How to cite: *Angew. Chem. Int. Ed.* **2021**, *60*, 10286–10294
International Edition: doi.org/10.1002/anie.202016801
German Edition: doi.org/10.1002/ange.202016801

i-DNA

Thermal and pH Stabilities of i-DNA: Confronting in vitro Experiments with Models and In-Cell NMR Data

Mingpan Cheng, Dehui Qiu, Liezel Tamon, Eva Ištvánková, Pavlína Víšková, Samir Amrane, Aurore Guédin, Jielin Chen, Laurent Lacroix, Huangxian Ju, Lukáš Trantírek, Aleksandr B. Sahakyan, Jun Zhou, and Jean-Louis Mergny*

Abstract: *Recent studies indicate that i-DNA, a four-stranded cytosine-rich DNA also known as the i-motif, is actually formed in vivo; however, a systematic study on sequence effects on stability has been missing. Herein, an unprecedented number of different sequences (271) bearing four runs of 3–6 cytosines with different spacer lengths has been tested. While i-DNA stability is nearly independent on total spacer length, the central spacer plays a special role on stability. Stability also depends on the length of the C-tracts at both acidic and neutral pHs. This study provides a global picture on i-DNA stability thanks to the large size of the introduced data set; it reveals unexpected features and allows to conclude that determinants of i-DNA stability do not mirror those of G-quadruplexes. Our results illustrate the structural roles of loops and C-tracts on i-DNA stability, confirm its formation in cells, and allow establishing rules to predict its stability.*

Introduction

i-DNA (also known as the i-motif) is a fascinating four-stranded structure discovered in the 1990s by M. Guéron and colleagues, and stemming from the interlocking of two equivalent parallel-stranded right-handed duplexes.[1] Such cytosine-rich structure, which relies on the formation of hemi-protonated C·C$^+$ base pairs (Figure 1 A),[2] can be formed with two or more independent strands, or be intramolecular, as depicted in Figure 1 B.[3] Different conformations are possible, but i-DNA is not as polymorphic as G-quadruplexes (G4s) as two diametrically distant strands must remain parallel to each other and adjacent strands are always running in opposite orientations (Figure 1 C).[3a,4] In addition, bi- or tetra-molecular complexes may coexist with intramolecular structures.[5]

i-DNA has long remained in the shadow of G4s formed by complementary G-rich sequences, given its limited stability at physiological pH. Formation of each C·C$^+$ base pair requires the protonation of one cytosine at its N3 position (p$K_a \approx 5$): as a consequence, the stability of this motif is optimal under mildly acidic conditions but remains questionable at neutral pH.[6] i-DNA extreme pH dependency can actually become an asset to design sensitive pH-responsive devices[7] and may be applicable to analytical chemistry,[8] nanotechnology,[7,9] and therapeutics.[10] Regarding its biological relevance, two recent independent studies indicate that i-DNA is actually present within human cells.[11] Similar to G4-prone sequences, i-DNA-prone motifs are widely distributed in genomes,[12] and have been found to modulate telomerase activity,[13] transcription of genes,[14] and DNA biosynthesis.[15]

Our understanding of i-DNA 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-DNA at room temperature and neutral pH.[5a,c,6] Additional interactions involving hydrogen bonding also stabilize i-DNA.[16] Burrows and colleagues analyzed dC homo-oligonucleotides, and found that pure cytosine tracts may adopt stable i-motif conformations.[17] These results somewhat mirror those found for G4 formation;[18] as a consequence, the complementary strand of a G4-forming sequence is generally prone to i-DNA formation. Besides C-tracts, the nature of the loops (length and base composition) also plays a role in i-DNA formation.[19] However, contradictory conclusions have been drawn upon how loop length influences i-DNA stability.[17a,19b,d,20] These results came from the investigations of a limited number of sequences; system-

[*] Dr. M. Cheng, D. Qiu, J. Chen, Prof. Dr. H. Ju, Prof. Dr. J. Zhou, Dr. J.-L. Mergny
State Key Laboratory of Analytical Chemistry for Life Science, School of Chemistry & Chemical Engineering, Nanjing University
Nanjing 210023 (China)
E-mail: jun.zhou@nju.edu.cn

Dr. M. Cheng, Dr. S. Amrane, A. Guédin, Dr. J.-L. Mergny
ARNA Laboratory, Université de Bordeaux, INSERM U 1212, CNRS UMR5320, IECB
33607 Pessac (France)

L. Tamon, Prof. Dr. A. B. Sahakyan
MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford
Oxford OX3 9DS (UK)

E. Ištvánková, P. Víšková, Prof. Dr. L. Trantírek
Central European Institute of Technology, Masaryk University
62500 Brno (Czech Republic)

Dr. L. Lacroix
IBENS, Ecole Normale Supérieure, CNRS, INSERM, PSL Research University
75005 Paris (France)

Dr. J.-L. Mergny
Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris
91128 Palaiseau (France)

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202016801.

