

# **i-motif structures in long cytosine-rich sequences found upstream of the promoter region of the SMARCA4 gene**

Sanae Benabou<sup>1</sup>, Anna Aviñó<sup>2</sup>, S. Lyonnais<sup>3</sup>, C. González<sup>4,5</sup>, Ramon Eritja<sup>2</sup>, Anna De Juan<sup>1</sup>, Raimundo Gargallo<sup>1,5\*</sup>

<sup>1</sup> Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Barcelona, Spain

<sup>2</sup> Institute for Advanced Chemistry of Catalonia (IQAC), CSIC, Networking Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain

<sup>3</sup> Department of Structural Biology, Molecular Biology Institute of Barcelona (IBMB-CSIC), Barcelona, Spain

<sup>4</sup> Institute of Physical Chemistry “Rocasolano”, CSIC, Madrid, Spain

<sup>5</sup> BIOESTRAN, associated unit UB-CSIC

\* To whom correspondence should be addressed.

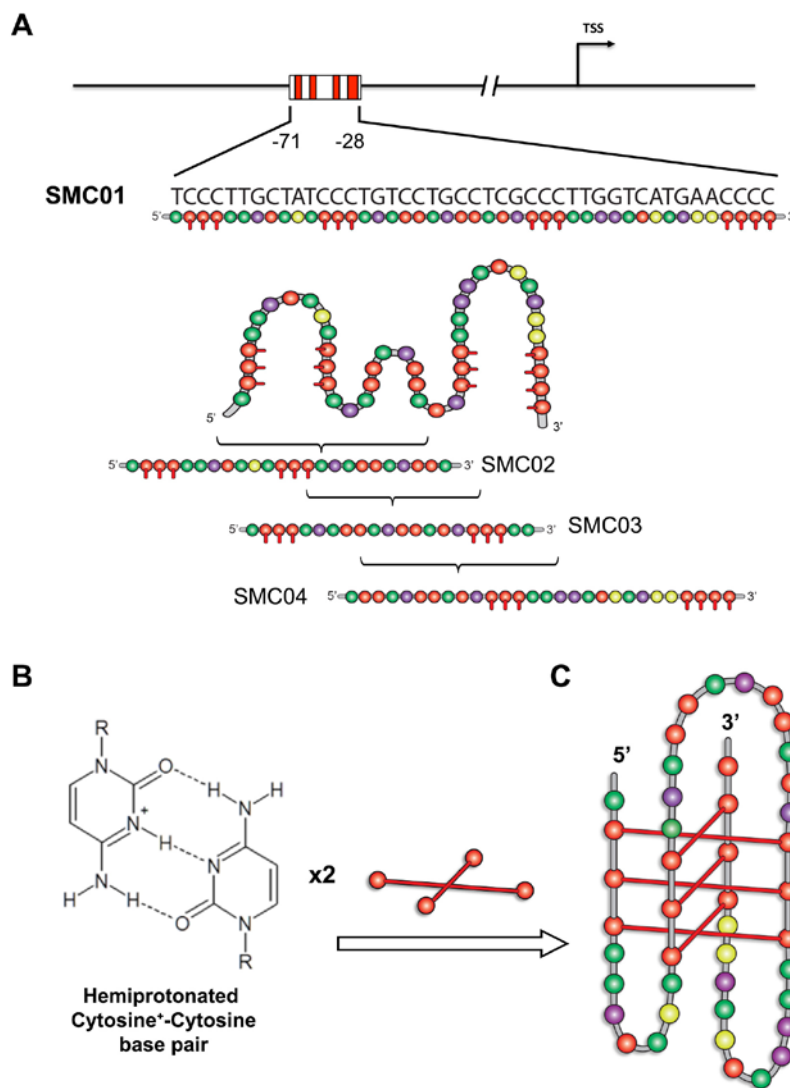
## Abstract

Cytosine-rich oligonucleotides are capable of forming complex structures known as i-motif with increasingly studied biological properties. The study of sequences prone to form i-motifs located near the promoter region of genes may be difficult because these sequences not only contain repeats of cytosine tracts of disparate length but also these may be separated by loops of varied nature and length. In this work, the formation of an intramolecular i-motif structures by a long sequence located upstream of the promoter region of the SMARCA4 gene has been demonstrated. Nuclear Magnetic Resonance, Circular Dichroism, Gel Electrophoresis, Size-Exclusion Chromatography, and multivariate analysis have been used. Not only the wild sequence (5'-TC<sub>3</sub>T<sub>2</sub>GCTATC<sub>3</sub>TGTC<sub>2</sub>TGC<sub>2</sub>TCGC<sub>3</sub>T<sub>2</sub>G<sub>2</sub>TCATGA<sub>2</sub>C<sub>4</sub>-3') has been studied but also several other truncated and mutated sequences. Despite the apparent complex sequence, the results showed that the wild sequence may form a relatively stable and homogeneous unimolecular i-motif structure, both in terms of pH or temperature. The model ligand TMPyP4 destabilizes the structure, whereas the presence of 20% (w/v) PEG200 stabilized it slightly. This finding opens the door to the study of the interaction of these kind of i-motif structures with stabilizing ligands or proteins.

Keywords: *i-motif; hairpin; ligand; SMARCA4; NMR; crowding*

# 1. Introduction

The study of transcriptional regulation by guanine- and cytosine-rich (GC-rich) sequences in promoter regions knows a growing interest for their possible implications in cell proliferation and human diseases [1-3]. G-rich sequences can give rise to G-quadruplex structures of stacked G-quartets associated through Hoogsteen base pairing. Similarly, the complementary C-rich strand can fold back on itself to form an i-motif secondary structure comprising two pairs of parallel strands associated through intercalated hemiprotonated C·C<sup>+</sup> base pairs (Figure 1). Both structures require tracts of repeated guanine or cytosine separated by nucleotide sequences of various lengths that influence their stability [4, 5]. They have the potential to form from local unwinding of B-DNA under negative supercoiling and torsional stress induced by transcription [6].



**Figure 1.** (A) Position of the cytosine-rich element located upstream of the SMARCA4 gene promoter and sequences of the oligonucleotides used in this study. Red, green, purple and yellow circles represent the deoxynucleotides cytosine, thymine, guanine and adenines respectively. Reactive cytosines assembled in tracts of three or four are highlighted in red with a bar. (B) Hemiprotonated cytosine<sup>+</sup>-cytosine base pair form the i-motif structure. (C) Example of a possible intramolecular i-motif from the SCM01 sequence studied in this work.

Bioinformatics and genomic studies have discovered that these dynamic GC-rich elements are concentrated in particular regions of nuclear DNA, specifically upstream of transcriptional start sites and 5'-untranslated regions (5'-UTR) elements or in telomeric regions [7-11]. In promoters, G-quadruplex are proposed to act as molecular switches that turn transcription on/off in association with transcriptional proteins that can facilitate their unwinding or their folding [12]. These alternative way of regulation involving G-quadruplexes may have connection with oncogenic transformation, where a critical protein or proteins can be found with a G-quadruplex in the core or proximal promoter of oncogene, for example *c-myc*, *c-Kit* and *KRAS* (self-sufficiency), *bcl-2* (evasion of apoptosis), *hTERT* (limitless replication), *PDGF-A* (matatstatis) or *VEGF-A* (angiogenesis).

So far, studies of GC-rich sequences have focused principally on the characterization of G-quadruplexes. The i-motif has received less attention due to its strong dependency on pH. The C·C<sup>+</sup> base pair needs the protonation of one of the cytosines at N3, the pK<sub>a</sub> value of which is around 4.5. For this reason the formation of a relatively stable i-motif structure needs a slightly acid environment [13]. At pH values close to 4.5, cytosine bases are partially protonated and the DNA folds into the i-motif structure. When the pH value is higher than the cytosine pK<sub>a</sub>, around 6, the C bases are deprotonated and the i-motif structure unfolds to an unstacked strand. However, negative supercoiling and molecular crowding conditions have been shown to minimize these pH limitations, supporting the possibility that i-motif may form opposite of the G-quadruplex *in vivo* [6, 14] in a mutually exclusive maneer [15, 16]. Indeed, i-motif forming structures have been characterized within several promoter regions, such as *c-myc*, *Rb*, *VEGF*, *EPM1*, or *bcl-2* [5, 17, 18]. In addition, recent studies have reported two transcription regulations mechanisms involving the recognition and unwinding of i-motifs in the *HRAS* and *Bcl-2* promoters by the hnRNP A1 and hnRNP L-like (hnRNP-LL) proteins, respectively [19-21]. Thus, i-motif may also play a role in the transcriptional process mediated by GC-rich sequences and in the associated cancers.

Germline and somatic mutations in the *SMARCA4* gene have been identified in small cell carcinoma of the ovary of hypercalcemic type (SCCOHT), a rare and aggressive cancer affecting children and young women [22]. *SMARCA4* mutations in SCCOHT lead to the loss of the *SMARCA4* SWI/SNF chromatin-remodelling protein, suggesting an important tumor suppressor role for *SMARCA4*. Inspection of the upstream region of the promoter of this gene reveals a wealth of tracts involving more than three cytosine bases that present characteristics of sequences that could fold into i-motif structures (Figure 1A). This prompted us to test such a capability. We chose for this a 44-nucleotides long C-rich sequence (SMC01, Figure 1) located between the positions -71 to -28 upstream of the promoter region of *SMARCA4*. SMC01 contains one tract of four cytosines, three tracts of three cytosines, and two tracts of two cytosines. Hence, it is expected the formation of multiple i-motif conformers in equilibrium including intramolecular structures with large loops. Because of this, three additional sequences (SMC02, SMC03, and SMC04), each one containing four cytosine tracts near the 5' end, in the middle, and near the 3' end of SMC01 sequence, respectively, have been considered (Figure 1 and Table 1). The effect of the presence of a long loop near the 3' end was also studied. Nuclear Magnetic Resonance, Circular dichroism and molecular absorption spectroscopies have been used to determine the conditions of pH and temperature under which i-motif structures are formed. Polyacrylamide Gel Electrophoresis

(PAGE) and Size-Exclusion Chromatography (SEC) have been used to assess the results obtained from spectroscopy. Multivariate data analysis has been used to recover qualitative and quantitative information about the species and conformations present in all experiments.

Name	Sequence (5' → 3')
SMC00	TCAACTTGCTATCAACTGT <u>CC</u> TG <u>CC</u> TCGCAACTTGGTCATGAACAAC
SMC01	<u>TCCC</u> TTGCTAT <u>CCC</u> TGT <u>CC</u> TG <u>CC</u> TCG <u>CCC</u> TTGGTCATGAAC <u>CCCC</u>
SMC01_4x4C	<u>TCCCC</u> TTGCTAT <u>CCCC</u> TGT <u>CC</u> TG <u>CC</u> TCG <u>CCCC</u> TTGGTCATGAAC <u>CCCC</u>
SMC01_4x4C_AA	<u>TCCCC</u> TTGCTAT <u>CCCC</u> TGTAATGAATCG <u>CCCC</u> TTGGTCATGAAC <u>CCCC</u>
SMC02	<u>TCCC</u> TTGCTAT <u>CCC</u> TGT <u>CC</u> TG <u>CC</u> T
SMC02s	<u>TCCC</u> TTGCTAT <u>CCC</u> T
SMC03	<u>TCCC</u> TGT <u>CC</u> TG <u>CC</u> TCG <u>CCC</u> TT
SMC04	<u>TCC</u> TG <u>CC</u> TCG <u>CCC</u> TTGGTCATGAAC <u>CCCC</u>
TT	TT <u>CCCC</u> TTT <u>CCCC</u> TTT <u>CCCC</u> TTT <u>CCCC</u> TT

**Table 1.** DNA sequences studied in this work.

## 2. Material and methods

### 2.1. Reagents

DNA sequences (Table 1) were synthesized on an Applied Biosystems 3400 DNA synthesizer using the 1  $\mu$ mol scale synthesis cycle. Standard phosphoramidites were used. Ammonia deprotection was performed overnight at 55 °C. The resulting products were purified using Glen-Pak Purification Cartridge (Glen Research, USA). Purified oligonucleotides were characterized by MALDI-TOF MS Spectrometry (Figure S1). DNA strand concentration was determined by absorbance measurements (260 nm) at 90 °C using the extinction coefficients calculated using the nearest-neighbor method as implemented on the IDT analyzer webpage [<https://eu.idtdna.com/calc/analyzer>]. Before any experiment, DNA solutions were first heated to 90 °C for 10 min and then allowed to reach room temperature. KCl, KH<sub>2</sub>PO<sub>4</sub>, K<sub>2</sub>HPO<sub>4</sub>, NaCH<sub>3</sub>COO, HCl and NaOH were purchased from Panreac (Spain). MilliQ® (Millipore, USA) water was used in all experiments.

## 2.2. Procedures

NMR spectra were acquired in a Bruker Advance spectrometer operating at 600 and 800 MHz and equipped with a cryoprobe. Water suppression was achieved by the inclusion of a WATERGATE [23] module in the pulse sequence prior to acquisition. Samples for NMR spectra were suspended in 9:1 H<sub>2</sub>O/D<sub>2</sub>O (20 mM buffer) in presence of 150 mM KCl. Absorbance spectra in the UV-VIS region were recorded on an Agilent 8453 diode array spectrophotometer. The temperature was controlled by means of an 89090A Agilent Peltier device. CD spectra were recorded on a Jasco J-810 spectropolarimeter equipped with a Julabo F-25/HD temperature control unit. Quartz cells (1 or 10 mm path length, 300, 1400 and 3000 µl volume) were used (Hellma, Germany).

Acid-base titrations were carried out by adjusting the pH of solutions containing the oligonucleotides. pH was measured using an Orion SA 720 pH/ISE meter and a micro-combination pH electrode (Thermo, USA). Titrations were carried out at 25 °C and 150mM KCl. For melting experiments, the DNA solution was transferred to a covered 10-mm-path-length cell and absorption spectra were recorded at 1°C intervals with a hold time of 3 min at each temperature, which yielded an average heating rate of approximately 0.3 °C·min<sup>-1</sup>. Buffer solutions were 20 mM phosphate or acetate, and 150 mM KCl. Each sample was heated at 90 °C for 5 min and let cool overnight (before starting the experiments).

For PAGE assays, the oligonucleotides were suspended in water and diluted at a concentration of 5 µM in a solution buffered at pH 5.0 (50 mM sodium acetate, 50 mM potassium acetate) or at pH 8.3 (50 mM Tris acetate, 50 mM potassium acetate). The solutions were next boiled for 5 min at 95 °C and cooled down to room temperature overnight. Glycerol was added (2.5 % v/v) and 5 µL of the samples were loaded on 10 x 10.5 cm – 11 % non-denaturing polyacrylamide gels (19/1 acrylamide / bisacrylamide, Sigma, USA) either buffered with 50 mM Tri-sodium citrate adjusted to pH 5.0 with citric acid or with 0.5 x Tris pH 8.3 Borate EDTA (TBE, Sigma-Aldrich, USA). Electrophoresis was run on ice with a miniVE apparatus (Hoeffer, USA) for 3 h at 11 V·cm<sup>-1</sup>. 0.1 M Tri-sodium citrate pH 5.0 was used a running buffer for the acidic gel and 0.5 x TBE for the other. The gel buffered at pH 5.0 was then incubated 15 min in 0.5 x TBE and both gels were stained with SybrGold (Molecular Probes, USA) and digitalized with a Typhoon 8600 system (Molecular Dynamics, USA).

The chromatographic system consisted of an Agilent 1100 Series HPLC instrument equipped with a G1311A quaternary pump, a G1379A degasser, a G1392A autosampler, a G1315B photodiode-array detector furnished with a 13 µL flow cell, and an Agilent Chemstation for data acquisition and analysis (Rev. A 10.02), all from Agilent Technologies (Waldbronn, Germany). A BioSep-SEC-S 3000 column (300 × 7.8 mm, particle size 5 µm and pore size 290 Å) from Phenomenex (Torrance, USA) was used for the chromatographic separation at room temperature. The mobile phase was 300 mM KCl and 20 mM buffer (phosphate or acetate) adjusted to desired pH value. The flow rate was set to 1.0 mL·min<sup>-1</sup>. The injection volume was 15 µL. Absorbance spectra were recorded between 200 and 500 nm. Poly d(T) markers (T<sub>15</sub>, T<sub>20</sub>, T<sub>25</sub>, T<sub>30</sub>, and T<sub>45</sub>) sequences were used as standards to construct the t<sub>R</sub> vs. log(MW) calibration plot

for unfolded sequences. Standards were injected twice to assess the reproducibility of the  $t_R$  values. At all pH values studied, the relative difference between  $t_R$  values for a given standard was lower than 0.5 %.

Measurements in media simulating crowding conditions were done using PEG200. In this case, the appropriate weight of PEG200 was added to the aqueous DNA solution containing 20 or 100 mM acetate or phosphate buffer and 150 mM KCl. pH was measured before and after the addition of PEG200 with electrodes previously calibrated using aqueous buffers. No strong acid or base were used to modify the pH after the addition of PEG200.

### 2.3. Data analysis

For melting experiments, absorbance data as a function of temperature were analyzed as described elsewhere [24]. The physico-chemical model is related to the thermodynamics of DNA unfolding. Hence, for the unfolding of intramolecular structures such as those studied here, the chemical equation and the corresponding equilibrium constant may be written as:



For melting experiments, the concentration of the folded and unfolded forms is temperature-dependent. Accordingly, the equilibrium constant depends on temperature according to the van't Hoff equation:

$$\ln K_{\text{unfolding}} = - \Delta H_{\text{vH}} / RT + \Delta S_{\text{vH}} / R \quad (4)$$

It is assumed that  $\Delta H_{\text{vH}}$  and  $\Delta S_{\text{vH}}$  will not change throughout the range of temperatures studied here.

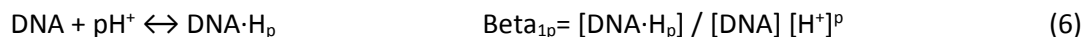
Spectra recorded during acid-base experiments were arranged in a table or data matrix **D**, with  $m$  rows (spectra recorded) and  $n$  columns (wavelengths measured). The goal of data analysis was the calculation of distribution diagrams and pure (individual) spectra for all  $nc$  spectroscopically-active species considered throughout an experiment. The distribution diagram provides information about the pH stability of the species considered, as well as about the cooperativity of the transitions. In addition, the shape and intensity of the pure spectra may provide qualitative information about the structure of the species. With this goal in mind, data matrix **D** was decomposed according to Beer-Lambert-Bouguer's law in matrix form:

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T + \mathbf{E} \quad (5)$$

where **C** is the matrix ( $m \times nc$ ) containing the distribution diagram, **S<sup>T</sup>** is the matrix ( $nc \times n$ ) containing the pure spectra, and **E** is the matrix of data ( $m \times n$ ) not explained by the proposed decomposition.

