on C-tract length: each oligonucleotide contained four runs of 3, 4, 5, or 6 cytosines; see sequences information provided in **Tables S1** and **S2**. Each sequence contains four identical C-tracts, which are separated by three spacers. Note that we will generally prefer the word "*spacer*" over "*loop*" as the identity of the bases participating in the loop does not always matches the spacer sequence: some cytosines thought to be involved in the i-DNA stem may rather participate in the loops, especially when spacers are short. For most sequences (212 out of 236), these spacer regions were consisting of thymidines only, ranging from 1 to 6 nucleotides. The nomenclature is shown in **Table S2**: a "T" prefix means that the three spacers are composed of thymine bases only; the three consecutive numbers refer to lengths of the three spacers in the 5' to 3' direction; while the "-3", "-4", "-5" or "-6"' suffix refers to sequences with four C_3, C_4, C_5, and C_6 tracts, respectively.

In order to compare the effects of spacer arrangement on i-DNA stability, the notion of sequence *group* was introduced (1). The sequences in the same group are only differing in the way spacers are permuted. A group is named after the first sequence in the group. For example, the T112-3 group is composed of three sequences T112-3, T121-3, and T211-3. All three sequences have the same length, the same overall base content with short spacers composed of one or two thymines separating four runs of three cytosines.

All oligonucleotides purified by ultra-PAGE were ordered from Sangon Biotech (Shanghai, China) and chemicals were purchased from Sigma-Aldrich (Shanghai, China), dissolved in distilled and deionized water (18.2 MΩ·cm). Concentration of sample stock was determined by ultraviolet (UV) absorbance at 260 nm using the molar extinction coefficients provided by manufacturer. Samples were then stored at 4 °C and used without further purification. Unless otherwise stated, Britton-Robinson buffers (B-R) contain four components: H_3BO_3/H_3PO_4/CH_3COOH/NaOH; they were chosen in this work considering their wide buffering range and small temperature coefficient, which is important for i-DNA studies (2). pH was adjusted after the addition of 140 mM KCl at room temperature. Prior to all following experiments, all oligonucleotides samples were prepared in 20 mM B-R buffer containing 140 mM KCl at the chosen pH, denatured at 95 °C for 3 min, slowly cooled down during 2 hours to room temperature, and then incubated at 4 °C overnight to ensure the completely equilibration of folding and unfolding processes.

Absorbance and circular dichroism (CD) measurements

Thermal difference spectra (TDS) (3)**.** 5.0 μM oligonucleotide samples were prepared in 20 mM B-R buffer containing 140 mM KCl (pH 5.0 or 7.0). Ultraviolet (UV) -Visible absorbance spectra (220-320 nm, Cary100, Agilent) were recorded at low temperature (5 °C for both pH 5.0 and 7.0) first and then at high temperature (95 and 65 °C for pH 5.0 and 7.0, respectively). Prior to the measurements, samples were incubated at the corresponding temperature for at least 5.0 min. During each scan, high speed dry air was used to flush the cuvette holder in order to prevent condensation. TDS spectra were calculated by subtraction of the spectrum recorded at low temperature from the one at high temperature, normalized by difference absorbance at 239 nm, to compare the curve shapes.

pH-dependent transition experiments. Experiments were performed by monitoring the UV-Visible absorbance (Carry 100, Agilent) and CD spectra (Applied Photophysics) in the 220-320 nm wavelength range at 25 °C. Oligonucleotides were dissolved at a final concentration of 5.0 μM in 20 mM B-R buffer at a pH varying from 5.0 to 8.0 with 0.25 pH unit intervals (*i.e.*, 13 different pH values were tested) in the presence of 140 mM KCl. All samples in the corresponding pH solutions were denatured at 95 °C for 3 min, slowly cooled down to room temperature, then stored at 4 °C

for at least overnight. All samples were then incubated at 25 °C for at least two hours prior to spectral measurements. Each sample scan was subtracted by the corresponding buffer scan before data processing. The changes in signal intensities at 295 and 288 nm for UV absorbance and CD ellipticity, respectively, were used to calculate the pH transition midpoint (pH_T) of the structure switching from stable i-DNA to random coil. pH_T were obtained by fitting the signals from UV or CD *vs.* pH values, by using a Boltzman sigmoidal function.

UV-melting/annealing experiments (4)***.*** Samples were prepared at 5.0 μM oligonucleotide concentration in 20 mM B-R buffers containing 140 mM KCl. UV-absorbance at 295 nm was recorded at pH 5.0 (for all sequences, 0.5 °C/min rate in 5 to 95 °C temperature range) or 7.0 (for sequences with C_5- and C_6-tracts, using a slower temperature gradient of 0.2 °C/min to limit hysteresis). Absorbance was normalized between 1 and 0 to compare the profiles. Half transition temperatures (T_m or $T_{1/2}$) were calculated by fitting the plot of UV absorbance *vs.* temperature with a Boltzman sigmoidal function. $T_{1/2}$ is used rather than T_m when hysteresis is present.

Differential scanning calorimetry (DSC)

DSC measurements were carried out using a Nano DSC equipment. Oligonucleotide was prepared as 100 μM strand concentration in 20 mM B-R buffer (pH 5.0 or 7.0). All heating and cooling scans were recorded at 1.0 °C/min rate, and in the 0-100 °C and 0-65 °C temperature ranges for pH 5.0 and 7.0 supplemented with 140 mM KCl, respectively. The DNA sample versus buffer scan was subtracted by the previously performed buffer versus buffer for all the scans. T_m or $T_{1/2}$ was calculated by using *TwoStateScaled* model to fit the heat capacity *vs.* temperature curve.

Gel electrophoresis

Non-denaturing polyacrylamide gel electrophoresis (PAGE) experiments were performed to check the molecularity of sequences with C_5-tract. Oligonucleotides were dissolved in B-R buffer (pH 5.0 or 7.0) at 100 μM strand concentration, denatured at 95 °C for 3 min, then slowly cooled to room temperature. Samples were stored at 4 °C overnight before incubation. Oligonucleotides were incubated for two hours at room temperature, then 30% (w/v) sucrose was added before loading and final oligo concentration was 25.0 μM. Gels (7 × 10 × 0.1 cm) were prepared with 15% acrylamide (acrylamide:bisacrylamide, 19:1) in 50 mM B-R buffer (pH 5.0 or 7.0) which was also used as the running buffer. Gels were run at room temperature (*ca.* 25 °C) with 80 V for 90 min and stained by Stains-all (Sigma, 95%). Oligothymidylate DNA single-strands (dT_n, n = 60, 30, 21, 15, 10) were used as internal migration standards, and bromophenol blue was added to act as an indicator of migration.

***In vitro* 1D ^{1}H NMR**

100 μM oligonucleotide was prepared in 20 mM pH 7.0 potassium phosphate (KPi) buffer supplemented with 10% (v/v) D_2O and KCl (total potassium in solution is 140 mM), then denatured at 95 °C for 3 min, slowly cooled down to room temperature, and stored at 4 °C for overnight. Prior to experiments, samples were incubated at room temperature for at least two hours. ^{1}H NMR experiments were carried out on a 400 (**Figure S2**) or 600 MHz (**Figure 7**) Bruker spectrometer at 20 °C (unless stated otherwise). The jump-and-return pulse program was used in recording proton spectra and suppressing the water signal.

In-cell 1D ^{1}H NMR

Preparation of DNA oligonucleotides. Oligonucleotides used for in-cell NMR experiments were ordered from Sigma-Aldrich (USA). Non-labelled oligonucleotides were purchased as pre-purified by desalting at a 10-μmol scale, while the fluorescently-labelled oligonucleotides were purified by HPLC and ordered at a 1-nmol scale. For oligonucleotide labelling, FAM dye was used at the 5'

end. The fluorescently (FAM) labelled oligonucleotides were dissolved in H_2O to yield 100.0 µM stock solutions. Desalted oligonucleotides were further subjected to *n*-butanol precipitation, to remove contaminants from the solid-state synthesis. At first, they were dissolved in 1 mL Milli-Q H_2O. Then, 30 mL of *n*-butanol was added, the samples were mixed thoroughly for ~ 10 min and transferred into centrifuge tubes (Beckman Coulter, USA). The centrifugation parameters were set to 30,000 x g, 4 °C and 1 h. After centrifugation, the supernatant was carefully drained and the samples were left open to dry at room temperature. The dried pellets were resuspended in 1 mL Milli-Q H_2O again and annealed by heating the solution for 5 min at 95 °C and allowing the samples to cool down to room temperature. Finally, the concentration was determined by UV absorbance, using NanoDrop 2000c (Thermo Fisher Scientific, USA).

Preparation of in-cell NMR samples. The in-cell NMR samples were prepared according to protocol by Krafcikova *et al.* (5). For purpose of in-cell NMR experiments, HeLa cell line (Sigma-Aldrich, USA) was used and cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco, USA) supplemented with 10 % fetal bovine serum (HyClone, GE Life Sciences) and penicillin-streptomycin solution (100 units penicillin and 0.10 mg streptomycin/mL) (Sigma-Aldrich, USA) at 37 °C in a 5 % CO_2 atmosphere.

The DNA was introduced inside the cells via electroporation using the BTX-ECM 830 system (Harvard Apparatus, USA). Prior to electroporation, cells were washed with 1 x Dulbecco's Phosphate Buffered Saline (PBS) (Sigma-Aldrich, USA) and harvested using 1 x Trypsin/EDTA (Sigma-Aldrich, USA) in PBS. Cells were then centrifuged at 1000 rpm for 5 min and resuspended in 1 x PBS. To estimate the number of cells per mL, cells were counted in a Bürker counting chamber and approximately 1.1 x 10^8 cells were used to prepare the NMR sample. The proper amount of cells was centrifuged (1000 rpm for 5 min) and resuspended in 2.8 mL of the Electroporation buffer (EC buffer) (140 mM NaPOi, 5 mM KCl, 10 mM $MgCl_2$, pH 7.0) containing 400 µM DNA and 10 µM FAM-labelled DNA. The cell suspension was then divided into 4-mm electroporation cuvettes (Cell Projects, UK) and incubated on ice for 5 min. The used electroporation procedure consisted of two square-wave pulses (100 µs/1000 V and 30 ms/350 V) separated by a 5 s interval. After electroporation, cells were incubated at room temperature for 2 min, then transferred into Leibovitz L15-/- medium (no FBS/no antibiotics) (Gibco, USA) and centrifuged (1000 rpm, 5 min). Cells were resuspended in fresh L15 -/- medium and a small portion of the suspension (~ 6 x 10^5 cells) was used for flow cytometry (FCM) and confocal microscopy analysis (see below) to evaluate the cell viability, electroporation efficiency and DNA localization. The rest of the suspension was centrifuged (1000 rpm, 5 min) and after removing the supernatant, the cell pellet was resuspended in 550 µL of Leibovitz L15 -/- medium containing 10 % D_2O and placed into a 5-mm Shigemi NMR tube (Shigemi Co., Tokyo, Japan). Prior to performing the NMR experiment, cells in the NMR tube were manually centrifuged using a "hand centrifuge" (CortecNet, France) to form a fluffy pellet at the bottom of the NMR tube. Finally, 450 µL of L15 -/- medium (with 10 % D2O) was carefully added into the tube.

Flow cytometry. For FCM analysis, ~ 10^5 cells were resuspended in 200 µL of PBS buffer (Sigma-Aldrich, USA) and 1 µL (stock solution was 1 mg/mL) of propidium iodide (PI) (Exbio, Czech Republic) was added for distinguishing the living cells from the apoptotic population, dead cells, or cells with compromised membrane integrity. Subsequently, 10^4 HeLa cells were analyzed using a BD FACSVerse flow cytometer with BD FACSuite software (BD Biosciences, San Jose, CA, USA). To detect the cell viability, excitation wavelength for PI was set to 488 nm, and the emission was detected at 700/54 nm. To evaluate the transfection efficiency, the fluorescently (FAM) labelled DNA was excited at 488 nm, and the emission was detected at 527/32 nm.

Confocal microscopy. For confocal microscopy, ~ 5 x 10^5 cells were placed in one drop onto a 35-mm glass dish (ibidi GmbH, Germany) pre-coated with 0.01 % poly-L-lysine (Sigma-Aldrich, USA). The cell drop was then immersed in 2 mL of Leibovitz L15 -/- medium containing 1 µg/mL

Hoechst dye (Sigma-Aldrich, USA) to stain the cell nuclei. All microscopy images were obtained using a Zeiss LSM 800 confocal microscope with a 63x/1.2 C-Apo-chromat objective.

In-cell NMR spectra acquisition. For in-cell NMR spectroscopy analysis, a 600 MHz Bruker Avance III HD spectrometer (Bruker, Corporation, Billerica, MA, USA) equipped with a quadruple-resonance cryogenic probe was used. In-cell 1D ^{1}H NMR spectra were acquired at 20 °C in Leibovitz L15 -/- medium containing 10 % D_2O, with 5 x 256 scans, using a 1D ^{1}H JR-echo (1-1 echo) pulse sequence (6) with zero excitation set to the resonance of water and the excitation maximum set to 13 ppm. The spectra were corrected for baseline and processed with the exponential apodization function with the line-broadening parameter set to 14. Data were processed using MNova v12.0.0 (Mestrelab Research, Spain).

Immediately after the acquisition of the in-cell NMR spectrum, 1D ^{1}H NMR spectrum of the supernatant was measured (using the same parameters as were used for acquiring the in-cell NMR spectra) to control for possible leakage of the transfected DNA from the cells. Meanwhile, a fraction of cells in the NMR tube was taken for FCM analysis to evaluate the cell mortality in the course of the NMR experiment, and the rest of the cells were subjected to lysation (see below) in order to control for possible DNA degradation and to acquire higher resolution spectra in a more homogenous sample (5, 7). Finally, 1D ^{1}H NMR spectrum of the cell lysate was measured using the same NMR parameters as mentioned above.

Cell lysation. Following the acquisition of the in-cell NMR spectrum, the cellular pellet was resuspended in 200 µL of the Lysation buffer (10 mM NaPOi, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 % NP-40, pH 6.8), sonicated on ice with a micro-tip (3 x 10 s at amplitude of 50 %; then 1 x 5 s at amplitude of 85 %), and heated at 95 °C for 2 min. The sample was then centrifuged at max speed (15 000 rpm) for 10 min and the pH of the resulting supernatant was adjusted to pH < 6.5. After 10 % D_2O enrichment, the sample was finally placed into a 5-mm Shigemi NMR tube (Shigemi Co., Tokyo, Japan) and taken for NMR measurement.

Modeling studies for stability prediction

Modeling studies have been constructed, to separately predict melting temperatures (T_m at pH 7.0) or the pH transition mid-points (pH_T), by adhering to three general strategies. Unless otherwise stated, all the calculations were done *via* custom scripts written in R programming language (5).

First, a model was generated by modifying the existing G4Hunter algorithm (8), adapting that to the given sequence space of C-based i-DNAs with T-only spacers. In the original G4Hunter, designated for G-quadruplex sequences, a sole G singleton acquires a score of 1 (a scoring coefficient), each G in a GG tract acquires 2 and so on. In the modified version, the base that adds positive scoring was set to be C, instead of G, with the maximum cutoff for the length of the tract set to 6. The contribution from the T bases, along with any other possible bases, was set to be 0. Furthermore, the individual non-0 scoring coefficients, for each C in CC tract, each C in CCC tract and so on, were optimized to values different from the conventional G4Hunter integer numbers. The optimization was done to fit the provided i-DNA dataset (T_m or pH_T), using the *Optimus* optimization engine (9), *via* an acceptance ratio annealing Monte Carlo technique. Acceptance ratio values were allowed to linearly reduce from 90 % to 5 % in 4 cycles, each using 250,000 optimization steps. In each step, a random scoring coefficient was selected, altering its value by 0.1, with a sign of alteration (*i.e.* whether adding or subtracting) also randomly determined. The new configuration was then either retained or rejected based on the Metropolis criterion, with the acceptance probabilities conforming the above-mentioned linear regiment of the acceptance ratio annealing through a special self-adjusting pseudo-temperature bath (9).

The second strategy was to develop a novel sequence-only machine learning model for the restricted C- and T-based sequence space used in this study. Due to the simplicity of the explored

sub-universe of i-DNA structures, we were able to use only four features to fully abstract the sequence in our dataset. Those features were C-tract length (denoted as C, same for all C-tracts in a given i-DNA candidate sequence), and the lengths of all three T-based loops (denoted as T_1, T_2 and T_3, from 5' to 3' direction). All features were next checked against the presence of a strong cross-correlation, and were centered and scaled. Those were then used for machine learning, by adopting the XGBoost (10) implementation of the Gradient Boosting Machines(11-13), as interfaced through R via the Caret library (14). Gradient boosting machines have been successfully applied for modelling G-quadruplex structures before (15), hence a similar strategy was used here, but with simple initial feature set. To tune the machine learning architecture, five hyperparameters (in the XGBoost implementation denoted as *eta* - learning rate or shrinkage, *max_depth* - interaction depth, *min_child_weight* - final leave characteristics in the trees, *subsample* - bag fraction or subsampling ratio, and *nrounds* - number of trees) were optimized on all 196 sequences data, using root mean squared errors of predictions (RMSE) in a repeated cross-validation (5-fold, repeated thrice) setup for the performance evaluation. Using the architecture-defining optimal hyperparameters separately identified for the modeling of T_m and pH_T, the GBM models were then trained on randomly chosen 80 % of data, further testing on the 20 % left out test set. This resulted in two models, one with 1.210 RMSE and 0.990 Pearson's R for T_m predictions, and the second with 0.053 RMSE and 0.973 Pearson's R for pH_T.

In the third approach, we searched for more transparent mathematical models to express T_m and pH_T measurements as a function of C-tract (C) and loop (T_1, T_2 and T_3) lengths. To search for such non-linear equations, we used Eureqa (16), a program for an unrestricted search for analytical forms in provided data. We used the default absolute error as a performance metric for the search, and, for the sake of the lucidity of the resulting equations, allowed only constant, input variable, addition, subtraction, multiplication, division and exponentiation terms and operations in the equations. All sequences were inputted to the program, making use of Eureqa's internal capability to split the data for training and validation.

Table S1 196 sequences information and BLAST results.

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
	C_3 Tract					
T111-3	CCCTCCCTCCCTCCC	15	3	110900	6.27	1-22,x,y
T222-3	CCCTTCCCTTCCCTTCCC	18	6	135200	6.09	1-22,x,y
T333-3	CCCTTTCCCTTTCCCTTTCCC	21	9	159500	6.16	1-13,17-20,22,x,y
T444-3	CCCTTTTCCCTTTTCCCTTTTCCC	24	12	183800	6.23	10,16
T112-3	CCCTCCCTCCCTTCCC	16	4	119000	6.07	1-22,x,y
T121-3	CCCTCCCTTCCCTCCC	16	4	119000	6.26	1-22,x,y
T211-3	CCCTTCCCTCCCTCCC	16	4	119000	6.10	1-22,x,y
T113-3	CCCTCCCTCCCTTTCCC	17	5	127100	6.12	1-22,x,y
T131-3	CCCTCCCTTTCCCTCCC	17	5	127100	6.10	1-7,8-17,19,20,y
T311-3	CCCTTTCCCTCCCTCCC	17	5	127100	6.11	1-20,22,y
T114-3	CCCTCCCTCCCTTTTCCC	18	6	135200	6.12	1,3,5-11,19,20,22
T141-3	CCCTCCCTTTTCCCTCCC	18	6	135200	6.11	3,5,10,12,16
T411-3	CCCTTTTCCCTCCCTCCC	18	6	135200	6.12	1,2,4,6,7,11-13,17,20,x
T115-3	CCCTCCCTCCCTTTTTCCC	19	7	143300	6.03	3,7,12-15,17,20,22
T151-3	CCCTCCCTTTTTCCCTCCC	19	7	143300	6.10	1,2,5-11,17,18,x,y
T511-3	CCCTTTTTCCCTCCCTCCC	19	7	143300	6.01	1,2,7,10,11,12
T116-3	CCCTCCCTCCCTTTTTTCCC	20	8	151400	6.30	5,12,15,17
T161-3	CCCTCCCTTTTTTCCCTCCC	20	8	151400	6.01	2,5,17
T611-3	CCCTTTTTTCCCTCCCTCCC	20	8	151400	6.08	7,10,12
					--	
T221-3	CCCTTCCCTTCCCTCCC	17	5	127100	6.14	1-21,x,y
T212-3	CCCTTCCCTCCCTTCCC	17	5	127100	6.19	1-19,22,x
T122-3	CCCTCCCTTCCCTTCCC	17	5	127100	6.14	1-21,x,y
					--	
T223-3	CCCTTCCCTTCCCTTTCCC	19	7	143300	6.16	1-12,15-17,19,10,x
T232-3	CCCTTCCCTTTCCCTTCCC	19	7	143300	6.21	1-7,10-16,19-21,x
T322-3	CCCTTTCCCTTCCCTTCCC	19	7	143300	6.19	1-12,16-20,x
					--	
T224-3	CCCTTCCCTTCCCTTTTCCC	20	8	151400	5.99	2,4,5,7,10,15,21,x,y
T242-3	CCCTTCCCTTTTCCCTTCCC	20	8	151400	6.13	1,2,4,5,7,10,11,15,x
T422-3	CCCTTTTCCCTTCCCTTCCC	20	8	151400	6.05	1,2,4,6,7,10,15,x
					--	
T225-3	CCCTTCCCTTCCCTTTTTCCC	21	9	159500	5.85	10,11
T252-3	CCCTTCCCTTTTTCCCTTCCC	21	9	159500	6.09	5,7,8,10
T522-3	CCCTTTTTCCCTTCCCTTCCC	21	9	159500	5.88	4,7,10

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mole^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
					--	
T226-3	CCCTTCCCTTCCCTTTTTTCCC	22	10	167600	5.93	none
T262-3	CCCTTCCCTTTTTTCCCTTCCC	22	10	167600	6.07	none
T622-3	CCCTTTTTTCCCTTCCCTTCCC	22	10	167600	5.84	8
					--	
T331-3	CCCTTTCCCTTTCCCTCCC	19	7	143300	6.07	2,6,7,10,11,16,19,20,x,y
T313-3	CCCTTTCCCTCCCTTTCCC	19	7	143300	6.00	1,2,3,5,9,12,13,17,20
T133-3	CCCTCCCTTTCCCTTTCCC	19	7	143300	6.10	1,11,15,18,19,20
					--	
T332-3	CCCTTTCCCTTTCCCTTCCC	20	8	151400	6.29	1-12,14,16,19,20,22,x
T323-3	CCCTTTCCCTTCCCTTTCCC	20	8	151400	6.25	7,9-11,15-17,20,22
T233-3	CCCTTCCCTTTCCCTTTCCC	20	8	151400	6.28	1-12,14,16,19,20,22,x
					--	
T334-3	CCCTTTCCCTTTCCCTTTTCCC	22	10	167600	6.20	1,13
T343-3	CCCTTTCCCTTTTCCCTTTCCC	22	10	167600	6.19	13
T433-3	CCCTTTTCCCTTTCCCTTTCCC	22	10	167600	6.19	3,8,x
					--	
T335-3	CCCTTTCCCTTTCCCTTTTTCCC	23	11	175700	6.11	none
T353-3	CCCTTTCCCTTTTTCCCTTTCCC	23	11	175700	6.18	6
T533-3	CCCTTTTTCCCTTTCCCTTTCCC	23	11	175700	6.11	9
					--	
T336-3	CCCTTTCCCTTTCCCTTTTTTCCC	24	12	183800	6.11	none
T363-3	CCCTTTCCCTTTTTTCCCTTTCCC	24	12	183800	6.15	none
T633-3	CCCTTTTTTCCCTTTCCCTTTCCC	24	12	183800	6.15	none
	C_4 Tract					
T111-4	CCCCTCCCCTCCCCTCCCC	19	3	139700	6.23	1-22,x,y
T222-4	CCCCTTCCCCTTCCCCTTCCCC	22	6	164000	6.27	1-12,14-22,x,y
T333-4	CCCCTTTCCCCTTTCCCCTTTCCCC	25	9	188300	6.37	15
T444-4	CCCCTTTTCCCCTTTTCCCCTTTTCCCC	28	12	212600	6.52	none
					--	
T112-4	CCCCTCCCCTCCCCTTCCCC	20	4	147800	6.25	1-10,12-14,16-22,x,y
T121-4	CCCCTCCCCTTCCCCTCCCC	20	4	147800	6.29	2,4-10,13,14,16,17,19,20,22,x,y
T211-4	CCCCTTCCCCTCCCCTCCCC	20	4	147800	6.26	1-17,19,20,x,y
					--	
T113-4	CCCCTCCCCTCCCCTTTCCCC	21	5	155900	6.22	2,3,8,13,20
T131-4	CCCCTCCCCTTTCCCCTCCCC	21	5	155900	6.37	2,5,8,15,21,22
T311-4	CCCCTTTCCCCTCCCCTCCCC	21	5	155900	6.30	2,16,21
					--	
T114-4	CCCCTCCCCTCCCCTTTTCCCC	22	6	164000	6.32	none
T141-4	CCCCTCCCCTTTTCCCCTCCCC	22	6	164000	6.32	none
T411-4	CCCCTTTTCCCCTCCCCTCCCC	22	6	164000	6.28	19

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
					--	
T115-4	CCCCTCCCCTCCCCTTTTTCCCC	23	7	172100	6.27	none
T151-4	CCCCTCCCCTTTTTCCCCTCCCC	23	7	172100	6.33	none
T511-4	CCCCTTTTTCCCCTCCCCTCCCC	23	7	172100	6.30	11
					--	
T116-4	CCCCTCCCCTCCCCTTTTTTCCCC	24	8	180200	6.14	none
T161-4	CCCCTCCCCTTTTTTCCCCTCCCC	24	8	180200	6.33	none
T611-4	CCCCTTTTTTCCCCTCCCCTCCCC	24	8	180200	6.15	none
					--	
T221-4	CCCCTTCCCCTTCCCCTCCCC	21	5	155900	6.25	1-5,7,8,11,16,19,x
T212-4	CCCCTTCCCCTCCCCTTCCCC	21	5	155900	6.22	2,4-7,9,10,13,15,16,19,21
T122-4	CCCCTCCCCTTCCCCTTCCCC	21	5	155900	6.28	2,3,7,8,10,13,16,17,19,21
					--	
T223-4	CCCCTTCCCCTTCCCCTTTCCCC	23	7	172100	6.35	2,7
T232-4	CCCCTTCCCCTTTCCCCTTCCCC	23	7	172100	6.36	12
T322-4	CCCCTTTCCCCTTCCCCTTCCCC	23	7	172100	6.29	3
T224-4	CCCCTTCCCCTTCCCCTTTTCCCC	24	8	180200	6.28	none
T242-4	CCCCTTCCCCTTTTCCCCTTCCCC	24	8	180200	6.34	none
T422-4	CCCCTTTTCCCCTTCCCCTTCCCC	24	8	180200	6.30	none
					--	
T225-4	CCCCTTCCCCTTCCCCTTTTTCCCC	24	9	188300	6.26	none
T252-4	CCCCTTCCCCTTTTTCCCCTTCCCC	25	9	188300	6.35	none
T522-4	CCCCTTTTTCCCCTTCCCCTTCCCC	25	9	188300	6.27	none
					--	
T226-4	CCCCTTCCCCTTCCCCTTTTTTCCCC	26	10	196400	6.23	none
T262-4	CCCCTTCCCCTTTTTTCCCCTTCCCC	26	10	196400	6.29	none
T622-4	CCCCTTTTTTCCCCTTCCCCTTCCCC	26	10	196400	6.22	none
					--	
T331-4	CCCCTTTCCCCTTTCCCCTCCCC	23	7	172100	6.35	none
T313-4	CCCCTTTCCCCTCCCCTTTCCCC	23	7	172100	6.26	none
T133-4	CCCCTCCCCTTTCCCCTTTCCCC	23	7	172100	6.33	none
					--	
T332-4	CCCCTTTCCCCTTTCCCCTTCCCC	24	8	180200	6.35	none
T323-4	CCCCTTTCCCCTTCCCCTTTCCCC	24	8	180200	6.36	none
T233-4	CCCCTTCCCCTTTCCCCTTTCCCC	24	8	180200	6.34	none
					--	
T334-4	CCCCTTTCCCCTTTCCCCTTTTCCCC	26	10	196400	6.40	None
T343-4	CCCCTTTCCCCTTTTCCCCTTTCCCC	26	10	196400	6.36	none
T433-4	CCCCTTTTCCCCTTTCCCCTTTCCCC	26	10	196400	6.39	none
					--	