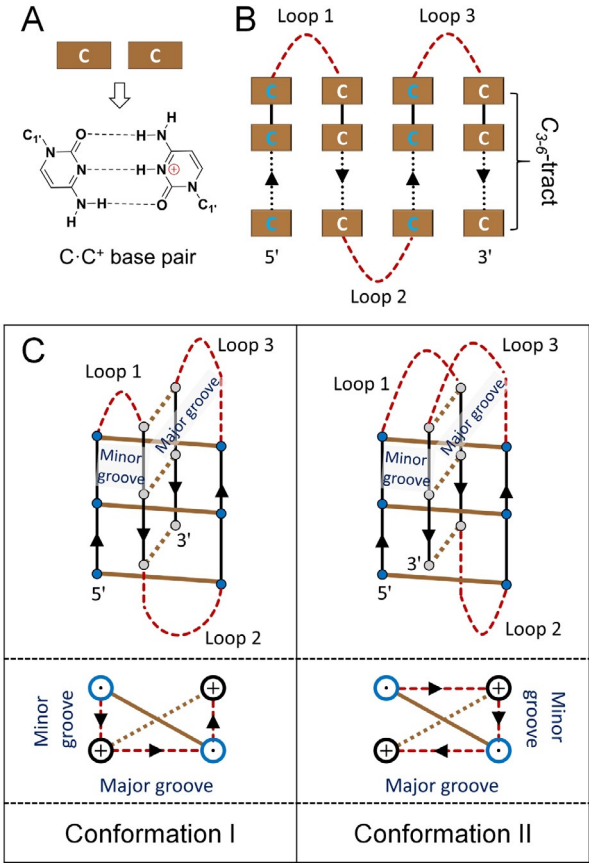

Figure 1. A) Hemi-protonated C·C$^+$ base pair. B) Example of a sequence prone to form intramolecular i-DNA. C) Possible loop arrangements in an i-DNA structure; Simplified diagram of two linking directions between strands: Central loop can across either major (*left*, conformation I) or minor (*right*, conformation II) groove.[3a].

atic studies based on large numbers of examples are needed to achieve an objective conclusion.

Herein, we systematically analyzed i-DNA stability on an unprecedented large selection of sequences (271 in total). This unique dataset unveiled important parameters governing the stability of i-DNA. Global trends were identified and more subtle effects were found using machine learning and other modeling approaches, 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.

Results

Sequences design and nomenclature

Sequences information and nomenclature are shown in Table S1. Each sequence bears four C-tracts, containing 3 to 6 cytosines (C_3 to C_6). These four C-tracts (which are generally of identical length) are separated by three spacer regions, which should allow the formation of an intramolecular structure.[6] Sequences with four non-equal C-tracts have also been considered in a limited number of cases, as discussed later. 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 is longer than six.[5a,c,6] To reduce the number of spacer arrangements, most sequence groups contain two identical spacers, which are generally composed of thymines only. Each spacer involves one to six thymine nucleotides, and total spacer length is capped at twelve nucleotides in most cases. Note that the term "*spacer*" corresponds here to the non-C nucleotides connecting C-tracts: 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.

The following nomenclature was chosen: unless otherwise stated, a "T" prefix means that the three spacers are composed of thymine bases only; the three consecutive numbers refer to three spacer length in the 5′ to 3′ direction; while the "-3", "-4", "-5" or "-6"′ suffix refers to four C-tracts of C_3, C_4, C_5, and C_6, respectively. To compare the effects of spacer arrangement on i-DNA stability, the notion of sequence *group* was introduced.[21] The sequences in the same group are only differing in the way that spacers are arranged. A group is named after the first sequence in the group. For example, the *T112-3* group is composed of 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. Tables S1 and S2 summarize the results obtained for 60 groups of three sequences with different spacer arrangements.

Evidence for i-DNA formation

First, i-DNA formation was checked under acidic (pH 5.0) or neutral (pH 7.0) conditions. Thermal difference spectra (TDS) are provided in Figure S1 and clearly showed that they fold into an i-motif at pH 5.0 (two major peaks around 239 and 294 nm).[22] In addition, i-DNA formation for 12 selected sequences was also proved by the presence of imino proton peaks from C·C$^+$ at 15 to 16 ppm in ^{1}H NMR spectra (Figure S2).[3a,d]

The molecularities of the 49 sequences with a C_5-tract in Table S2 were checked at both pH 5.0 and 7.0 by native PAGE (Figures S3 and S4). All sequences tested mainly fold into intramolecular species, in agreement with previous studies.[5a,c] 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 its stability.

In contrast to TDS recorded at pH 5.0, the situation was more diverse 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 structure at neutral pH (Figure 2, dashed lines):

I. None of the three sequences in a given group fold into an i-DNA at neutral pH. This category includes all groups with C_3-tract (Figures 2A,B and S1A), *T336-4* group (Figure S1B) and *T336-5* group (Figure S1C);