The mathematical decomposition of **D** into matrices **C**, **S<sup>T</sup>**, and **E** may be conducted in two different ways, depending on whether a physico-chemical model is initially proposed (hard-modeling approach) or not (soft-modeling approach) [25]. For hard-modeling approaches, the proposed model depends on the nature of the process under study. Hence, for acid-base experiments the model will include a set of chemical equations describing the formation of the different

acid-base species from the neutral species, together with approximate values for the stability constants, such as the following:



In this equation, the parameter  $p$  is related to the Hill coefficient and describes qualitatively the cooperativity of the equilibrium. Values of  $p$  greater than 1 indicate the existence of a cooperative process. Equation 6 is applied to each one of the acid-base transitions present in the pH range studied. From the calculated equilibrium constants and  $p$  parameters, the pH-transition midpoints are calculated. Whenever a physico-chemical model is applied, the distribution diagram in **C** complies with the proposed model. Accordingly, the proposed values for the equilibrium constants and the shape of the pure spectra in **S<sup>T</sup>** are refined to explain satisfactorily data in **D**, whereas residuals in **E** are minimized. In this study, hard-modeling analysis of acid-base experiments used the EQUISPEC program [26].

### 2.3.1. Mathematical comparison of calculated spectra

Appart from visually, it is possible to calculate mathematically the similarity of the calculated pure spectra in **S<sup>T</sup>** matrix with those of known structures, like an homogeneous i-motif, a partially stacked single strand, or even a random coil. In this work, the similarity is given by the sine of the angle between two vectors according to the following equation:

$$\sin \alpha = (1 - \cos^2 \alpha)^{1/2} \quad (7)$$

where:

$$\cos \alpha = S_{\text{calculated}} \cdot S_{\text{known}} / ( \|S_{\text{calculated}}\| \cdot \|S_{\text{known}}\| ) \quad (8)$$

The expression “ $\cdot$ ” denotes the scalar product between the calculated spectrum and that of the known structure. The expression “ $\|S_{\text{calculated}}\|$ ” denotes the norm of the calculated spectrum, computed as the square root of the sum of all squared elements in  $S_{\text{calculated}}$ . Using the dissimilarity value ( $\sin \alpha$ ) instead of the similarity value ( $\cos \alpha$ ) provides higher discrimination power for very similar spectra. When the calculated pure spectrum and the known one have exactly the same value,  $\cos \alpha$  is equal to one and  $\sin \alpha$  is equal to zero. A good correlation in shape between two spectra is obtained when the dissimilarity value is lower than 0.0141, which corresponds to a correlation between them greater than 0.99990 [27].



## 3. Results

### 3.1. Preliminary *in silico* study

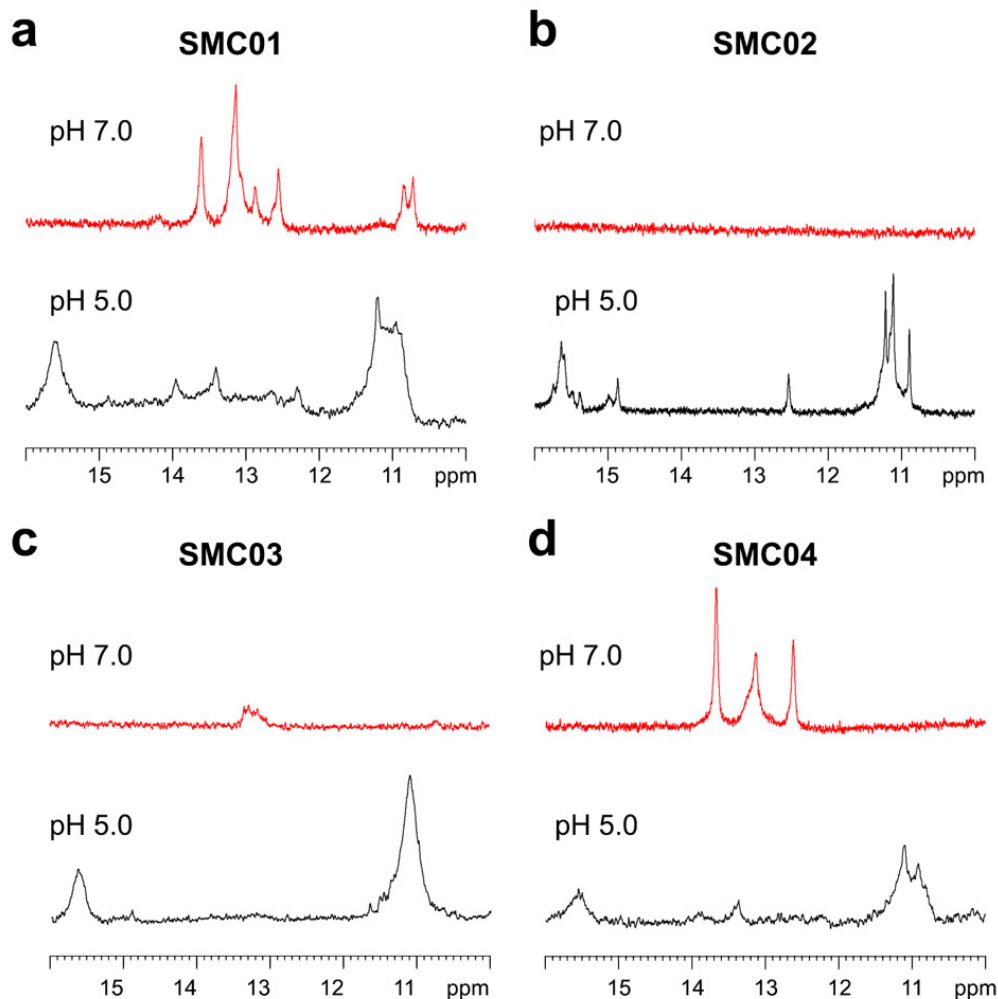
It is possible to predict the formation of intramolecular or intermolecular structures associated by Watson-Crick base pairs by using *in silico* calculations. Therefore, previously to the experimental part of this work, an estimation of the formation of potential intra- and intermolecular structures at neutral pH values was done by using the *mfold* method[28] stored in a web server [<http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form>]. The summary of the predictions is shown in Table S1.

The proposed intermolecular structure for SMC01 at neutral pH shows the potential formation of six base pairs. It should be noted that all these pairs involve bases located near the 3' end, i.e., bases that are also present in the truncated sequence SMC04, but not in SMC02 nor SMC03 (Table 1). Accordingly, SMC04 could also potentially form intermolecular structures. On the other hand, all sequences could form intramolecular structures. As expected, those formed by SMC01 and SMC04 sequences are the most stable because they involve the formation of three Watson-Crick base pairs located near the 3' end. Hairpins formed by SMC02 and SMC03 are very unstable because they are only stabilized by two base pairs. At the DNA concentrations used in typical CD and molecular absorbance measurements, intramolecular structures are expected to be predominant, whereas intermolecular structures could be potentially formed in NMR measurements.

### 3.2. Evidences of i-motif formation at acidic pH values

Figure 2 shows the exchangeable proton regions of the NMR spectra of SMC01-SMC04 sequences at pH 7.0 and 5.0. At pH 7.0, SMC01 and SMC04 showed imino peaks between 12 and 14 ppm, indicative of Watson-Crick base pairs. The number of peaks was consistent with the number of Watson-Crick base pairs in the hairpin structures suggested by *in silico* calculations. In addition, the presence of mismatched base pairs (probably G-T) was observed in the spectra of SMC01 (signals around 10-11 ppm). It must be noted the absence of signals around 15.5 ppm, indicative of protonated imino groups in C·C<sup>+</sup> base pairs, the building block of the i-motif structure. On the contrary, at pH 5.0, all sequences showed signals around 15.5 ppm, suggesting the formation of that structure. The sequence SMC02 showed sharper signals in this region, suggesting the formation of a homogeneous structure. On the other side, SMC01 and SMC03 showed broad peaks around 15.5 ppm, indicating the presence of multiple conformers in equilibrium. All sequences showed signals around 11 ppm that could be indicative of T·T or G·T base pairs. These base pairs are common in the loop residues connecting C-tracts in i-motif structures [29, 30]. Finally, the presence of potential Watson-Crick base pairs (signals between 12.5-14.0 ppm) in the folded structure at pH 5.0 was observed for SMC02 and, in minor extension, for SMC01 and SMC04.

NMR spectra recorded at different temperatures indicated that the i-motifs are not very stable (Figure S2), being the SMC02 more stable than the others, and SMC04 only marginally stable (i-motif signals are only observed at T= 5°C, data not shown).

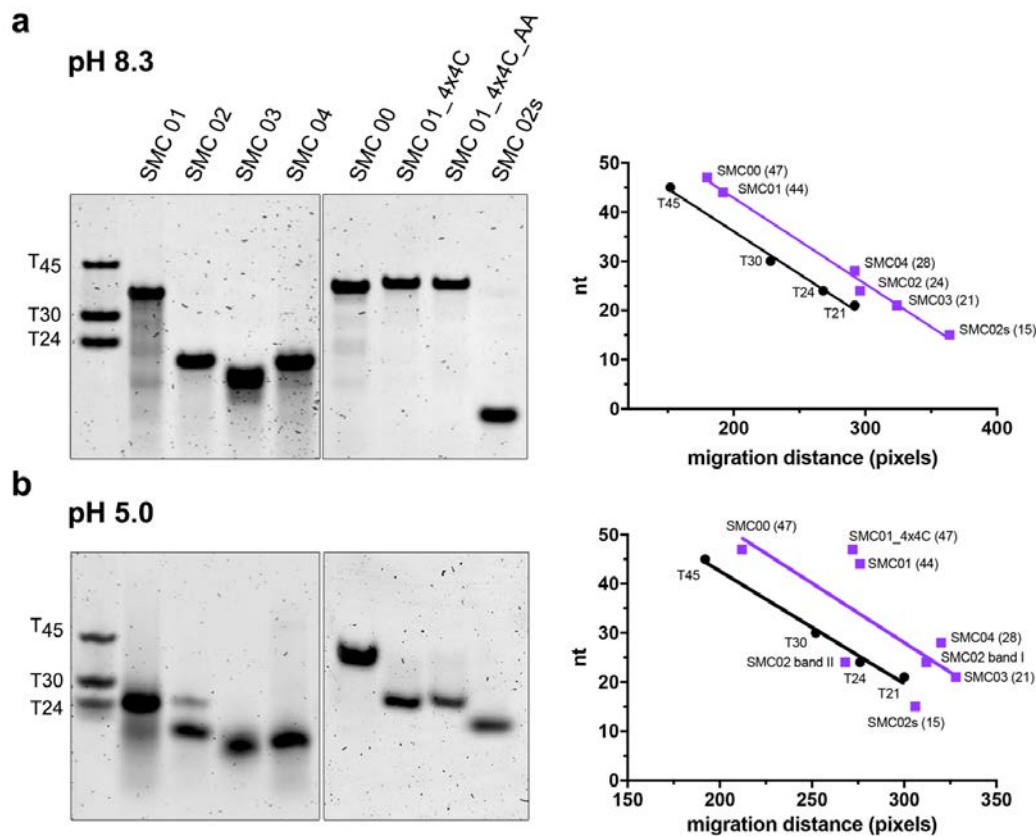


**Figure 2.**  $^1\text{H}$  NMR spectra of SMC01 (a), SMC02 (b), SMC03 (c), and SMC04 (d) sequences at pH 7.0 (in red) and 5.0 (in black). The experimental conditions were 0.3 mM DNA, 100 mM KCl, 20 mM phosphate or acetate buffer, 5 °C.

### 3.3. Molecularity of the formed species

PAGE and SEC experiments were next performed to gain information about the molecularity and structure of SMC01 and variants. The PAGE mobility of the oligonucleotides was compared at pH 5.0 and pH 8.3 after incubation in buffers at the same pH (Figure 3). Poly-d(T) sequences were used as migration markers. At pH 8.3, the four SMC0x sequences showed a marked different mobility with the polyd(T) homopolymers, as noted [31], but migrated linearly as a function of their size in parallel with the calibration line of the polyd(T) markers. This result indicates the absence of folding for these sequences at neutral pH. SMC01 showed additional low intensity bands of faster mobility in a trail of blurry species, suggesting a minor population of intramolecular folded species as predicted *in silico*. At pH 5.0, the poly d(T) markers mobility remained linear as a function of length, as expected (Figure 3B). The SMC01 and SMC04 sequences showed a faster mobility indicative of a compact, folded structure characteristic of i-motifs in these pH conditions [32]. The major band of SMC01 also associated smearing species and a low intensity band migrating slightly faster, confirming the structural fluctuation detected by NMR. The intramolecular folding of SMC01 was further confirmed by substituting the C<sub>3</sub> tracts of the sequence by series of C<sub>4</sub> in order to increase the

stability of the i-motif (SMC01\_4x4C, see Table 1). As expected, SMC01\_4x4C showed a similar behavior than SMC01, characterized by a faster mobility at pH 5 as compared with the unfolded sequence at pH 8.3. The stabilized structure also presented less structural fluctuation at pH 5.0. At the opposite, replacement of the C<sub>4</sub> tracts by CAAC in SMC00 (Table 1) cancelled the i-motif formation and restored the mobility of a non-folded structure. Substitution of the two CC repetitions in the central loop in SMC01\_4x4C\_AA showed a migration pattern similar to those of SMC01 and SMC01\_4x4C, suggesting that the CC tracts in the central loop participate marginally in the stabilization of the folding. The mobility of SMC00, SMC02 and SMC03 was seen linear and parallel with the polyd(T) markers, like the experiment at pH 8.3. This indicates an absence of folding or/and very weak secondary structures for both SMC02 and SMC03 in these conditions. A minor band of higher mobility was detectable for SMC02, suggesting possible dimeric structures. This dimer was confirmed by the migration of a truncated SMC02s (5'-TCCCTTGCTATCCT-3'), which showed a unique band of retarded migration at pH 5.0, close to the position of the unfolded 24-mer SMC02.



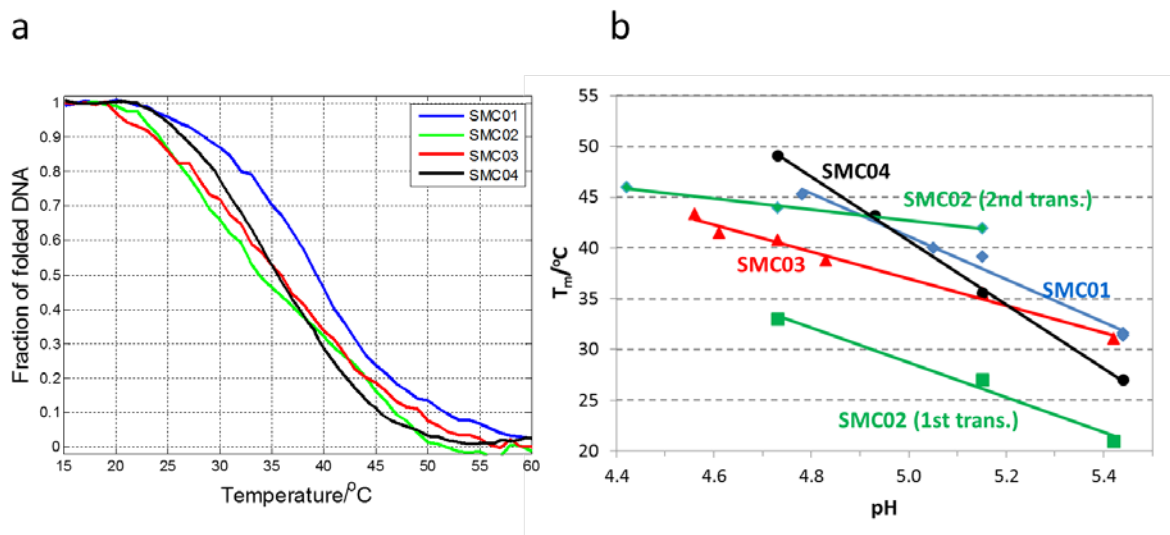
**Figure 3.** PAGE of SMC01 and variants in native conditions at pH 8.3 (A) and 5.0 (B). The oligonucleotides sequences are reported in Table 1. The electrophoresis was run on ice, the bands stained with Sybr Gold. T<sub>45</sub>, T<sub>30</sub> and T<sub>24</sub> were used as migration ladder. The average migration of each oligonucleotide *versus*. nucleotide size is reported on the right from two independent experiments including Poly-d(T)<sub>45</sub>, 30, 24 and 21 markers. The line between points was fitted using a linear regression model.

Size-Exclusion Chromatography (SEC) measurements were carried out at in 300 mM KCl concentration to reduce the ionic exclusion mechanism at pH values higher than the  $pK_a$  of silanol groups (around 4-6 [33]). The experiment was calibrated using the retention times ( $t_R$ ) values of the polyd(T) oligonucleotides ( $T_{15}$ ,  $T_{20}$ ,  $T_{25}$ ,  $T_{30}$  and  $T_{45}$ ). At pH 7.0 all the samples eluted mainly as single bands with  $t_R$  between 8.95 (for SMC01, the longest sequence) and 9.68 (for SMC03, the shortest sequence) minutes (Figure S3). Reported to the calibration plot, the experimental and calculated  $t_R$  values of the SMC0x sequences matched with a relative difference around 1 % (Table S2), confirming the unfolded state of these sequences at pH 7. At pH 5.3, both  $T_x$  standards and samples eluted at slightly higher  $t_R$  values than at pH 7.0. All SMC0x sequences eluted mainly as a single and broad band. The comparison of the measured  $t_R$  values with those calculated for the unfolded strands using the calibration plot revealed unambiguously that SMC01 eluted as a more compact structure, confirming its folding into an intramolecular i-motif structure [34]. We also noted that SMC01 eluted as a broad band at both pH values, confirming the existence of several topologies in equilibrium suggested by the broad band observed around 15.5 ppm by NMR and with the smear bands observed by PAGE. The SMC02-04 sequences also eluted as slightly more compact structures, suggesting possible i-motif structures in agreement with the presence of C·C<sup>+</sup> base pairs detected by NMR.