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
T335-4	CCCCTTTCCCCTTTCCCCTTTTTCCCC	27	11	204500	6.33	none
T353-4	CCCCTTTCCCCTTTTTCCCCTTTCCCC	27	11	204500	6.35	none
T533-4	CCCCTTTTTCCCCTTTCCCCTTTCCCC	27	11	204500	6.32	none
					--	
T336-4	CCCCTTTCCCCTTTCCCCTTTTTTCCCC	28	12	212600	6.31	none
T363-4	CCCCTTTCCCCTTTTTTCCCCTTTCCCC	28	12	212600	6.41	none
T633-4	CCCCTTTTTTCCCCTTTCCCCTTTCCCC	28	12	212600	6.32	none
	C_5 Tract					
T111-5	CCCCCTCCCCCTCCCCCTCCCCC	23	3	168500	6.44	1-20,22,x
T222-5	CCCCCTTCCCCCTTCCCCCTTCCCCC	26	6	192800	6.48	4,10,11
T333-5	CCCCCTTTCCCCCTTTCCCCCTTTCCCCC	29	9	217100	6.56	none
T444-5	CCCCCTTTTCCCCCTTTTCCCCCTTTTCCCCC	32	12	241400	6.71	none
					--	
T112-5	CCCCCTCCCCCTCCCCCTTCCCCC	24	4	176600	6.41	21
T121-5	CCCCCTCCCCCTTCCCCCTCCCCC	24	4	176600	6.44	7,9,21
T211-5	CCCCCTTCCCCCTCCCCCTCCCCC	24	4	176600	6.43	8,9,15
					--	
T113-5	CCCCCTCCCCCTCCCCCTTTCCCCC	25	5	184700	6.62	none
T131-5	CCCCCTCCCCCTTTCCCCCTCCCCC	25	5	184700	6.75	1,13
T311-5	CCCCCTTTCCCCCTCCCCCTCCCCC	25	5	184700	6.62	none
					--	
T114-5	CCCCCTCCCCCTCCCCCTTTTCCCCC	26	6	192800	6.41	none
T141-5	CCCCCTCCCCCTTTTCCCCCTCCCCC	26	6	192800	6.57	none
T411-5	CCCCCTTTTCCCCCTCCCCCTCCCCC	26	6	192800	6.39	none
					--	
T115-5	CCCCCTCCCCCTCCCCCTTTTTCCCCC	27	7	200900	6.34	none
T151-5	CCCCCTCCCCCTTTTTCCCCCTCCCCC	27	7	200900	6.45	none
T511-5	CCCCCTTTTTCCCCCTCCCCCTCCCCC	27	7	200900	6.37	none
					--	
T116-5	CCCCCTCCCCCTCCCCCTTTTTTCCCCC	28	8	209000	6.40	none
T161-5	CCCCCTCCCCCTTTTTTCCCCCTCCCCC	28	8	209000	6.41	none
T611-5	CCCCCTTTTTTCCCCCTCCCCCTCCCCC	28	8	209000	6.32	none
					--	
T221-5	CCCCCTTCCCCCTTCCCCCTCCCCC	25	5	184700	6.65	none
T212-5	CCCCCTTCCCCCTCCCCCTTCCCCC	25	5	184700	6.64	1,21
T122-5	CCCCCTCCCCCTTCCCCCTTCCCCC	25	5	184700	6.80	15
					--	
T223-5	CCCCCTTCCCCCTTCCCCCTTTCCCCC	27	7	200900	6.64	none
T232-5	CCCCCTTCCCCCTTTCCCCCTTCCCCC	27	7	200900	6.54	none
T322-5	CCCCCTTTCCCCCTTCCCCCTTCCCCC	27	7	200900	6.51	none
					--	
T224-5	CCCCCTTCCCCCTTCCCCCTTTTCCCCC	28	8	209000	6.57	none

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
T242-5	CCCCCTTCCCCCTTTTCCCCCTTCCCCC	28	8	209000	6.44	none
T422-5	CCCCCTTTTCCCCCTTCCCCCTTCCCCC	28	8	209000	6.51	none
					--	
T225-5	CCCCCTTCCCCCTTCCCCCTTTTTCCCCC	29	9	217100	6.55	none
T252-5	CCCCCTTCCCCCTTTTTCCCCCTTCCCCC	29	9	217100	6.56	none
T522-5	CCCCCTTTTTCCCCCTTCCCCCTTCCCCC	29	9	217100	6.47	none
					--	
T226-5	CCCCCTTCCCCCTTCCCCCTTTTTTCCCCC	30	10	225200	6.40	none
T262-5	CCCCCTTCCCCCTTTTTTCCCCCTTCCCCC	30	10	225200	6.52	none
T622-5	CCCCCTTTTTTCCCCCTTCCCCCTTCCCCC	30	10	225200	6.44	none
					--	
T331-5	CCCCCTTTCCCCCTTTCCCCCTCCCCC	27	7	200900	6.49	none
T313-5	CCCCCTTTCCCCCTCCCCCTTTCCCCC	27	7	200900	6.40	none
T133-5	CCCCCTCCCCCTTTCCCCCTTTCCCCC	27	7	200900	6.42	none
					--	
T332-5	CCCCCTTTCCCCCTTTCCCCCTTCCCCC	28	8	209000	6.48	none
T323-5	CCCCCTTTCCCCCTTCCCCCTTTCCCCC	28	8	209000	6.56	none
T233-5	CCCCCTTCCCCCTTTCCCCCTTTCCCCC	28	8	209000	6.56	none
					--	
T334-5	CCCCCTTTCCCCCTTTCCCCCTTTTCCCCC	30	10	225200	6.69	none
T343-5	CCCCCTTTCCCCCTTTTCCCCCTTTCCCCC	30	10	225200	6.54	none
T433-5	CCCCCTTTTCCCCCTTTCCCCCTTTCCCCC	30	10	225200	6.60	none
					--	
T335-5	CCCCCTTTCCCCCTTTCCCCCTTTTTCCCCC	31	11	233300	6.50	none
T353-5	CCCCCTTTCCCCCTTTTTCCCCCTTTCCCCC	31	11	233300	6.66	none
T533-5	CCCCCTTTTTCCCCCTTTCCCCCTTTCCCCC	31	11	233300	6.58	none
					--	
T336-5	CCCCCTTTCCCCCTTTCCCCCTTTTTTCCCCC	32	12	241400	6.74	none
T363-5	CCCCCTTTCCCCCTTTTTTCCCCCTTTCCCCC	32	12	241400	6.77	none
T633-5	CCCCCTTTTTTCCCCCTTTCCCCCTTTCCCCC	32	12	241400	6.67	none
	C_6 Tract					
T111-6	CCCCCCTCCCCCCTCCCCCCTCCCCCC	27	3	197300	6.57	1,3,5,7,8,12
T222-6	CCCCCCTTCCCCCCTTCCCCCCTTCCCCCC	30	6	221600	6.66	none
T333-6	CCCCCCTTTCCCCCCTTTCCCCCCTTTCCCCCC	33	9	245900	6.79	none
T444-6	CCCCCCTTTTCCCCCCTTTTCCCCCCTTTTCCCCCC	36	12	270200	6.84	none
					--	
T112-6	CCCCCCTCCCCCCTCCCCCCTTCCCCCC	28	4	205400	6.66	none
T121-6	CCCCCCTCCCCCCTTCCCCCCTCCCCCC	28	4	205400	6.65	4
T211-6	CCCCCCTTCCCCCCTCCCCCCTCCCCCC	28	4	205400	6.60	none
					--	
T113-6	CCCCCCTCCCCCCTCCCCCCTTTCCCCCC	29	5	213500	6.51	none
T131-6	CCCCCCTCCCCCCTTTCCCCCCTCCCCCC	29	5	213500	6.68	none

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
T311-6	CCCCCCTTTCCCCCCTCCCCCCTCCCCCC	29	5	213500	6.64	none
					--	
T114-6	CCCCCCTCCCCCCTCCCCCCTTTTCCCCCC	30	6	221600	6.57	none
T141-6	CCCCCCTCCCCCCTTTTCCCCCCTCCCCCC	30	6	221600	6.69	none
T411-6	CCCCCCTTTTCCCCCCTCCCCCCTCCCCCC	30	6	221600	6.59	none
					--	
T115-6	CCCCCCTCCCCCCTCCCCCCTTTTTCCCCCC	31	7	229700	6.59	none
T151-6	CCCCCCTCCCCCCTTTTTCCCCCCTCCCCCC	31	7	229700	6.63	none
T511-6	CCCCCCTTTTTCCCCCCTCCCCCCTCCCCCC	31	7	229700	6.50	none
					--	
T116-6	CCCCCCTCCCCCCTCCCCCCTTTTTTCCCCCC	32	8	237800	6.58	none
T161-6	CCCCCCTCCCCCCTTTTTTCCCCCCTCCCCCC	32	8	237800	6.63	none
T611-6	CCCCCCTTTTTTCCCCCCTCCCCCCTCCCCCC	32	8	237800	6.53	none
					--	
T221-6	CCCCCCTTCCCCCCTTCCCCCCTCCCCCC	29	5	221600	6.69	none
T212-6	CCCCCCTTCCCCCCTCCCCCCTTCCCCCC	29	5	221600	6.64	none
T122-6	CCCCCCTCCCCCCTTCCCCCCTTCCCCCC	29	5	221600	6.65	none
					--	
T223-6	CCCCCCTTCCCCCCTTCCCCCCTTTCCCCCC	31	7	229700	6.88	none
T232-6	CCCCCCTTCCCCCCTTTCCCCCCTTCCCCCC	31	7	229700	6.84	none
T322-6	CCCCCCTTTCCCCCCTTCCCCCCTTCCCCCC	31	7	229700	6.81	none
					--	
T224-6	CCCCCCTTCCCCCCTTCCCCCCTTTTCCCCCC	32	8	237800	6.86	none
T242-6	CCCCCCTTCCCCCCTTTTCCCCCCTTCCCCCC	32	8	237800	6.79	none
T422-6	CCCCCCTTTTCCCCCCTTCCCCCCTTCCCCCC	32	8	237800	6.80	none
					--	
T225-6	CCCCCCTTCCCCCCTTCCCCCCTTTTTCCCCCC	33	9	245900	6.71	none
T252-6	CCCCCCTTCCCCCCTTTTTCCCCCCTTCCCCCC	33	9	245900	6.75	none
T522-6	CCCCCCTTTTTCCCCCCTTCCCCCCTTCCCCCC	33	9	245900	6.70	none
					--	
T226-6	CCCCCCTTCCCCCCTTCCCCCCTTTTTTCCCCCC	34	10	254000	6.66	none
T262-6	CCCCCCTTCCCCCCTTTTTTCCCCCCTTCCCCCC	34	10	254000	6.71	none
T622-6	CCCCCCTTTTTTCCCCCCTTCCCCCCTTCCCCCC	34	10	254000	6.67	none
					--	
T331-6	CCCCCCTTTCCCCCCTTTCCCCCCTCCCCCC	31	7	229700	6.76	none
T313-6	CCCCCCTTTCCCCCCTCCCCCCTTTCCCCCC	31	7	229700	6.72	none
T133-6	CCCCCCTCCCCCCTTTCCCCCCTTTCCCCCC	31	7	229700	6.74	none
					--	
T332-6	CCCCCCTTTCCCCCCTTTCCCCCCTTCCCCCC	32	8	237800	6.75	none
T323-6	CCCCCCTTTCCCCCCTTCCCCCCTTTCCCCCC	32	8	237800	6.75	none
T233-6	CCCCCCTTCCCCCCTTTCCCCCCTTTCCCCCC	32	8	237800	6.73	none
					--	

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mol e^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
T334-6	CCCCCCTTTCCCCCCTTTCCCCCCTTTTCCCCCC	34	10	254000	6.82	none
T343-6	CCCCCCTTTCCCCCCTTTTCCCCCCTTTCCCCCC	34	10	254000	6.80	none
T433-6	CCCCCCTTTTCCCCCCTTTCCCCCCTTTCCCCCC	34	10	254000	6.81	none
					--	
T335-6	CCCCCCTTTCCCCCCTTTCCCCCCTTTTTCCCCCC	35	11	262100	6.77	none
T353-6	CCCCCCTTTCCCCCCTTTTTCCCCCCTTTCCCCCC	35	11	262100	6.92	none
T533-6	CCCCCCTTTTTCCCCCCTTTCCCCCCTTTCCCCCC	35	11	262100	6.81	none
					--	
T336-6	CCCCCCTTTCCCCCCTTTCCCCCCTTTTTTCCCCCC	36	12	270200	6.77	none
T363-6	CCCCCCTTTCCCCCCTTTTTTCCCCCCTTTCCCCCC	36	12	270200	6.81	none
T633-6	CCCCCCTTTTTTCCCCCCTTTCCCCCCTTTCCCCCC	36	12	270200	6.74	none
	40 extended sequences with C_5-tract					
	Longer (7-15) central loop					
T1Z1-5	CCCCCTCCCCCTTTTTTTCCCCCTCCCCC	29	9	217100		none
T181-5	CCCCCTCCCCCTTTTTTTTCCCCCTCCCCC	30	10	225200		none
T1101-5	CCCCCTCCCCC T_{10} CCCCCTCCCCC	32	12	241400		none
T1151-5	CCCCCTCCCCC T_{15} CCCCCTCCCCC	37	17	281900		none
	Adenine in loop					
AA115-5	CCCCCACCCCCACCCCCTTTTTCCCCC	27	7	209500		none
AA151-5	CCCCCACCCCCTTTTTCCCCCACCCCC	27	7	209500		none
AA511-5	CCCCCTTTTTCCCCCACCCCCACCCCC	27	7	209500		none
--						
1A15-5	CCCCCACCCCCTCCCCCTTTTTCCCCC	27	7	205200		none
11A5-5	CCCCCTCCCCCACCCCCTTTTTCCCCC	27	7	205200		none
1A51-5	CCCCCACCCCCTTTTTCCCCCTCCCCC	27	7	205200		none
151A-5	CCCCCTCCCCCTTTTTCCCCCACCCCC	27	7	205200		none
51A1-5	CCCCCTTTTTCCCCCACCCCCTCCCCC	27	7	205200		none
511A-5	CCCCCTTTTTCCCCCTCCCCCACCCCC	27	7	205200		none
--						
115_1A-5	CCCCCTCCCCCTCCCCCATTTTCCCCC	27	7	206200		none
151_1A-5	CCCCCTCCCCCATTTTCCCCCTCCCCC	27	7	206200		none
511_1A-5	CCCCCATTTTCCCCCTCCCCCTCCCCC	27	7	206200		none
--						
115_2A-5	CCCCCTCCCCCTCCCCCTATTTCCCCC	27	7	206800		none
151_2A-5	CCCCCTCCCCCTATTTCCCCCTCCCCC	27	7	206800		none
511_2A-5	CCCCCTATTTCCCCCTCCCCCTCCCCC	27	7	206800		none
--						
115_3A-5	CCCCCTCCCCCTCCCCCTTATTCCCCC	27	7	206800		none
151_3A-5	CCCCCTCCCCCTTATTCCCCCTCCCCC	27	7	206800		none
511_3A-5	CCCCCTTATTCCCCCTCCCCCTCCCCC	27	7	206800		none

Name	Sequence (5'→3')	nt	Total Loop Length	ε_{260}/L·mole^{-1}·cm^{-1}	pH_T^{UV}	Human Genome (Chromosome) [a]
--						
115_4A-5	CCCCCTCCCCCTCCCCCTTTATCCCCC	27	7	206800		none
151_4A-5	CCCCCTCCCCCTTTATCCCCCTCCCCC	27	7	206800		none
511_4A-5	CCCCCTTTATCCCCCTCCCCCTCCCCC	27	7	206800		none
--						
115_5A-5	CCCCCTCCCCCTCCCCCTTTTACCCCC	27	7	206800		none
151_5A-5	CCCCCTCCCCCTTTTACCCCCTCCCCC	27	7	206800		none
511_5A-5	CCCCCTTTTACCCCCTCCCCCTCCCCC	27	7	206800		none
	Two short loops of different length					
T152-5	CCCCCTCCCCCTTTTTCCCCCTTCCCCC	28	8	209000		none
T251-5	CCCCCTTCCCCCTTTTTCCCCCTCCCCC	28	8	209000		none
T153-5	CCCCCTCCCCCTTTTTCCCCCTTTCCCCC	29	9	217100		none
T351-5	CCCCCTTTCCCCCTTTTTCCCCCTCCCCC	29	9	217100		none
T253-5	CCCCCTTCCCCCTTTTTCCCCCTTTCCCCC	30	10	225200		none
T352-5	CCCCCTTTCCCCCTTTTTCCCCCTTCCCCC	30	10	225200		none
T162-5	CCCCCTCCCCCTTTTTTCCCCCTTCCCCC	29	9	217100		none
T261-5	CCCCCTTCCCCCTTTTTTCCCCCTCCCCC	29	9	217100		none
T163-5	CCCCCTCCCCCTTTTTTCCCCCTTTCCCCC	30	10	225200		none
T361-5	CCCCCTTTCCCCCTTTTTTCCCCCTCCCCC	30	10	225200		none
T263-5	CCCCCTTCCCCCTTTTTTCCCCCTTTCCCCC	31	11	233300		none
T362-5	CCCCCTTTCCCCCTTTTTTCCCCCTTCCCCC	31	11	233300		none

[a] If the sequence is found in human genome, chromosome number is given here; 'None' means that the sequence does not been found in human genomes by BLAST (17).

Table S2 pH transition midpoint and thermal stability of 196 pyrimidine sequences containing thymidine spacers (T_1 to T_6) and C-tract (C_3 to C_6) of variable lengths. [a]

Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	$T_m^{pH\ 7.0}$ [d]		
								$T_{heating}$	$T_{Cooling}$	T_m
	C_3 Tract				***C_5 Tract***					
T111-3	--	6.22	57.8	T111-5	--	6.48	72.0	14.6	13.7	14.1
T222-3	--	6.16	53.2	T222-5	--	6.58	73.3	18.0	15.8	16.9
T333-3	--	6.22	56.9	T333-5	--	6.68	73.4	22.3	14.3	18.3
T444-3	--	6.27	58.8	T444-5	--	6.71	73.4	25.8	9.3	17.6
			--	--		--	--	--	--	--
T112-3	*SSL*	6.11	53.5	T112-5	*SSL*	6.51	71.1	14.8	13.8	14.3
T121-3	*SLS*	6.30	58.0	T121-5	*SLS*	6.56	73.6	16.9	15.6	16.2
T211-3	*LSS*	6.12	54.5	T211-5	*LSS*	6.50	71.9	15.4	13.9	14.7
			--	--		--	--	--	--	--
T113-3	*SSL*	6.15	54.9	T113-5	*SSL*	6.56	74.4	17.6	15.5	16.5
T131-3	*SLS*	6.20	54.3	T131-5	*SLS*	6.58	74.7	17.7	16.1	16.9
T311-3	*LSS*	6.10	58.3	T311-5	*LSS*	6.50	73.9	15.7	14.2	15.0
			--	--		--	--	--	--	--
T114-3	*SSL*	6.24	54.3	T114-5	*SSL*	6.52	69.5	15.2	12.1	13.6
T141-3	*SLS*	6.11	59.2	T141-5	*SLS*	6.60	74.2	18.2	15.9	17.0
T411-3	*LSS*	6.12	54.3	T411-5	*LSS*	6.47	70.6	15.7	13.8	14.8
			--	--		--	--	--	--	--
T115-3	*SSL*	6.07	52.2	T115-5	*SSL*	6.40	69.3	13.7	11.6	12.7
T151-3	*SLS*	6.11	57.5	T151-5	*SLS*	6.60	74.4	18.4	15.2	16.8
T511-3	*LSS*	6.03	51.7	T511-5	*LSS*	6.45	70.5	14.4	12.0	13.2
			--	--		--	--	--	--	--
T116-3	*SSL*	6.24	50.3	T116-5	*SSL*	6.45	66.6	13.1	10.6	11.8
T161-3	*SLS*	6.00	57.0	T161-5	*SLS*	6.60	73.1	17.9	14.3	16.1
T611-3	*LSS*	6.00	50.0	T611-5	*LSS*	6.45	68.0	14.0	11.4	12.7
		--	--	--		--	--	--	--	--
T221-3	*LLS*	6.07	54.6	T221-5	*LLS*	6.56	76.3	17.9	16.2	17.0
T212-3	*LSL*	6.05	54.9	T212-5	*LSL*	6.52	74.3	16.1	14.6	15.4
T122-3	*SLL*	6.07	54.6	T122-5	*SLL*	6.55	75.7	17.3	15.8	16.6
		--	--	--		--	--	--	--	--
T223-3	*SSL*	6.13	52.5	T223-5	*SSL*	6.62	72.3	19.0	15.4	17.2
T232-3	*SLS*	6.04	57.3	T232-5	*SLS*	6.57	74.0	19.2	15.6	17.4
T322-3	*LSS*	6.07	53.7	T322-5	*LSS*	6.58	73.2	19.5	15.8	17.6
		--	--	--		--	--	--	--	--
T224-3	*SSL*	5.97	52.2	T224-5	*SSL*	6.59	72.8	20.8	14.8	17.8
T242-3	*SLS*	6.07	59.1	T242-5	*SLS*	6.60	74.7	20.6	15.4	18.0
T422-3	*LSS*	6.00	53.3	T422-5	*LSS*	6.58	74.2	20.3	14.0	17.2
		--	--	--		--	--	--	--	--
T225-3	*SSL*	5.96	48.3	T225-5	*SSL*	6.55	69.1	18.4	12.7	15.6

Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	$T_m^{pH\ 7.0}$ [d]		
								$T_{heating}$	$T_{Cooling}$	T_m
T252-3	*SLS*	6.16	58.3	T252-5	*SLS*	6.59	73.4	20.8	14.0	17.4
T522-3	*LSS*	5.98	49.5	T522-5	*LSS*	6.53	70.5	18.8	13.3	16.1
		--	--	--		--	--	--	--	--
T226-3	*SSL*	6.02	48.0	T226-5	*SSL*	6.50	70.8	18.3	11.7	15.0
T262-3	*SLS*	6.16	57.1	T262-5	*SLS*	6.59	73.7	20.5	12.9	16.7
T622-3	*LSS*	5.97	48.8	T622-5	*LSS*	6.50	66.5	18.3	11.2	14.7
		--	--	--		--	--	--	--	--
T331-3	*LLS*	6.09	56.4	T331-5	*LLS*	6.50	73.5	19.7	15.9	17.8
T313-3	*LSL*	6.08	53.2	T313-5	*LSL*	6.47	71.8	18.2	14.4	16.3
T133-3	*SLL*	6.14	56.7	T133-5	*SLL*	6.50	73.7	20.6	16.3	18.4
		--	--	--		--	--	--	--	--
T332-3	*LLS*	6.16	59.1	T332-5	*LLS*	6.54	75.4	20.7	15.4	18.1
T323-3	*LSL*	6.05	55.3	T323-5	*LSL*	6.60	73.7	21.6	15.8	18.7
T233-3	*SLL*	6.08	58.7	T233-5	*SLL*	6.62	76.5	20.9	15.5	18.2
		--	--	--		--	--	--	--	--
T334-3	*SSL*	6.23	56.6	T334-5	*SSL*	6.69	72.3	23.9	13.1	18.5
T343-3	*SLS*	6.23	57.6	T343-5	*SLS*	6.67	73.5	23.7	13.0	18.4
T433-3	*LSS*	6.21	57.0	T433-5	*LSS*	6.67	73.2	23.7	13.1	18.4
		--	--	--		--	--	--	--	--
T335-3	*SSL*	6.12	55.5	T335-5	*SSL*	6.57	73.0	22.7	11.4	17.0
T353-3	*SLS*	6.16	57.5	T353-5	*SLS*	6.65	73.5	24.0	11.6	17.8
T533-3	*LSS*	6.10	55.6	T533-5	*LSS*	6.62	73.1	22.5	11.3	16.9
		--	--	--		--	--	--	--	--
T336-3	*SSL*	6.11	52.6	T336-5	*SSL*	6.57	69.2	22.8	10.0	16.4
T363-3	*SLS*	6.15	56.1	T363-5	*SLS*	6.63	71.2	24.1	10.3	17.2
T633-3	*LSS*	6.08	53.0	T633-5	*LSS*	6.56	69.7	22.6	10.0	16.3
	C_4 Tract				***C_6 Tract***					
T111-4	--	6.44	66.4	T111-6	--	6.64	77.5	20.9	18.7	19.8
T222-4	--	6.28	66.0	T222-6	--	6.68	78.2	24.1	17.5	20.8
T333-4	--	6.45	68.0	T333-6	--	6.76	77.9	28.4	13.1	20.7
T444-4	--	6.52	68.6	T444-6	--	6.75	75.3	29.7	9.0	19.4
		--	--	--		--	--	--	--	--
T112-4	*SSL*	6.42	64.0	T112-6	*SSL*	6.64	76.9	21.6	18.4	20.0
T121-4	*SLS*	6.52	67.0	T121-6	*SLS*	6.66	79.1	22.7	19.3	21.0
T211-4	*LSS*	6.48	65.0	T211-6	*LSS*	6.59	77.9	21.8	18.6	20.2
		--	--	--		--	--	--	--	--
T113-4	*SSL*	6.33	65.5	T113-6	*SSL*	6.60	77.9	21.9	17.3	19.6
T131-4	*SLS*	6.35	68.5	T131-6	*SLS*	6.67	79.4	23.8	18.9	21.4
T311-4	*LSS*	6.38	65.6	T311-6	*LSS*	6.64	77.9	22.4	17.8	20.1
		--	--	--		--	--	--	--	--
T114-4	*SSL*	6.49	62.1	T114-6	*SSL*	6.47	75.2	21.9	15.9	18.9

Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	$T_m^{pH\ 7.0}$ [d]		
								$T_{heating}$	$T_{Cooling}$	T_m
T141-4	*SLS*	6.32	68.6	T141-6	*SLS*	6.67	78.1	24.1	17.8	20.9
T411-4	*LSS*	6.40	63.3	T411-6	*LSS*	6.60	76.2	22.4	16.5	19.5
		--	--	--		--	--	--	--	--
T115-4	*SSL*	6.35	62.5	T115-6	*SSL*	6.58	76.5	20.8	14.7	17.7
T151-4	*SLS*	6.41	68.8	T151-6	*SLS*	6.66	78.7	24.2	16.9	20.6
T511-4	*LSS*	6.30	60.7	T511-6	*LSS*	6.54	76.0	21.4	15.1	18.2
		--	--	--		--	--	--	--	--
T116-4	*SSL*	6.24	59.0	T116-6	*SSL*	6.54	73.1	20.3	13.1	16.7
T161-4	*SLS*	6.38	66.5	T161-6	*SLS*	6.62	77.4	23.7	15.8	19.7
T611-4	*LSS*	6.25	59.7	T611-6	*LSS*	6.54	74.1	21.3	13.7	17.5
		--	--	--		--	--	--	--	--
T221-4	*LLS*	6.34	67.6	T221-6	*LLS*	6.67	81.9	23.5	18.6	21.0
T212-4	*LSL*	6.29	65.0	T212-6	*LSL*	6.67	78.7	22.3	17.4	19.8
T122-4	*SLL*	6.37	67.4	T122-6	*SLL*	6.69	79.8	23.4	18.4	20.9
		--	--	--		--	--	--	--	--
T223-4	*SSL*	6.38	64.9	T223-6	*SSL*	6.66	77.5	25.5	15.9	20.7
T232-4	*SLS*	6.39	67.4	T232-6	*SLS*	6.74	78.4	25.9	16.6	21.2
T322-4	*LSS*	6.33	67.5	T322-6	*LSS*	6.74	78.4	25.9	16.2	21.1
		--	--	--		--	--	--	--	--
T224-4	*SSL*	6.33	65.2	T224-6	*SSL*	6.74	77.1	26.6	14.4	20.5
T242-4	*SLS*	6.40	68.5	T242-6	*SLS*	6.74	79.1	25.9	14.6	20.2
T422-4	*LSS*	6.37	66.2	T422-6	*LSS*	6.78	78.4	26.4	15.2	20.8
		--	--	--		--	--	--	--	--
T225-4	*SSL*	6.30	60.6	T225-6	*SSL*	6.59	75.1	25.0	12.8	18.9
T252-4	*SLS*	6.41	67.4	T252-6	*SLS*	6.68	77.2	26.2	13.3	19.8
T522-4	*LSS*	6.33	63.1	T522-6	*LSS*	6.63	76.1	25.4	12.9	19.2
		--	--	--		--	--	--	--	--
T226-4	*SSL*	6.27	61.6	T226-6	*SSL*	6.64	75.0	24.8	12.1	18.4
T262-4	*SLS*	6.35	68.4	T262-6	*SLS*	6.67	77.3	25.7	12.7	19.2
T622-4	*LSS*	6.27	63.1	T622-6	*LSS*	6.65	76.0	25.3	11.8	18.5
		--	--	--		--	--	--	--	--
T331-4	*LLS*	6.43	67.2	T331-6	*LLS*	6.74	78.4	25.9	16.6	21.2
T313-4	*LSL*	6.30	67.4	T313-6	*LSL*	6.71	76.4	25.8	16.0	20.9
T133-4	*SLL*	6.38	64.4	T133-6	*SLL*	6.73	78.4	26.9	16.5	21.7
		--	--	--		--	--	--	--	--
T332-4	*LLS*	6.42	69.5	T332-6	*LLS*	6.72	78.6	26.8	14.8	20.8
T323-4	*LSL*	6.39	66.9	T323-6	*LSL*	6.73	78.3	27.4	15.0	21.2
T233-4	*SLL*	6.41	69.7	T233-6	*SLL*	6.70	79.1	27.1	14.8	20.9
		--	--	--		--	--	--	--	--
T334-4	*SSL*	6.53	66.4	T334-6	*SSL*	6.74	76.5	28.8	11.9	20.4
T343-4	*SLS*	6.59	67.1	T343-6	*SLS*	6.74	77.3	29.1	11.3	20.2

Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	$T_m^{pH\ 7.0}$ [d]		
								$T_{heating}$	$T_{Cooling}$	T_m
T433-4	*LSS*	6.53	66.7	T433-6	*LSS*	6.70	77.0	28.7	11.4	20.1
		--	--	--		--	--	--	--	--
T335-4	*SSL*	6.44	66.2	T335-6	*SSL*	6.74	76.1	27.7	10.9	19.3
T353-4	*SLS*	6.50	67.6	T353-6	*SLS*	6.77	77.4	28.7	10.4	19.6
T533-4	*LSS*	6.43	67.1	T533-6	*LSS*	6.71	77.4	27.8	9.5	18.6
		--	--	--		--	--	--	--	--
T336-4	*SSL*	6.39	65.7	T336-6	*SSL*	6.76	74.6	27.9	8.8	18.4
T363-4	*SLS*	6.49	66.7	T363-6	*SLS*	6.90	75.8	28.7	9.6	19.2
T633-4	*LSS*	6.39	66.4	T633-6	*LSS*	6.78	75.5	27.5	8.5	18.0

[a] Detailed information for all sequences is presented in **Table S1**. pH transition midpoints were identified by pH-dependent CD (pH_T^{CD}) spectra at 288 nm and pH-dependent UV absorption (pH_T^{UV}, given in **Table S1**) spectra at 295 nm. Thermal stabilities (T_m, °C) were characterized by UV-melting curves at 295 nm. Standard deviations of pH_T and T_m of two independent measurements were less than 0.2 and 1.0 °C, respectively. pH_T obtained by CD and UV absorbance were in excellent agreement ($pH_T^{CD} - pH_T^{UV} < 0.25$). No melting experiment was performed for the C_3 and C_4 tracts since most of these sequences do not form an i-DNA at pH 7.0 (see TDS in Figure S1), or their melting temperatures is too low to be measured accurately.

[b] Name: The first 'T' letter means that all spacers are composed of thymine bases only; three consecutive numbers refer to lengths of the three spacers in the 5' to 3' direction; '-3, -4, -5 and -6' refer to sequences with four C_3, C_4, C_5, and C_6 tracts (all of equal length), respectively. For example, the T112-3 sequence is 5'-CCCTCCCTCCCTTCCC-3' (four repeats of 3 cytosines separated by one, one, and two thymines). Each *group* is composed of sequences which differ only in spacer permutation; it is named after the first sequence in the group. For example, the *T112-3* group is composed of three sequences: T112-3, T121-3, and T211-3.

[c] Spacer permutation is defined as the swap between two sequences of an intramolecular i-DNA keeping length and overall base composition constant (1). These sequences belong to the same group. Each group contains three sequences. In this study, as two spacers are of identical length by design, the spacer pattern can either be *SSL*, *SLS*, *LLS*, or *LLS*, *LSL*, *SLL*. Herein, *S* and *L* are short for relatively *short* and *long* spacers, respectively.

[d] As a first approximation (18), T_m at pH 7.0 is assumed to be equal to the average of half-transition values for heating and cooling curves, provided by the heating and cooling profiles which are recorded with the same temperature gradient: $T_m^{pH\ 7.0} = \frac{T_{heating} + T_{cooling}}{2}$.

Table S3 pH transition and thermal stability at pH 5.0 of extended sequences with four C_5-tracts. [a]

Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$	Name [b]	Spacer Permutation [c]	pH_T^{CD}	$T_m^{pH\ 5.0}$
Longer (7-15) central spacer				***Adenine in spacer***			
T171-5	*SLS*	6.54	72.7	AA115-5	*SSL*	6.38	69.5
T181-5	*SLS*	6.52	72.3	AA151-5	*SLS*	6.43	74.6
T1101-5	*SLS*	6.47	71.0	AA511-5	*LSS*	6.42	69.3
T1151-5	*SLS*	6.34	69.2	--		--	--
Two short spacers of different length				1A15-5	*SSL*	6.46	69.8
T152-5	*SLM*	6.71	75.3	11A5-5	*SSL*	6.41	70.1
T251-5	*MLS*	6.62	75.7	1A51-5	*SLS*	6.50	75.1
		--		151A-5	*SLS*	6.59	74.8
T153-5	*SLM*	6.60	75.6	51A1-5	*LSS*	6.43	71.1
T351-5	*MLS*	6.86	73.5	511A-5	*LSS*	6.50	70.5
		--		--		--	--
T253-5	*SLM*	6.84	73.6	115_1A-5	*SSL*	6.70	69.2
T352-5	*MLS*	6.79	74.6	151_1A-5	*SLS*	6.59	73.5
		--		511_1A-5	*LSS*	6.49	70.5
T162-5	*SLM*	6.70	74.2	--		--	--
T261-5	*MLS*	6.70	76.1	115_2A-5	*SSL*	6.36	68.7
		--		151_2A-5	*SLS*	6.49	76.2
T163-5	*SLM*	6.71	73.3	511_2A-5	*LSS*	6.36	69.2
T361-5	*MLS*	6.72	72.6	--		--	--
		--		115_3A-5	*SSL*	6.42	67.2
T263-5	*SLM*	6.74	72.8	151_3A-5	*SLS*	6.57	73.2
T362-5	*MLS*	6.70	74.2	511_3A-5	*LSS*	6.35	68.5
				--		--	--
				115_4A-5	*SSL*	6.48	68.5
				151_4A-5	*SLS*	6.69	76.1
				511_4A-5	*LSS*	6.33	70.8
				--		--	--
				115_5A-5	*SSL*	6.49	70.7
				151_5A-5	*SLS*	6.75	75.1
				511_5A-5	*LSS*	6.43	70.3

[a] Detailed information for all sequences is provided in **Table S1**. pH transition midpoints were identified by pH-dependent CD (pH_T^{CD}) spectra at 288 nm. Thermal stabilities (T_m) were characterized by UV-melting curves at 295 nm. Standard deviations of pH_T and T_m of two independent measurements were no more than 0.2 and 1.0 ºC, respectively.

[b] Name: 'T' means that all spacers are composed of thymine base only; 'A' and 'AA' means that one or two thymine(s) in the spacer are replaced by or two adenines, respectively; three consecutive numbers refer to lengths of the three spacers in the 5' to 3' direction; '-5' refers to sequences with four C_5-tracts.

[c] Loop permutation is defined as the swap between two sequences of an intramolecular i-DNA keeping length and overall base composition constant. These sequences belong to a same group. Their spacer pattern can be either *SSL*, *SLS*, *LLS*, or *SLM, MLS*. Herein, *S, M* and *L* are short for relatively *short*, *middle*, *long* spacer length, respectively.

Table S4 Thermal stability measured by DSC-melting and annealing experiments (**Figure S19**).

Sequence	pH 5.0	pH 7.0			
	T_m	$T_{heating}$	$T_{cooling}$	T_m	$T_{hysteresis}$
T112-5	75.6	21.2	15.6	18.4	5.6
T121-5	77.6	23.9	17.2	20.6	6.7
T211-5	71.8	22.6	17.0	19.8	5.6
T225-5	71.8	28.3	5.4	16.9	22.9
T252-5	76.1	32.1	7.4	19.8	24.6
T522-5	72.0	30.0	5.4	17.7	24.7
T112-6	81.0	33.0	16.1	24.6	16.9
T121-6	82.4	32.4	14.5	23.5	17.9
T211-6	80.2	31.4	14.5	23.0	16.9
T225-6	76.9	35.8	7.2	21.5	28.6
T252-6	79.8	38.8	8.2	23.5	30.6
T522-6	77.5	36.4	7.5	22.0	28.9

Figure S1 Thermal difference spectra (TDS).

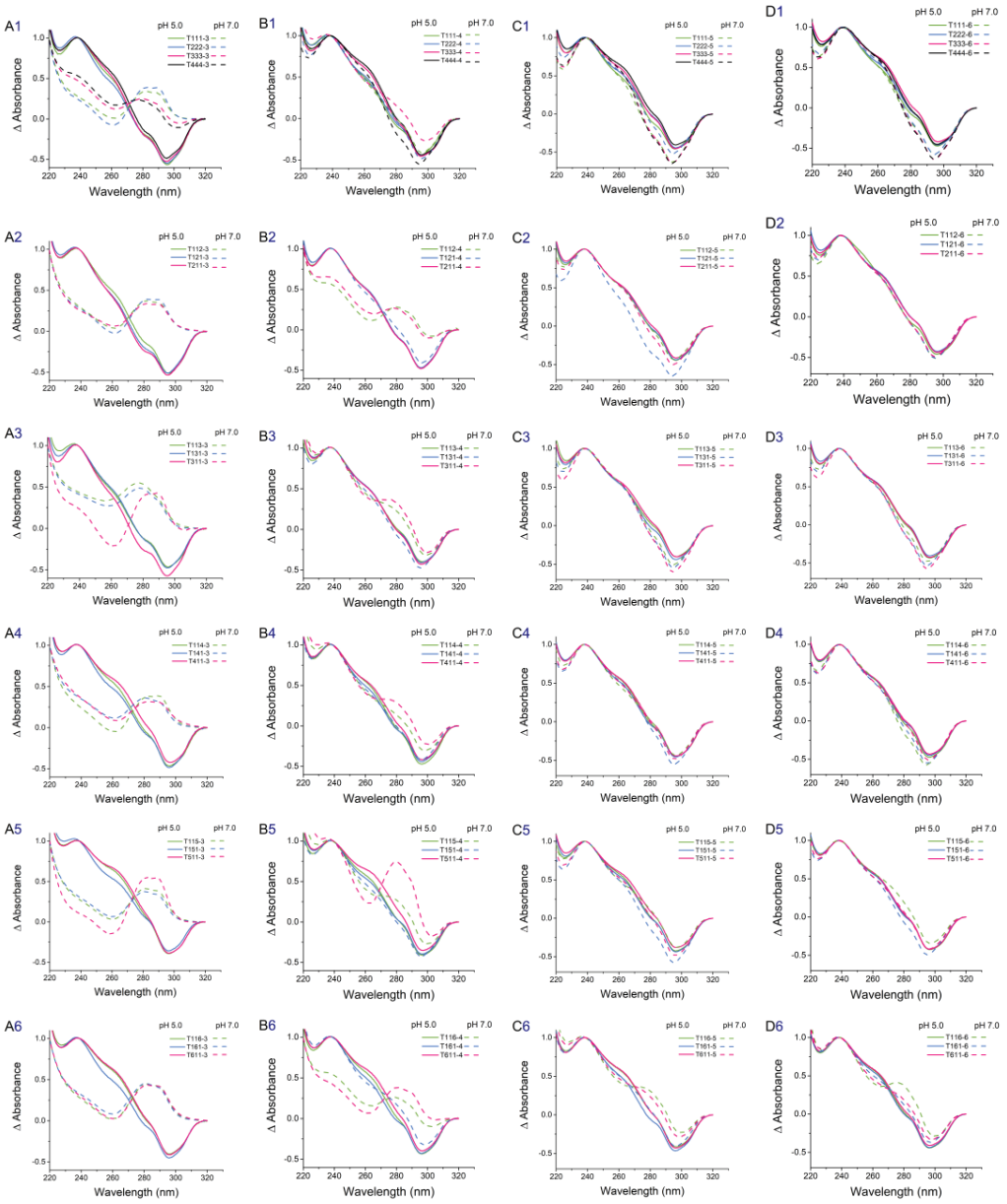


Figure S1 Thermal difference spectra (TDS). (*Continued_01*)

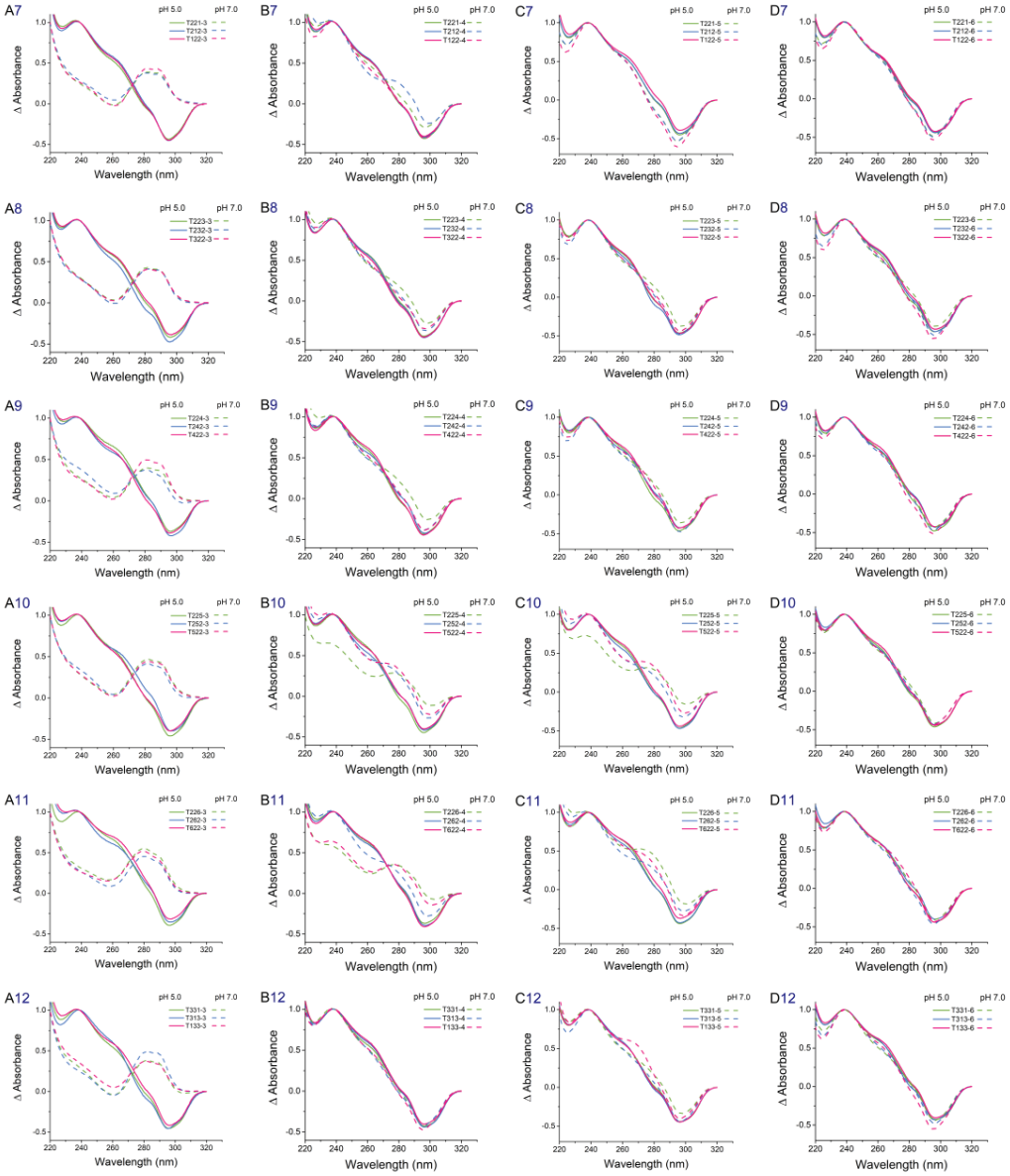


Figure S1 Thermal difference spectra (TDS). (*Continued_02*)

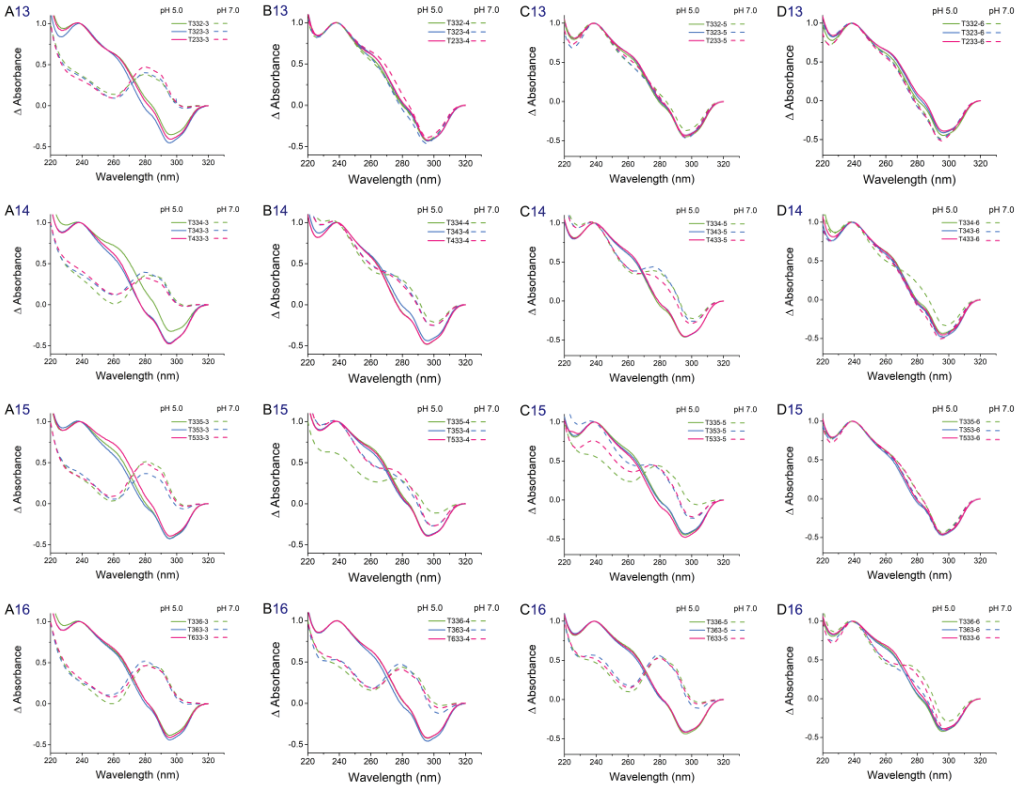


Figure S1 Normalized thermal difference spectra (TDS) of (**A**) i-DNAs with C_3 tract (first column, A1~A16), (**B**) i-DNAs with C_4 tract (second column, B1~B16), (**C**) i-DNAs with C_5 tract (third column, C1~C16), and (**D**) i-DNAs with C_6 tract (fourth column, D1~D16), resulting from the subtraction of UV absorption spectra at 5 °C from the spectra at 95 °C (for pH 5.0) or 65 °C (for pH 7.0). 5 μM DNA in pH 5.0 (solid line) and 7.0 (dashed line).

Figure S2 ^{1}H 1D NMR spectra.

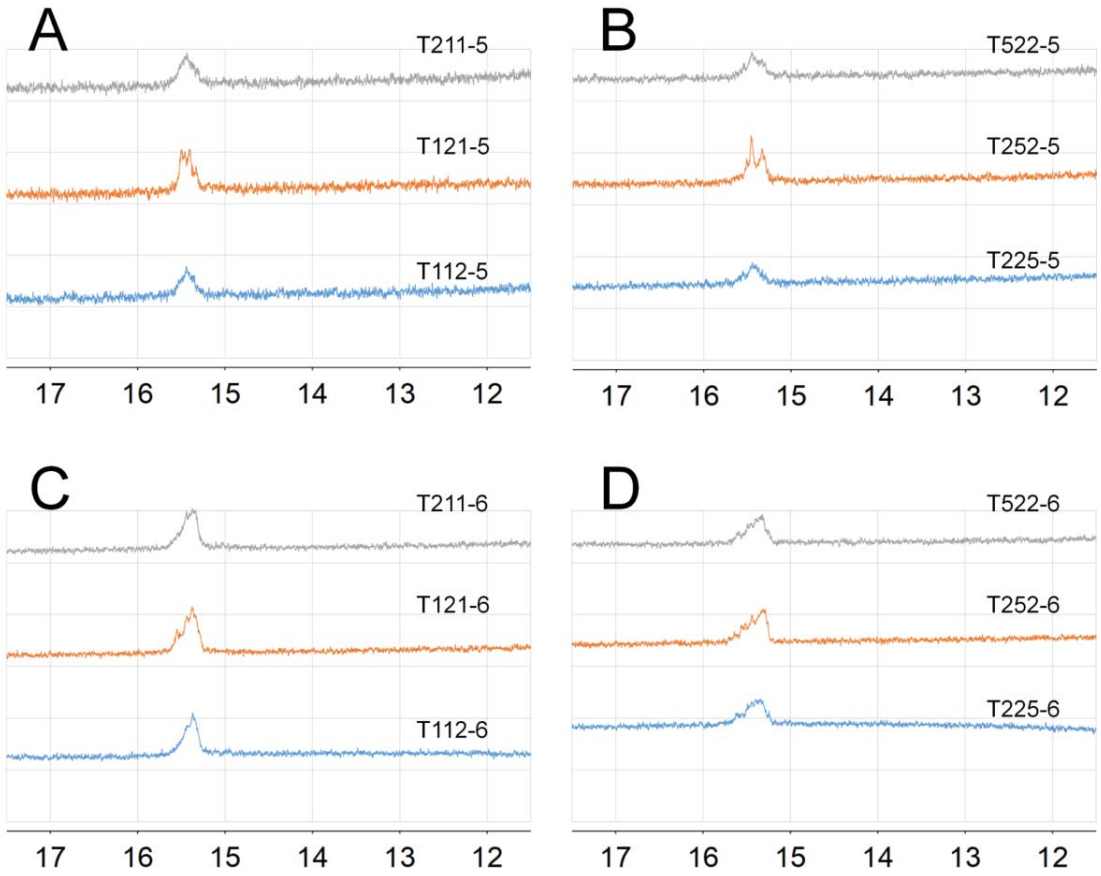


Figure S2 1D ^{1}H NMR spectra of 12 selected sequences at pH 7.0 in the region assigned to imino protons of protonated cytosines at 20 ^{o}C. (**A**) T112-5, T121-5 and T211-5 sequences; (**B**) T225-5, T252-5 and T522-5 sequences; (**C**) T112-6, T121-6 and T211-6 sequences; (**D**) T225-6, T252-6 and T522-6 sequences.

Figures S3-S4 Non-denaturing PAGEs.

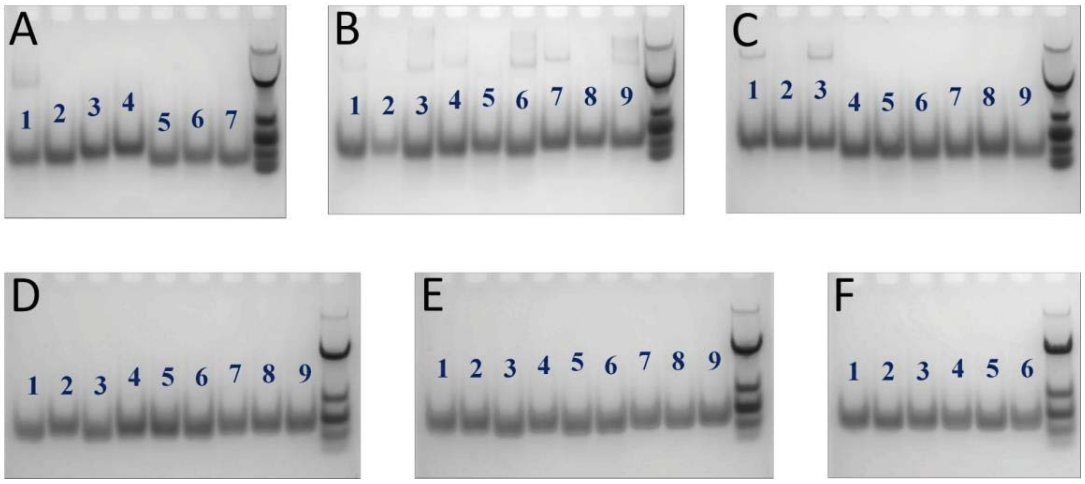


Figure S3 Non-denaturing PAGE of i-DNA with C_5-tract at pH 5.0. Samples concentration is 25 μM. (**A**-**F**) 49 sequences with C_5-tract in Table 1. Last lane in each gel is dTn (n=60, 30, 21, 10) ladder.