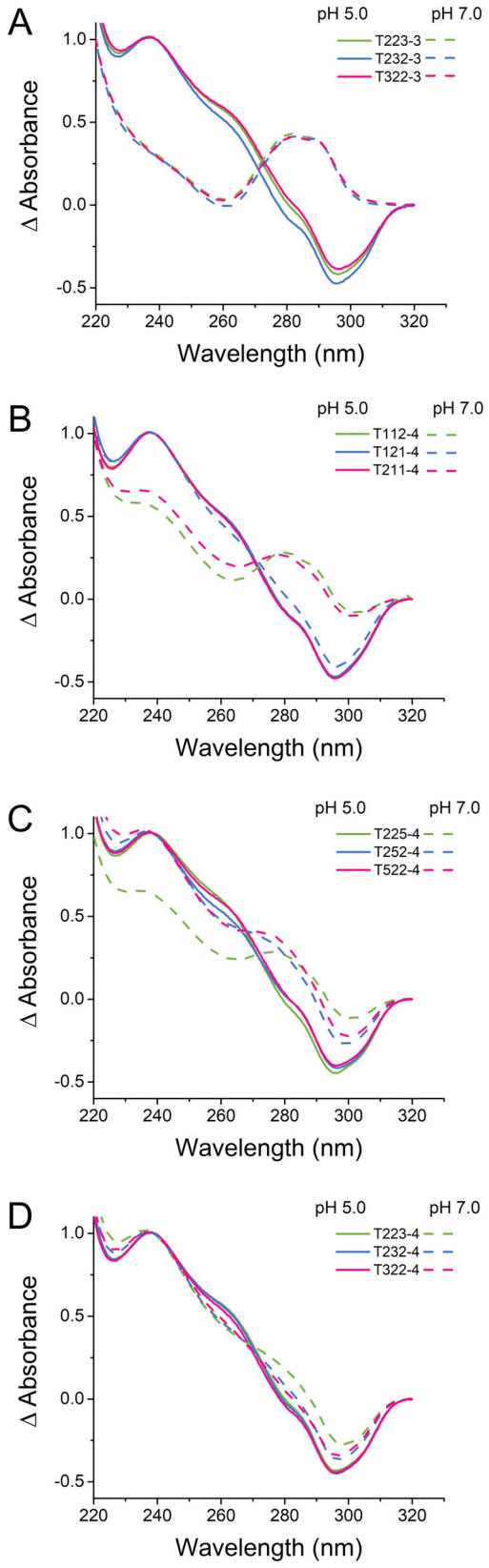

Figure 2. Normalized TDS of selected groups at pH 5.0 (solid lines) and 7.0 (dashed lines). A) Zero, B) one, C) two, D) all three sequences in a group fold into i-DNA completely at neutral pH.

II. Only one of three sequences within the same group folds into an i-DNA. This category includes *T121-4* (Figure 2B), *T161-4* (Figure S1B), *T262-4* (Figure S1B) as well as *T353-5* groups (Figure S1C);
III. Two of the three sequences in the same group fold into an i-DNA. This category includes *T225-4* (Figure 2C), *T335-4* (Figure S1B), and *T225-5* groups (Figure S1C);
IV. All three sequences in the same group fold into an i-DNA (Figures 2D and S1B,C).

An interesting trend emerges from this classification. In types II and III categories (for which some, but not all, group members form an i-DNA 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 exhibit higher pH and thermal stabilities

To confirm the differences in stability inferred from TDS at pH 7.0, we performed CD and UV measurements (Figures S5–S14). pH transition mid-point (pH_T) was 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 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. 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 spectra, 48 or 46 of the 60 groups (80% and 77%) follow this tendency, respectively.

Then thermal denaturation of i-DNAs at pH 5.0 and pH 7.0 was tracked by UV-absorbance at 295 nm (Figures S16–S18).[23] 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 low stabilities ($T_m <$ 12 °C) preventing accurate determinations. Folding and unfolding processes follow relatively fast kinetics under

Table 1: Enumeration of groups which obey 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.

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 processes.[5a,6] For some sequences, such as T444-6, T336-6, T363-6 and T633-6, this difference in melting/cooling T_ms 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.[24]

The analysis of T_m values further confirmed the *"long central spacer is better"* rule: 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_ms 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.