### 3.4. Thermally-induced unfolding

Spectrophotometrically-monitored melting experiments were carried out to gain information about the stability of the folded structures in front of pH and temperature changes. Surprisingly, all sequences, except SMC02, showed a two-state unfolding process in the pH range studied (4.5 – 5.5, approximately) (Figure 4a). Annealing experiments showed a 1 °C difference in the folding and unfolding traces, which was considered as a proof the reversibility of the process (Figure S4). These rather simple unfolding processes are reflected in the plot of the first derivative of  $A_{295}$  (Figures S5-S8). Even in the case of SMC01 sequence, which was shown to form minor conformers by NMR and PAGE, multivariate analysis revealed that the unfolding can be successfully modelled using a two-state mechanism (Figure S5). This fact could be related to a similar thermal stability of the i-motif conformers formed by this sequence. Consequently,  $T_m$  values and thermodynamic parameters such as the changes in enthalpy and entropy could be calculated assuming a simple two-state process for SMC01, SMC03 and SMC04 sequences (Figure 4b and Table 2).

In general, all structures showed lower thermal stabilities in comparison to other i-motif structures previously described, such as those observed in genomic sequences as *bcl-2* or *c-myc* (this will be discussed below). At pH 5.4, SMC01 and SMC03 showed similar  $T_m$  values, whereas SMC04 showed a slightly lower  $T_m$  value. At pH values 4.7, however, SMC04 showed the highest  $T_m$  values, whereas SMC03 showed the lowest values. Overall, the SMC01 and SMC03 sequences showed a rather similar dependence of  $T_m$  in front of pH, being the folded structure formed by SMC01 slightly more stable than that of SMC03 in terms of  $T_m$  values. The SMC04 sequence, on the other hand, showed a higher dependence of  $T_m$  with pH. This stronger dependence could be due to the formation of additional C·C<sup>+</sup> base pairs when the pH was decreased gradually from 5.4 to 4.7, reflecting a low cooperativity in the formation of these base pairs. In terms of  $\Delta G^0_{370C}$ , the most stable structure at pH 5.2 was that formed by SMC01 sequence (Table 2). At pH 4.7, both SMC01 and SMC04 showed similar  $\Delta G^0_{370C}$  values.



**Figure 4.** Thermally-induced unfolding. (a) Fraction of folded DNA calculated from the absorbance trace at 295 nm. The experimental conditions were 2  $\mu$ M DNA, 150 mM KCl, 20 mM acetate buffer, pH 5.2. (b) Plot of determined  $T_m$  values versus pH.

Sequence	pH 4.7				pH 5.2			
	$T_m$ (°C)	$\Delta H^0$ (kcal·mol <sup>-1</sup> )	$\Delta S^0$ (cal·K <sup>-1</sup> ·mol <sup>-1</sup> )	$\Delta G^0_{370c}$ (kcal·mol <sup>-1</sup> )	$T_m$ (°C)	$\Delta H^0$ (kcal·mol <sup>-1</sup> )	$\Delta S^0$ (cal·K <sup>-1</sup> ·mol <sup>-1</sup> )	$\Delta G^0_{370c}$ (kcal·mol <sup>-1</sup> )
SMC01	45	-40.5	-127.1	-1.0	38	-39.0	-124.9	-0.3
SMC02	33 / 44	n.d.	n.d.	n.d.	27 / 42	n.d.	n.d.	n.d.
SMC03	41	-27.9	-88.9	-0.3	36	-30.7	-99.5	0.1
SMC04	49	-26.1	-80.7	-1.0	36	-32.6	-105.7	0.1

n.d.: Not determined

Uncertainty in  $T_m$  values was  $\pm 0.7^\circ\text{C}$ . Uncertainties in  $\Delta H^0$  and  $\Delta S^0$  values were lower than  $\pm 5\%$ . Uncertainty in  $\Delta G^0$  values was lower than  $\pm 10\%$ .

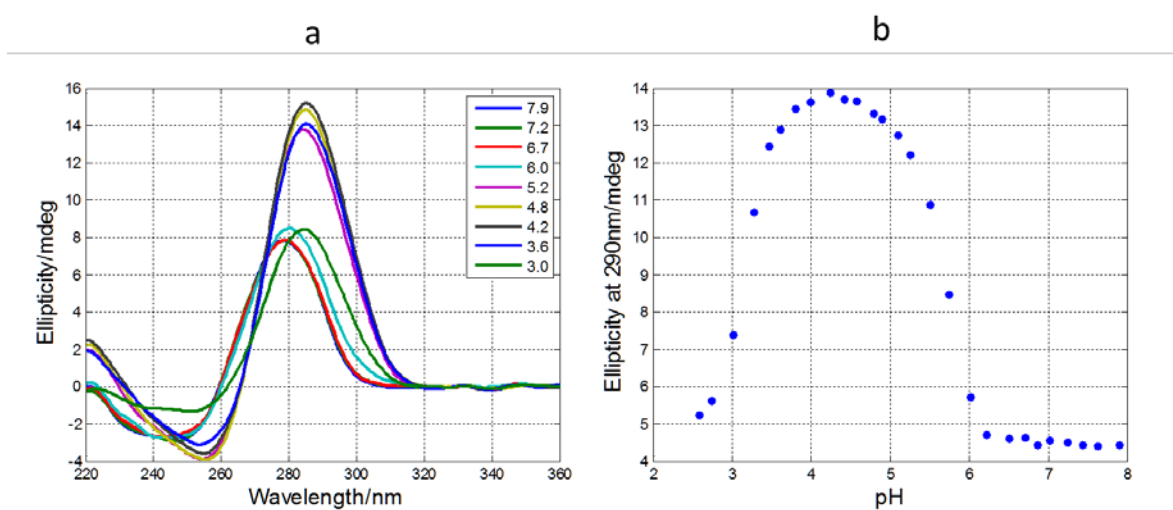
**Table 2.** Thermodynamic parameters calculated for the unfolding processes. A two-state model has been applied.

The SMC02 sequence showed a different behavior. Recently prepared samples showed two transitions with  $T_m$  values in the ranges 22-33 °C and 40-50 °C (for both of them,  $T_m$  values were dependent on pH). The relative magnitude of these two unfolding processes was also pH dependent: whereas the transition described by the lower  $T_m$  was predominant at pH 5.2 and 5.4, the transition described by the higher  $T_m$  was more clear at pH 4.4 (Figure S6). Also, when a sample containing 2  $\mu$ M DNA was kept at 4 °C and pH 5.2, time-dependent disappearance of the first transition together with an enhancement of the second transition were observed. The molecularity of this transition was studied from the melting studies of samples whose DNA concentration ranged from 2 to 60  $\mu$ M at pH 5.2. The small variation of  $T_m$  in front of DNA concentration did not bring light to the multimer nature of the transition. To clarify this point, the melting curves of the SMC02s sequence were also recorded. Now, melting curves showed a monophasic transition with  $T_m = 40^\circ\text{C}$  at pH 5.2. As PAGE experiments already pointed to the formation of

a dimer, it was finally proposed that SMC02 forms a mixture of dimer ( $T_m = 45\text{ }^\circ\text{C}$ ) and monomer ( $T_m = 27\text{ }^\circ\text{C}$ ) at pH 5.2. At pH values higher than 5.2, the dimer predominates, whereas the monomer is the major form at pH values lower than 4.4. The higher  $T_m$  value for the dimer in SMC02 ( $45\text{ }^\circ\text{C}$ ) in relation to that in the shorter sequence (SMC02s,  $40\text{ }^\circ\text{C}$ ) has been related to the existence of additional hydrogen bonds in the overhangs (5'-GTCCTGCCT-3') present in SMC02.

### 3.5. Circular dichroism and pH-dependent formation

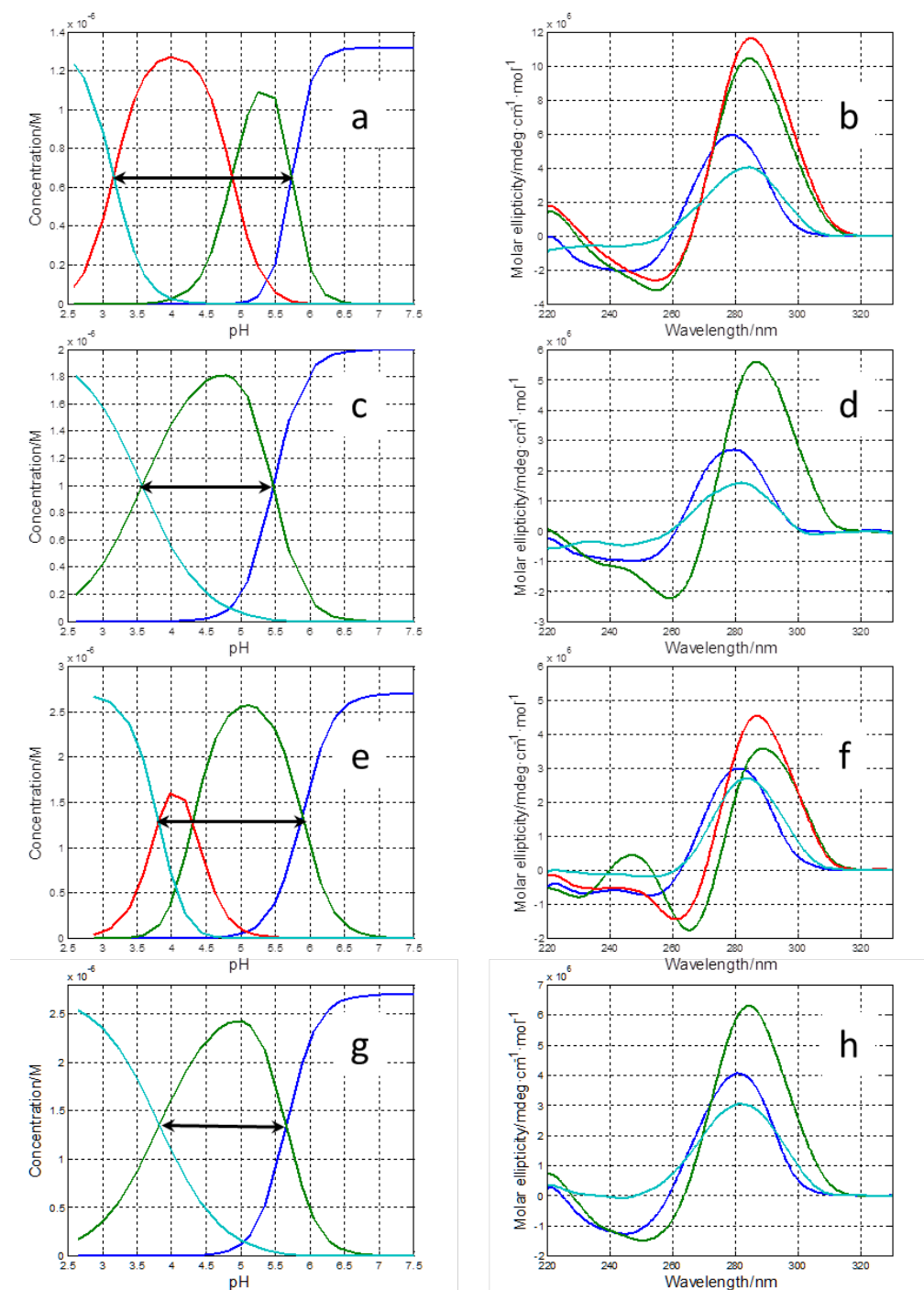
The typical CD spectrum of an i-motif structure shows two characteristic negative and positive bands centered at  $\sim 260$  and  $\sim 285$  nm, respectively, being the second band two times more intense than the first one [35]. In the case of SMC01, the position of the negative and positive bands (at 255 and 286 nm, respectively) confirmed the formation of an i-motif structure at pH 5.2 (Figure 5a). To study the influence of pH on the conformational equilibria of SMC0X sequences, we next carried out CD-monitored acid–base titrations at  $25\text{ }^\circ\text{C}$  in the pH range 2.5–7.5. Figure 5a shows some of the experimental spectra recorded along the acid-base titration of SMC01. The complete set of experimental spectra are shown in Figure S9b. The plot of ellipticity versus pH (Figure 5b) showed two steeper slopes at pH values around 5.5 and 3.5, respectively, which denoted folding processes characterized by a strong positive cooperativity. The pH range where the formation of i-motif takes place was narrow, being its maximum formation at pH equal to 4.2.



**Figure 5.** Acid-base titration of SMC01 sequence. (a) Selected experimental CD spectra. The complete set of CD spectra measured along the titration are included as Supplementary Information. (b) Variation with pH of ellipticity values measured at 290nm. The experimental conditions were  $2\text{ }\mu\text{M}$  DNA,  $150\text{ mM}$  KCl,  $25\text{ }^\circ\text{C}$ .

A multivariate analysis method was applied to gain more information about the acid-base properties of this sequence [26]. This approach has been shown previously to be an excellent tool to identify the number of acid-base species present in a given pH range, to determine the pH regions where these folded structures do exist, as well as to characterize the cooperativity accompanying their formation [36, 37]. Mathematically, the set of experimental spectra (matrix **D** in equation 5) is decomposed in matrices **C** (pH diagram of species distribution), **S** (pure CD spectra), and **E** (experimental data not explained by the proposed model by **C** and **S**). The calculated pH-transition

midpoints and  $p$  parameters that rule the pH diagram of species distribution are summarized in Table 3. Figure 6 shows the pH diagrams of species distribution, as well as the calculated pure CD spectra. Plots of residuals in matrix E are shown in Figure S9c-f and discussed below.



**Figure 6.** pH diagrams of species distribution (matrices C) and pure spectra (matrices S) calculated from the acid-base titrations of the SMC01 (a, b), SMC02 (c, d), SMC03 (e, f) and SMC04 (g, h) sequences after application of multivariate analysis. The titrations were carried out in 150 mM KCl and at 25 °C. Blue line: neutral form; green line: i-motif 1; red line: i-motif 2; cyan line: protonated form. The black line denotes the pH range where i-motif structures predominate (Table 3).

In the case of SMC01, the first acid-base species (depicted in blue in Figure 6a) was present at pH values higher than 5.0, reaching its maximum concentration at pH values higher than 6.5, approximately. Accordingly, this acid-base species contains all nitrogenated bases in their respective neutral forms. The calculated CD pure spectrum for this acid-base species suggested a certain degree of ordenation at low temperatures because both, intensity and shape are slightly modified upon heating (Figure S9a). Therefore, its secondary structure was probably that of a hairpin, like that proposed from previous *in silico* calculations. There are two acid-base species that reach maximum concentrations at pH values around 5.2 (green) and 4.0 (red), respectively. The calculated pure CD spectra of these two acid-base species showed negative and positive bands around 255 and 285 nm, approximately. These features were similar, but not equal, to those traditionally assigned to i-motif structures. Accordingly, and considering the observed NMR signal around 15.5 ppm, these two acid-base species would correspond to folded forms stabilized by C·C<sup>+</sup> base pairs. Taking into account the low T<sub>m</sub> values determined at pH 4.8 (Table 2) and the temperature at which the acid-base titration was carried out (25 °C), these “pure” species were, in fact, a mixture of folded i-motif and unfolded strands. The differences between both “pure” species would be due to the protonation of some bases in the red species (the apparent pK<sub>a</sub> being equal to 4.9) that did not affect to the overall secondary structure. Finally, the fourth acid-base species appearing at pH values lower than 4.5 (cyan) was assigned to that species in which all (or almost all) cytosine and adenine (whose pK<sub>a</sub> value is around 3.5) bases are protonated. The secondary structure is almost completely lost, as denoted by the calculated CD spectrum.

Sequence	pH-transition midpoint ( $\rho$ parameter) <sup>a</sup>	pH range where i-motif structures predominate <sup>b</sup>
SMC01	5.7 ± 0.1 (3)	5.7-3.2
	4.9 ± 0.1 (2)	
	3.2 ± 0.1 (2)	
SMC02	5.5 ± 0.1 (2)	5.5-3.6
	3.6 ± 0.1 (1)	
SMC03	5.9 ± 0.1 (2)	5.9-3.8
	4.3 ± 0.1 (2)	
	3.8 ± 0.1 (2)	
SMC04	5.7 ± 0.1 (2)	5.7-3.8
	3.8 ± 0.1 (1)	

<sup>a</sup> This parameter describes the cooperativity of the transition according to Equation 6.

<sup>b</sup> Calculated from the difference between the first and last pH-transition midpoints in Figures 6a, 6c, 6e and 6g, respectively.

**Table 3.** Calculated parameters associated to the acid-base titrations of the studied sequences.

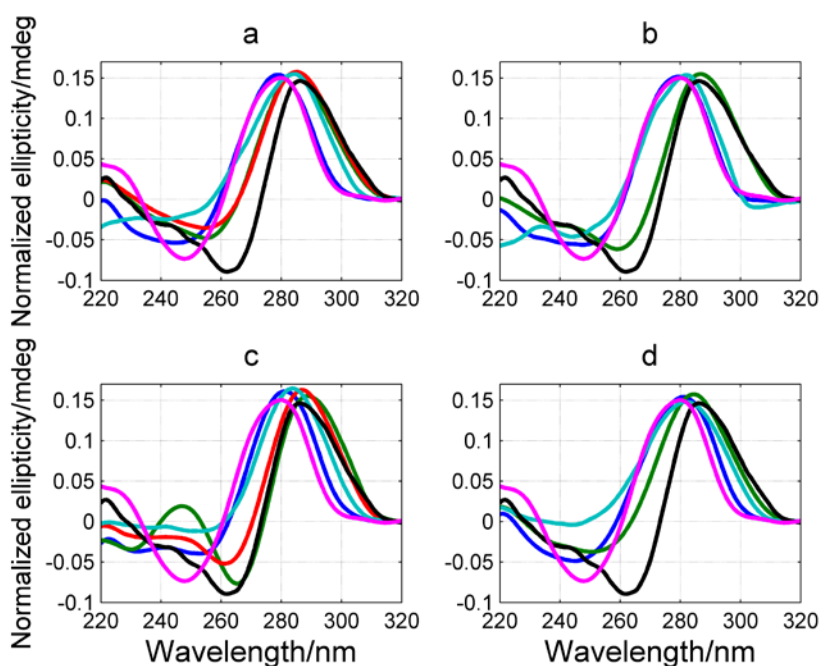
These results depend strongly on the number of acid–base species considered in the calculation. As discussed above, four acid-base species were proposed in the case of SMC01. The calculated fits are shown in Figure S9c-f, overlapped to the experimental ellipticity values at the considered wavelengths. The analysis assuming only three acid-base



species clearly yielded worse fits at 290 nm. Therefore, after a careful study of fitted curves and of the resulting residuals, the presence of four acid–base species was proposed for the SMC01 sequence.