(**A**) Lane A1: T111-5; Lane A2: T222-5; Lane A3: T333-5; Lane A4: T444-5; Lane A5: T112-5; Lane A6: T121-5; Lane A7: T211-5.

(**B**) Lane B1: T113-5; Lane B2: T131-5; Lane B3: T311-5; Lane B4: T114-5; Lane B5: T141-5; Lane B6: T411-5; Lane B7: T115-5; Lane B8: T151-5; Lane B9: T511-5.

(**C**) Lane C1: T116-5; Lane C2: T161-5; Lane C3: T611-5; Lane C4: T221-5; Lane C5: T212-5; Lane C6: T122-5; Lane C7: T223-5; Lane C8: T232-5; Lane C9: T322-5.

(**D**) Lane D1: T224-5; Lane D2: T242-5; Lane D3: T422-5; Lane D4: T225-5; Lane D5: T252-5; Lane D6: T522-5; Lane D7: T226-5; Lane D8: T262-5; Lane D9: T622-5.

(**E**) Lane E1: T331-5; Lane E2: T313-5; Lane E3: T133-5; Lane E4: T332-5; Lane E5: T323-5; Lane E6: T233-5; Lane E7: T334-5; Lane E8: T343-5; Lane E9: T433-5.

(**F**) Lane F1: T335-5; Lane F2: T353-5; Lane F3: T533-5; Lane F4: T336-5; Lane F5: T363-5; Lane F6: T633-5.

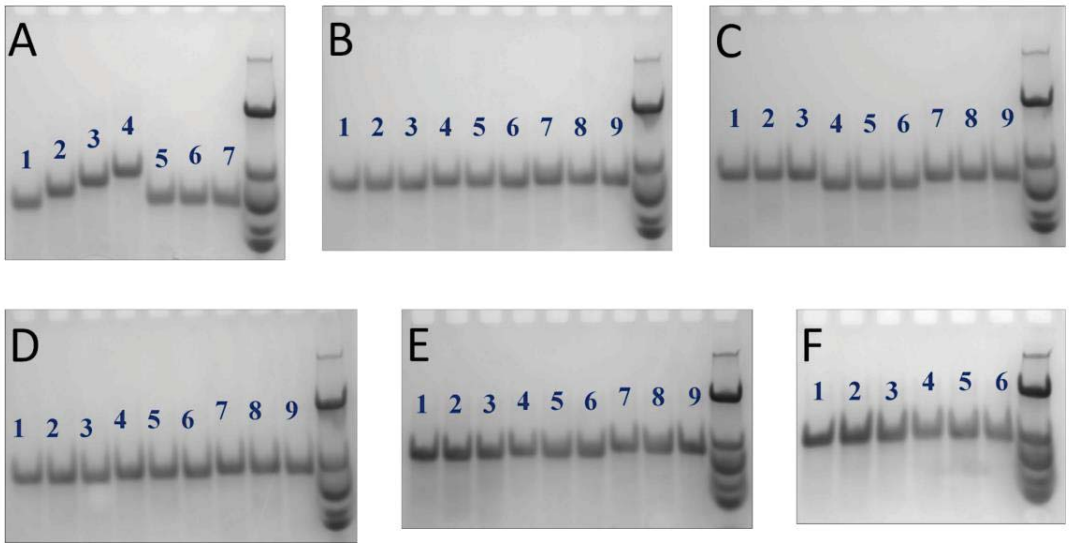


Figure S4 Non-denaturing PAGE of i-DNA with C_5-tract at pH 7.0. Samples concentration is 25 μM. (**A-F**) 49 sequences with C_5-tract in **Table 1**. Last lane in each gel is dTn (n=60, 30, 21, 10) ladder.

(**A**) Lane A1: T111-5; Lane A2: T222-5; Lane A3: T333-5; Lane A4: T444-5; Lane A5: T112-5; Lane A6: T121-5; Lane A7: T211-5.

(**B**) Lane B1: T113-5; Lane B2: T131-5; Lane B3: T311-5; Lane B4: T114-5; Lane B5: T141-5; Lane B6: T411-5; Lane B7: T115-5; Lane B8: T151-5; Lane B9: T511-5.

(**C**) Lane C1: T116-5; Lane C2: T161-5; Lane C3: T611-5; Lane C4: T221-5; Lane C5: T212-5; Lane C6: T122-5; Lane C7: T223-5; Lane C8: T232-5; Lane C9: T322-5.

(**D**) Lane D1: T224-5; Lane D2: T242-5; Lane D3: T422-5; Lane D4: T225-5; Lane D5: T252-5; Lane D6: T522-5; Lane D7: T226-5; Lane D8: T262-5; Lane D9: T622-5.

(**E**) Lane E1: T331-5; Lane E2: T313-5; Lane E3: T133-5; Lane E4: T332-5; Lane E5: T323-5; Lane E6: T233-5; Lane E7: T334-5; Lane E8: T343-5; Lane E9: T433-5.

(**F**) Lane F1: T335-5; Lane F2: T353-5; Lane F3: T533-5; Lane F4: T336-5; Lane F5: T363-5; Lane F6: T633-5.

Figure S5 pH-dependent normalized ellipticities at 288 nm for sequences with C_3 tract.

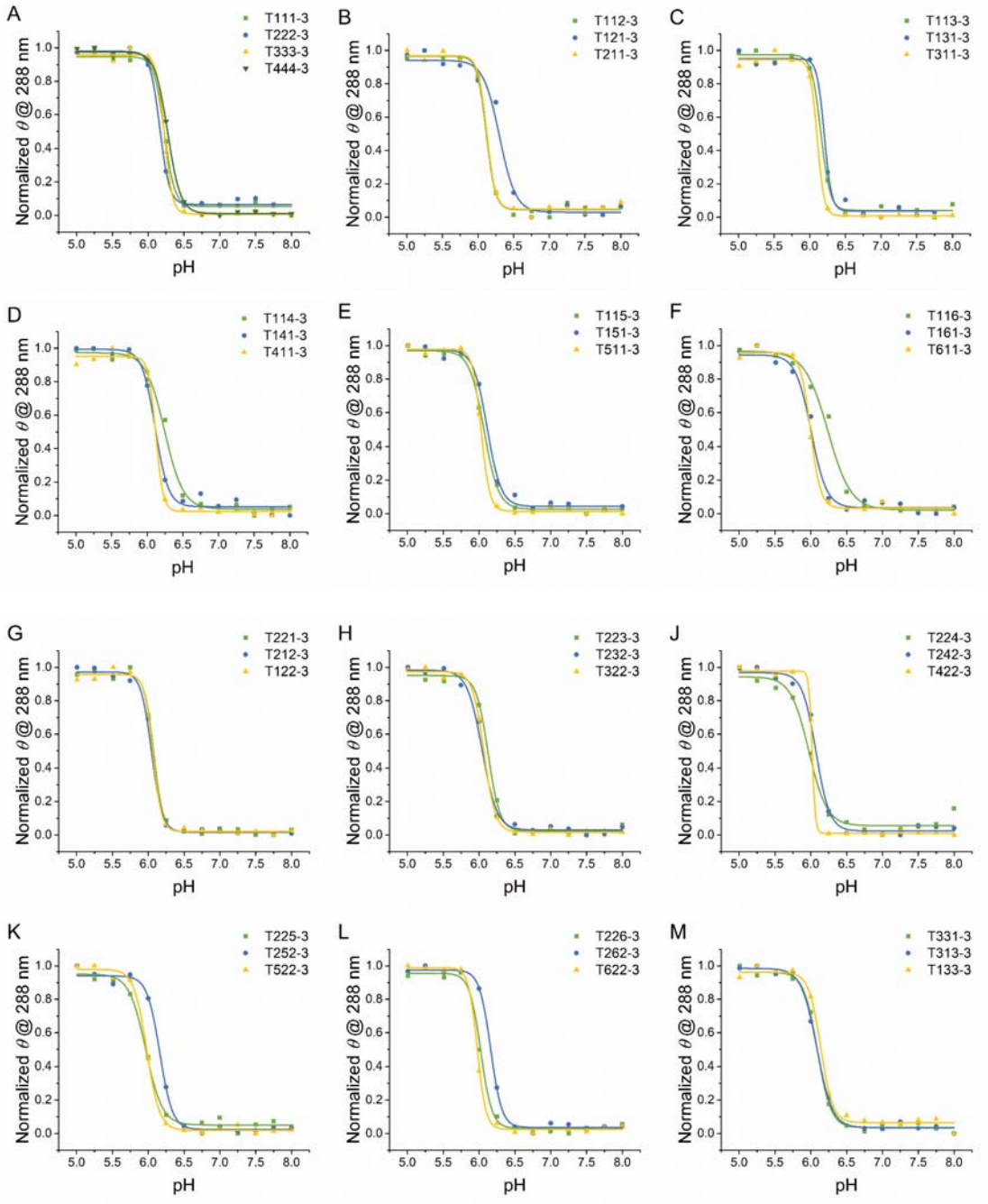


Figure S5 pH-dependent normalized ellipticities at 288 nm of sequences with C_3 tract. (*Continued_01*)

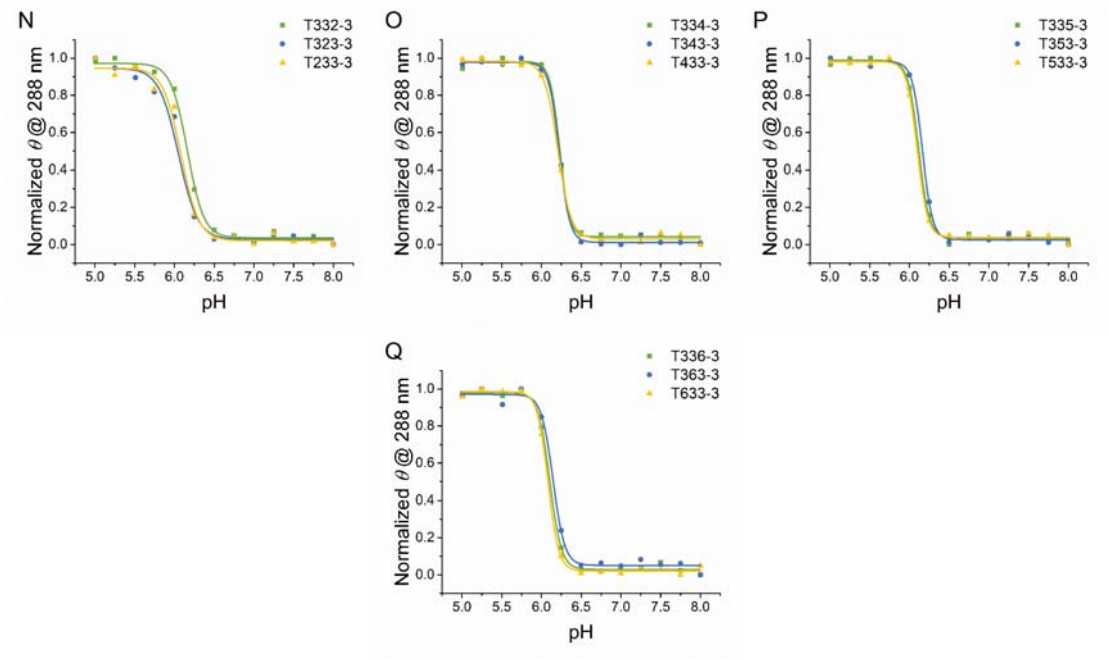


Figure S5 pH-transition by CD spectra at 288 nm of i-DNAs with C_3 tract. (**A**) *T111-3* group, (**B**) *T112-3* group, (**C**) *T113-3* group, (**D**) *T114-3* group, (**E**) *T115-3* group, (**F**) *T116-3* group, (**G**) *T221-3* group, (**H**) *T223-3* group, (**J**) *T224-3* group, (**K**) *T225-3* group, (**L**) *T226-3* group, (**M**) *T331-3* group, (**N**) *T332-3* group, (**O**) *T334-3* group, (**P**) *T335-3* group, and (**Q**) *T336-3* group.

Figure S6 pH-dependent normalized ellipticities at 288 nm of sequences with C_4 tract.

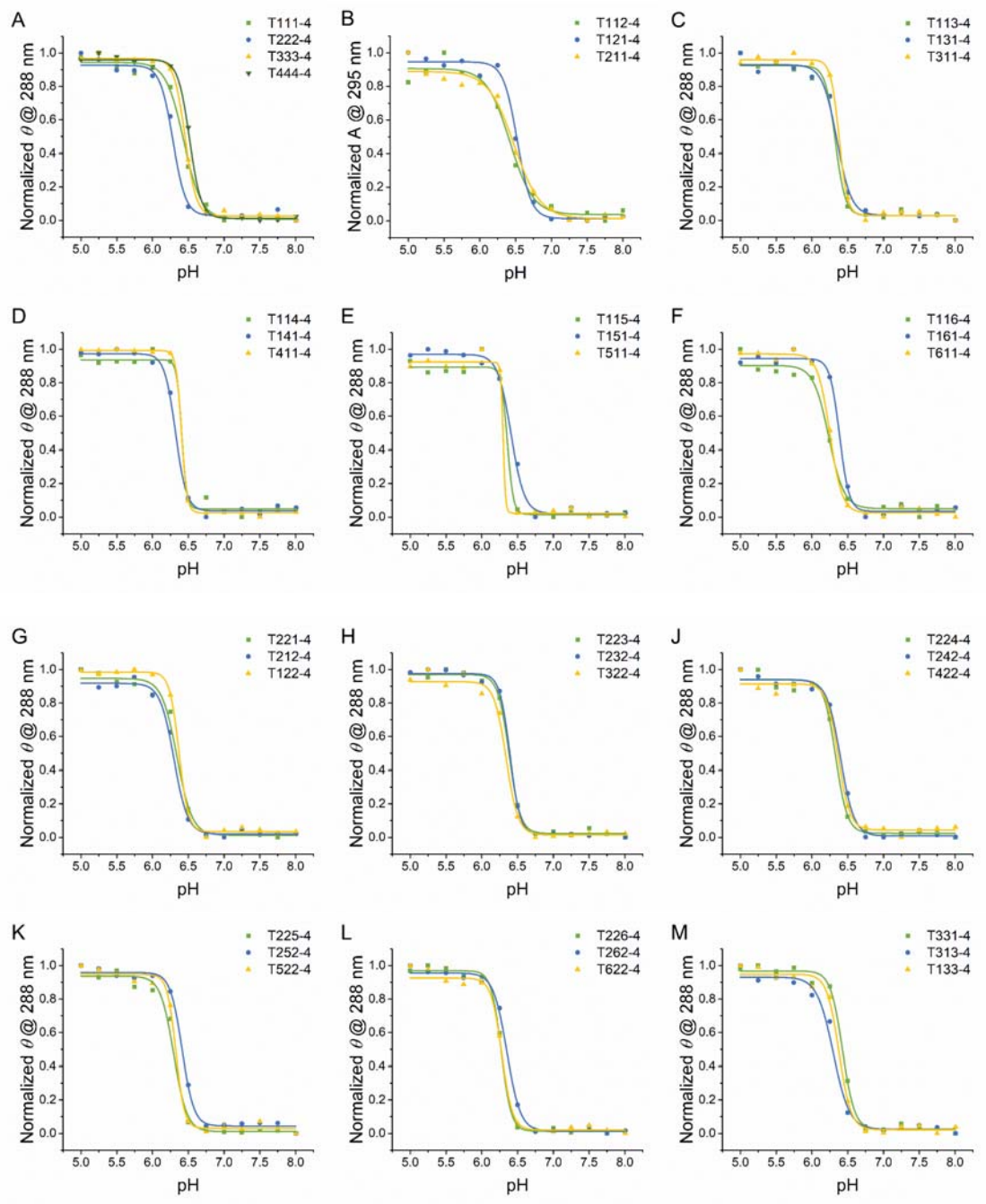


Figure S6 pH-dependent CD spectra of sequences with C_4 tract. (*Continued_01*)

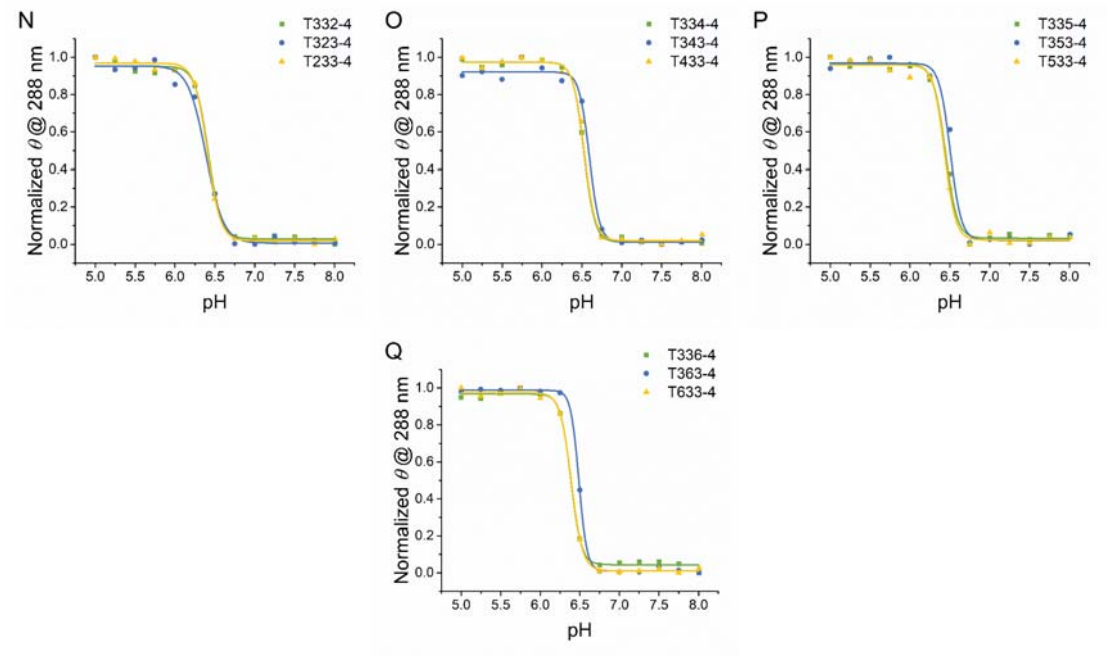


Figure S6 pH-transition by CD spectra at 288 nm of i-DNAs with C_4 tract. (**A**) *T111-4* group, (**B**) *T112-4* group, (**C**) *T113-4* group, (**D**) *T114-4* group, (**E**) *T115-4* group, (**F**) *T116-4* group, (**G**) *T221-4* group, (**H**) *T223-4* group, (**J**) *T224-4* group, (**K**) *T225-4* group, (**L**) *T226-4* group, (**M**) *T331-4* group, (**N**) *T332-4* group, (**O**) *T334-4* group, (**P**) *T335-4* group, and (**Q**) *T336-4* group.

Figure S7 pH-dependent normalized ellipticities at 288 nm of sequences with C_5 tract.

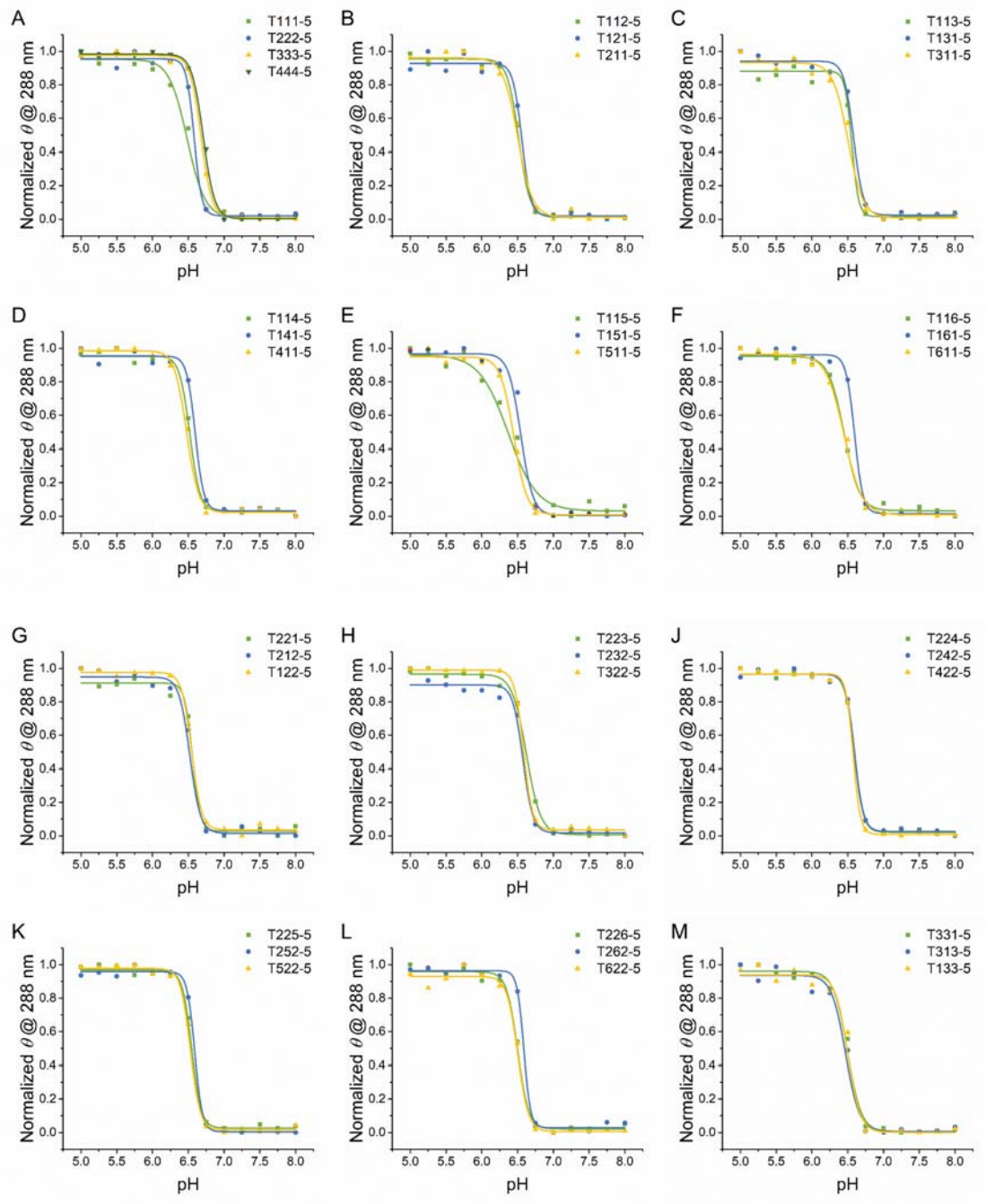


Figure S7 pH-dependent normalized ellipticities at 288 nm of sequences with C_5 tract. (*Continued_01*)

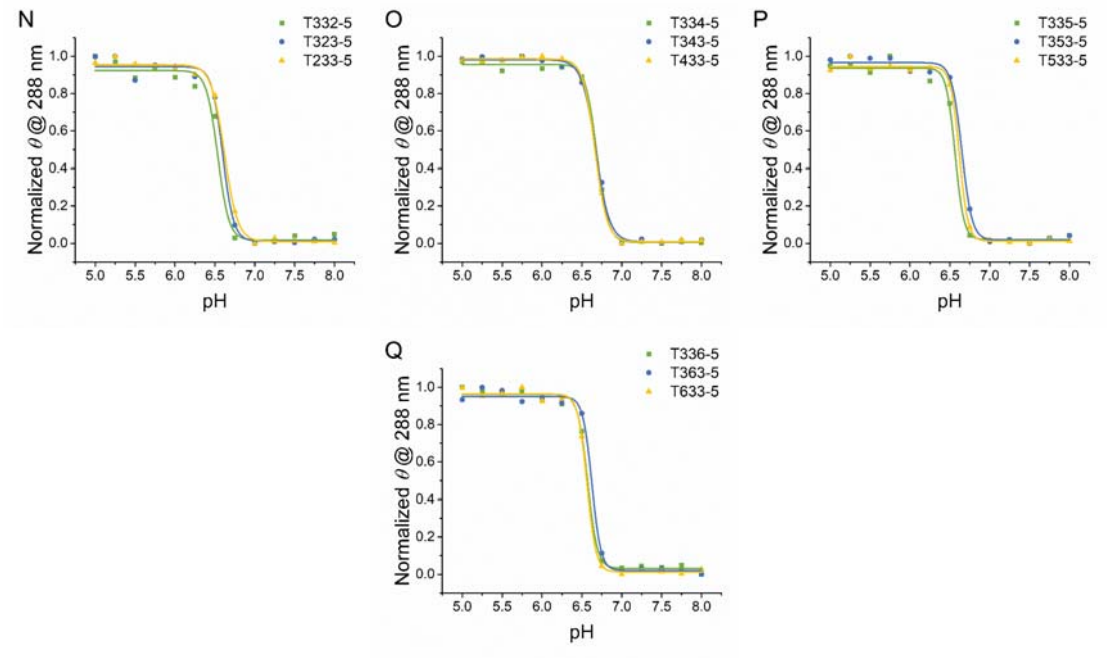


Figure S7 pH-transition by CD spectra at 288 nm of i-DNAs with C_5 tract. (**A**) *T111-5* group, (**B**) *T112-5* group, (**C**) *T113-5* group, (**D**) *T114-5* group, (**E**) *T115-5* group, (**F**) *T116-5* group, (**G**) *T221-5* group, (**H**) *T223-5* group, (**J**) *T224-5* group, (**K**) *T225-5* group, (**L**) *T226-5* group, (**M**) *T331-5* group, (**N**) *T332-5* group, (**O**) *T334-5* group, (**P**) *T335-5* group, and (**Q**) *T336-5* group.

Figure S8 pH-dependent normalized ellipticities at 288 nm of sequences with C_6 tract.