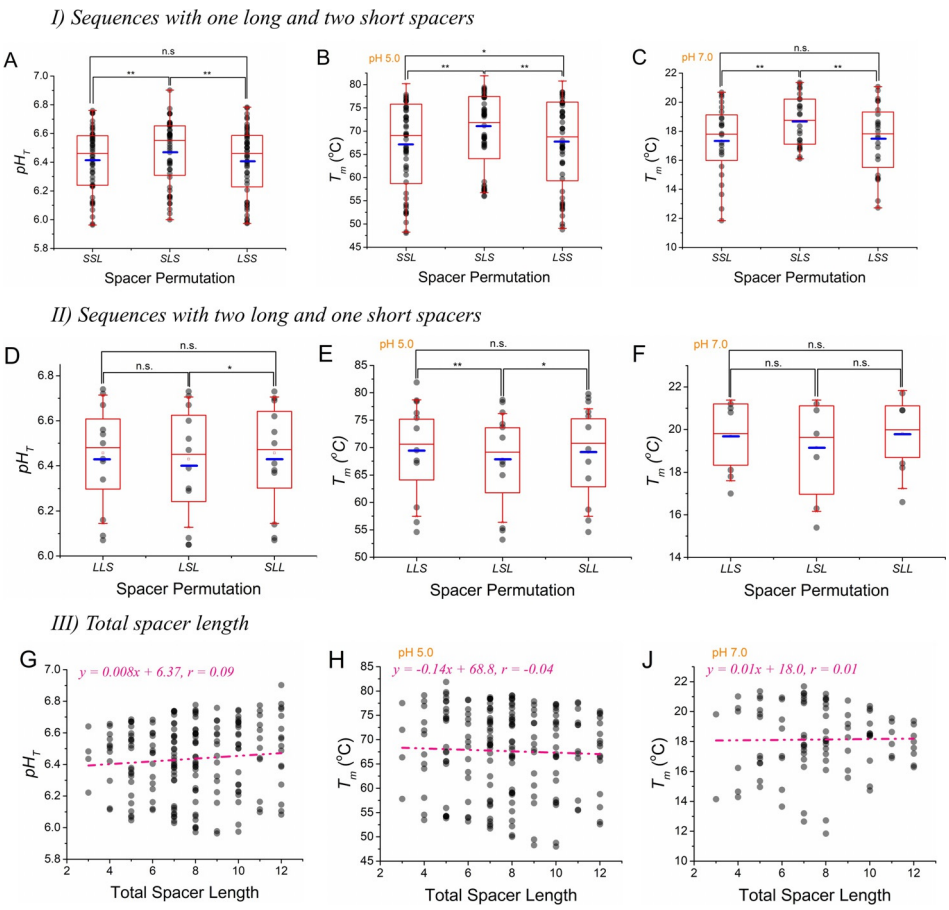

Figure 3. Effects of total spacer length and individual spacer permutation on pH_T and T_m. A&D) pH_T versus spacer permutation; T_ms at B&E) pH 5.0 or at C&F) pH 7.0 versus spacer permutation. Sequences come from Table S2. Blue and red lines in the red box are the average and median values for each spacer combination, respectively. Hypotheses of pair-sample *t*-test are performed between every two spacer combinations: not significant (n.s.); $p > 0.05$; *, $p < 0.05$; **, $p < 0.0005$. pH_T (G), T_ms at (H) pH 5.0 and (J) pH 7.0 as a function of total spacer length. The dashed line corresponds to a linear fit (equation shown above).

Analyses of effects of spacer permutation are presented in Figures 3 A–F. Sequences are divided into two categories: *i)* sequences with two relatively long (L) and one relatively 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 relatively longer central spacer, including *SLS* (Figures 3 A–C), *LLS* and *SLL* (Figures 3 D–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 3 F, and *LLS* versus *LSL* in Figure 3 D), 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 3 A–C; *LLS* or *SLL* versus *LSL* in Figures 3 D–E). In addition, except for one comparison (*SSL* versus *LSS* in Figure 3 B), 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 proposed previously[3a] and depicted in Figure 1 C. Three loops stretch and pass through either minor-major-minor grooves (conformation I) or major-minor-major grooves (conformation II). Given the results obtained here, assuming spacer length would allow both possibilities, conformation II generally appears less stable than conformation I.

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 T_m values and hysteresis are summarized in Table S4. These results 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 C-tract but not total spacer length. pH_T and T_m (at pH 7.0, only the sequences with C_5 and C_6-tracts are used) of 196 sequences are presented in Table S2

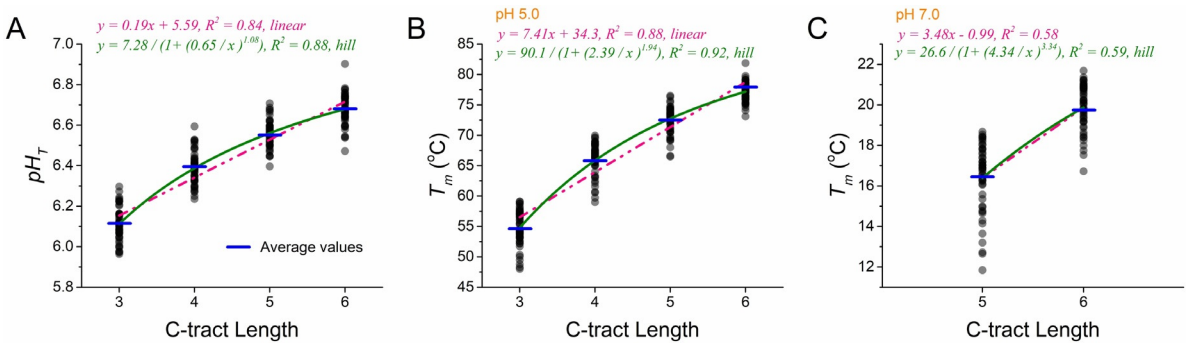

Figure 4. A) pH_T, T_ms at B) pH 5.0 and C) pH 7.0 as a function of C-tract length. Averages of pH_T and T_m are indicated in blue short lines. Sequences in Table S2 are used.

and plotted as a function of total spacer length in Figures 3 G–J. 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, indicating that it has a limited effect on the i-DNA stability at both acidic and neutral pHs. This maybe the reason why previously reported studies about the effects of loop length on i-DNA are contradictory or negligible.[5a,c,17a,19b,d]

Quantitative analyses of the relationships between pH_T or T_m vs. C-tract length were previously missing.[5a,c,6] The assessment of C-tract role on i-DNA stability is depicted in Figure 4. Stability increases with C-tracts length: averaged pH_Ts with C_3, C_4, C_5 and C_6-tracts are 6.11, 6.39, 6.56 and 6.68, respectively (Figure 4 A). A similar relationship was found between T_m and C-tract length under both acidic and neutral conditions (Figures 4 B 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.[5a,c]

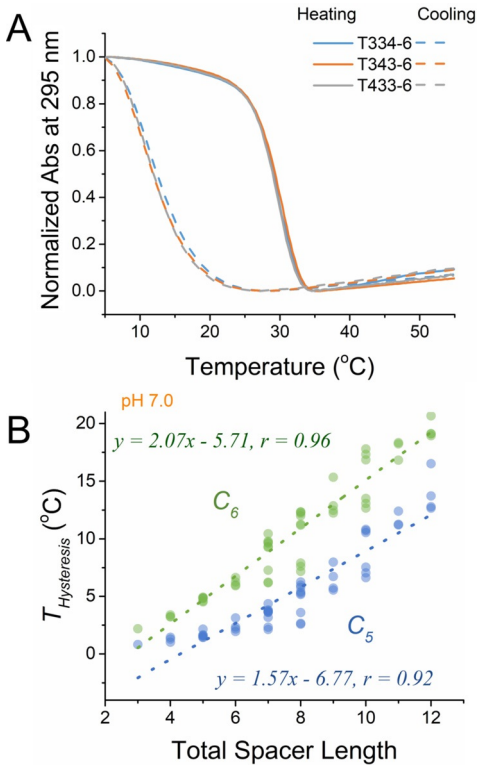

Figure 5. A) Hysteresis for the *T334-6* group in the UV-melting (solid line) and annealing (dashed line) processes at pH 7.0. Curves of other sequences are provided in Figure S17. B) Hysteresis as a function of total spacer length ($T_{Hysteresis} = T_{Heating} - T_{Cooling}$).