In the case of SMC03 sequence, the proposed model was similar to that described previously for SMC01. Analysis of residuals reinforced the assumption of a four-component model (Figure S10). The pH diagram of species distribution showed small differences with respect to that of SMC01, being the main difference the low concentration of the red species at pH 4.0. In relation to the calculated pure CD spectra, it should be noted the positive CD signal around 250nm for the green species, which is not present in the red one. The spectral features of the CD spectrum of the green species resemble more those assigned to a canonical *i*-motif structure.

In the case of SMC02 and SMC04, only three acid-base species were needed to fit the experimental data. For SMC02, the first species, present at a pH higher than 6, would contain all bases in their neutral form. The major species at pH 4.7 would correspond to an *i*-motif or *i*-motif-like structure. Finally, the major species at pH values around 2.5 would correspond to protonated form. For SMC04, the main trends were similar to those described for the acid-base titration of SMC02 apart from the assignation of the structure for the major species at pH 5. In this case, the calculated pure CD spectrum was not completely characteristic of *i*-motif structure.



**Figure 7.** Comparison of the calculated CD pure spectra of sequences (a) SMC01, (b) SMC02, (c) SMC03, (d) SMC04 with those of T<sub>30</sub> (magenta) and TT (black) sequences. All spectra have been normalized in such a way that their length (i.e., the sum of squares of ellipticities at all wavelengths) is equal to one. The calculated CD pure spectra for SMC01-SMC04 are those shown in Figure 6b, 6d, 6f and 6g, respectively.

For all sequences, the first pH-transition midpoint was located between pH values 5.5 and 5.9, and was accompanied by a relatively high cooperativity (*p* parameter equal or great than 2). Therefore, all sequences adopted low

structured conformations at pH 7.4. In other words, canonical i-motif structures seemed not being formed at neutral pH values at the experimental conditions used in this work, 25 °C and 150 mM KCl.

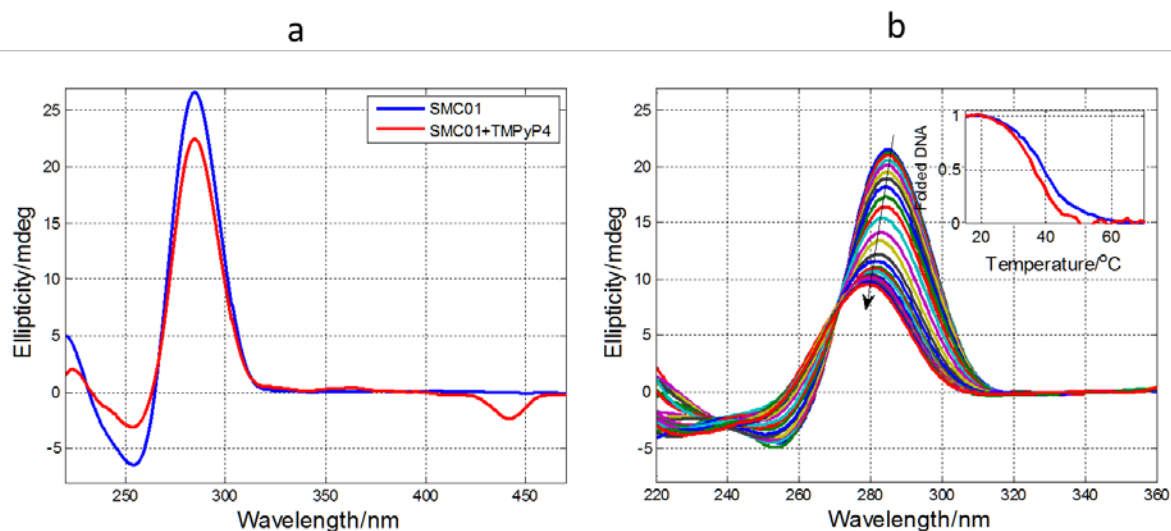
The calculated CD spectra for all the acid-base species shown in Figure 6 have been also compared visually and mathematically with the CD spectra of two “standard” structures in order to identify those acid-base species more similar to a “canonical” i-motif (Figure 7). The first “standard” spectrum corresponds to that of the i-motif formed by the TT sequence (Table 1), which is expected to form a conformationally restricted i-motif structure stabilized by six C·C<sup>+</sup> base pairs at pH 5.1 [37]. The second “standard” CD spectrum corresponds to that of T<sub>30</sub> sequence, which is expected to exist as a partially stacked, single strand at neutral pH values. To remove artifacts due to errors in the determination of DNA concentration, and to facilitate the visual inspection, all spectra were normalized to equal length. It is observed that the species whose CD spectra are more similar to that of a “standard” i-motif are those present around pH 5 in SMC02 and SMC03 sequences. Mathematically, the similarity among species were calculated according to equations 7 and 8, and are shown in Table S3. Again, the structure formed by SMC02 and SMC03 at pH ~5 showed the lower dissimilarity with the i-motif formed by TT sequence. On the other hand, the structures formed by SMC01 and SMC04 appeared more similar to the single, partially stacked strand adopted by the T<sub>30</sub> sequence.

### 3.6. Interaction of SMC01 with TMPyP4

The stability of i-motif structures may be tuned by the interaction with appropriate ligands [17, 20]. We thus studied the interaction of a model ligand, the cationic porphyrin meso-tetra (N-methyl-4-pyridyl) porphine (TMPyP<sub>4</sub>), with the considered sequences by means of CD and melting measurements. This ligand was previously proposed to stabilize the tetramolecular i-motif structure formed by the 5'-AACCC-3' sequence because of a favourable interaction at 5' and 3' ends [38]. On the other hand, the interaction of this ligand with the intramolecular i-motif formed by the 5'-(CCCTAA)<sub>4</sub>-3' sequence showed little stabilization of the structure at pH 5 [36].

Significant changes were observed in CD spectra of SMC01 sequence recorded in absence and presence of the ligand (Figure 8a). The decrease in the intensity of the positive and negative bands at 285 nm and 255 nm, respectively, indicated a partial unfolding in the presence of the ligand. On the other hand, a weak induced CD signal appeared around 442 nm, which may be related with the interaction of the ligand. The rather low intensity of this induced CD signal may indicate a mechanism of interaction not involving intercalation, a fact that would be coherent with the compactness of the i-motif structure [39, 40]. This was confirmed by NMR spectra, which did not exhibit significant changes in the DNA imino region upon addition of TMPyP<sub>4</sub> (Figure S11).

Molecular-absorption monitored titrations of DNA sequences with TMPyP<sub>4</sub> were carried out to determine the binding stoichiometry. For SMC01, two complexes were found with stoichiometries 1:2 and 1:4 (DNA:TMPyP<sub>4</sub>) (Figure S12). The logarithms of the overall binding constant were  $13.8 \pm 0.3 \text{ M}^{-2}$  and  $26.0 \pm 0.4 \text{ M}^{-4}$ , respectively. These stoichiometries and binding constants are in accordance with values previously reported [36, 40].



**Figure 8.** Interaction of SMC01 with TMPyP4. (a) CD spectra of SMC01 sequence in absence (blue) and in presence of TMPyP4 (red). The experimental conditions were 2  $\mu$ M DNA concentration, 4  $\mu$ M TMPyP4 concentration, 150 mM KCl, 20 mM acetate buffer, pH 5.2, 25  $^{\circ}$ C. (b) CD spectra recorded along the melting of the previous mixture from 15 to 70  $^{\circ}$ C. Inset shows the fraction of folded DNA in absence (blue) and in presence of ligand (red).

The potential ability of TMPyP4 to modify the i-motif structure was assessed by means of melting experiments, which showed a decrease of the  $T_m$  values upon addition of TMPyP4 (at 1:2 DNA:ligand ratio) at pH 5.2 (Figure 8b and Figure S13). This decrease was dependent on TMPyP4 concentration (Figure S14). No thermodynamic parameters were calculated from the melting curves in presence of ligand because of the unreliable baseline determination at low temperatures (Table S4 and Figure S13). Hence, a destabilization effect has been observed at the experimental conditions used in this work. Interestingly, the lower destabilizing effect was observed for the complete SMC01 sequence, whereas dramatic decreases in  $T_m$  were observed for SMC03 (Figure S15) and SMC04. The destabilizing effect of TMPyP4 on i-motif structure was also observed for another sequence, TT, that forms a “standard” intramolecular i-motif structure.

### 3.7. Effect of molecular crowding on the stability of SMC01

The stability of i-motifs structures may be affected by the conditions of nuclear cell solution environment. The presence of macromolecules makes that water accounts for only 60% of the solution mass in the nucleus. Therefore, this crowded environment may affect the solution equilibria of macromolecules in such a way that they could be very different to those studied in absence of crowding agents. The use of polyethylene glycol (PEG) having molecular weights up to 12,000 g·mol<sup>-1</sup> as molecular crowding co-solute has been reported widely in the literature [41]. Here, melting experiments in presence of PEG200 (20 % w/v) were carried out to determine its potential stabilizing effect on the i-motif structure formed by SMC01.

In order to study the effect of PEG200 on the stability of SMC01, the buffer concentration is a critical variable because a low concentration of buffer may produce artefacts due to the presence of the co-solute. Hence, addition

of PEG200 (20 % w/v) to a 20 mM buffer produced a strong stabilization of the i-motif (Figure S16), which was related to the pH decrease due to the low buffer capacity. This decrease was more clear in the case of phosphate buffer (a reduction from 6.6 in aqueous medium to 4.7 after addition of PEG200) than in the case of acetate buffer (from 5.2 to 4.6). As a result of this pH reduction, melting experiments carried out in solutions containing PEG200 and low concentrated buffers showed a great stabilization of the SMC01 i-motif. As example, the thermal stability of the i-motif structure increased dramatically: in absence of PEG200 at pH 6.6 no i-motif structure is formed, whereas the  $T_m$  value increased to 35 °C, approximately, in presence of PEG200 (Figure S16).

The same experiments done in a 100 mM buffer yielded different results (Figure S17). At pH 5.2, addition of PEG200 (20 % w/v) practically did not affect the thermal stability of the i-motif formed by SMC01 in aqueous medium as the  $T_m$  did not vary significantly (36.3 vs. 37.6 °C). On the contrary, at pH 6.5, addition of PEG200 produced a slight stabilization of the structure ( $T_m \sim 20$  °C), as deduced from the incipient sigmoidal shape of the melting curve. Finally, at pH 7.2, addition of PEG200 only produced a slight stabilization of the stacked single strand, as observed in the linear shape of the melting curve.

## 4. Discussion

Studies have showed that C-rich sequences can form stable i-motif structures at neutral pH values and relatively low [8, 11, 30, 42, 43]. Recently, it has been reported that a mutant promoter sequence of the *c-myc* gene can form an unstructured conformation, a classical i-motif, and an intramolecular folded “i-motif-like” structure *in vitro* as a function of pH, with the latter being the prevalent species at neutral pH [44]. A similar folding landscape has been proposed for C-rich sequences corresponding to the *bcl-2* gene [45]. All these experimental evidences imply that i-motif and/or “i-motif-like” structures may exist in living cells and participate in biological processes, such as replication, regulation, and transcription [17]. In this sense, it has been shown recently the interaction of C-rich sequences, prone to form i-motif structures, with proteins [19, 21]. We describe here a new C-rich sequence upstream of an oncogene promoter potentially capable of forming a i-motif structure. These findings open the door to the study of the interaction of this new i-motif structure with stabilizing ligands or proteins. In this context, the potential formation *in vivo* of i-motifs by these sequences has to be discussed in light of previous knowledge about pH and thermal stability of these structures.

Recently, the acid-base and thermal properties of a set of i-motifs with the sequence 5'-T<sub>2</sub>C<sub>3</sub>TXTC<sub>3</sub>T<sub>3</sub>C<sub>3</sub>TXTC<sub>3</sub>T<sub>2</sub>-3' (where X are T, A, C or G) was studied [37]. The sequence where X=T (TT, Table 1) showed the maximum symmetry in its primary structure and the i-motif formed, which was stabilized by six C· C<sup>+</sup> base pairs, could be considered as a sort of “standard” i-motif structure for later comparisons. It was observed that the TT sequence showed two acid-base transitions with pH-transition midpoints around 6.5 and 2.6, approximately. Other sequences, where X=A, C or G, resulted in the appearance of an additional acid-base transition with a pH-transition midpoint around 4.5, approximately (Figure S18). This third transition, which is also present in pH diagrams for SMC01 and SMC03 (Figure

6) was related to the existence of an additional protonation process in bases involved in the i-motif core. In light of these results, it was clear that the sequences studied in the present work formed i-motif structures at pH values lower than those corresponding to a “standard” i-motif structure. The structure formed by the SMC03 sequence showed the highest apparent  $pK_a$  value (5.9), whereas that formed by the SMC02 sequence showed the lowest value (5.5). Xu and Sugiyama described the formation of an i-motif structure by a sequence located at the Rb gene [46], being the pH-transition midpoint equal to  $5.9 \pm 0.2$  at  $25\text{ }^\circ\text{C}$ , in concordance with the value determined in this work for SMC01 and SMC03. Also, the values of the  $p$  parameter, which characterizes the cooperativity of the process, were higher for the “standard” i-motif structure ( $p=3$ ) than for those studied here ( $p=2$ ). Finally, the pH range of existence of the folded structures observed in this work were also smaller (around 2 pH units) than those described for the “standard” i-motif structures (3.9 pH units). Interestingly, the i-motif formed by the wild sequence SMC01 shows the longest pH range of existence. All together, nevertheless, these facts pointed to a low stability against pH of the folded structures formed by the SMC0x sequences. *In vivo*, i-motifs not only could be stabilized by transcriptionally induced DNA superhelicity [6], but also for the presence of crowding conditions [41] or by a decrease in cellular pH, as that generated by the glucose-lactate flux [47].



[48]. The high flexibility of these loops may reduce the stability of the whole structure. It should be noted that these loops hardly produce Watson-Crick base pairs that could restraint that flexibility, as denoted by the scarce NMR signals between 12 and 14 ppm at pH 5.

Overall, the thermal stability of the structures formed by the SMC01-SMC04 sequences is low in comparison with those reported for other intramolecular i-motif structures formed by C-rich sequences in promoter regions. Hence, the  $T_m$  values for the 5'-AC<sub>5</sub>TGC ATC TGC ATGC<sub>5</sub>TC<sub>3</sub>AC<sub>5</sub>T-3' sequence (that was found in the *n-myc* promoter [49]) was 65 °C at pH 4.8, almost 30 °C higher than the value determined for SMC01. In this case, the sequence contained at least four cytosine tracts with a minimum length of three cytosines. Guo *et al.* also reported a  $T_m$  value near 70 °C at pH 4.4 for a C-rich sequence found at the RET gene promoter, 5'-C<sub>2</sub>GC<sub>5</sub>GC<sub>4</sub>GC<sub>4</sub>GC<sub>4</sub>TA-3' [50]. From the examination of the thermodynamic values calculated, it may be deduced that SMC01 produced the i-motif the most stable in terms of  $T_m$ ,  $\Delta H^0$  and  $\Delta S^0$ , which is in agreement with the length of the C-tracts involved in the i-motif core.

The interaction of a model ligand, TMPyP4, and the addition of a crowding co-solute, PEG200, have been studied as ways to increase the stability of the i-motifs. Addition of the ligand clearly produced a destabilization of the folded form, in contrast to previous reports where this ligand was proposed as a stabilizing agent [38]. The reason of this discrepancy could be on the tetrameric nature of the i-motifs studied in this work that produce an increase of negatively charged surface at both 3' and 5' ends. In fact, it was postulated that TMPyP4 stacked on these DNA ends. On the other hand, addition of PEG200 20 % (w/v) showed a slight stabilization of the structure in a media that mimic cellular environment. It should be noted, however, the importance of the buffer capacity on these kind of studies. It is often found in literature reports on the influence of co-solutes that use low concentrated buffers. In this situation, addition of co-solutes may produce a dramatic decrease of pH that is reflected in the stabilization of the i-motif structure.

Finally, at neutral pH values it was observed the presence of NMR signals that were assigned not only to canonical Watson-Crick base pairs (12-14 ppm), but also to T·T or G·T base pairs (around 11 ppm). These signals may be due to intramolecular hairpins, as the ones proposed from theoretical calculations. The existence of these structures has been proposed previously. Some authors described them as "i-motif-like" structures, where the relatively weak C·C pairing is present [44]. The potential importance of these hairpins has been envisaged recently as they may provide a mechanism for modulation of gene expression by ligands that may bind selectively to the hairpin or to the i-motif [20]. In this sense, the neutral forms, rather than the i-motifs, could be a drug target for cancer therapy.

## 5. Conclusions

The formation of intramolecular i-motif structures by a long sequence located upstream of the promoter region of the SMARCA4 gene has been demonstrated. Despite the apparent complex sequence, the results showed that the wild sequence may form a relatively stable and homogeneous intramolecular i-motif structure, both in terms of pH or temperature. The model ligand TMPyP4 destabilizes the structure, whereas the presence of 20 % (w/v) PEG200

stabilized it. On the other hand, folded structures have been observed at neutral pH values. These findings may represent novel pharmacological targets for treatment of SMARCA4-associated disorders.