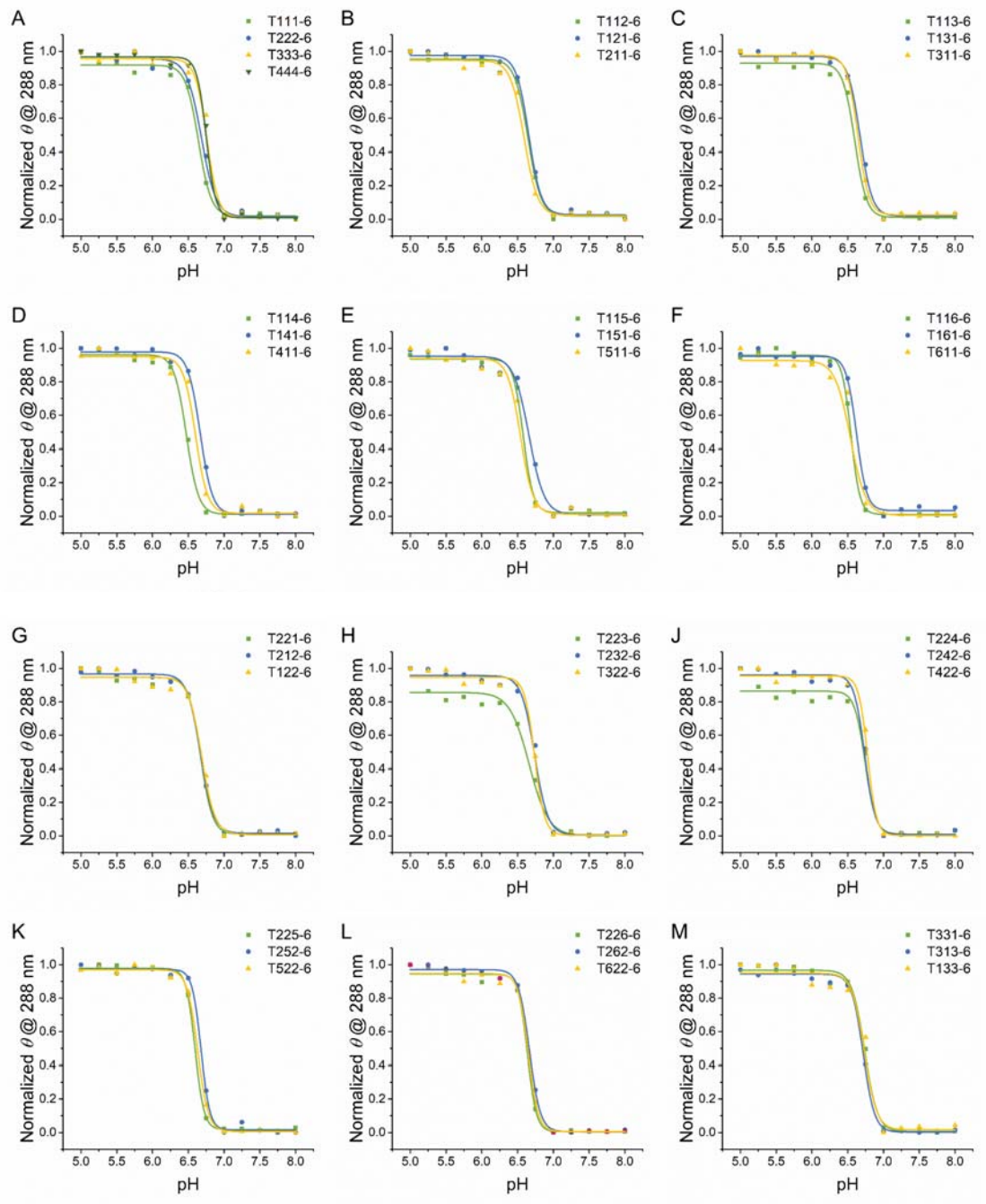


Figure S8 pH-dependent normalized ellipticities at 288 nm of sequences with C_6 tract. (*Continued_01*)

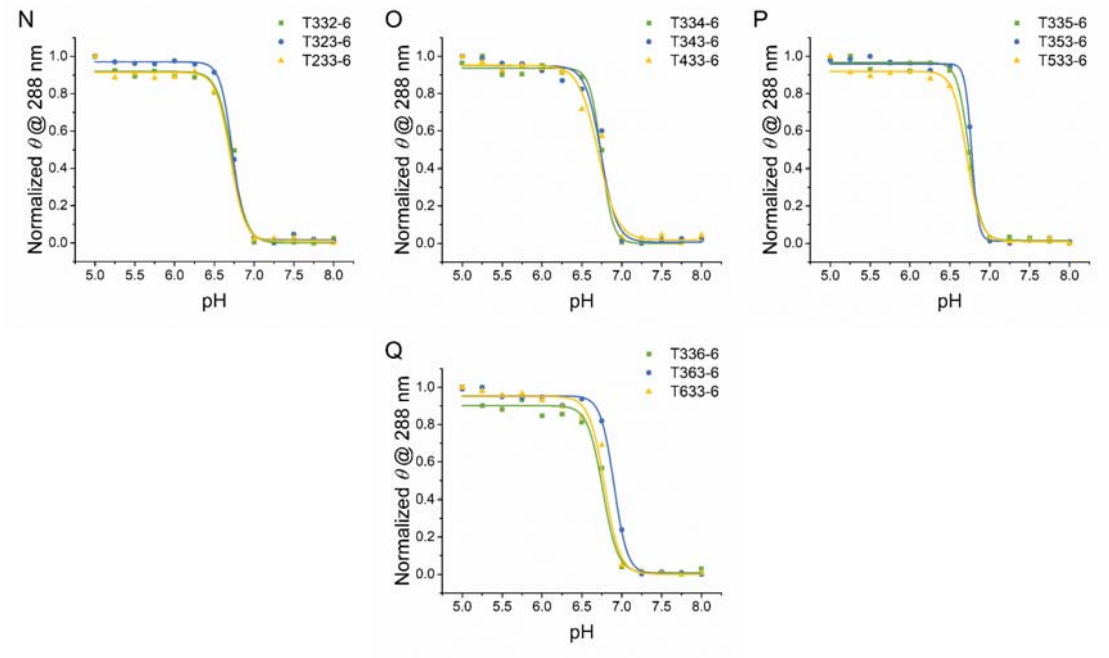


Figure S8 pH-transition by CD spectra at 288 nm of i-DNAs with C_6 tract. (**A**) *T111-6* group, (**B**) *T112-6* group, (**C**) *T113-6* group, (**D**) *T114-6* group, (**E**) *T115-6* group, (**F**) *T116-6* group, (**G**) *T221-6* group, (**H**) *T223-6* group, (**J**) *T224-6* group, (**K**) *T225-6* group, (**L**) *T226-6* group, (**M**) *T331-6* group, (**N**) *T332-6* group, (**O**) *T334-6* group, (**P**) *T335-6* group, and (**Q**) *T336-6* group.

Figure S9 pH transition midpoint (pH_T) of i-DNA determined by CD.

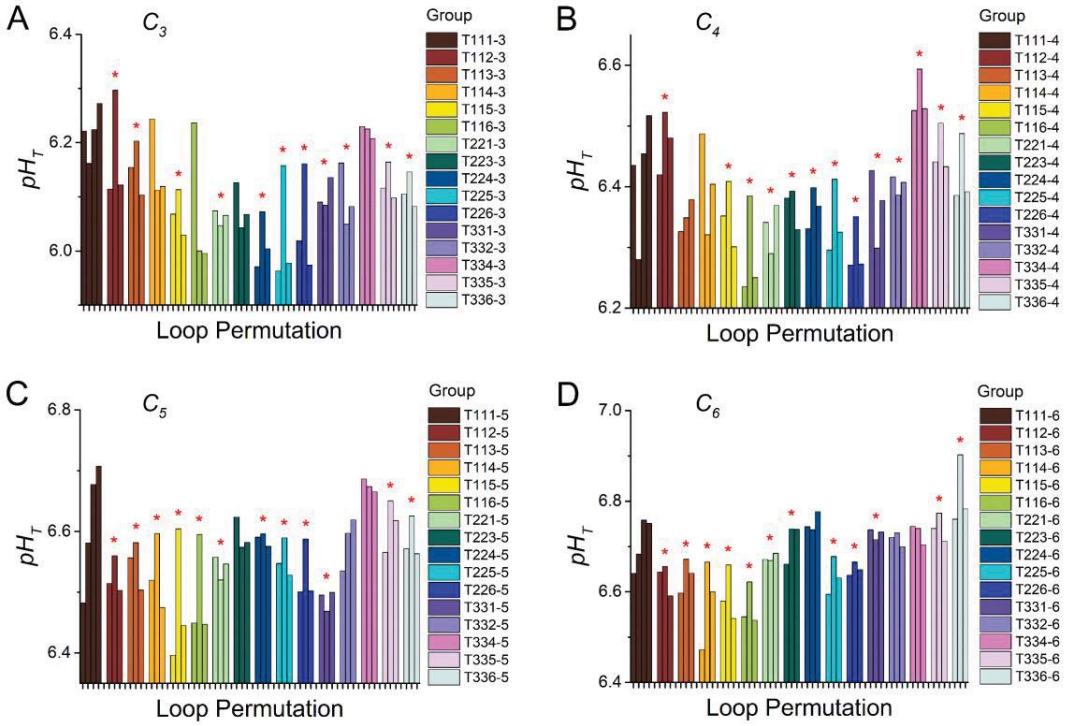


Figure S9 pH transition midpoint (pH_T) of i-DNA determined by CD: (**A**) I-DNAs with four C_3 tracts; (**B**) I-DNAs with four C_4 tracts; (**C**) I-DNAs with four C_5 tracts; (**D**) I-DNAs with four C_6 tracts. The experiments were carried out in 20 mM Britton-Robinson buffer with 140 mM KCl and 20 mM NaCl at room temperature (25 °C). pH titrations are shown in the supplementary information; pH varied from 5.00 to 8.00 with 0.25 pH unit intervals. All oligonucleotide strand concentrations were 5 μM. Symbol * at top of the bar stands for that the group obeys the rule that sequence with longer central loop exhibits higher pH_T.

Figure S10 pH-dependent normalized absorbances at 295 nm for sequences with C_3 tract.

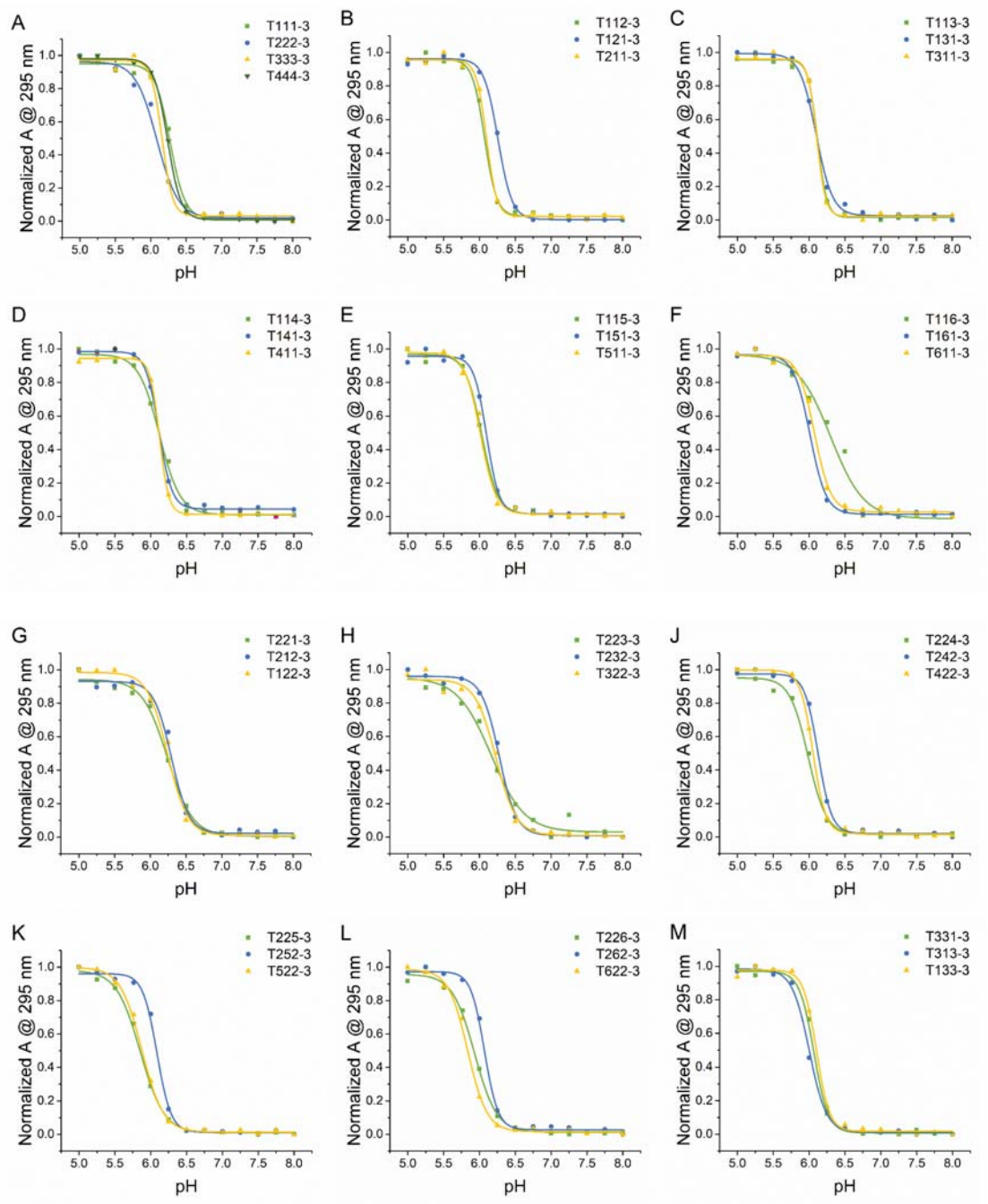


Figure S10 pH-dependent normalized absorbances at 295 nm for sequences with C_3 tract. (*Continued_01*)

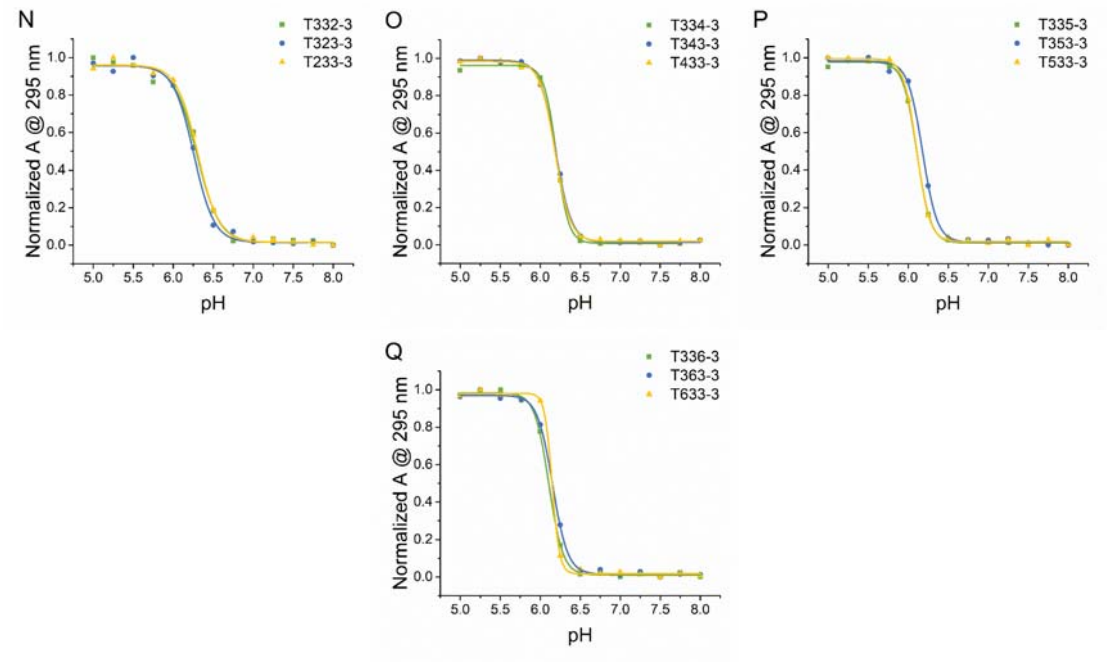


Figure S10 pH-transition by UV absorption spectra at 295 nm of i-DNAs with C_3 tract. (**A**) *T111-3* group, (**B**) *T112-3* group, (**C**) *T113-3* group, (**D**) *T114-3* group, (**E**) *T115-3* group, (**F**) *T116-3* group, (**G**) *T221-3* group, (**H**) *T223-3* group, (**J**) *T224-3* group, (**K**) *T225-3* group, (**L**) *T226-3* group, (**M**) *T331-3* group, (**N**) *T332-3* group, (**O**) *T334-3* group, (**P**) *T335-3* group, and (**Q**) *T336-3* group.

Figure S11 pH-dependent normalized absorbances at 295 nm for sequences with C_4 tract.

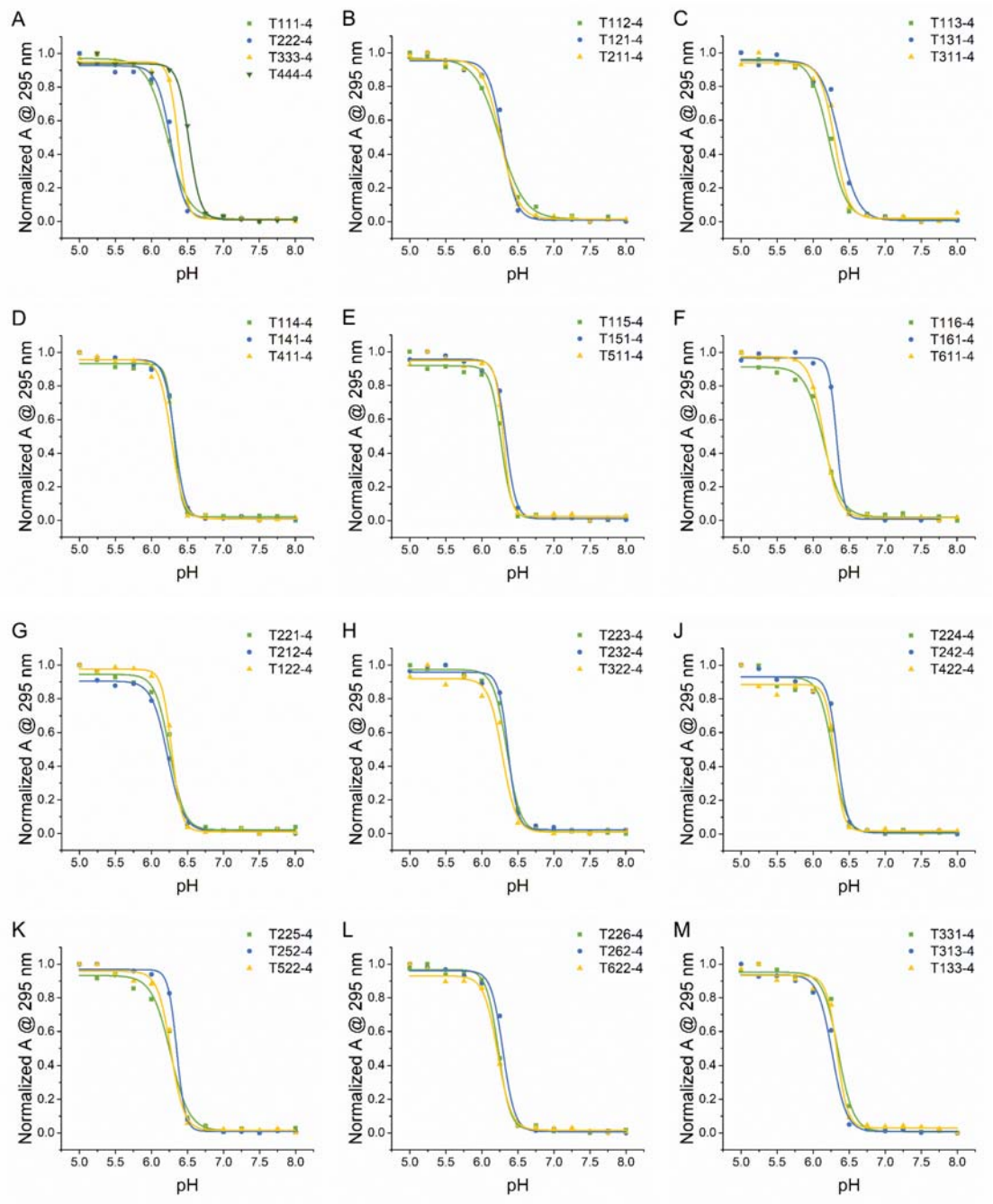


Figure S11 pH-dependent normalized absorbances at 295 nm for sequences with C_4 tract. (*Continued_01*)

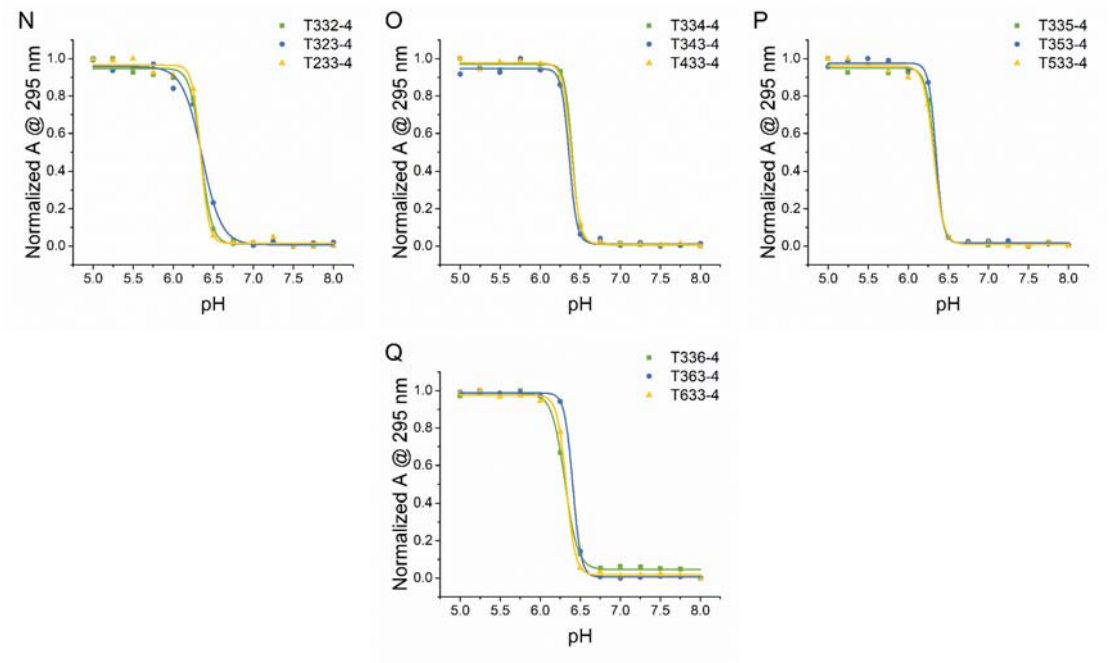


Figure S11 pH-transition by UV absorption spectra at 295 nm of i-DNAs with C_4 tract. (**A**) *T111-4* group, (**B**) *T112-4* group, (**C**) *T113-4* group, (**D**) *T114-4* group, (**E**) *T115-4* group, (**F**) *T116-4* group, (**G**) *T221-4* group, (**H**) *T223-4* group, (**J**) *T224-4* group, (**K**) *T225-4* group, (**L**) *T226-4* group, (**M**) *T331-4* group, (**N**) *T332-4* group, (**O**) *T334-4* group, (**P**) *T335-4* group, and (**Q**) *T336-4* group.

Figure S12 pH-dependent normalized absorbances at 295 nm for sequences with C_5 tract.

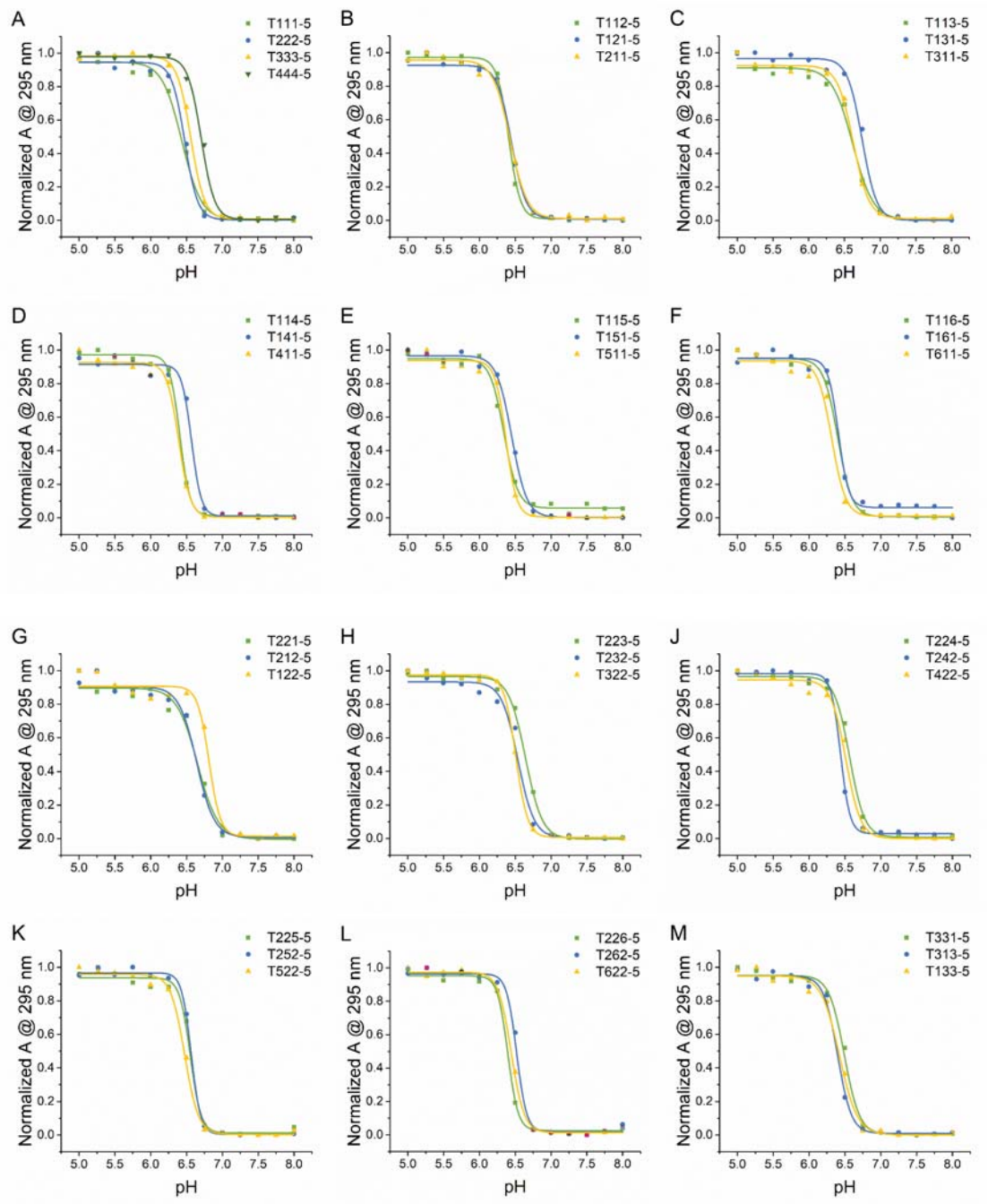


Figure S12 pH-dependent normalized absorbances at 295 nm for sequences with C_5 tract. (*Continued_01*)

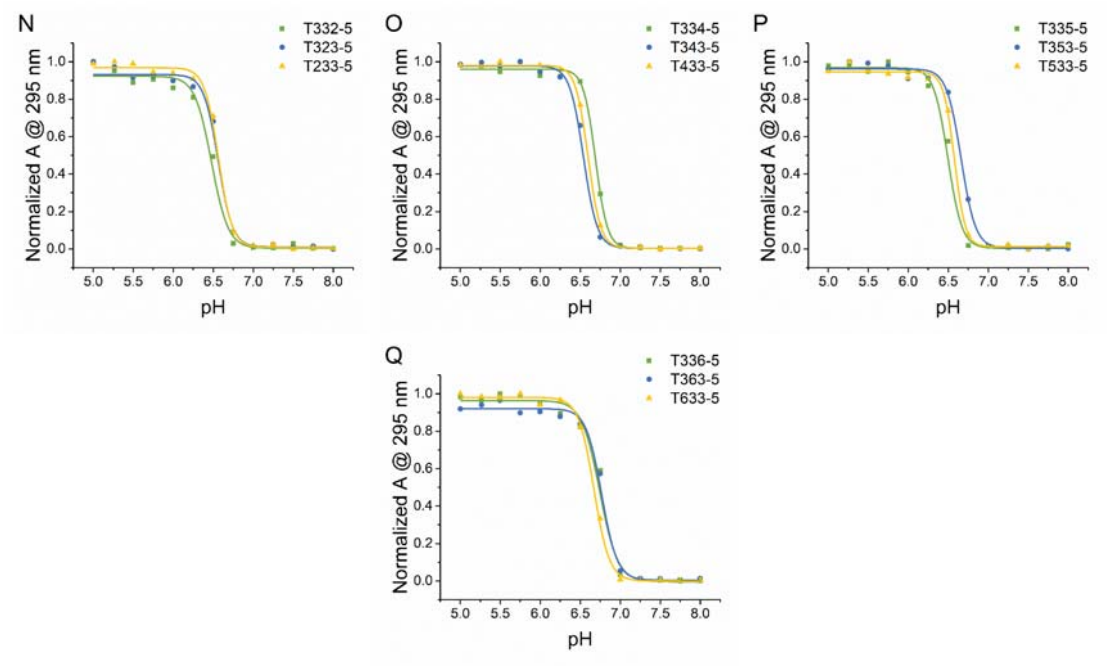


Figure S12 pH-transition by UV absorption spectra at 295 nm of i-DNAs with C_5 tract. (**A**) *T111-5* group, (**B**) *T112-5* group, (C) *T113-5* group, (**D**) *T114-5* group, (**E**) *T115-5* group, (**F**) *T116-5* group, (**G**) *T221-5* group, (**H**) *T223-5* group, (**J**) *T224-5* group, (**K**) *T225-5* group, (**L**) *T226-5* group, (**M**) *T331-5* group, (**N**) *T332-5* group, (**O**) *T334-5* group, (**P**) *T335-5* group, and (**Q**) *T336-5* group.