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

As noted before, a hysteresis phenomenon is observed at near-neutral pH: the apparent melting transition is shifted towards higher temperatures than the value deduced from cooling profiles (Figures 5 A and S17). The analysis of melting (heating) profiles alone would 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; hysteresis being larger when fast temperature changes are implemented, as expected (not shown). 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 B): in other words, sequences with longer T-loops fold and unfold slower than motifs with shorter ones. The hysteresis, induced by longer sequence length and higher pH value, is also observed in the DSC experiments (Figure S19 and Table S4). For this reason, the analysis of 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, and also inaccurate conclusion.

Expanding the "long central spacer is better" rule

All sequences 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. 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 (sequences listed in Table S1). i-DNA formation was confirmed by TDS (Figure S20). pH_T and T_m were also evaluated (Figure S21, data 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 results 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* refer to the relatively *short*, *middle*, *long* loop length for the two sequences in a group, 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.

We further expanded the sequence space, including additional terminal nucleotides, spacer variants and odd numbers of C·C$^+$ base pairs. We designed sequence variants based on T252-5. The results showed that the presence of a thymine, adenine or guanine at one (5′ or 3′) or both ends do not strongly affect the thermodynamic stability but influence the hysteresis of i-DNAs (Figures S29–S31, results summarized in Table S5). Interestingly, long adenine or guanine spacers result in the destabilization of i-DNA. Substitution of a single thymine by adenine or guanine in the second spacer increases the thermal stability, whereas the opposite effect is found in the first and third spacers. Significantly, the *"long central spacer is better"* rule can be extended to i-DNAs with odd numbers of C·C$^+$ base pairs.

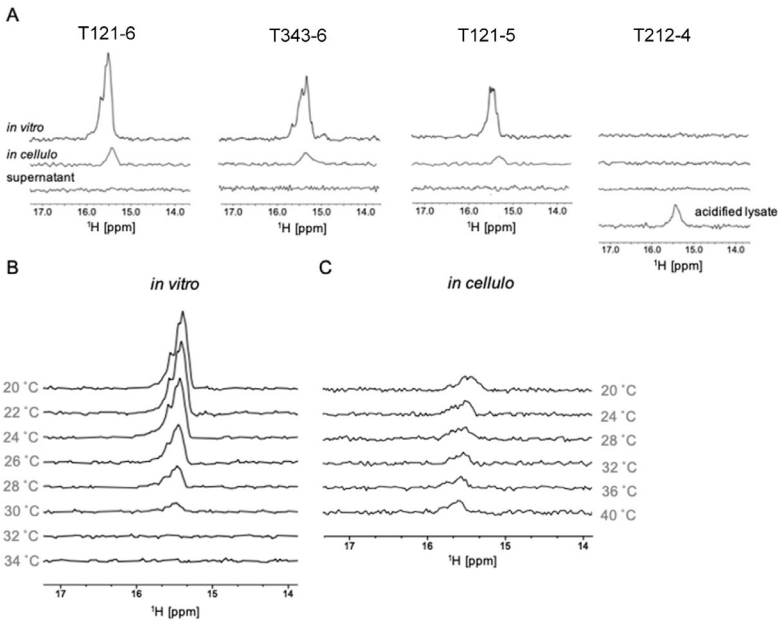

Figure 6. *in vitro* and in-cell NMR. A) ^{1}H NMR spectra acquired at 20 °C *in vitro*, in cellulo (living HeLa cells), and in supernatant (medium) collected from in-cell NMR sample post in-cell NMR spectra acquisition, respectively. Absence of signals in the "supernatant" spectra evidences that the signals observed in in-cell NMR spectra originates from DNA localized in cells. Presence of i-DNA 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. B, C) ^{1}H NMR spectra of T121-6 acquired at various temperatures under (B) *in vitro* and (C) in living cells. The prerequisite flow cytometry plots and confocal images are shown in Figure S32.

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

To assess whether the rules we uncovered for the *in vitro* stability of i-DNA are applicable *in vivo*, 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 °C (Figure 6 A). As evidenced from confocal microscopy images, all transfected constructs were localized in the nuclei (Figure S33). Observation of signals in region of the in-cell NMR spectra specific for imino protons involved in C·C$^+$ base pairs (15–16 ppm) corroborated i-DNA formation for T121-5, T121-6, and T343-6, while absence of signals indicated no i-DNA formation for T212-4 (Figure 6 A). Notably, the order of relative intensities of the imino signals in the in-cell NMR spectra (T121-6 > T343-6 > T121-5 ⩾ T212-4) essentially paralleled that obtained *in vitro*. Altogether, these data suggest that the rules derived on the basis of *in vitro* data are reasonably accurate to predict the behavior of i-DNAs in cells.