## Acknowledgments

S.L. thanks Maria Solà (IBMB-CSIC, Barcelona, Spain) for support. Xavier Subirats (University of Barcelona, Spain) is thanked for helpful discussions. **Joan Josep Alba (University of Barcelona, Spain) is thanked for carrying out some of the experiments.** Funding from Spanish government (CTQ2014-61758-EXP, CTQ2014-52588-R and CTQ2015-66254-C2-2-P) and recognition from the Autonomous Catalan government (2014SGR1106) are acknowledged.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at

## References

- [1] Maizels N., G4-associated human diseases, *EMBO Rep* 16 (2015) 910-922.
- [2] Rhodes D., Lipps H.J., G-quadruplexes and their regulatory roles in biology, *Nucleic Acids Research* 43 (2015) 8627-8637.
- [3] Brooks T.A., Kendrick S., Hurley L., Making sense of G-quadruplex and i-motif functions in oncogene promoters, *FEBS J* 277 (2010) 3459-3469.
- [4] Huppert J.L., Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes, *Chem Soc Rev* 37 (2008) 1375-1384.
- [5] Benabou S., Avino A., Eritja R., Gonzalez C., Gargallo R., Fundamental aspects of the nucleic acid i-motif structures, *RSC Advances* 4 (2014) 26956-26980.
- [6] Sun D., Hurley L.H., The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression, *J Med Chem* 52 (2009) 2863-2874.
- [7] Bedrat A., Lacroix L., Mergny J.L., Re-evaluation of G-quadruplex propensity with G4Hunter, *Nucleic Acids Res* 44 (2016) 1746-1759.
- [8] Fleming A.M., Ding Y., Rogers R.A., Zhu J., Zhu J., Burton A.D., Carlisle C.B., Burrows C.J., 4n-1 Is a "Sweet Spot" in DNA i-Motif Folding of 2'-Deoxycytidine Homopolymers, *J Am Chem Soc* 139 (2017) 4682-4689.
- [9] Huppert J.L., Balasubramanian S., G-quadruplexes in promoters throughout the human genome, *Nucleic Acids Res* 35 (2007) 406-413.
- [10] Huppert J.L., Bugaut A., Kumari S., Balasubramanian S., G-quadruplexes: the beginning and end of UTRs, *Nucleic Acids Res* 36 (2008) 6260-6268.
- [11] Wright E.P., Huppert J.L., Waller Z.A.E., Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH, *Nucleic Acids Res* 45 (2017) 2951-2959.



- [12] Sutherland C., Cui Y., Mao H., Hurley L.H., A Mechanosensor Mechanism Controls the G-Quadruplex/i-Motif Molecular Switch in the MYC Promoter NHE III1, *J Am Chem Soc* 138 (2016) 14138-14151.
- [13] Mergny J.L., Lacroix L., Han X., Leroy J.L., Helene C., Intramolecular folding of pyrimidine oligodeoxynucleotides into a i-DNA motif, *Journal of the American Chemical Society* 117 (1995) 8887-8898.
- [14] Kovanda A., Zalar M., Sket P., Plavec J., Rogelj B., Anti-sense DNA d(GGCCCC)n expansions in C9ORF72 form i-motifs and protonated hairpins, *Sci Rep* 5 (2015) 17944.
- [15] Cui Y., Kong D., Ghimire C., Xu C., Mao H., Mutually Exclusive Formation of G-Quadruplex and i-Motif Is a General Phenomenon Governed by Steric Hindrance in Duplex DNA, *Biochemistry* 55 (2016) 2291-2299.
- [16] Dhakal S., Yu Z., Konik R., Cui Y., Koirala D., Mao H., G-quadruplex and i-motif are mutually exclusive in ILPR double-stranded DNA, *Biophys J* 102 (2012) 2575-2584.
- [17] Amato J., Iaccarino N., Randazzo A., Novellino E., Pagano B., Noncanonical DNA secondary structures as drug targets: the prospect of the i-motif, *ChemMedChem* 9 (2014) 2026-2030.
- [18] Day H.A., Pavlou P., Waller Z.A., i-Motif DNA: structure, stability and targeting with ligands, *Bioorg Med Chem* 22 (2014) 4407-4418.
- [19] Miglietta G., Cogo S., Pedersen E.B., Xodo L.E., GC-elements controlling HRAS transcription form i-motif structures unfolded by heterogeneous ribonucleoprotein particle A1, *Sci Rep* 5 (2015) 18097.
- [20] Kendrick S., Kang H.J., Alam M.P., Madathil M.M., Agrawal P., Gokhale V., Yang D., Hecht S.M., Hurley L.H., The dynamic character of the BCL2 promoter i-motif provides a mechanism for modulation of gene expression by compounds that bind selectively to the alternative DNA hairpin structure, *J Am Chem Soc* 136 (2014) 4161-4171.
- [21] Roy B., Talukder P., Kang H.J., Tsuen S.S., Alam M.P., Hurley L.H., Hecht S.M., Interaction of Individual Structural Domains of hnRNP LL with the BCL2 Promoter i-Motif DNA, *J Am Chem Soc* 138 (2016) 10950-10962.
- [22] Witkowski L., Carrot-Zhang J., Albrecht S., Fahiminiya S., Hamel N., Tomiak E., Grynspan D., Saloustros E., Nadaf J., Rivera B., Gilpin C., Castellsague E., Silva-Smith R., Plourde F., Wu M., Saskin A., Arseneault M., Karabakhtsian R.G., Reilly E.A., Ueland F.R., Margiolaki A., Pavlakis K., Castellino S.M., Lamovec J., Mackay H.J., Roth L.M., Ulbright T.M., Bender T.A., Georgoulis V., Longy M., Berchuck A., Tischkowitz M., Nagel I., Siebert R., Stewart C.J., Arseneau J., McCluggage W.G., Clarke B.A., Riazalhosseini Y., Hasselblatt M., Majewski J., Foulkes W.D., Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type, *Nat Genet* 46 (2014) 438-443.
- [23] Piotto M., Saudek V., Sklenar V., Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions, *J Biomol NMR* 2 (1992) 661-665.
- [24] Breslauer K.J., Extracting thermodynamic data from equilibrium melting curves for oligonucleotide order-disorder transitions, *Methods in Enzymology*, Academic Press, 1995, pp. 221-242.
- [25] Jaumot J., Vives M., Gargallo R., Application of multivariate resolution methods to the study of biochemical and biophysical processes, *Anal Biochem* 327 (2004) 1-13.
- [26] Dyson R., Kaderli S., Lawrence G.A., Maeder M., Zuberbühler A.D., Second order global analysis: the evaluation of series of spectrophotometric titrations for improved determination of equilibrium constants, *Analytica Chimica Acta* 353 (1997) 381-393.
- [27] Gargallo R., Tauler R., CuestaSanchez F., Massart D.L., Validation of alternating least-squares multivariate curve resolution for chromatographic resolution and quantitation, *Trac-Trends in Analytical Chemistry* 15 (1996) 279-286.

- [28] Zuker M., Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res* 31 (2003) 3406-3415.
- [29] Canalia M., Leroy J.L., Structure, internal motions and association-dissociation kinetics of the i-motif dimer of d(5mCCTCACTCC), *Nucleic Acids Res* 33 (2005) 5471-5481.
- [30] Escaja N., Viladoms J., Garavis M., Villasante A., Pedroso E., Gonzalez C., A minimal i-motif stabilized by minor groove G:T:G:T tetrads, *Nucleic Acids Res* 40 (2012) 11737-11747.
- [31] Kejnovska I., Kypr J., Vorlickova M., Oligo(dT) is not a correct native PAGE marker for single-stranded DNA, *Biochem Biophys Res Commun* 353 (2007) 776-779.
- [32] Mathur V., Verma A., Maiti S., Chowdhury S., Thermodynamics of i-tetraplex formation in the nuclease hypersensitive element of human c-myc promoter, *Biochem Biophys Res Commun* 320 (2004) 1220-1227.
- [33] Méndez A., Bosch E., Rosés M., Neue U.D., Comparison of the acidity of residual silanol groups in several liquid chromatography columns, *Journal of Chromatography A* 986 (2003) 33-44.
- [34] Guittet E., Renciuik D., Leroy J.L., Junctions between i-motif tetramers in supramolecular structures, *Nucleic Acids Res* 40 (2012) 5162-5170.
- [35] Manzini G., Yathindra N., Xodo L.E., Evidence for intramolecularly folded i-DNA structures in biologically relevant CCC-repeat sequences, *Nucleic Acids Res* 22 (1994) 4634-4640.
- [36] Fernandez S., Eritja R., Aviño A., Jaumot J., Gargallo R., Influence of pH, temperature and the cationic porphyrin TMPyP4 on the stability of the i-motif formed by the 5'-(C3TA2)4-3' sequence of the human telomere, *International Journal of Biological Macromolecules* 49 (2011) 729-736.
- [37] Benabou S., Garavis M., Lyonnais S., Eritja R., Gonzalez C., Gargallo R., 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 (2016) 7997-8004.
- [38] Fedoroff O.Y., Rangan A., Chemeris V.V., Hurley L.H., Cationic porphyrins promote the formation of i-motif DNA and bind peripherally by a nonintercalative mechanism, *Biochemistry* 39 (2000) 15083-15090.
- [39] White E.W., Tanious F., Ismail M.A., Reszka A.P., Neidle S., Boykin D.W., Wilson W.D., Structure-specific recognition of quadruplex DNA by organic cations: influence of shape, substituents and charge, *Biophys Chem* 126 (2007) 140-153.
- [40] Khan N., Avino A., Tauler R., Gonzalez C., Eritja R., Gargallo R., Solution equilibria of the i-motif-forming region upstream of the B-cell lymphoma-2 P1 promoter, *Biochimie* 89 (2007) 1562-1572.
- [41] Miyoshi D., Matsumura S., Nakano S., Sugimoto N., Duplex dissociation of telomere DNAs induced by molecular crowding, *J Am Chem Soc* 126 (2004) 165-169.
- [42] Zhou J., Wei C., Jia G., Wang X., Feng Z., Li C., Formation of i-motif structure at neutral and slightly alkaline pH, *Molecular BioSystems* 6 (2010) 580-586.
- [43] Brazier J.A., Shah A., Brown G.D., I-motif formation in gene promoters: unusually stable formation in sequences complementary to known G-quadruplexes, *Chem Commun (Camb)* 48 (2012) 10739-10741.
- [44] Dettler J.M., Buscaglia R., Cui J., Cashman D., Blynn M., Lewis E.A., Biophysical characterization of an ensemble of intramolecular i-motifs formed by the human c-MYC NHE III1 P1 promoter mutant sequence, *Biophys J* 99 (2010) 561-567.
- [45] Kendrick S., Akiyama Y., Hecht S.M., Hurley L.H., 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 (2009) 17667-17676.

- [46] Xu Y., Sugiyama H., Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Res* 34 (2006) 949-954.
- [47] Hanahan D., Weinberg R.A., Hallmarks of cancer: the next generation, *Cell* 144 (2011) 646-674.
- [48] Reilly S.M., Morgan R.K., Brooks T.A., Wadkins R.M., Effect of interior loop length on the thermal stability and pK(a) of i-motif DNA, *Biochemistry* 54 (2015) 1364-1370.
- [49] Benabou S., Ferreira R., Avino A., Gonzalez C., Lyonnais S., Sola M., Eritja R., Jaumot J., Gargallo R., Solution equilibria of cytosine- and guanine-rich sequences near the promoter region of the n-myc gene that contain stable hairpins within lateral loops, *Biochim Biophys Acta* 1840 (2014) 41-52.
- [50] Guo K., Pourpak A., Beetz-Rogers K., Gokhale V., Sun D., Hurley L.H., Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene, *J Am Chem Soc* 129 (2007) 10220-10228.

## **i-motif structures in long cytosine-rich sequences found upstream of the promoter region of the SMARCA4 gen**

Sanae Benabou, Anna Aviñó, S. Lyonnais, C. González, Ramon Eritja, Anna De Juan, Raimundo Gargallo\*

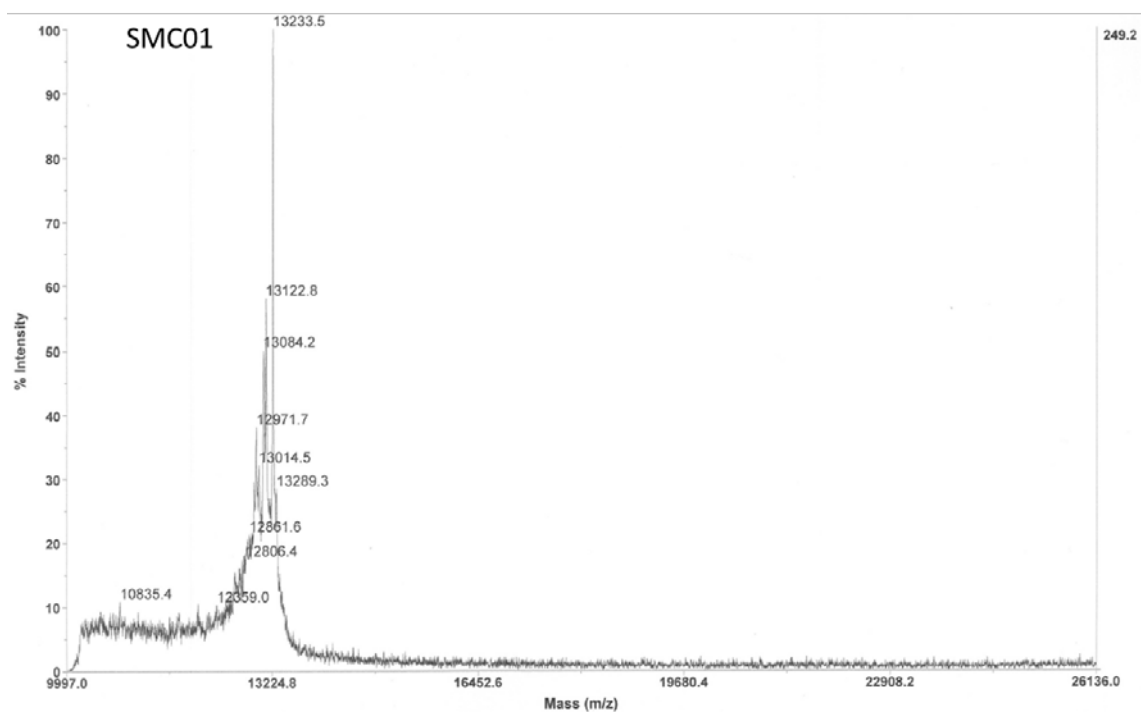
### **Contents:**

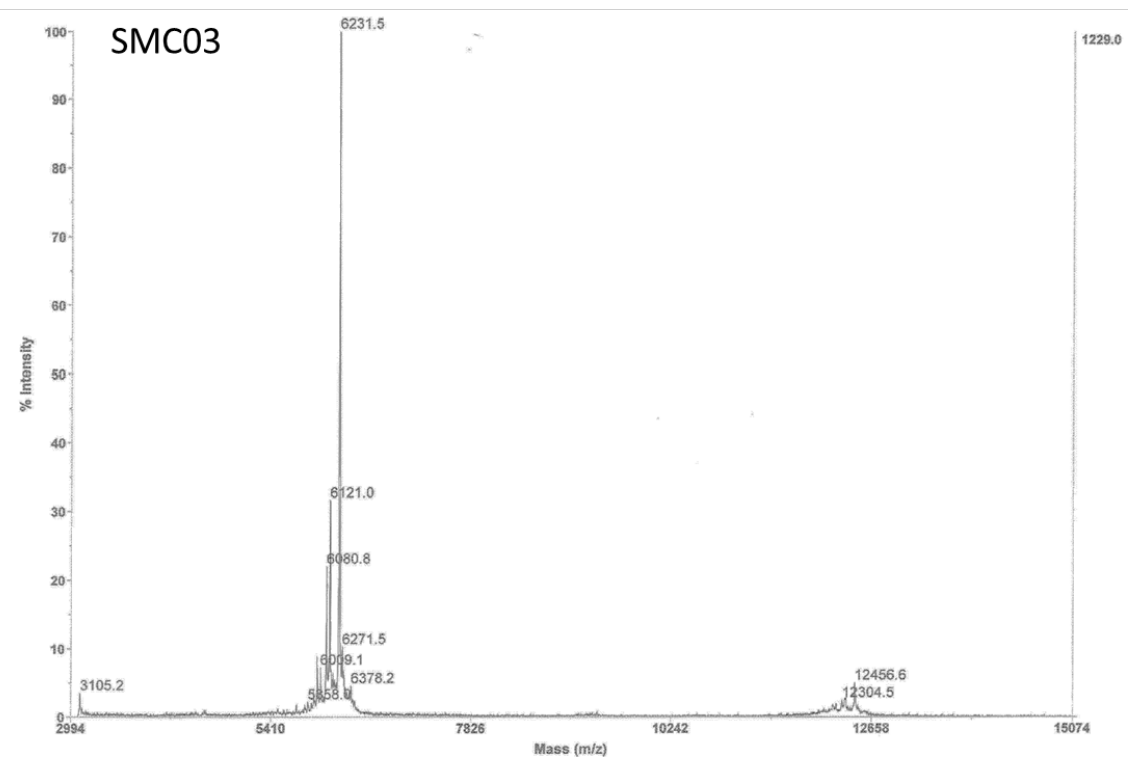
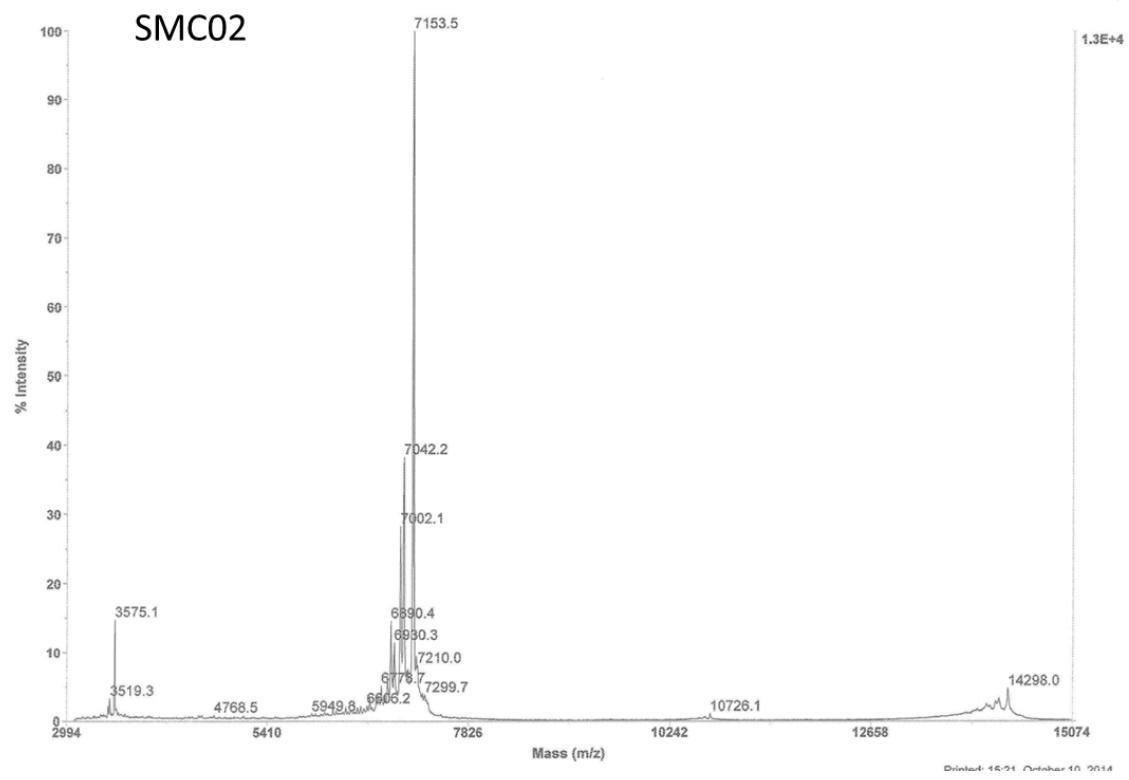
<b>Figure S1</b>	<b>MALDI-TOF MS spectra of oligonucleotides</b>	3
<b>Table S1</b>	Folding predictions based on <i>in silico</i> calculations	6
<b>Figure S2</b>	1D <sup>1</sup> H-NMR spectra of SMC01, SMC02 and SMC03	7
<b>Figure S3</b>	Normalized SEC profiles for SMC01-SMC04 sequences at pH 7.0 and 5.3	8
<b>Table S2</b>	SEC data and calibration plots at pH 7.0 and pH 5.3	9
<b>Figure S4</b>	<b>Melting and annealing experiments of SMC01-SMC04 sequences at pH 5.2</b>	11
<b>Figure S5</b>	First derivative plots for SMC01 sequence	12
<b>Figure S6</b>	First derivative plots for SMC02 sequence	13
<b>Figure S7</b>	First derivative plots for SMC03 sequence	14
<b>Figure S8</b>	First derivative plots for SMC04 sequence	15
<b>Figure S9</b>	CD-monitored melting and acid-base titration of SMC01	16
<b>Figure S10</b>	Acid-base titration of SMC02, SMC03 and SMC04	17
<b>Table S3</b>	Similarity of the calculated CD pure spectra for i-motif species with those of a partially stacked single strand (T <sub>30</sub> sequence) and of an “standard” i-motif structure (TT sequence)	18
<b>Figure S11</b>	1D <sup>1</sup> H-NMR spectra of SMC01, SMC02, SMC03 and SMC04 in presence of TMPyP4	19
<b>Figure S12</b>	<b>Figure S12. Determination of the SMC01:TMPyP4 stoichiometry and binding constant.</b>	20
<b>Figure S13</b>	Melting experiment of mixtures of TMPyP4 with SMC02, SMC03 and SMC04	21
<b>Figure S14</b>	<b>Melting experiment of mixtures of TMPyP4 with SMC01 at different DNA:ligand ratios, pH 5.2</b>	22
<b>Table S4</b>	T <sub>m</sub> values determined for the unfolding of DNA in absence and in presence of TMPyP4	23
<b>Figure S15</b>	<b>Fraction of folded SMC03 in absence and in presence of ligand</b>	23