Figure S13 pH-dependent normalized absorbances at 295 nm for sequences with C_6 tract.

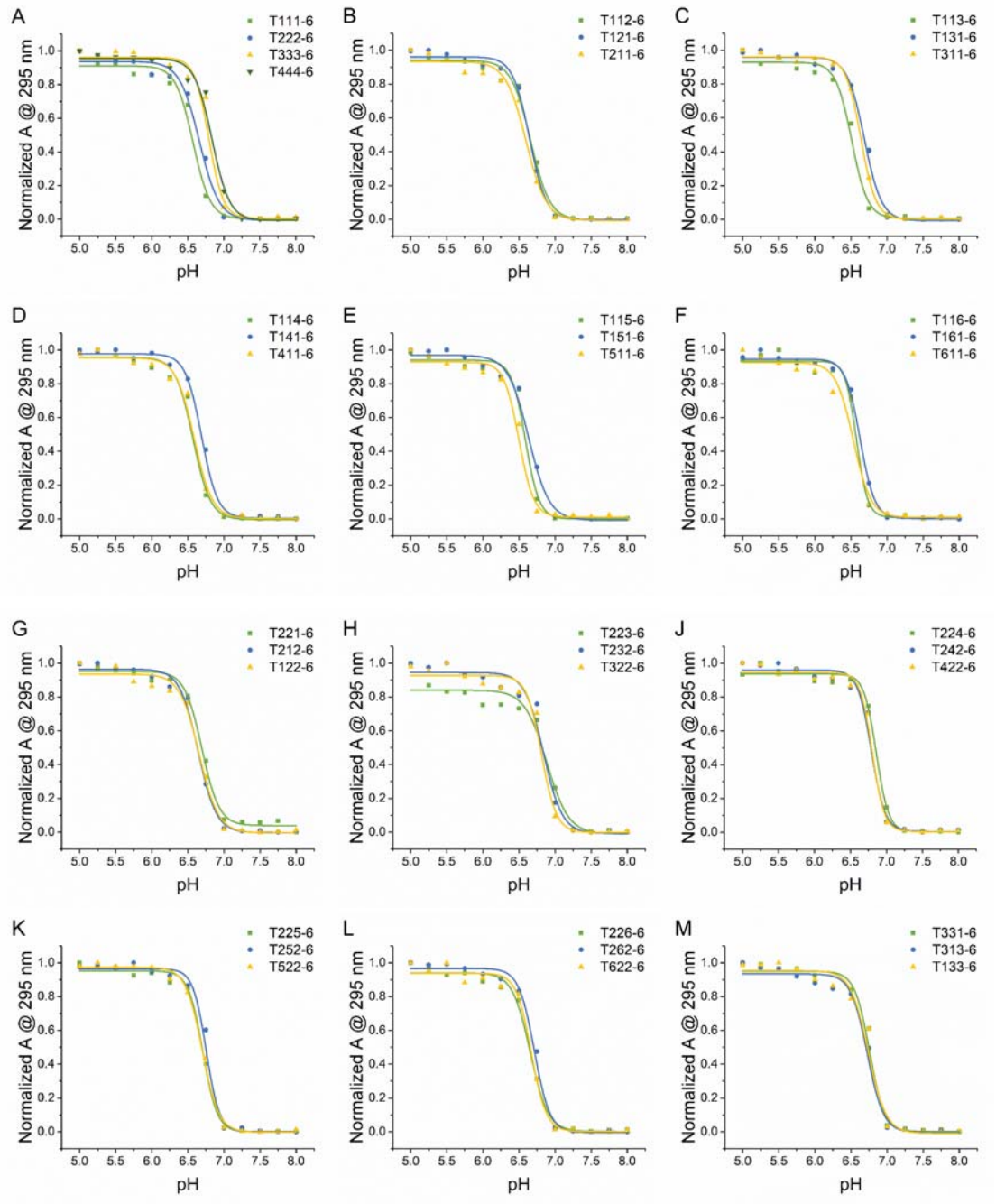


Figure S13 pH-dependent normalized absorbances at 295 nm for sequences with C_6 tract. (*Continued_01*)

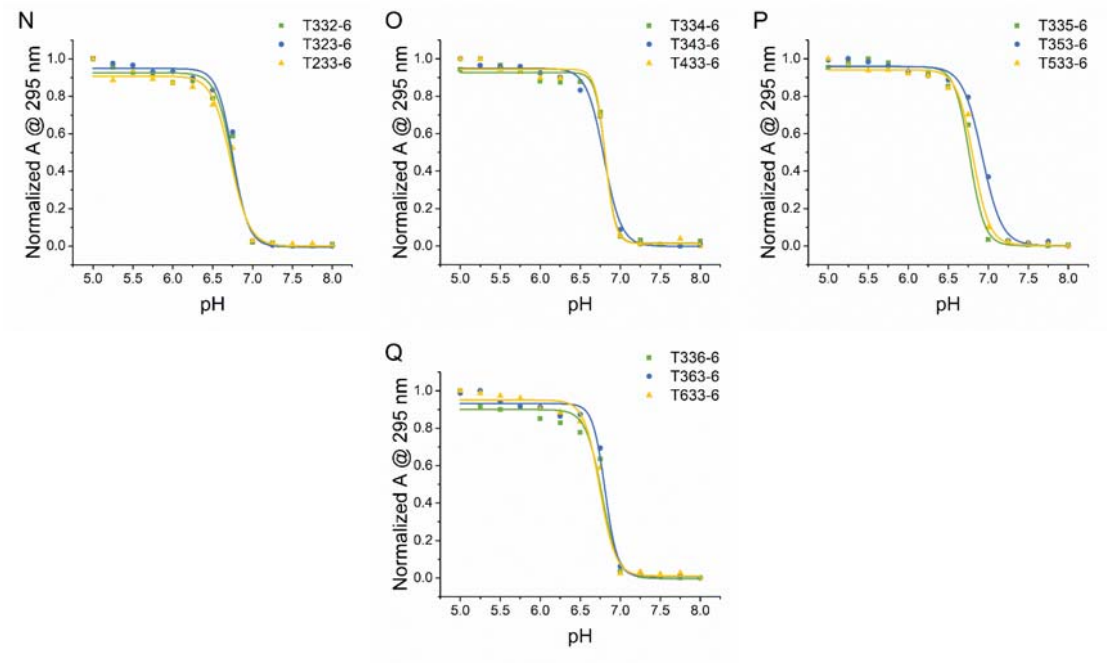


Figure S13 pH-transition by UV absorption spectra at 295 nm of i-DNAs with C_6 tract. (**A**) *T111-6* group, (**B**) *T112-6* group, (**C**) *T113-6* group, (**D**) *T114-6* group, (**E**) *T115-6* group, (**F**) *T116-6* group, (**G**) *T221-6* group, (**H**) *T223-6* group, (**J**) *T224-6* group, (**K**) *T225-6* group, (**L**) *T226-6* group, (**M**) *T331-6* group, (**N**) *T332-6* group, (**O**) *T334-6* group, (**P**) *T335-6* group, and (**Q**) *T336-6* group.

Figure S14 pH of mid-transition (pH_T) determined by UV absorbance data.

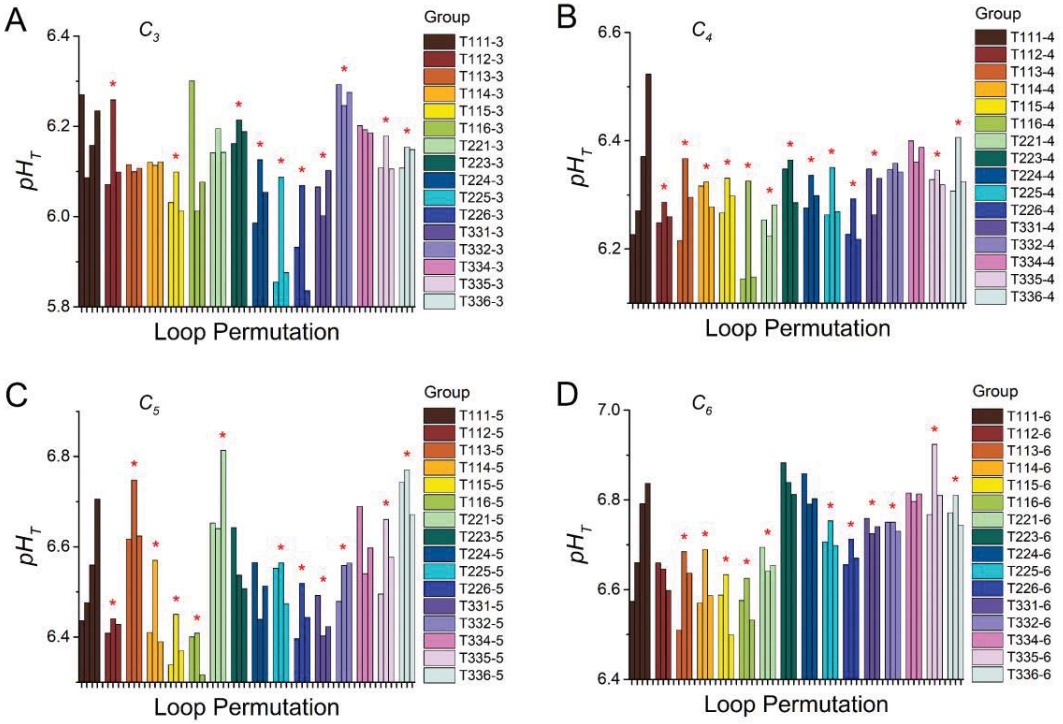


Figure S14 pH transition midpoint (pH_T) of i-DNAs identified by UV absorption spectra: (**A**) i-DNAs with four C_3 tracts; (**B**) i-DNAs with four C_4 tracts; (**C**) i-DNAs with four C_5 tracts; (**D**) i-DNAs with four C_6 tracts. The experiments were carried out in 20 mM Britton-Robinson buffer with 140 mM KCl and 20 mM NaCl at room temperature (25 ^{o}C). The pH varied from 5.00 to 8.00 at the interval of 0.25 pH unit and strand concentrations of oligonucleotides were 5 μM. Symbol * at top of the bar indicates that this group obeys the rule that sequence with longer central loop exhibits higher pH_T.

Figure S15 Comparison of pH_T obtained by pH-dependent CD and UV absorbance spectra.

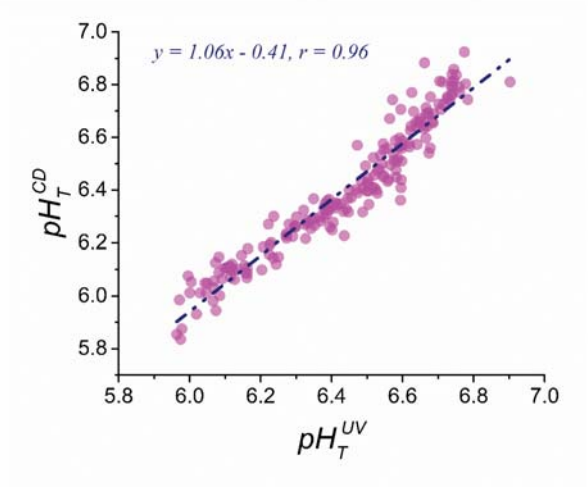


Figure S15 pH_T obtained by CD spectra as a function of that by UV absorbance spectra.

Figure S16 UV-melting curves at pH 5.0.

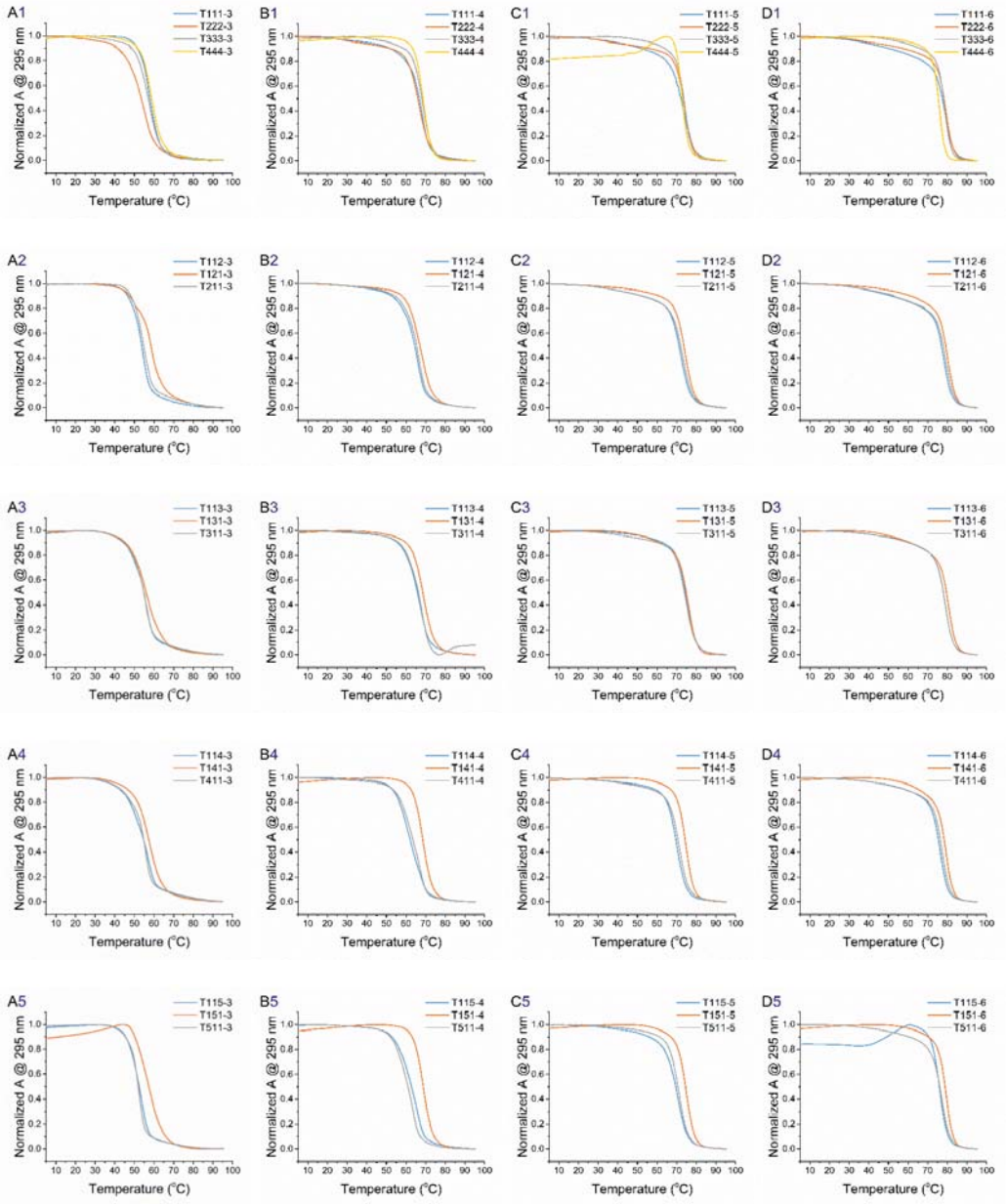


Figure S16 UV-melting curves at pH 5.0. (*Continued_01*)

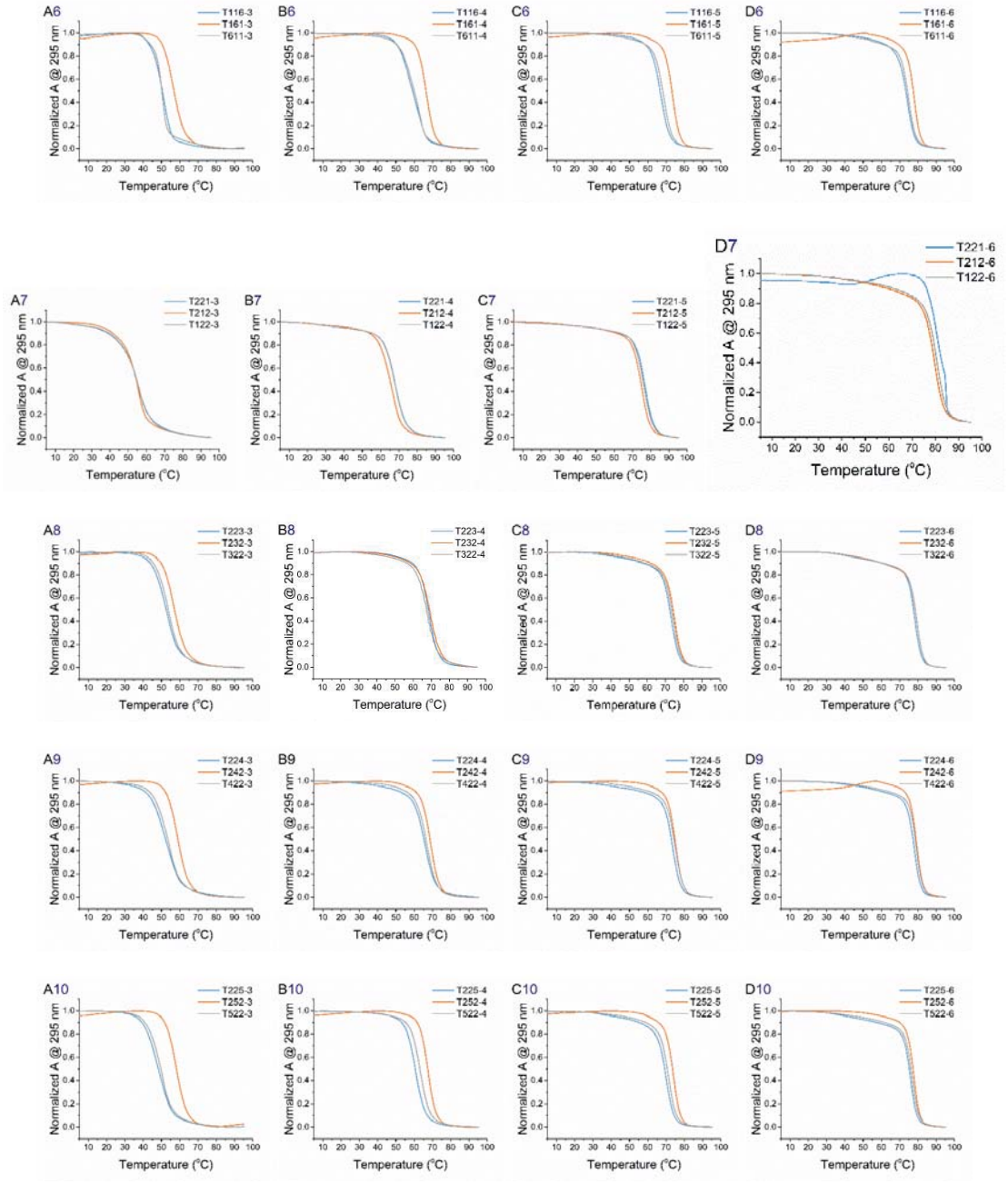


Figure S16 UV-melting curves at pH 5.0. (*Continued_02*)

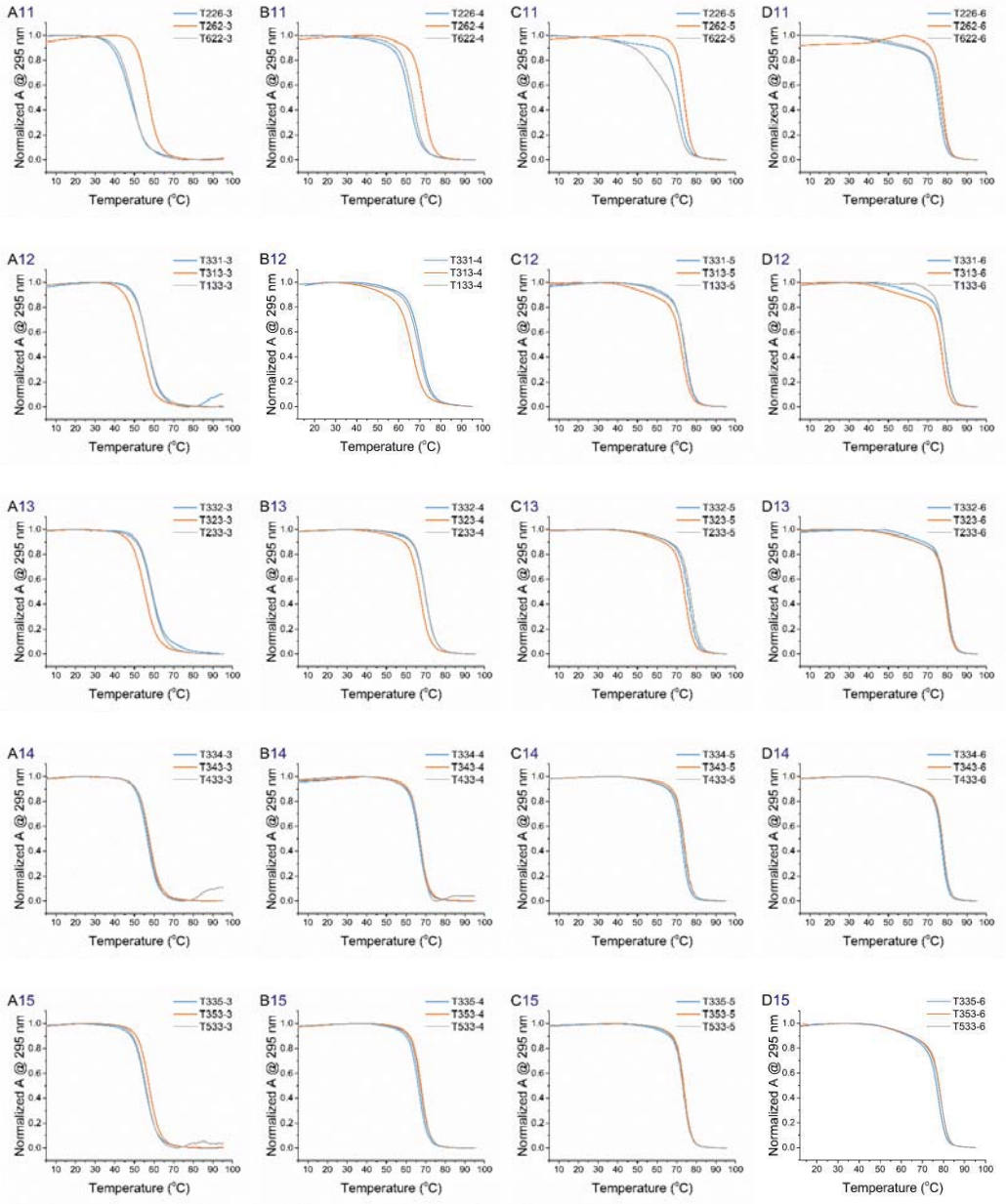


Figure S16 UV-melting curves at pH 5.0. (*Continued_03*)

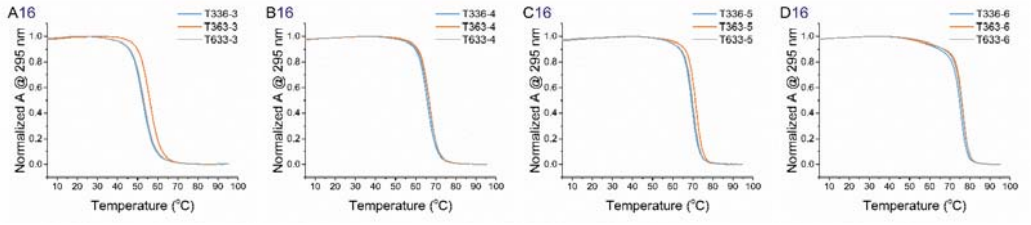


Figure S16 UV-melting curves at pH 5.0 of (**A**) i-DNAs with C_3 tract (first column, A1~A16), (**B**) i-DNAs with C_4 tract (second column, B1~B16), (**C**) i-DNAs with C_5 tract (third column, C1~C16), and (**D**) i-DNAs with C_6 tract (fourth column, D1~D16). Temperatures varied from 5 to 95 °C.

Figure S17 UV-melting and annealing curves at pH 7.0.