The absolute i-DNA stabilities in cells may differ from those observed *in vitro*.[11a] The intensities of imino signals in in-cell NMR spectra 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 32 °C, the detectable in the corresponding in-cell NMR spectrum measured at 36 °C (and even 40 °C), demonstrating i-DNAs may be more stable in cells (Figures 6 B and C).[11a]

Predicting i-DNA stability

Models for i-DNA stability were generated using three distinct approaches: G4Hunter-based,[25] machine learning

based,[26] and through a development of an analytical equation[27] via an increasingly popular symbolic regression that has recently been shown to correctly discover physical laws as tested on known phenomena.[28] The specifics of the approaches are detailed in the Supporting Information. 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 T_ms (pH 5.0) or pH_Ts 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 arise 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 results and discussion of the G4Hunter-based and analytical equation based methods are described in Supporting Information (Figure S34). We focus on the machine learning based approach here only.

Gradient boosting machines (GBM) as machine learning framework (Supporting Information), resulted in models that capture the T_m and pH_T measurements with great performance (data from the 20% left-out validation dataset, Figure 7). The restricted feature set, necessary to comprehensively describe the C/T-based i-DNA candidates, compensated the relatively small (for machine learning standards) dataset used in this initiative, hence arriving to a good model performance in the validation trials. 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. 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.

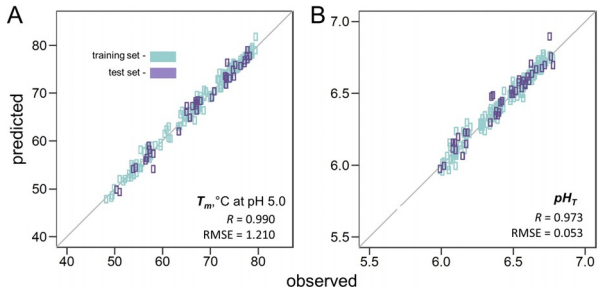

Figure 7. Correlation plots between the experimental stability measures (T_m at pH 5.0 and pH_T) and the i-DNA stability predicted via machine learning models obtained using gradient boosting machines. Plots are brought for both T_m (A) and pH_T (B) dependencies. The Pearson's correlation coefficients (R) and root mean squared errors (RMSE) are calculated on the test set brought on the individual plots.

Discussion

Our work on i-DNA sequence requirements is of unprecedented magnitude, with 271 sequences tested. Multivariate data analysis for the interpretation of TDS and CD spectra for these sequences are provided in our companion paper.[29] Even if impressive, this dataset does not allow to explore the full sequence space of i-DNA-prone sequences. Despite these restrictions, and because we tested a few sequences escaping this sequence space, our data already provides key information on i-DNA stability.

pH_T or T_m are useful to monitor i-DNA stability. As we found inappropriate to discard one of these parameters, both were used here, and it is difficult to conclude that one is superior to the other. If biological applications are contemplated, T_m and pH_T under physiological conditions would be recommended, although the accurate determination of intracellular (intranuclear) pH may prove harder than expected (see below). 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.[5b] Hysteresis is determined for each melting/annealing 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.[23] For pH_T determination, each sample was allowed to anneal at a given pH for a long period (> 12 hours), allowing thermodynamic equilibrium.

We determined how well correlated these values are. The analyses of pH_T versus T_m (Figures 8A,B), and T_m at pH 7.0 versus 5.0 (Figure 8C) revealed good but not perfect positive correlations between these figures (Pearson's R 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 reveals several trends for i-DNA sequence requirements: (*i*) Stability increases with the length of the cytosine tract (Figure 4). (*ii*) The nature of the spacer regions does not play a critical role on stability. Correlation coefficients of pH_T and T_m versus total spacer length are close to zero, indicating that total *spacer* length, assumed to reflect total *loop* length, does not affect the i-DNA stability at both acidic and neutral pH. (*iii*) The "long central spacer is better" rule seems to hold for both G4[21] 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 shared property, a duplex 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.

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. In addition, loop effects were more pronounced for G4 forming sequences, with large differences in T_m (and

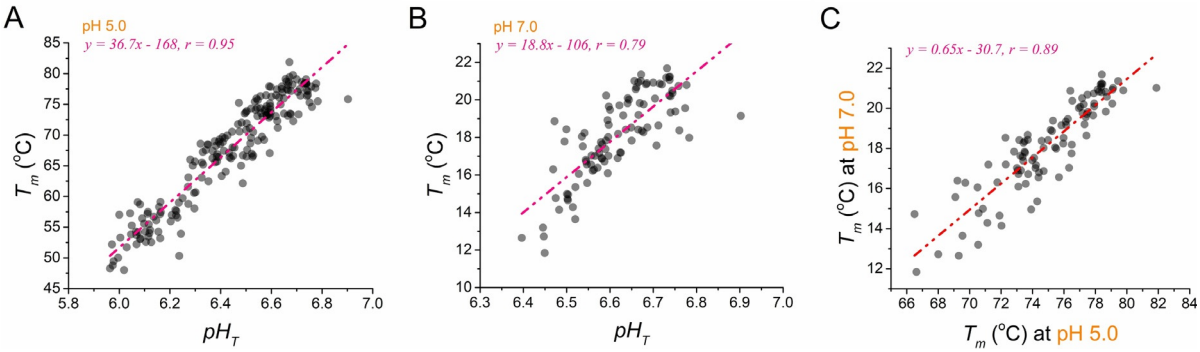

Figure 8. pH_T as function of T_m at A) pH 5.0 or B) pH 7.0. C) T_m at pH 7.0 as a function of T_m at pH 5.0. Linear fits are presented as a red dashed line. Sequences in Table S2 were used.

topology).[21] 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[25,30] is not optimized for i-DNA formation and should be recalibrated for this motif.