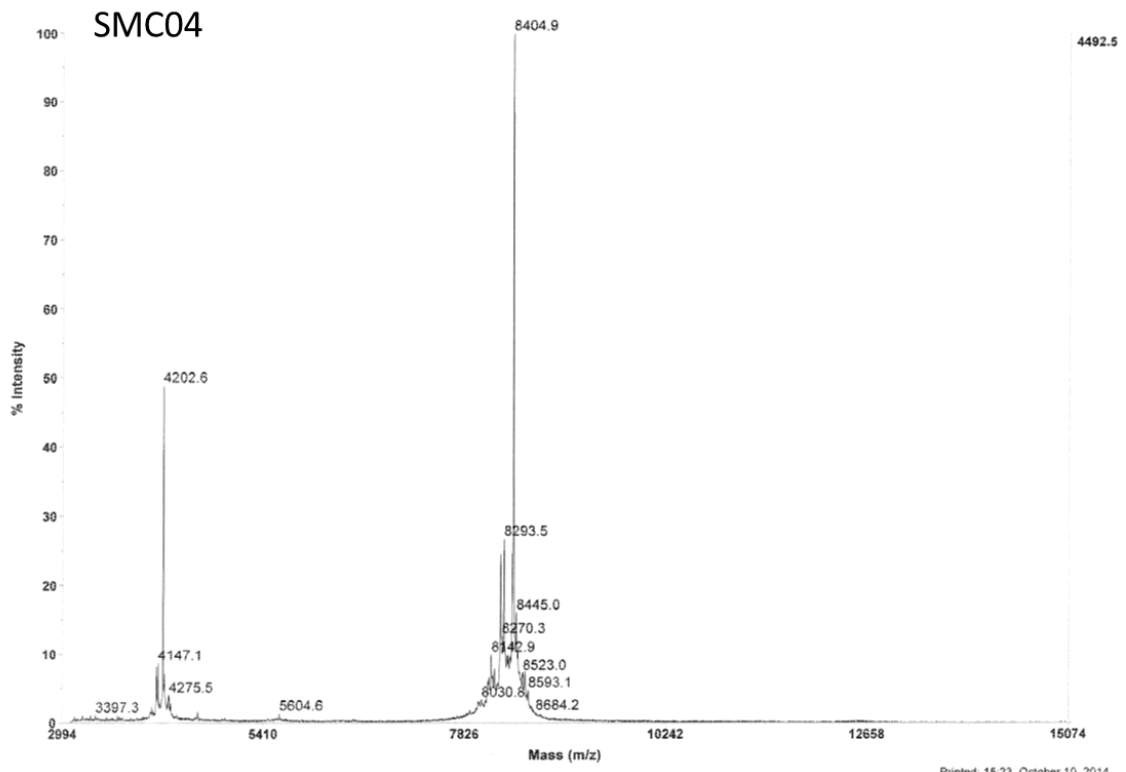
<b>Figure S16</b>	Thermally-induced unfolding of SMC01 in absence and presence of PEG200 (20% w/v) and 20mM buffer	24
<b>Figure S17</b>	Thermally-induced unfolding of SMC01 in absence and presence of PEG200 (20% w/v) and 100mM buffer	25
<b>Figure S18</b>	pH diagrams of species distribution for 5'-TT CCC TTT CCC TTT CCC TTT CCC TT-3' (TT sequence) and 5'-TT CCC TAT CCC TTT CCC TAT CCC TT-3' (AA sequence)	26

Figure S1. MALDI-TOF MS spectra of oligonucleotides. To analyse Oligonucleotides mass spectra were recorded on a Matrix-assisted laser desorption ionization (MALDI) Voyager DETM RP time of-flight (TOF) spectrometer (Applied Biosystems, USA).

Sequence	MW <sub>calculated</sub>	MW <sub>determined</sub>
SMC01	13232.8	13233.5
SMC02	7154.8	7153.5
SMC03	6235.2	6231.5
SMC04	8409.6	8409.9

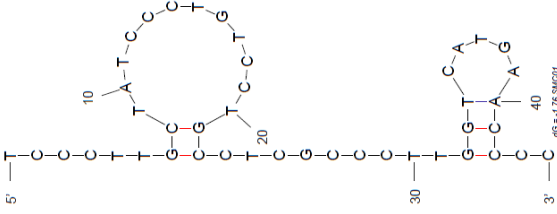
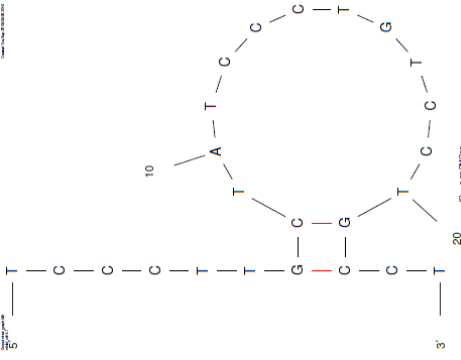
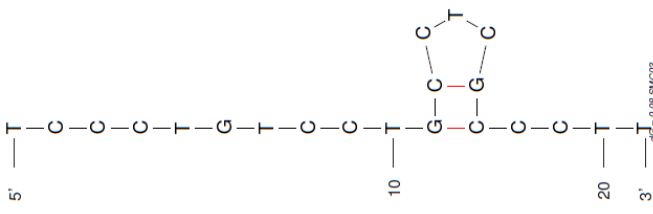
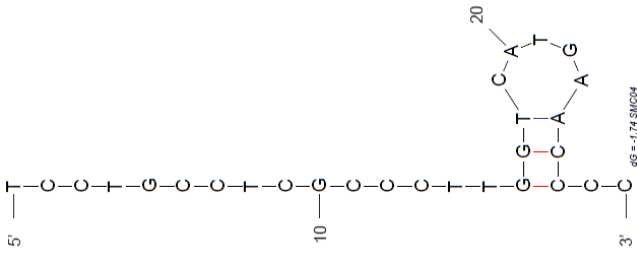




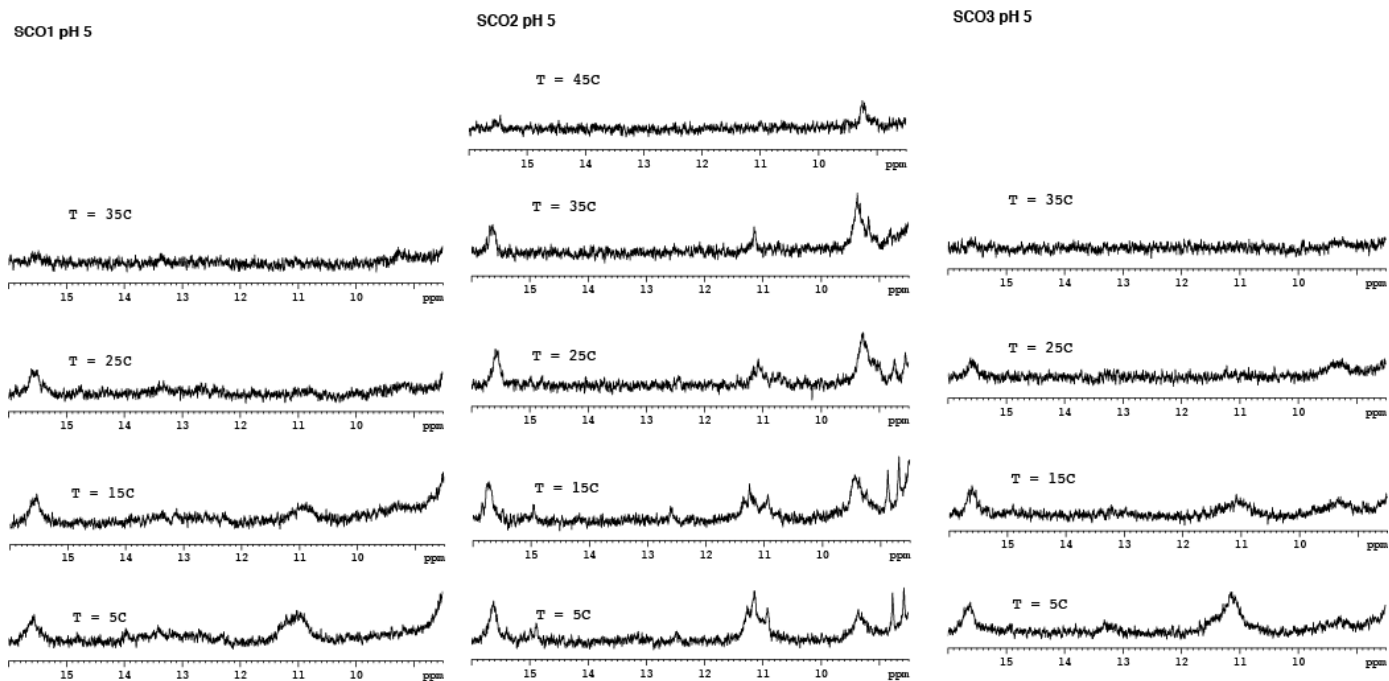




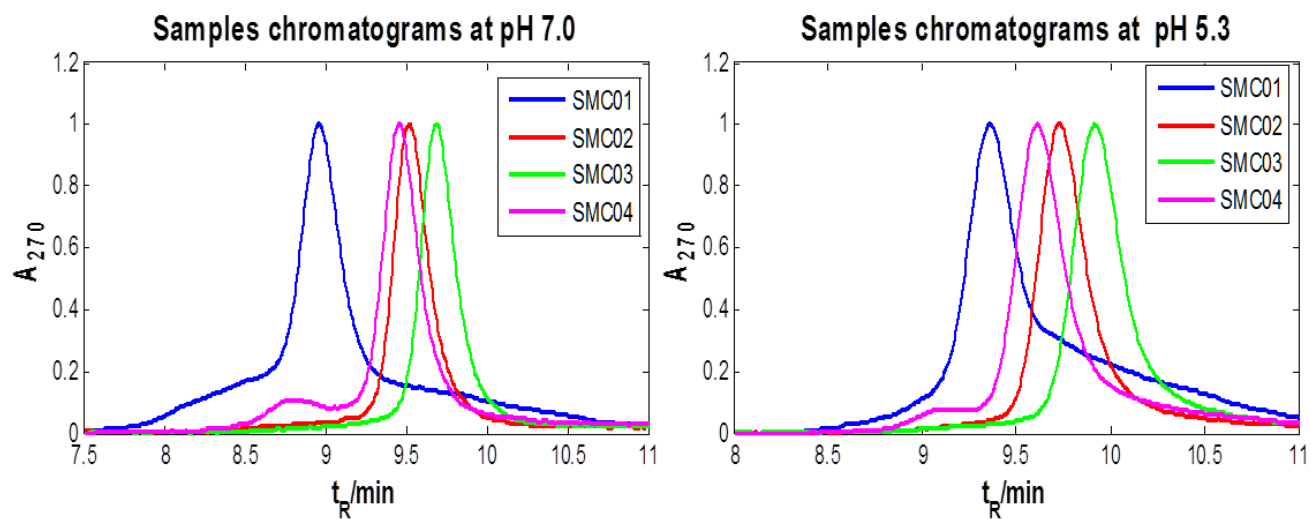
**Table S1. Folding predictions based on *in silico* calculations.** The calculations were done using the web server located in <http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form>.

Name	Proposed structure	$\Delta G_{250C}$ (kcal·mol <sup>-1</sup> )
SMCO 1	<p>5' -TCCCTTGCTATCCCTGTCCTGCCTCGCCCTTGGTCATGAACCCC-3'</p> <p style="text-align: center;">:       :</p> <p>3' - CCCCCAAGTACTGGTTCCCGCTCCGTCCTGTCCTATCGTTCCCT-5</p> 	-8.5  -1.7
SMCO 2		-0.1
SMCO 3		0.1
SMCO 4		-1.7

**Figure S2.** 1D  $^1\text{H}$ -NMR spectra of SMC01 (left), SMC02 (center) and SMC03 (right) at pH 5.0 and different temperatures. Experimental conditions: 20 mM potassium phosphate, 150 mM KCl,  $C_{\text{DNA}} = 0.1$  mM.



**Figure S3.** Normalized SEC profiles for SMC01-SMC04 sequences at pH 7.0 (phosphate buffer) and 5.3 (acetate buffer), 300 mM KCl, 25 °C.

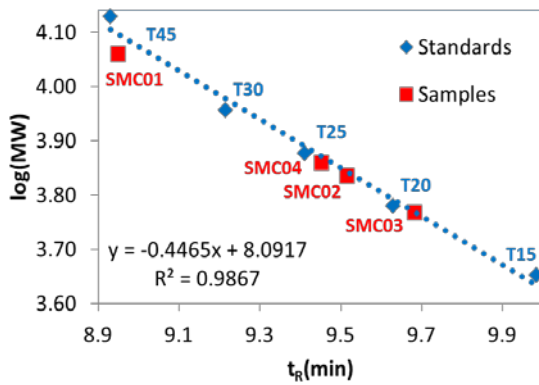


**Table S2.** SEC data and calibration plots at pH 7.0 (phosphate buffer) and pH 5.3 (acetate buffer).

Standards	MW (g·mol <sup>-1</sup> )	log (MW)	Measured t <sub>R</sub> at pH 7.0 (min)	Measured t <sub>R</sub> at pH 5.3 (min)
<b>T15</b>	4,501.0	3.65	9.98	9.88
<b>T20</b>	6,022.0	3.78	9.63	9.64
<b>T25</b>	7,542.9	3.88	9.41	9.43
<b>T30</b>	9,063.9	3.96	9.22	9.24
<b>T45</b>	13626.9	4.13	8.93	8.80

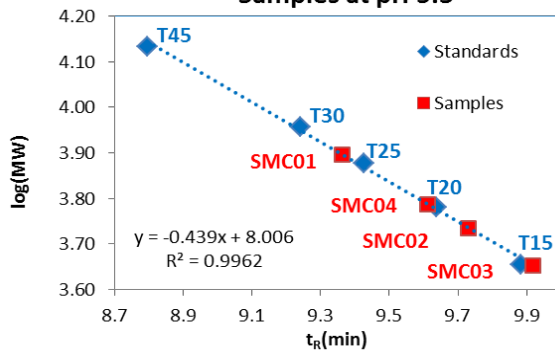
Name	MW (g·mol <sup>-1</sup> )	log(MW)	pH 7.0				pH 5.3			
			Measured t <sub>R</sub> (min)	Calculated t <sub>R</sub> (min)	Diff(%)	Proposed structure	Measured t <sub>R</sub> (min)	Calculated t <sub>R</sub> (min)	Diff(%)	Proposed structure
<b>SMC01</b>	13,233.6	4,09	8.95	8.89	0.6	Unfolded sequence	9.36	8.85	5.8	Folded sequence
<b>SMC02</b>	7,154.8	3,85	9.51	9.44	0.8	Unfolded sequence	9.73	9.41	3.4	Folded sequence
<b>SMC03</b>	6,235.2	3,79	9.68	9.62	0.6	Unfolded sequence	9.92	9.59	3.4	Folded sequence
<b>SMC04</b>	8,409.6	3,92	9.45	9.33	1.3	Unfolded sequence	9.61	9.30	3.3	Folded sequence