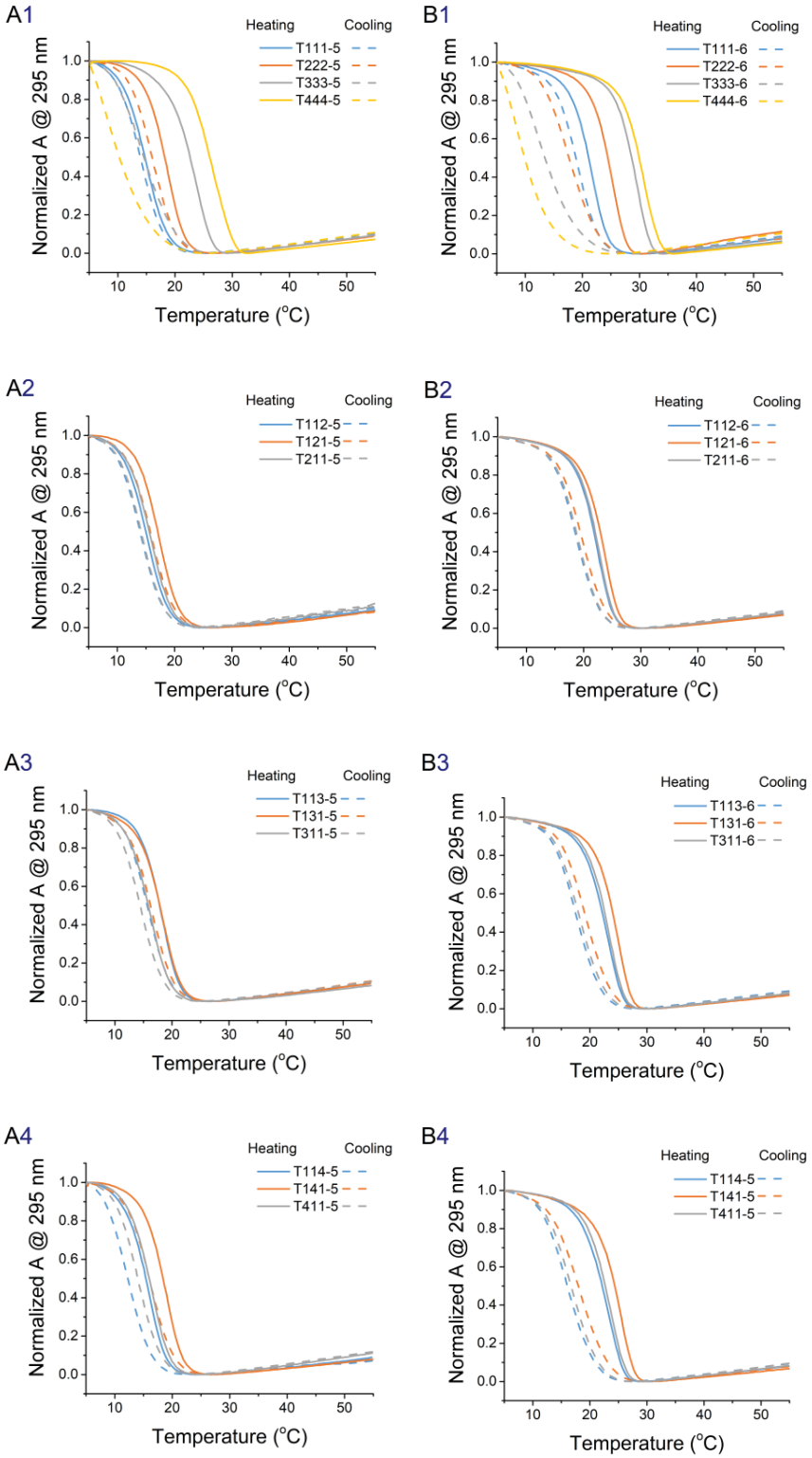


Figure S17 UV-melting and annealing curves at pH 7.0. (*Continued_01*)

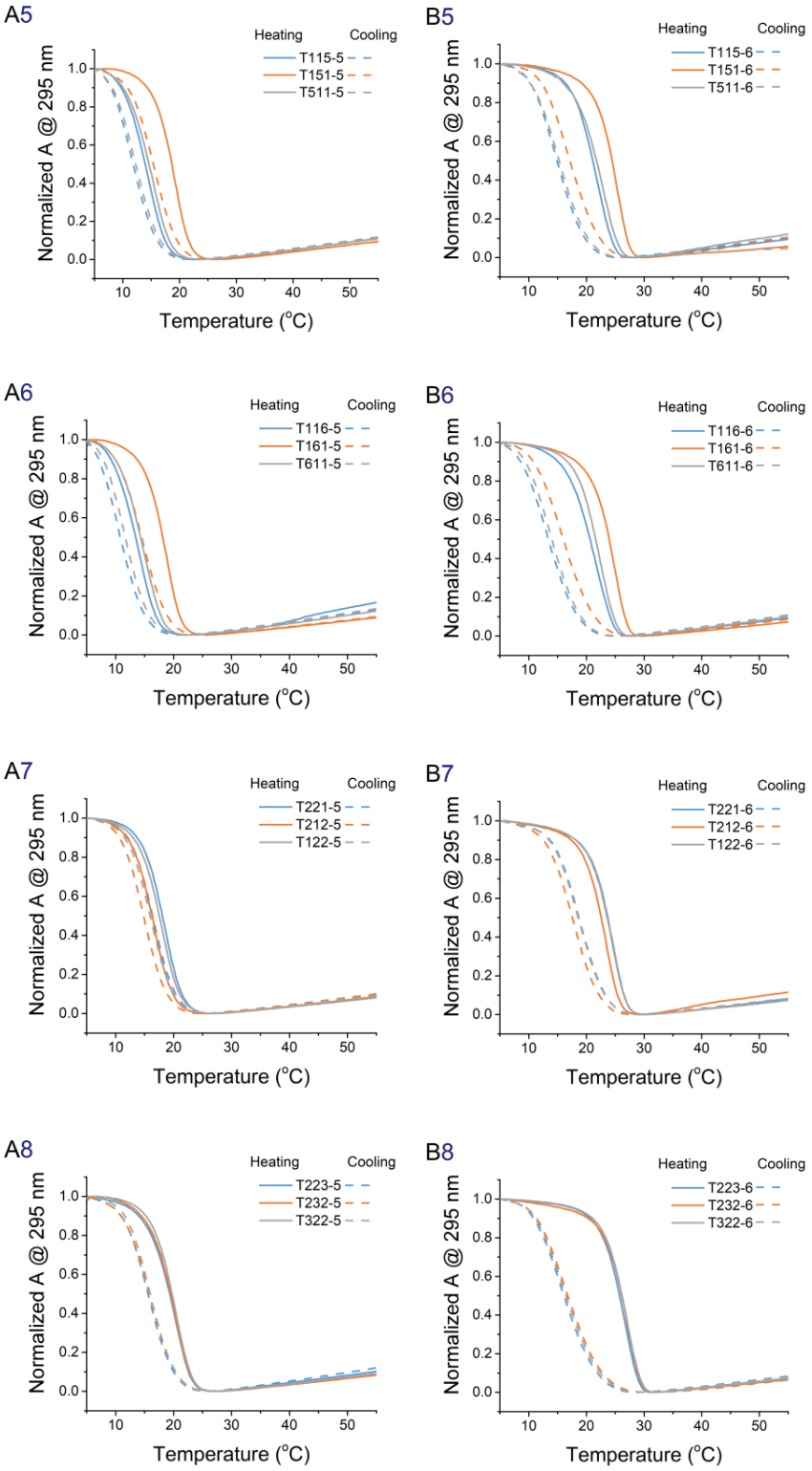


Figure S17 UV-melting and annealing curves at pH 7.0. (*Continued_02*)

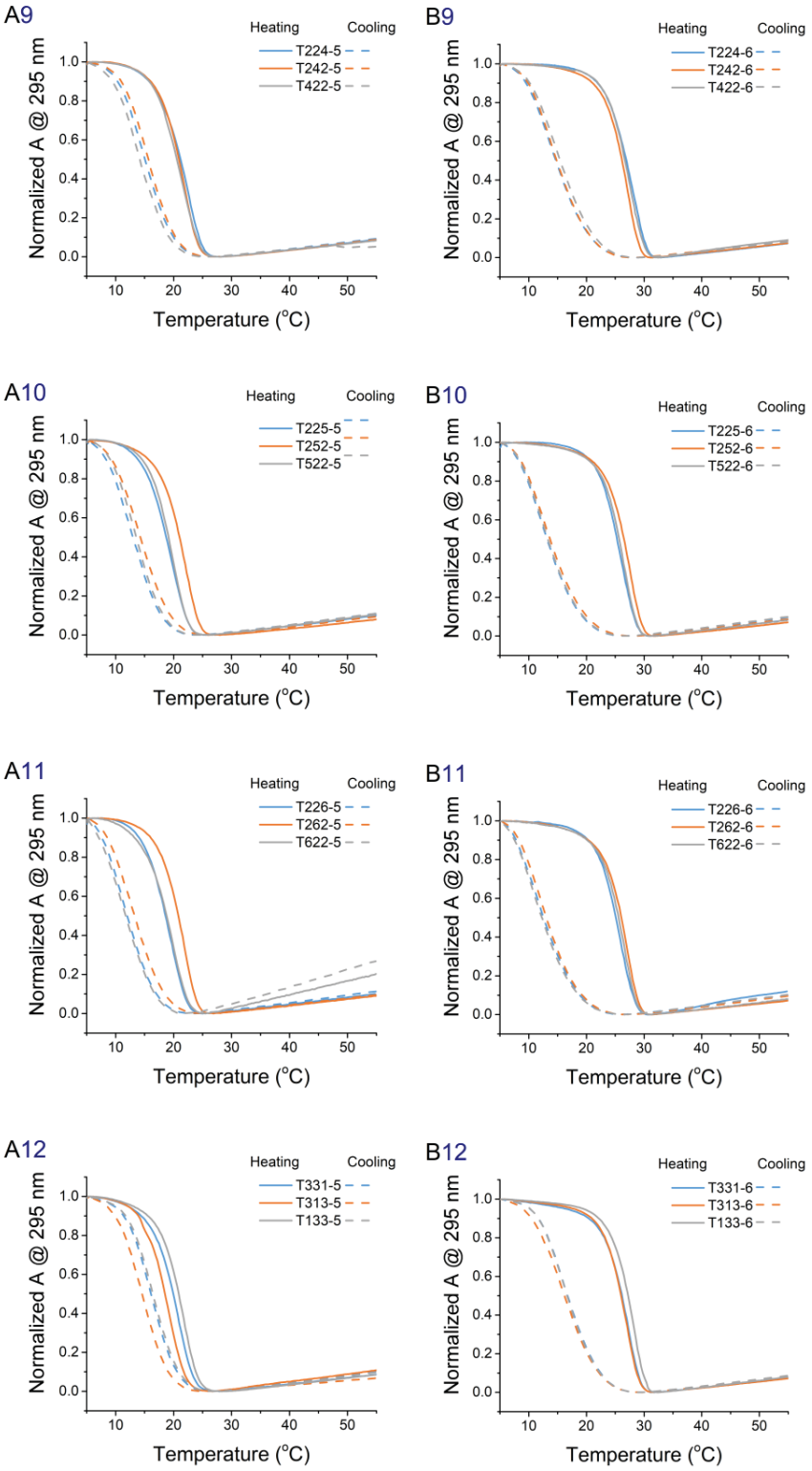


Figure S17 UV-melting and annealing curves at pH 7.0. (*Continued_03*)

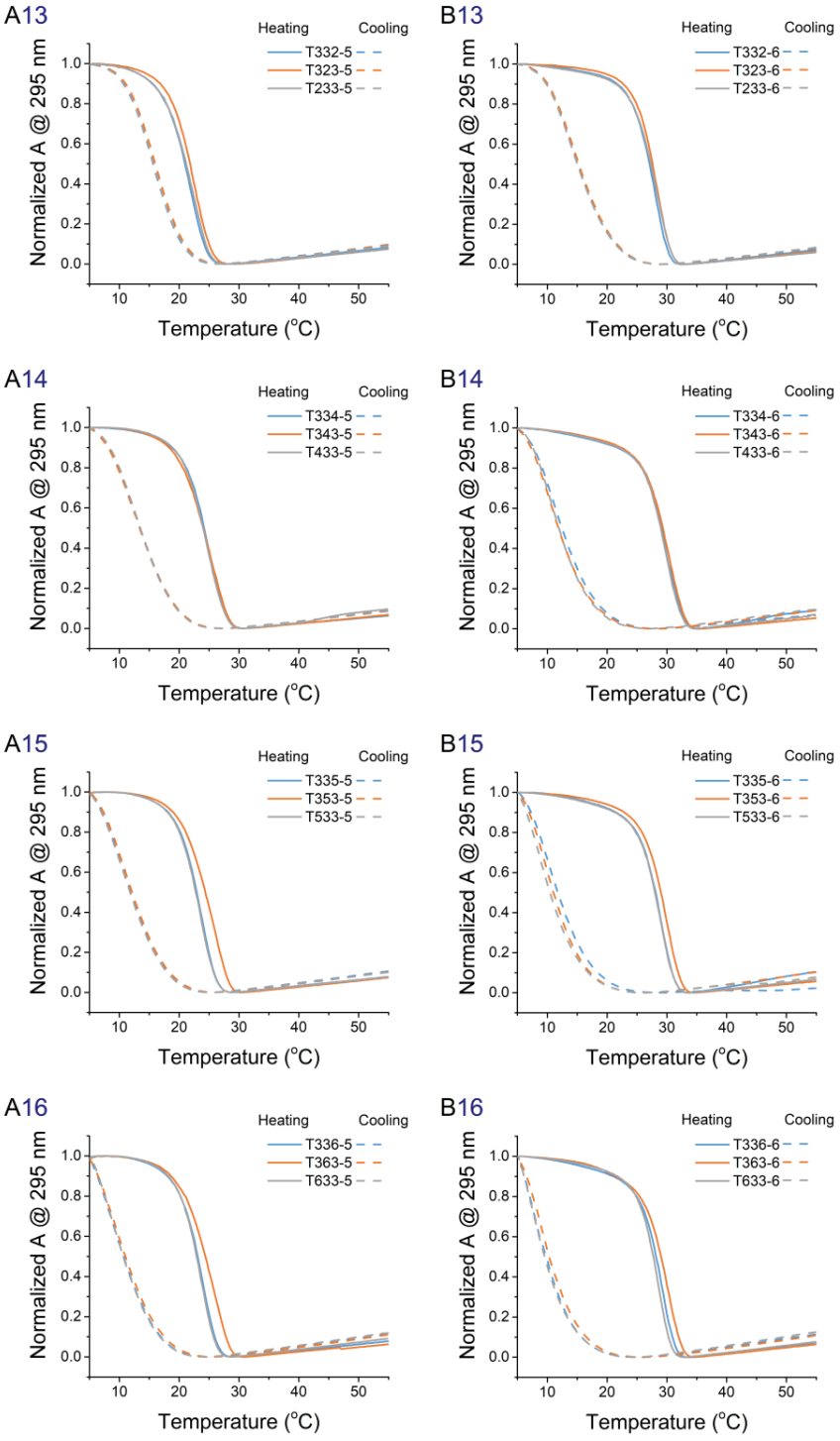


Figure S17 UV-melting (solid line) and annealing (dash line) curves at pH 7.0 of (**A**) i-DNAs with C_5 tract (first column,A1~A16), (**B**) i-DNAs with C_6 tract (second column,B1~B16). Temperatures varied between 5 and 55 °C.

Figure S18 Melting temperature (T_m) at pH 5.0 and 7.0.

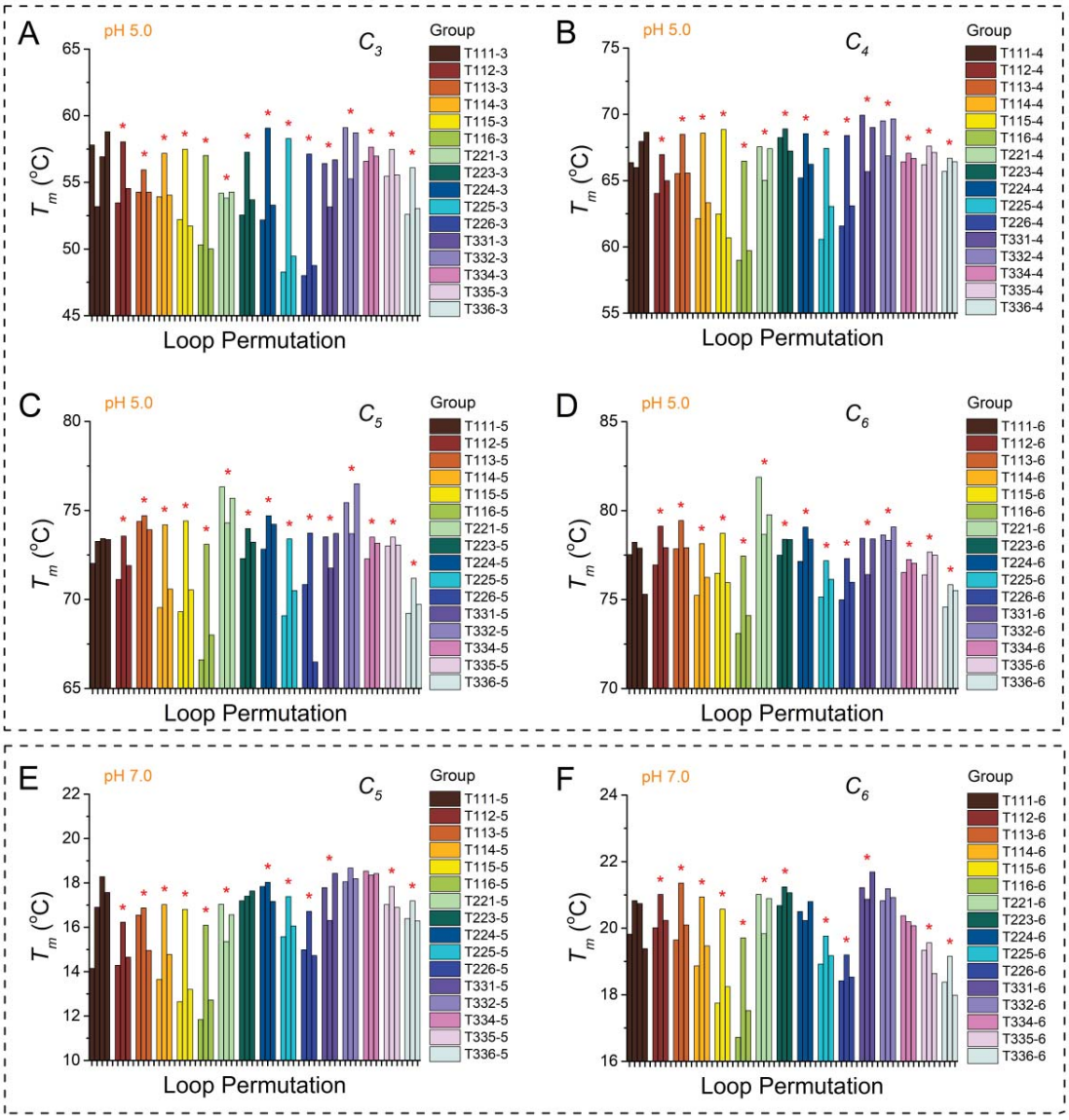


Figure S18 Melting temperature: (**A**) I-DNAs with four C_3 tracts at pH 5.0; **(B)** I-DNAs with four C_4 tracts at pH 5.0; (**C**) I-DNAs with four C_5 tracts at pH 5.0; (**D**) I-DNAs with four C_6 tracts at pH 5.0; (**E**) I-DNAs with four C_5 tracts at pH 7.0; (**F**) I-DNAs with four C_6 tracts at pH 7.0. The experiments were carried out in 20 mM Britton-Robinson buffer with 140 mM KCl and 20 mM NaCl. Data in this figure was acquired by UV-melting experiment. The temperatures of samples were recorded from 5 to 95 °C (for sequences in pH 5.0, **A**-**D**) at rate of 0.5 °C/min or 5 to 55 °C (for sequence in pH 7.0, **E** & **F**) at rate of 0.2 °C/min. Note the differences in Y-axis limits between panels. All oligonucleotide strand concentrations were 5 μM. Symbol * at top of the bar indicates that in this group the sequence with a longer central loop shows a higher thermal stability.

Figure S19 DSC-melting and annealing profiles of selected sequences.

Several results draw from UV-melting/annealing experiments are validated by DSC experiments here.

- In the same group, sequences with longer central loop show higher thermal stability. However, one group is an exception: T112-6 group at pH 7.0
- Melting and annealing processes at pH 5.0 are reversible, but show an obvious hysteresis at pH 7.0.
- Hysteresis is positively correlated to the lengths of total loop length and C-tract.

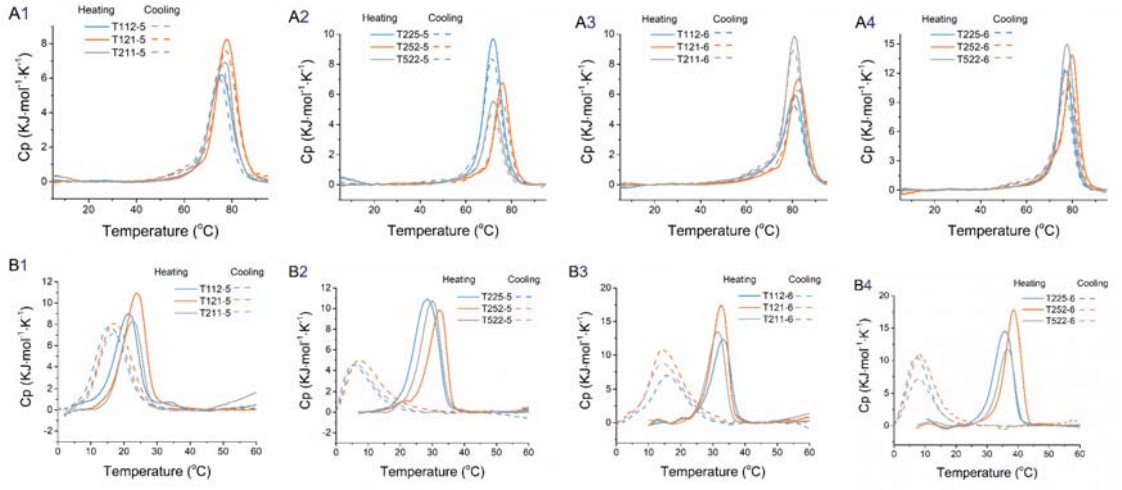
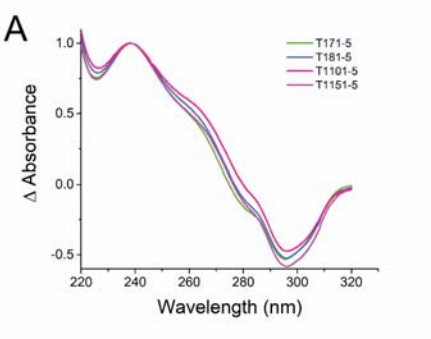


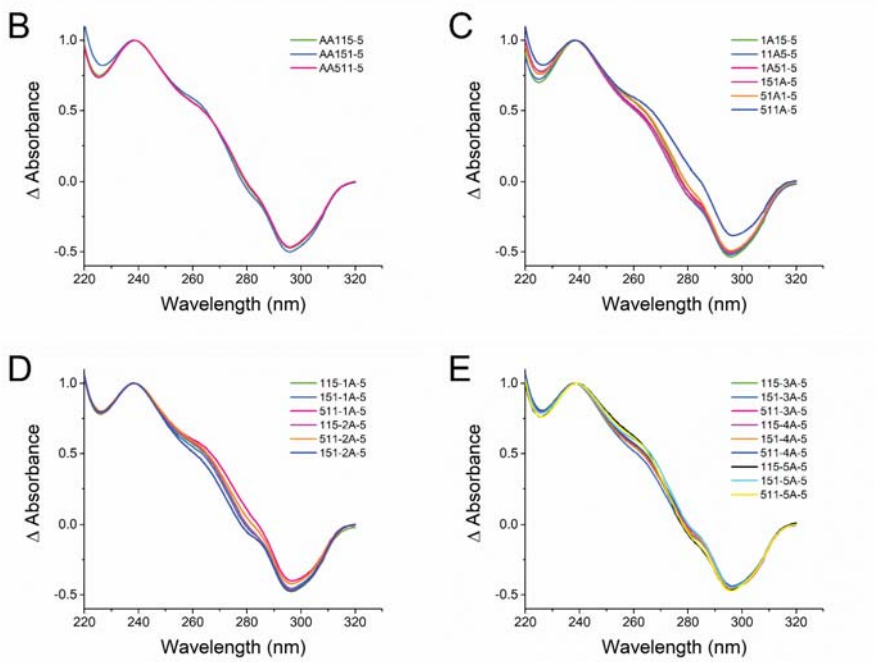
Figure S19 DSC-melting and annealing profiles of 12 selected sequences. (**A1-A4**) at pH 5.0; (**B1-B4**) at pH 7.0. Stand concentration is 100 μM. All scans are performed at 1.0 °C/min.

Figure S20 TDS of 40 extended sequences with C_5-tract.

A) Sequences with longer (7-15) central loop.



B-E) Sequences with adenine in loop.



F-G) Sequences with two different short loops.

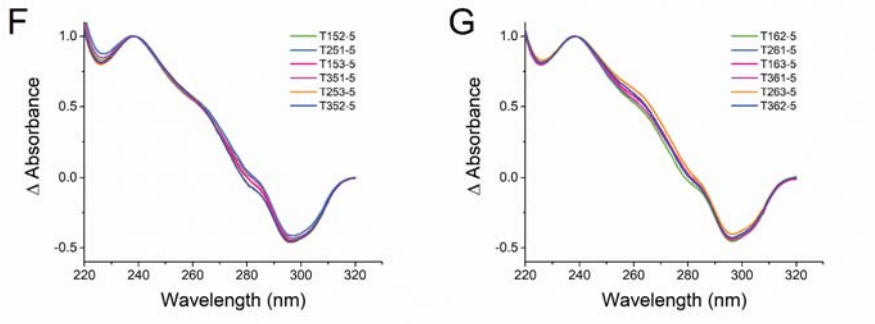


Figure S20 TDS of 40 extended sequences with C_5-tract at pH 5.0.

Figure S21 pH-dependent ellipticities and UV-melting curves at pH 5.0 of sequences with C_5-tract and longer central loop.

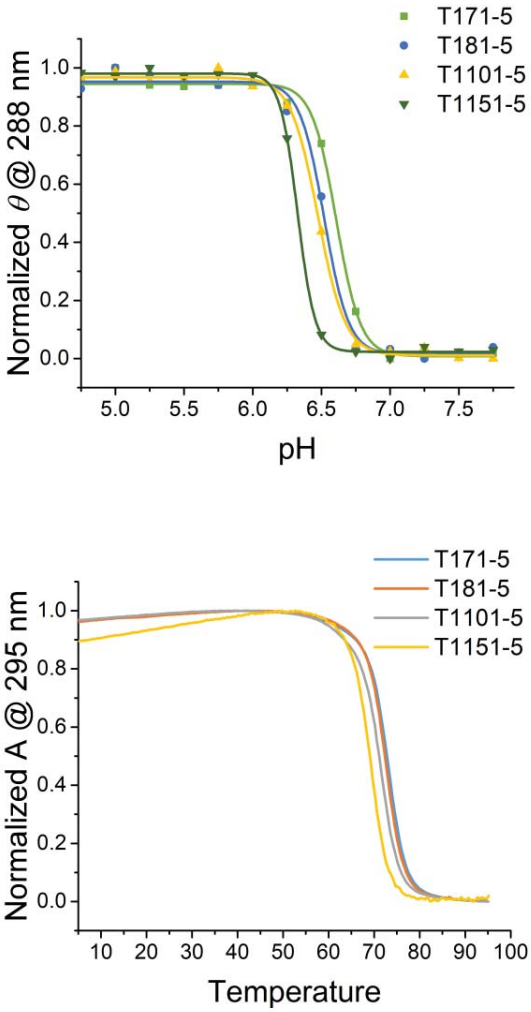


Figure S21 pH-dependent CD spectra (upper) and UV-melting curves (lower) at pH 5.0 of sequences with C_5-tract and longer (7-15) central loop.

Figure S22 Effect of central spacer length on pH_T and T_m of T1N1-5 sequences.