Gradient boosting machines (GBM). We built a *de novo* machine learning model to predict the experimental T_m and pH_T, 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 both in terms of T_m and pH_T is the C-tract length. For T_m prediction, the length of the third 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). Unsurprisingly, Eureqa results (see Supporting Information) agree with GBM's in that C-tract length is far more important in predicting both T_m and pH_T of this sub-universe of i-DNAs, as already visible from the plots shown in Figure 4. 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!

Implications for biology. The NMR results suggesting 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 intracellular pH has been reported to vary between 7.0 and 7.4, depending on tissues and phase of the cell cycle.[31] Invasive tumor cells tend to acidify their extracellular environment while keeping their pH_i more alkaline.[32] 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*. The 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.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (21977045), Fundamental Research Funds for the Central Universities (02051430210), and funding from NJU (020514912216). J.L.M. acknowledges ERDF (CZ.02.1.01/0.0/0.0/15_003/0000477) and dedicates this manuscript to the memory of Jean-Louis Leroy, one of the i-DNA pioneers in Ecole Polytechnique. M.C. acknowledges the China Postdoctoral Science Foundation (2019M661793), Lukáš T. acknowledges the Czech Science Foundation (19–26041X), the Ministry of Education, Youth and Sports of the Czech Republic (CIISB research infrastructure project LM2015043; CEITEC 2020, LQ1601). Liezel T. is grateful to the Jardine Foundation and A.B.S. thanks Medical Research Council (UK).

Conflict of interest

The authors declare no conflict of interest.

Keywords: DNA · i-motif · intracellular stability · pH transition · thermal stability

[1] K. Gehring, J.-L. Leroy, M. Guéron, *Nature* **1993**, *363*, 561 – 565.

[2] A. L. Lieblein, M. Kramer, A. Dreuw, B. Furtig, H. Schwalbe, *Angew. Chem. Int. Ed.* **2012**, *51*, 4067 – 4070; *Angew. Chem.* **2012**, *124*, 4143 – 4146.

[3] a) J.-L. Leroy, M. Guéron, J.-L. Mergny, C. Hélène, *Nucleic Acids Res.* **1994**, *22*, 1600 – 1606; b) S. Nonin-Lecomte, J.-L. Leroy, *J. Mol. Biol.* **2001**, *309*, 491 – 506; c) A. T. Phan, M. Guéron, J.-L. Leroy, *J. Mol. Biol.* **2000**, *299*, 123 – 144; d) X. Han, J.-L. Leroy, M. Guéron, *J. Mol. Biol.* **1998**, *278*, 949 – 965; e) K. Snoussi, S. Nonin-Lecomte, J.-L. Leroy, *J. Mol. Biol.* **2001**, *309*, 139 – 153.

[4] a) A. L. Lieblein, J. Buck, K. Schlepckow, B. Furtig, H. Schwalbe, *Angew. Chem. Int. Ed.* **2012**, *51*, 250 – 253; *Angew. Chem.* **2012**, *124*, 255 – 259; b) A. L. Lieblein, B. Furtig, H. Schwalbe, *ChemBioChem* **2013**, *14*, 1226 – 1230.

[5] a) E. P. Wright, J. L. Huppert, Z. A. E. Waller, *Nucleic Acids Res.* **2017**, *45*, 2951 – 2959; b) R. A. Rogers, A. M. Fleming, C. J. Burrows, *Biophys. J.* **2018**, *114*, 1804 – 1815; c) P. Školáková, D. Renčiuk, J. Palacký, D. Krafčík, Z. Dvořáková, I. Kejnovská, K. Bednářová, M. Vorlíčková, *Nucleic Acids Res.* **2019**, *47*, 2177 – 2189.

[6] J.-L. Mergny, L. Lacroix, X. Han, J.-L. Leroy, C. Hélène, *J. Am. Chem. Soc.* **1995**, *117*, 8887 – 8898.

[7] J.-L. Mergny, D. Sen, *Chem. Rev.* **2019**, *119*, 6290 – 6325.

[8] J. J. Alba, A. Sadurni, R. Gargallo, *Crit. Rev. Anal. Chem.* **2016**, *46*, 443 – 454.

[9] K. Leung, K. Chakraborty, A. Saminathan, Y. Krishnan, *Nat. Nanotechnol.* **2019**, *14*, 176 – 183.

[10] M. Debnath, K. Fatma, J. Dash, *Angew. Chem. Int. Ed.* **2019**, *58*, 2942 – 2957; *Angew. Chem.* **2019**, *131*, 2968 – 2983.

[11] a) S. Dzatko, M. Krafcikova, R. Hansel-Hertsch, T. Fessl, R. Fiala, T. Loja, D. Krafcik, J.-L. Mergny, S. Foldynova-Trantirkova, L. Trantirek, *Angew. Chem. Int. Ed.* **2018**, *57*, 2165 – 2169; *Angew. Chem.* **2018**, *130*, 2187 – 2191; b) M. Zeraati, D. B. Langley, P. Schofield, A. L. Moye, R. Rouet, W. E. Hughes, T. M. Bryan, M. E. Dinger, D. Christ, *Nat. Chem.* **2018**, *10*, 631 – 637.