**Samples at pH 7**



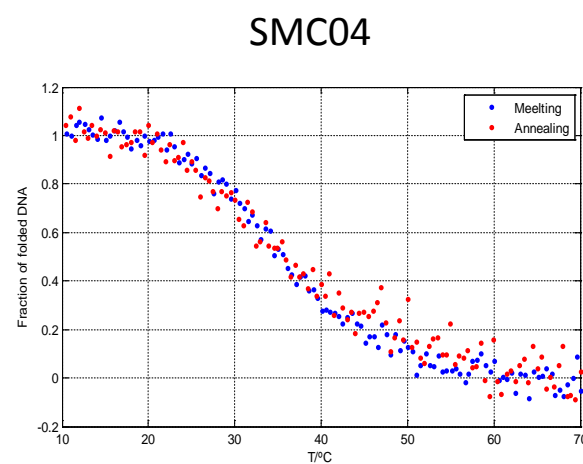
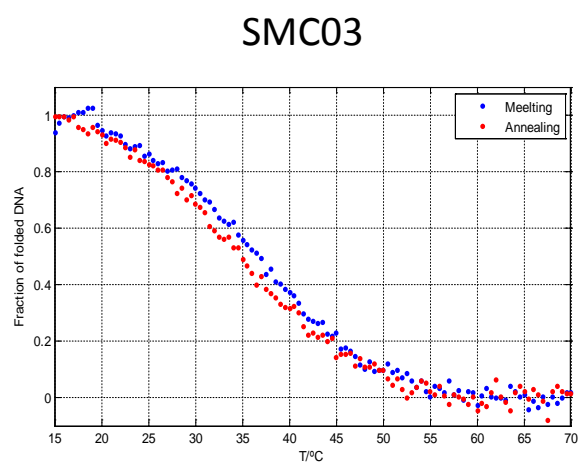
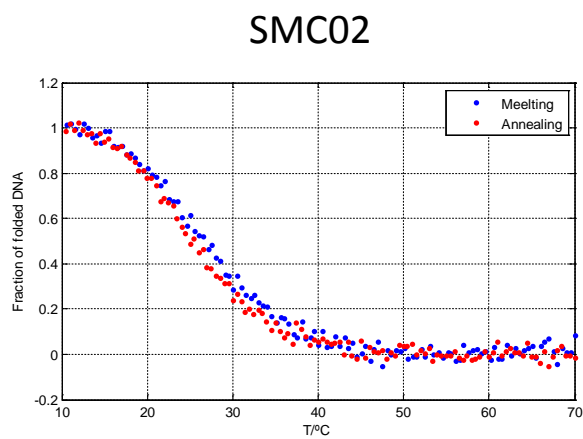
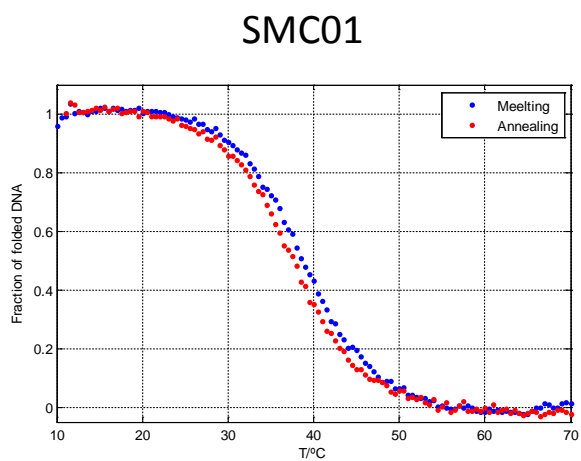
This graph represents the dependence of  $\log(\text{MW})$  with  $t_R$  for a series of five poly(dT) standards (T<sub>15</sub>, T<sub>20</sub>, T<sub>25</sub>, T<sub>30</sub> and T<sub>45</sub>) that do not show a folded structure in these experimental conditions. The variation of  $\log(\text{MW})$  with  $t_R$  shows a linear dependence. The samples eluted at  $t_R$  values equal (or very similar) to those calculated from the regression line, which means that these samples are not folded in this experimental conditions.

**Samples at pH 5.3**

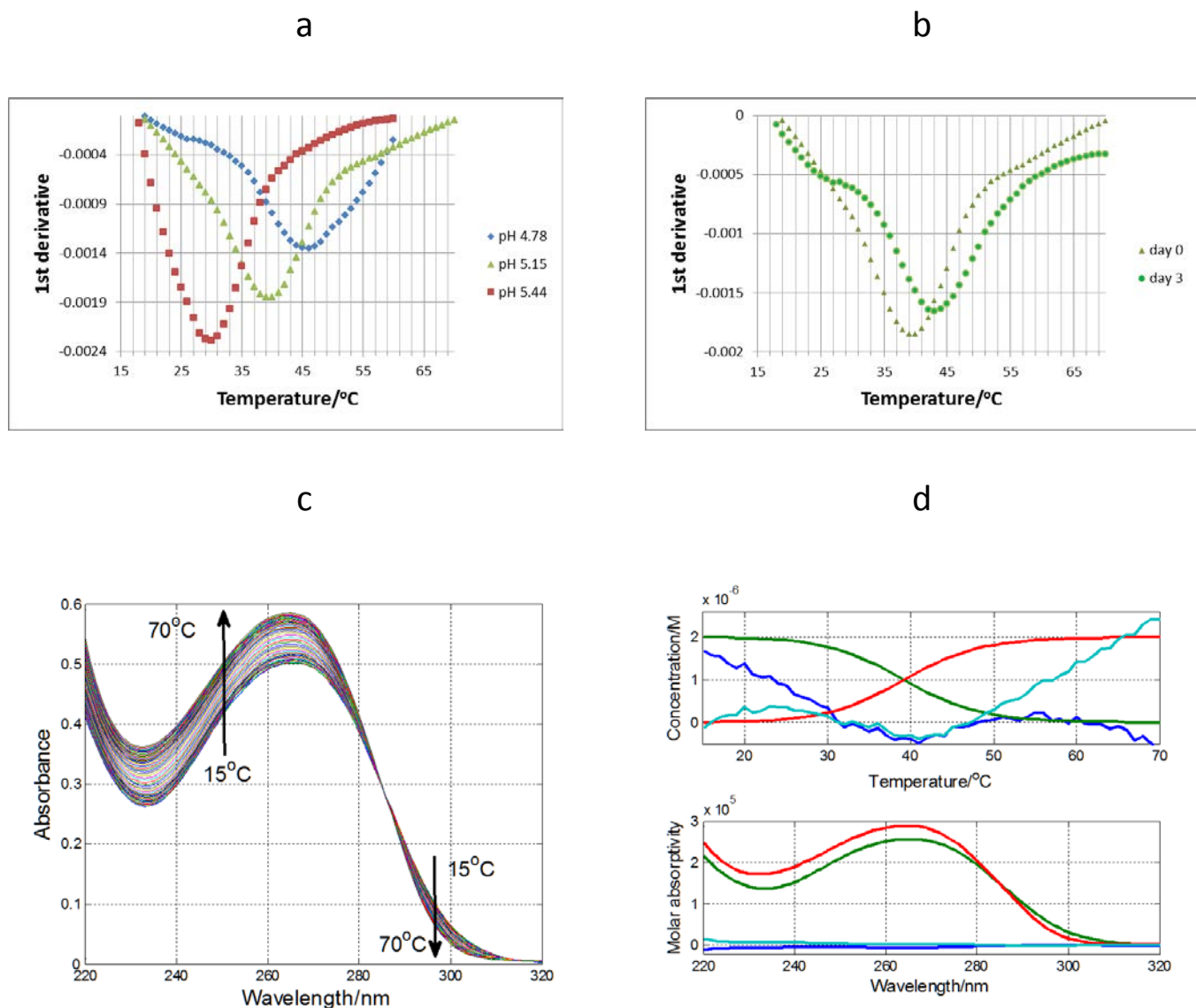


As in the previous graph (pH 7.0), poly(dT) standards do not show a folded structure in these experimental conditions, and the variation of  $\log(\text{MW})$  with  $t_R$  shows a linear dependence. However, at pH 5.3, the samples eluted at  $t_R$  values greater than those expected for unfolded strands, which means that the hydrodynamic volume of samples is smaller than at pH 7.0. This fact is related to the folding into i-motif structures.

Figure S4. Melting and annealing experiments of SMC01-SMC04 sequences. Experimental conditions: 150mM KCl, 20mM acetate buffer, pH 5.2.

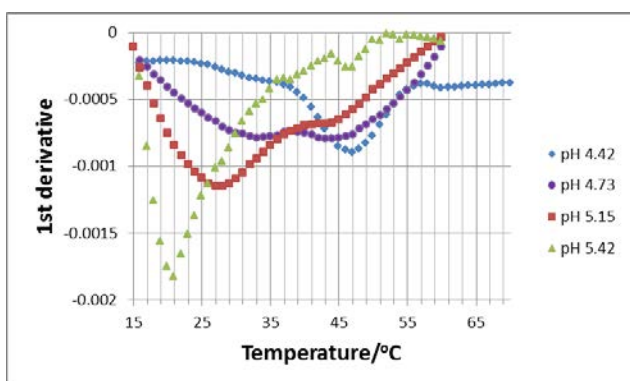


**Figure S5.** First derivative plots for SMC01 sequence. (a) First derivative of absorbance trace at 295 nm versus temperature at several pH values. All curves show clearly monophasic transitions. (b) First derivative of absorbance trace at 295 nm versus temperature at pH 5.2 of a recently prepared sample and of the same sample kept for 3 days at 4 °C. The variation of  $T_m$  value from 40 to 44 °C was not related to a dimerization process. (c) Spectral data collected at pH 5.2. (d) The data shown in (c) has been analyzed with multivariate analysis in order to confirm the proposal of a two-step unfolding process. Four components were observed, of which only two are directly correlated with the i-motif (green) and unfolded strand (red). Blue and cyan components are related with baseline drifts at low and high temperatures, respectively [R. Gargallo, Analytical Biochemistry 2014, 466, 4-15].

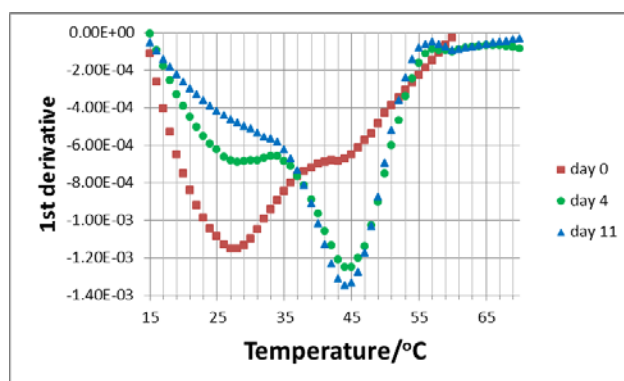


**Figure S6.** First derivative plots for SMC02 sequence. (a) First derivative of absorbance trace at 295 nm versus temperature at several pH values. Curves recorded at pH 4.7 and 5.2 show that the transitions were not a two-state process. (b) First derivative of absorbance trace at 295 nm versus temperature at pH 5.2 and different incubation times at 4 °C. (c) First derivative of absorbance trace at 295 nm versus temperature at three different concentrations, and after an incubation time of four days at 4 °C. The gradual disappearance of the shoulder at 27 °C was observed.

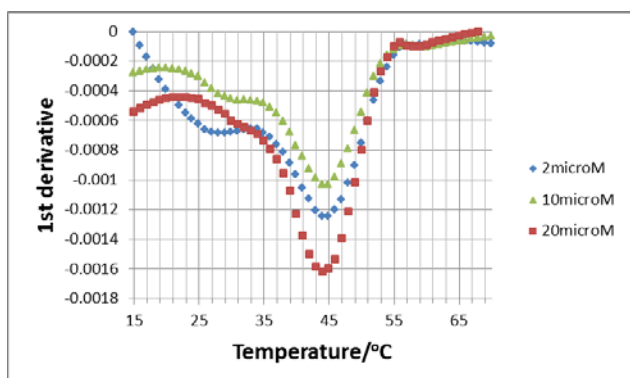
a



b



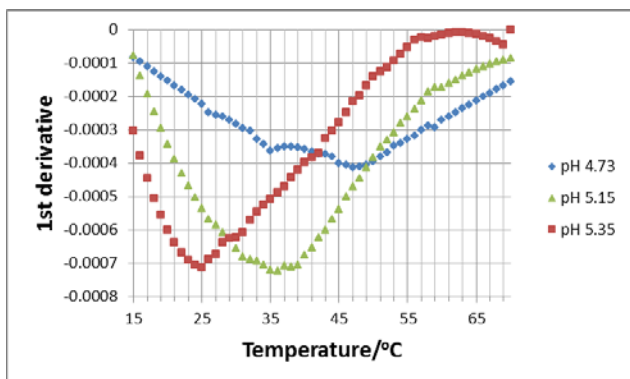
c



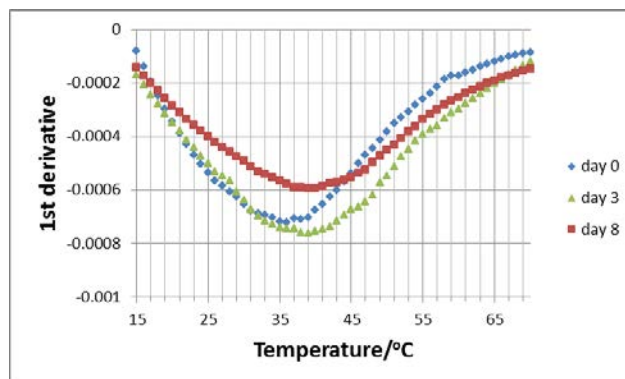


**Figure S7.** First derivative plots for SMC03 sequence. (a) First derivative of absorbance trace at 295 nm versus temperature at several pH values. (b) First derivative of absorbance trace at 295 nm versus temperature at pH 5.2 and different incubation times at 4 °C. No clear changes were observed.

a

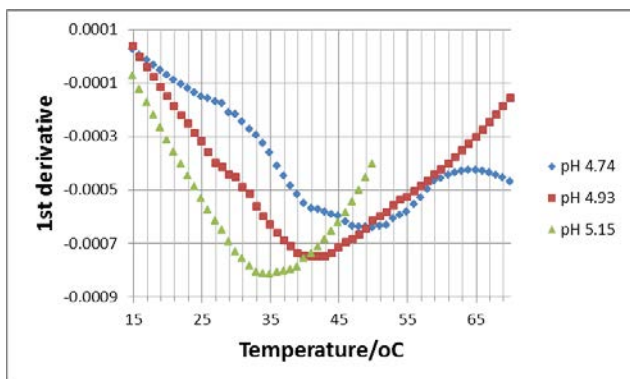


b

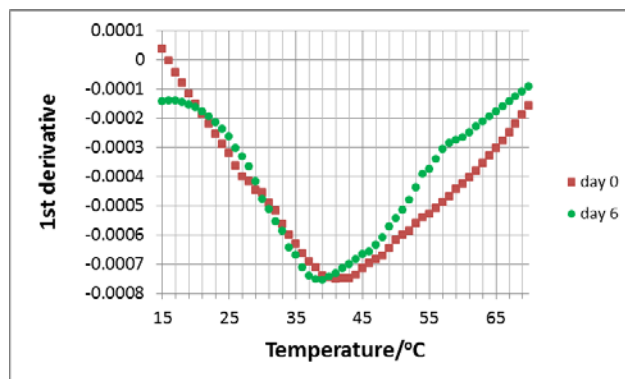


**Figure S8.** First derivative plots for SMC04 sequence. (a) First derivative of absorbance trace at 295 nm versus temperature at several pH values. (b) First derivative of absorbance trace at 295 nm versus temperature at pH 4.93 and different incubation times at 4 °C.