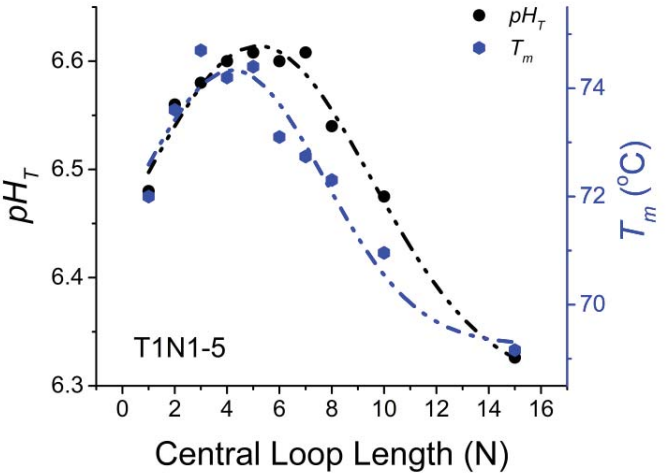


Figure S22 Effect of central spacer length on pH_T and T_m of T1N1-5 sequences. T1N1-5 represents the sequences with C_5-tract and two single nucleotide spacers, where N is a central spacer of variable length. Sequences are provided in **Tables S1** and **S3**. Gauss functions were used to fit the data.

Figure S23 pH-dependent CD spectra of sequences with two short loops of different length.

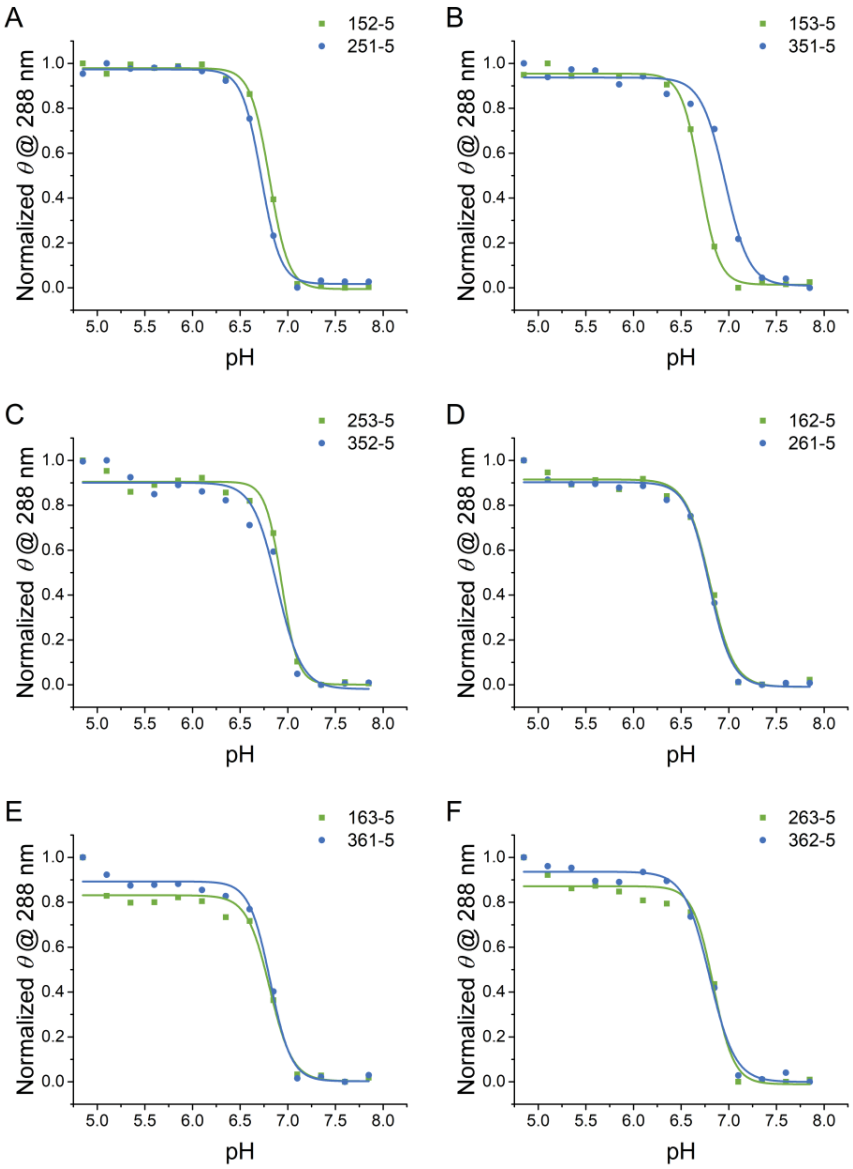


Figure S23 pH-dependent CD spectra of sequences with two short loops in different length. (**A**) *152-5* group; (**B**) *153-5* group; (**C**) *253-5* group; (**D**) *162-5* group; (**E**) *163-5* group; (**F**) *263-5* group.

Figure S24 UV-melting curves at pH 5.0 of sequences with two short loops in different length.

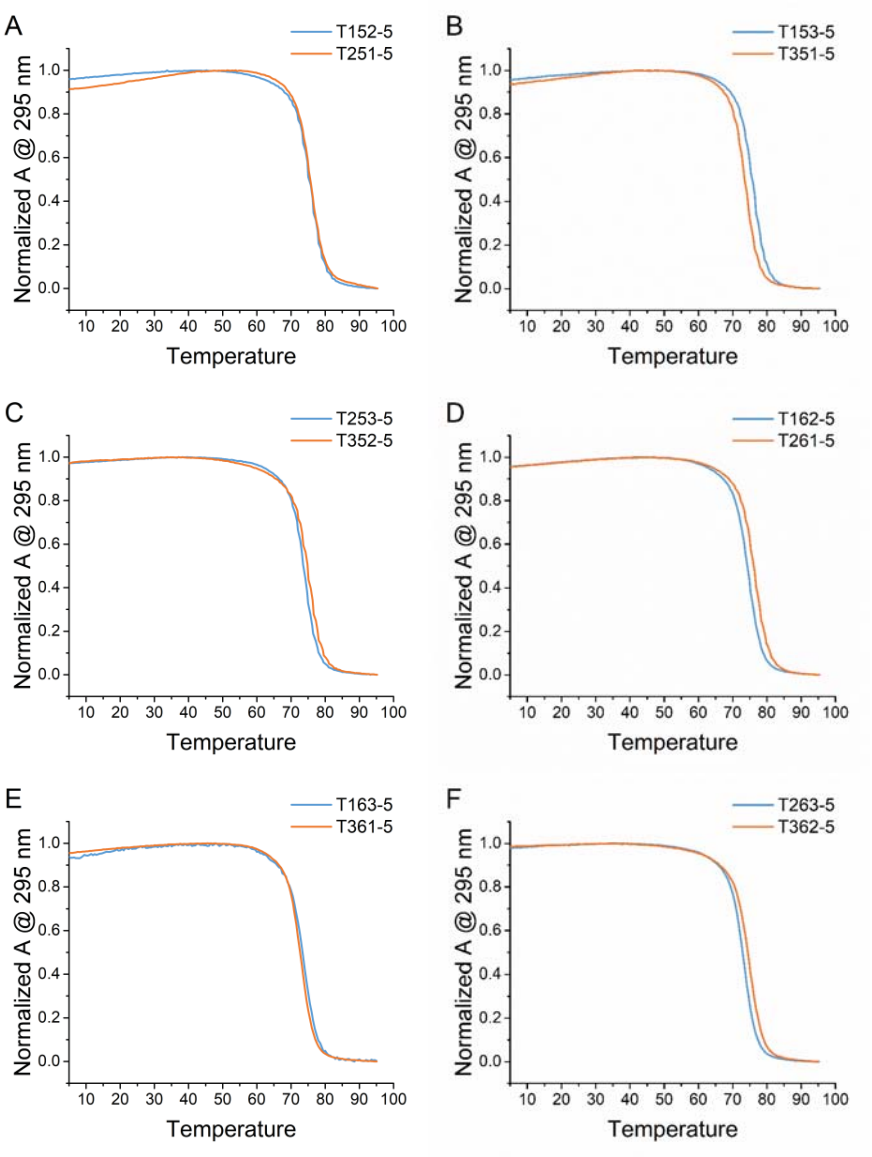


Figure S24 UV-melting curves at pH 5.0 of sequences with two short loops in different length. (**A**) *152-5* group; (**B**) *153-5* group; (**C**) *253-5* group; (**D**) *162-5* group; (**E**) *163-5* group; (**F**) *263-5* group.

Figure S25 Hypothesis of pair-sample *t*-test between *SLM* and *MLS* loop permutations.

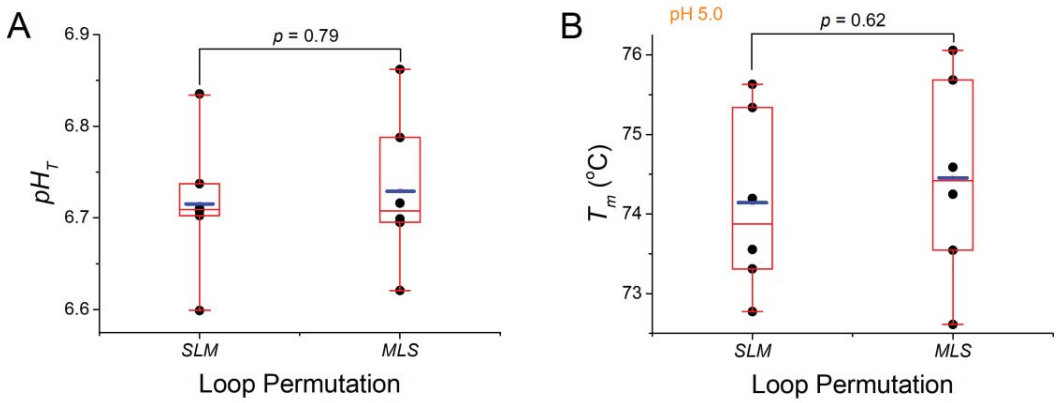


Figure S25 Hypothesis of pair-sample *t*-test between *SLM* and *MLS* loop permutations of 12 sequences with C_5-tract and two different short loops. Two sequences from the same group are paired samples. (**A**) pH_T and (**B**) T_m.

Figure S26 pH-dependent ellipticities of sequences with C_5-tract and one or two adenines in loop.

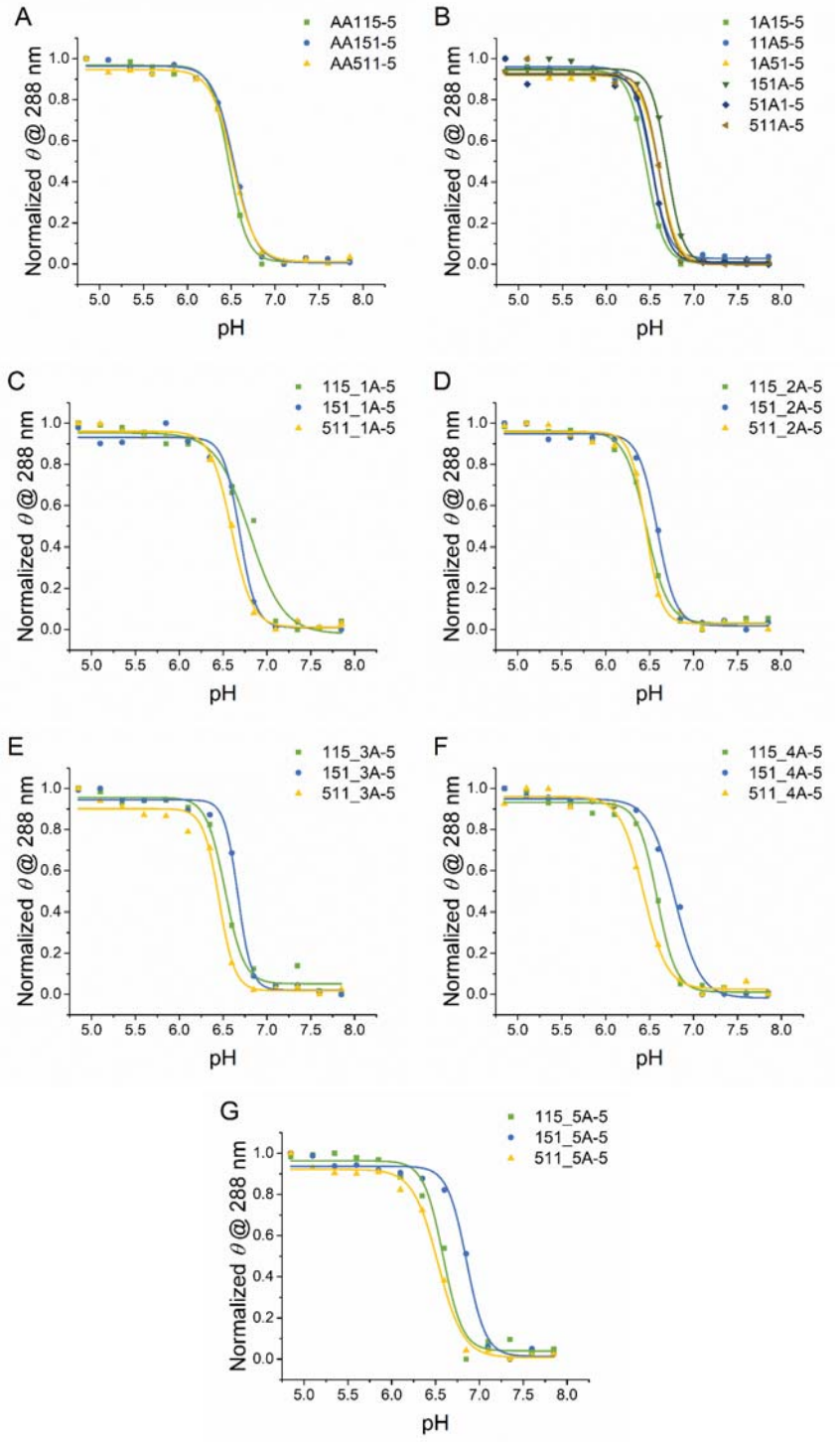


Figure S26 pH-dependent CD spectra of sequences with C_5-tract and one or two adenines in loop. (**A**) *AA115-5* group; (**B**) *1A15-5* group; (**C**) *115_1A-5* group; (**D**) *115_2A-5* group; (**E**) *115_3A-5* group; (**F**) *115_4A-5* group; (**G**) *115_5A-5* group.

Figure S27 UV-melting curves at pH 5.0 of sequences with C_5-tract and one / two adenines in loop.

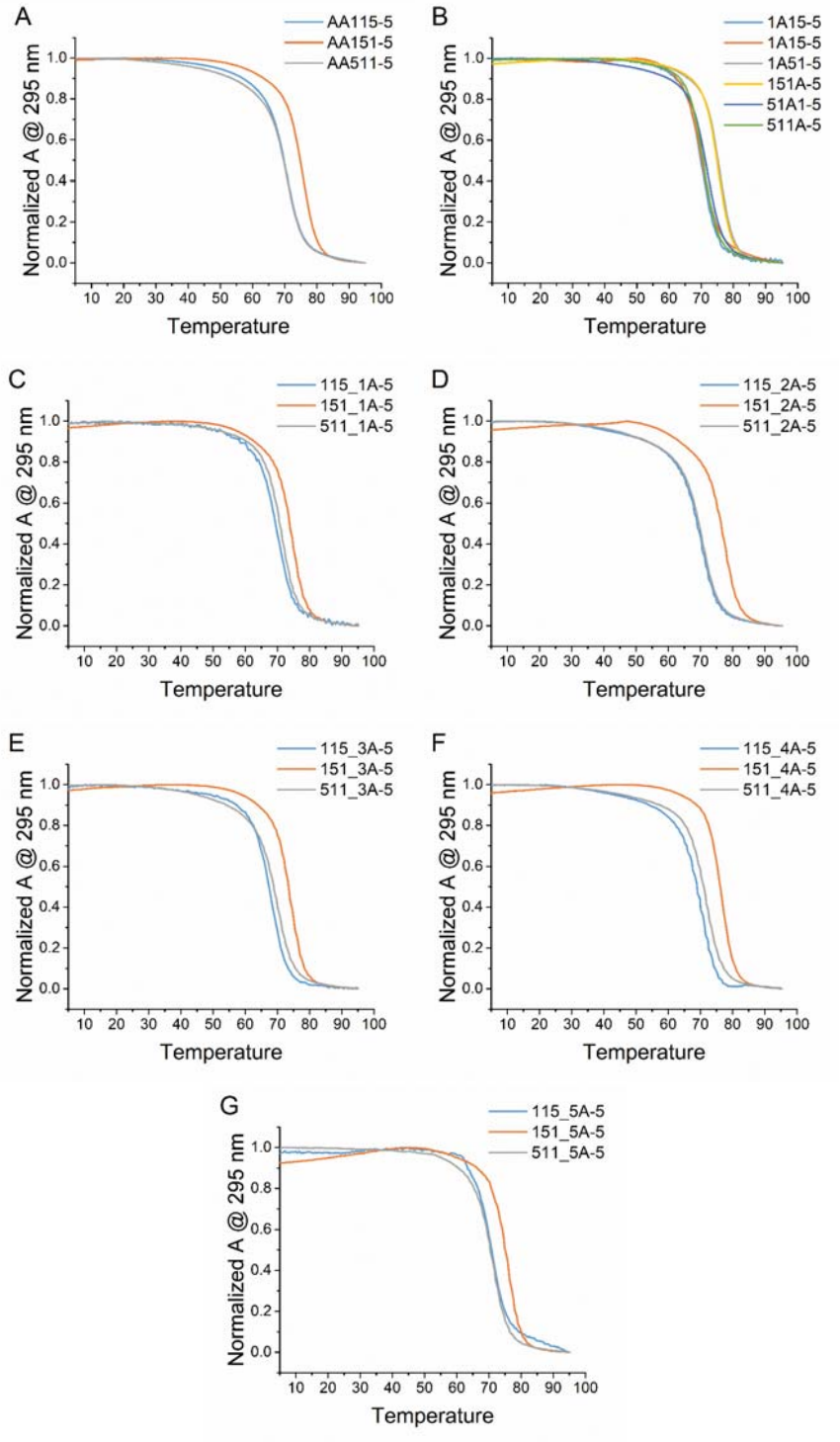


Figure S27 UV-melting curves at pH 5.0 of sequences with C_5-tract and one or two adenines in loop. (**A**) *AA115-5* group; (**B**) *1A15-5* group; (**C**) *115_1A-5* group; (**D**) *115_2A-5* group; (**E**) *115_3A-5* group; (**F**) *115_4A-5* group; (**G**) *115_5A-5* group.

Figure 28 Spacer permutation in sequences with different spacer compositions.

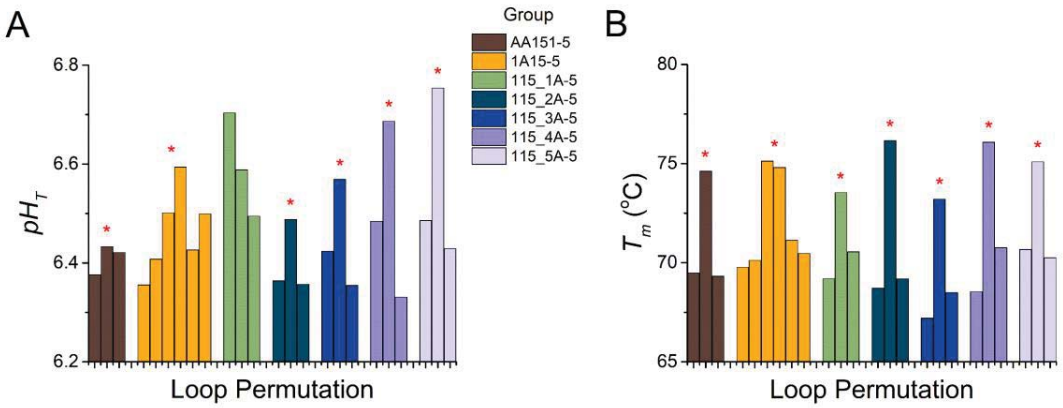


Figure 28 Spacer permutation in sequences with different spacer compositions: (**A**) pH_T and (**B**) T_m. Sequences information are given in **Tables S1** and **S3**. Symbol asterisk * at top of the bar stands for that the group obeys the rule that a sequence with a longer central spacer has a higher pH transition midpoint or thermal stability.

Figures S29-30 Cells viability, level of DNA transfection, and intracellular localization of transfected DNAs for in-cell NMR experiments.

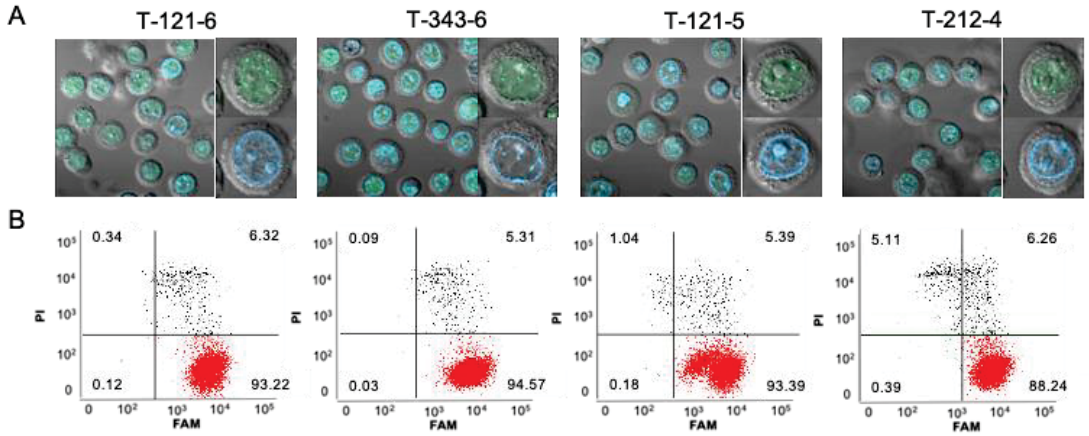


Figure S29 (**A**) Double-staining (PI/FAM) FCM analysis (post in-cell NMR spectra acquisition) and (**B**) confocal microscopy images of cells cotransfected with the (FAM)-T121-6, T343-6, T121-5, and T212-4 constructs. In the FCM plots, the percentages of viable nontransfected cells, viable DNA-containing cells, dead/compromised nontransfected cells, and dead/compromised cells transfected with DNA are indicated in the bottom-left, bottom-right, top-left, and top-right quadrants, respectively. In the confocal images, the green color marks the localization of (FAM)-DNA, while the blue color marks cell nuclei stained with Hoechst 33342.

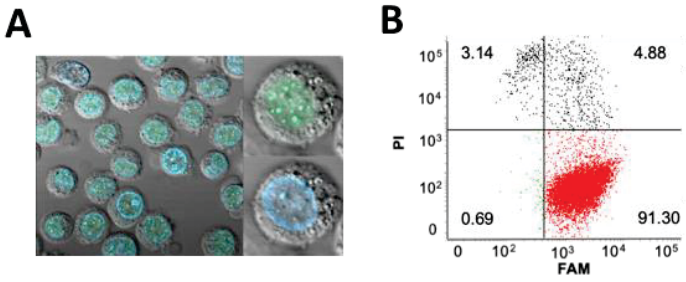


Figure S30 (**A**) Confocal microscopy image and (**B**) double-staining (PI/FAM) FCM analysis post temperature resolved in-cell NMR spectra acquisition of cells cotransfected with the (FAM)-T121-6. For meaning of colors and description of quadrants in the confocal image and FCM plot see legend of **Figure S29**.

Figure S31 Correlation plots between the experimental stability measures and the i-DNA stability scores obtained via optimized models analogous to G4Hunter.

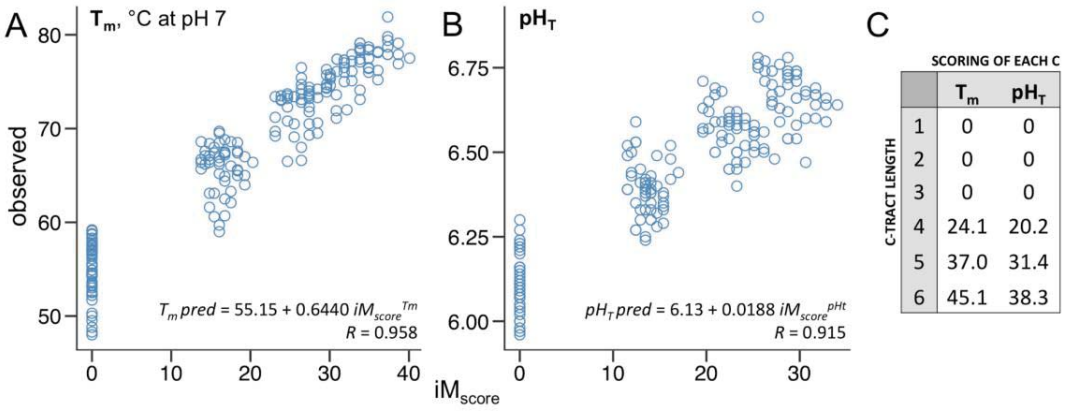


Figure S31 Correlation plots between the experimental stability measures (***T_m*** at pH 7.0 and ***pH_T***) and the i-DNA stability scores (iM_{score}) obtained via optimized models analogous to G4Hunter. Plots are brought for both T_m *vs.* iM_{score}^{Tm} (**A**) and pH_T *vs.* iM_{score}^{pHt} (**B**) dependencies. The correlation equations and the Pearson's correlation coefficients (R) are brought on the individual plots (**A**, **B**). The table in (**C**) shows the optimized positive scoring coefficients of each cytosine (counterpart of guanine in the case of G4s) in a C-tract of a given length, brought for both T_m and pH_T.

Supplementary References

1. M. Cheng *et al.*, Loop permutation affects the topology and stability of G-quadruplexes. *Nucleic Acids Res.* **46**, 9264-9275 (2018).
2. A. M. Fleming *et al.*, 4n-1 is a "sweet spot" in DNA i-motif folding of 2'-deoxycytidine homopolymers. *J. Am. Chem. Soc.* **139**, 4682-4689 (2017).
3. J.-L. Mergny, J. Li, L. Lacroix, S. Amrane, J. B. Chaires, Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.* **33**, e138 (2005).
4. J.-L. Mergny, L. Lacroix, Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.* **26**, 4797-4803 (1998).
5. P. Viskova, D. Krafcik, L. Trantirek, S. Foldynova-Trantirkova, In-cell NMR spectroscopy of nucleic acids in human cells. *Curr. Protoc. Nucleic Acid Chem.* **76**, e71 (2019).
6. V. Sklenář, A. Bax, Spin-echo water suppression for the generation of pure-phase two-dimensional NMR spectra. *J. Magn. Reson. (1969-1992)* **74**, 469-479 (1987).
7. R. Hansel *et al.*, Evaluation of parameters critical for observing nucleic acids inside living Xenopus laevis oocytes by in-cell NMR spectroscopy. *J. Am. Chem. Soc.* **131**, 15761-15768 (2009).
8. A. Bedrat, L. Lacroix, J. L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* **44**, 1746-1759 (2016).
9. N. A. G. Johnson, L. Tamon, A. B. Sahakyan (Optimus: a general purpose adaptive optimisation engine, GitHub link to the code: http://github.com/SahakyanLab/Optimus, accessed in November 2019.
10. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system. *arXiv* DOI: 10.1145/2939672.2939785, 1-13 (2016).
11. A. Natekin, A. Knoll, Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, 21 (2013).
12. J. H. Friedman, Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**, 367-378 (2002).
13. J. H. Friedman, Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189-1232 (2001).
14. M. Kuhn, K. Johnson, *Applied predictive modeling* (Springer, New York, USA, 2013), 10.1007/978-1-4614-6849-3.
15. A. B. Sahakyan *et al.*, Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).
16. M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data. *Science* **324**, 81-85 (2009).
17. M. Johnson *et al.*, NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5-9 (2008).
18. J.-L. Mergny, L. Lacroix, Analysis of thermal melting curves. *Oligonucleotides* **13**, 515-537 (2003).