[12] E. Belmonte-Reche, J. C. Morales, *NAR Genom. Bioinform.* **2020**, *2*, lqz005.

[13] X. Li, Y. Peng, J. Ren, X. Qu, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 19658 – 19663.

[14] a) K. Niu, X. Zhang, H. Deng, F. Wu, Y. Ren, H. Xiang, S. Zheng, L. Liu, L. Huang, B. Zeng, S. Li, Q. Xia, Q. Song, S. R. Palli, Q. Feng, *Nucleic Acids Res.* **2018**, *46*, 1710 – 1723; b) H. J. Kang, S. Kendrick, S. M. Hecht, L. H. Hurley, *J. Am. Chem. Soc.* **2014**, *136*, 4172 – 4185.

[15] S. Takahashi, J. A. Brazier, N. Sugimoto, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 9605 – 9610.

[16] a) B. Mir, I. Serrano, D. Buitrago, M. Orozco, N. Escaja, C. Gonzalez, *J. Am. Chem. Soc.* **2017**, *139*, 13985 – 13988; b) I. V. Nesterova, E. E. Nesterov, *J. Am. Chem. Soc.* **2014**, *136*, 8843 – 8846.

[17] a) A. M. Fleming, K. M. Stewart, G. M. Eyring, T. E. Ball, C. J. Burrows, *Org. Biomol. Chem.* **2018**, *16*, 4537 – 4546; b) A. M. Fleming, Y. Ding, R. A. Rogers, J. Zhu, J. Zhu, A. D. Burton, C. B. Carlisle, C. J. Burrows, *J. Am. Chem. Soc.* **2017**, *139*, 4682 – 4689.

[18] A. Sengar, B. Heddi, A. T. Phan, *Biochemistry* **2014**, *53*, 7718 – 7723.

[19] a) S. Benabou, M. Garavis, S. Lyonnais, R. Eritja, C. Gonzalez, R. Gargallo, *Phys. Chem. Chem. Phys.* **2016**, *18*, 7997 – 8004; b) S. M. Reilly, R. K. Morgan, T. A. Brooks, R. M. Wadkins, *Biochemistry* **2015**, *54*, 1364 – 1370; c) I. V. Nesterova, J. R. Briscoe, E. E. Nesterov, *J. Am. Chem. Soc.* **2015**, *137*, 11234 – 11237; d) S. P. Gurung, C. Schwarz, J. P. Hall, C. J. Cardin, J. A. Brazier, *Chem. Commun.* **2015**, *51*, 5630 – 5632; e) T. Fujii, N. Sugimoto, *Phys. Chem. Chem. Phys.* **2015**, *17*, 16719 – 16722; f) M. McKim, A. Buxton, C. Johnson, A. Metz, R. D. Sheardy, *J. Phys. Chem. B* **2016**, *120*, 7652 – 7661.

[20] S. Kendrick, Y. Akiyama, S. M. Hecht, L. H. Hurley, *J. Am. Chem. Soc.* **2009**, *131*, 17667 – 17676.

[21] M. Cheng, Y. Cheng, J. Hao, G. Jia, J. Zhou, J.-L. Mergny, C. Li, *Nucleic Acids Res.* **2018**, *46*, 9264 – 9275.

[22] J.-L. Mergny, J. Li, L. Lacroix, S. Amrane, J. B. Chaires, *Nucleic Acids Res.* **2005**, *33*, e138.

[23] J.-L. Mergny, L. Lacroix, *Nucleic Acids Res.* **1998**, *26*, 4797 – 4803.

[24] J.-L. Mergny, L. Lacroix, *Oligonucleotides* **2003**, *13*, 515 – 537.

[25] A. Bedrat, L. Lacroix, J.-L. Mergny, *Nucleic Acids Res.* **2016**, *44*, 1746 – 1759.

[26] A. B. Sahakyan, V. S. Chambers, G. Marsico, T. Santner, M. Di Antonio, S. Balasubramanian, *Sci. Rep.* **2017**, *7*, 14535.

[27] M. Schmidt, H. Lipson, *Science* **2009**, *324*, 81 – 85.

[28] S. M. Udrescu, M. Tegmark, *Sci. Adv.* **2020**, *6*, eaay2631.

[29] N. Iaccarino, M. Cheng, D. Qiu, B. Pagano, J. Amato, A. D. Porzio, J. Zhou, A. Randazzo, J.-L. Mergny, *Angew. Chem. Int. Ed.* **2021**, https://doi.org/10.1002/anie.202016822; *Angew. Chem.* **2021**, https://doi.org/10.1002/ange.202016822.

[30] a) V. Brázda, J. Kolomaznik, J. Lysek, M. Bartas, M. Fojta, J. Stastny, J.-L. Mergny, *Bioinformatics* **2019**, *35*, 3493 – 3495; b) L. Lacroix, *Bioinformatics* **2019**, *35*, 2311 – 2312.

[31] J. R. Casey, S. Grinstein, J. Orlowski, *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 50 – 61.

[32] a) E. Persi, M. Duran-Frigola, M. Damaghi, W. R. Roush, P. Aloy, J. L. Cleveland, R. J. Gillies, E. Ruppin, *Nat. Commun.* **2018**, *9*, 2997; b) B. A. Webb, M. Chimenti, M. P. Jacobson, D. L. Barber, *Nat. Rev. Cancer* **2011**, *11*, 671 – 677.

Manuscript received: December 18, 2020
Accepted manuscript online: February 18, 2021
Version of record online: March 24, 2021