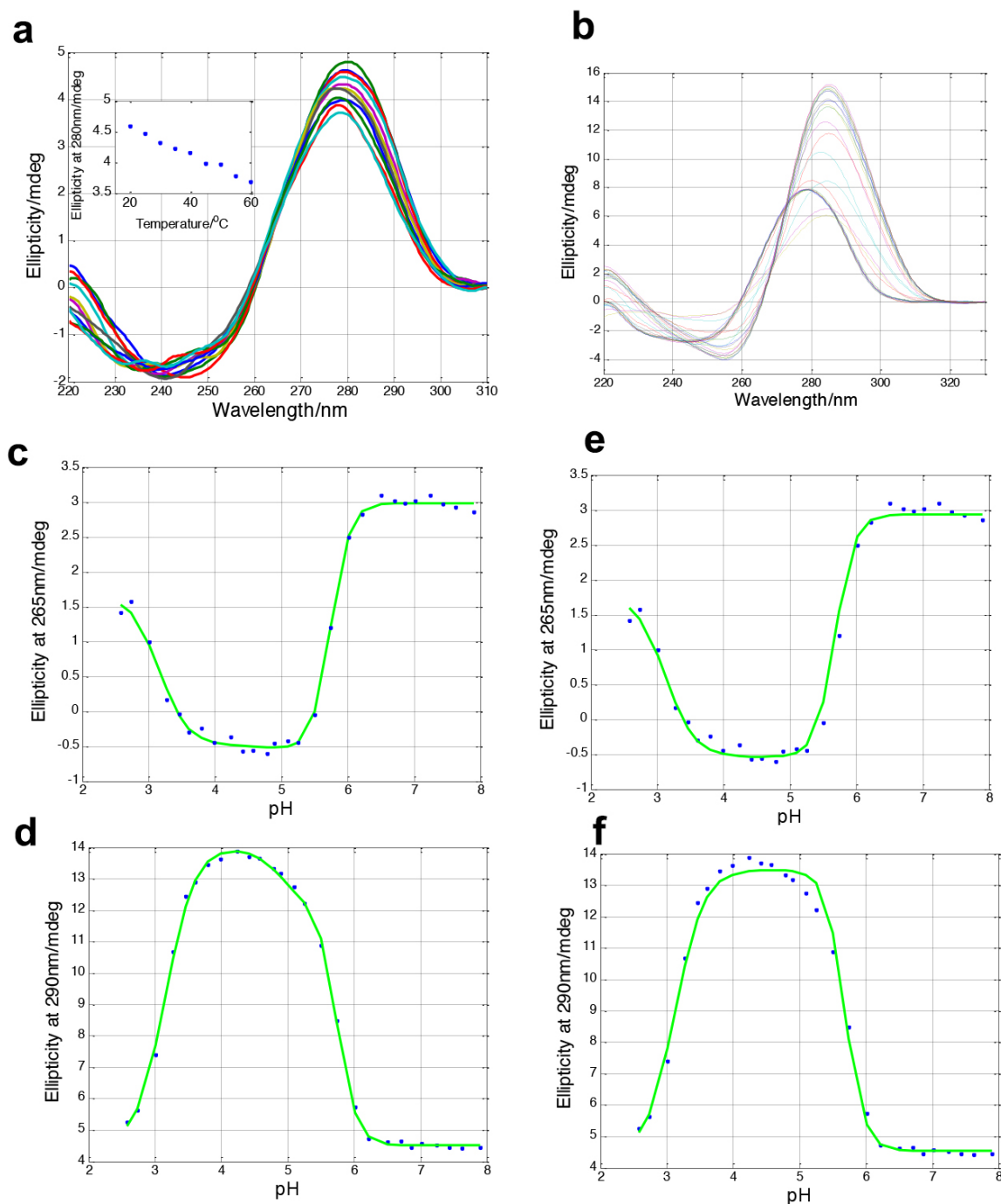
a



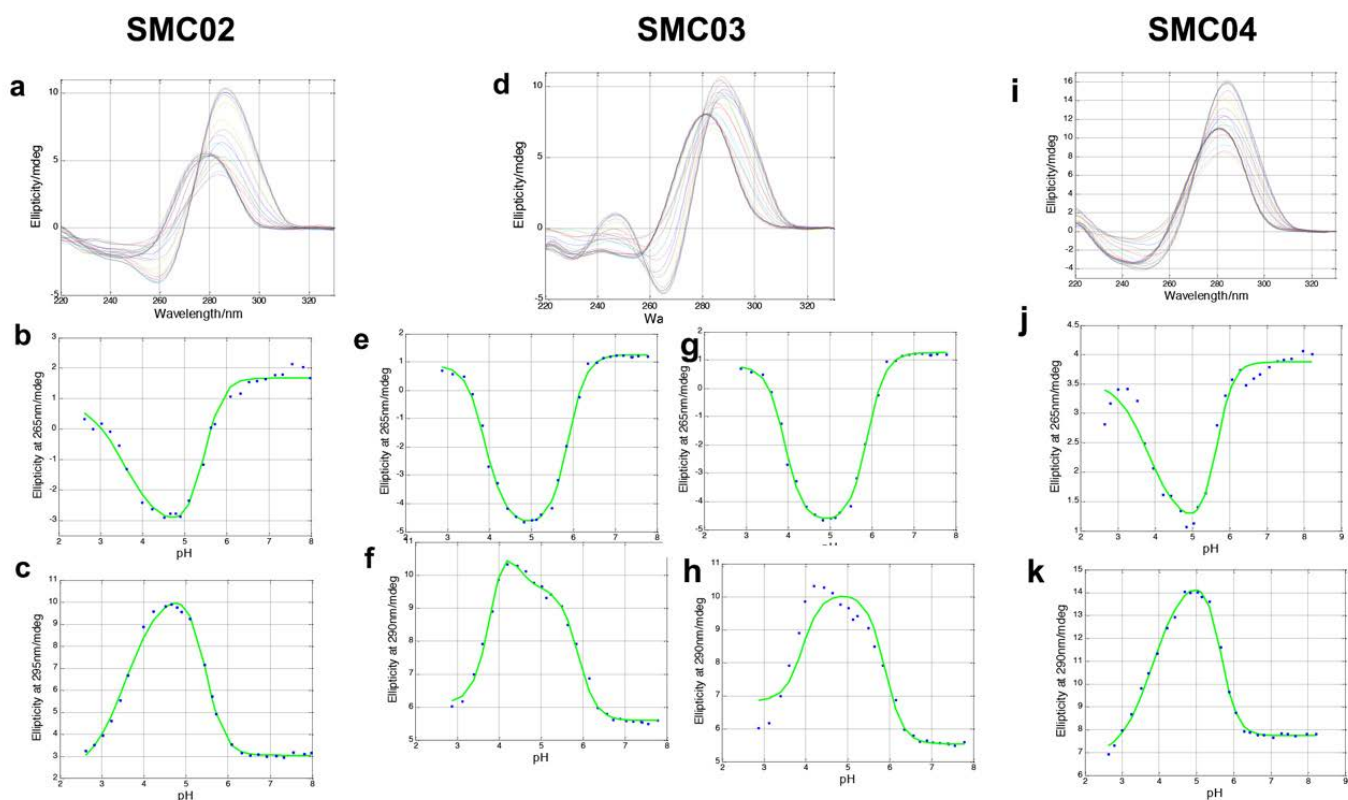
b



**Figure S9.** CD-monitored melting and acid-base titration of SMC01. (a) CD-monitored melting of SMC01. The experimental conditions were 2  $\mu$ M DNA, 150 mM KCl, 20 mM phosphate buffer, pH 7.3. Upon heating, a gradual decrease of the intensity of the band at 280nm, together with a shift to shorter wavelengths is observed. (b) Raw spectroscopic data corresponding to the acid-base titration of SMC01. (c) and (d) show the fits at 265 and 290 nm, respectively, considering the four-components model discussed in the main text. Experimental and calculated values of ellipticity are represented by blue symbols and continuous green line, respectively. (e) and (f) show the fits at 265 and 290 nm, respectively, considering the three-components model.



**Figure S10.** Acid-base titration of SMC02, SMC03 and SMC04. Experimental and calculated values of ellipticity are represented by blue symbols and continuous green line, respectively. SMC02 (a) Raw spectroscopic data, (b) and (c) show the fits at 265 and 290 nm, respectively, considering the three-components model discussed in the main text. SMC03 (d) Raw spectroscopic data, (e) and (f) show the fits at 265 and 290 nm, respectively, considering the four-components model discussed in the main text. (g) and (h) show the fits at 265 and 290 nm, respectively, considering the three-components model. SMC04 (i) Raw spectroscopic data, (j) and (h) show the fits at 265 and 290 nm, respectively, considering the three-components model.

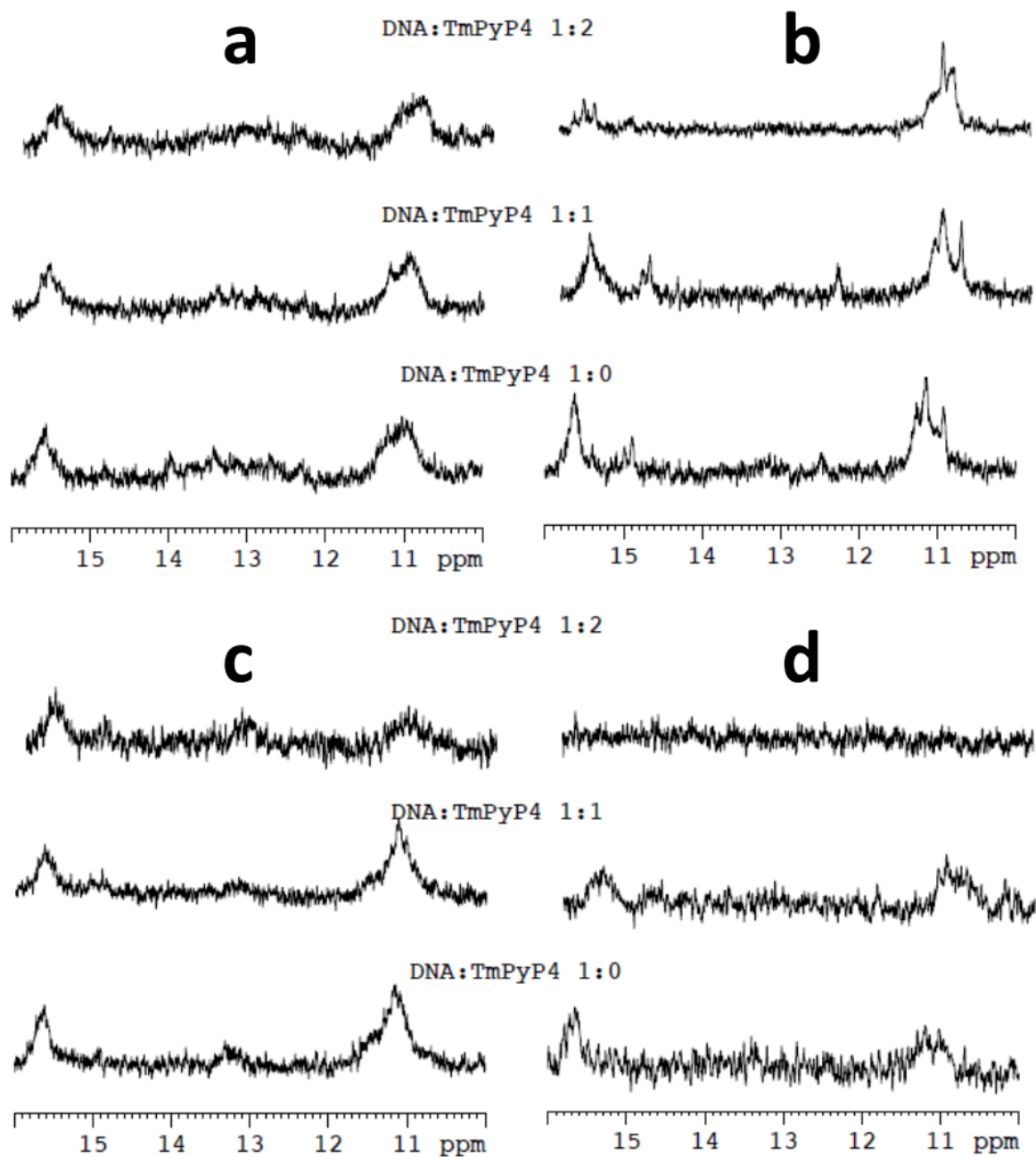


**Table S3.** Similarity of the calculated CD pure spectra for i-motif species with those of a partially stacked single strand (T<sub>30</sub> sequence) and of an “standard” i-motif structure (TT sequence). The dissimilarity values were calculated according to Equations 7 and 8.

Sequence	Species whose CD spectrum is compared	pH value at which the concentration of the species reaches its maximum value <sup>a</sup>	Dissimilarity with the pure CD spectrum of T <sub>30</sub> sequence	Dissimilarity with pure CD spectrum of TT sequence
SMC01	2	5.2	0.5112	0.4791
	3	4.0	0.5523	0.4860
SMC02	2	4.7	0.7122	0.2611
SMC03	2	5.1	0.9192	0.4126
	3	4.0	0.7366	0.3139
SMC04	2	4.9	0.4797	0.5530

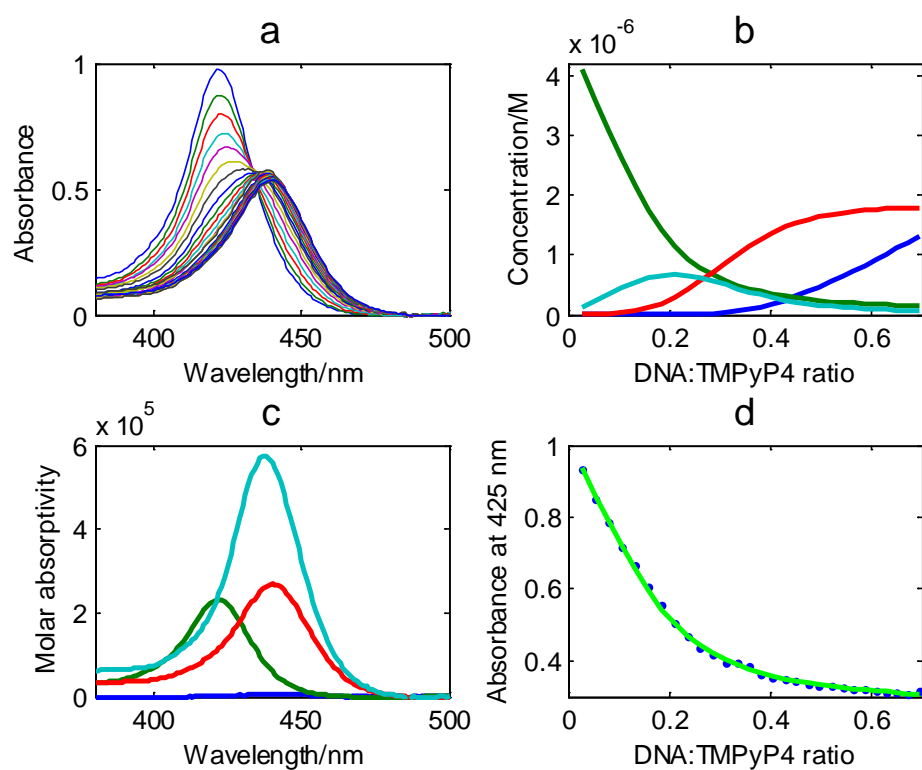
<sup>a</sup> According to Figure 6a, 6c, 6e, and 6g, respectively.

**Figure S11.** 1D  $^1\text{H}$ -NMR spectra of SMC01 (a), SMC02 (b), SMC03 (c) and SMC04 (d) and in the presence of 1:1 and 1:2 DNA:TMPyP<sub>4</sub> ratios. Experimental conditions: 20 mM potassium phosphate, 150 mM KCl,  $C_{\text{DNA}} = 0.1$  mM,  $T = 5$  °C, pH 5.0.



**Figure S12. Determination of the SMC01:TMPyP4 stoichiometry and binding constant.** (a) Experimental molecular absorbance data. (b) Calculated distribution diagram. (c) Calculated pure spectra. (d) Experimental (blue circles) and calculated (green line) absorbance data at 425 nm. In (b) and (c) blue line corresponds to free SMC01, green line corresponds to free TMPyP4, red line corresponds to the 1:2 SMC01:TMPyP4 complex, and the cyan line corresponds to the 1:4 SMC01:TMPyP4 complex. The experiment was carried out at 25 °C, 150 mM KCl, 20 mM acetate buffer, pH 5.2.

Experimental data in (a) were analyzed with a previously described multivariate analysis method that enables the calculation of the binding constants for the proposed model of species, and the corresponding pure spectra [Dyson *et al. Anal. Chim. Acta.* **1997**, 353, 381-393].



**Figure S13.** Melting experiment of mixtures of TMPyP4 with SMC02 (a), SMC03 (b) and SMC04 (c).  $C_{\text{DNA}} = 2 \mu\text{M}$ ,  $C_{\text{TMPyP4}} = 4 \mu\text{M}$ , pH 5.2.

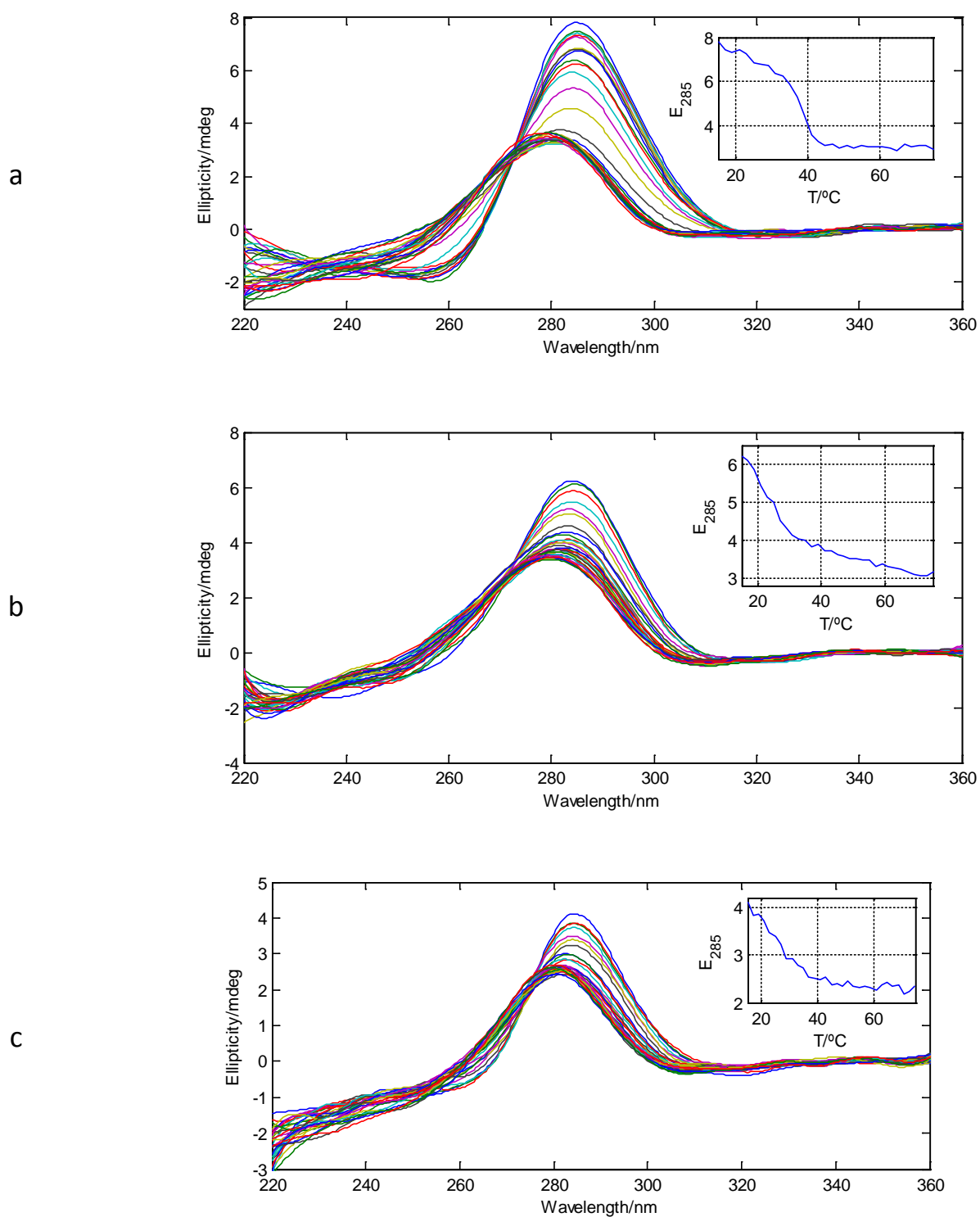
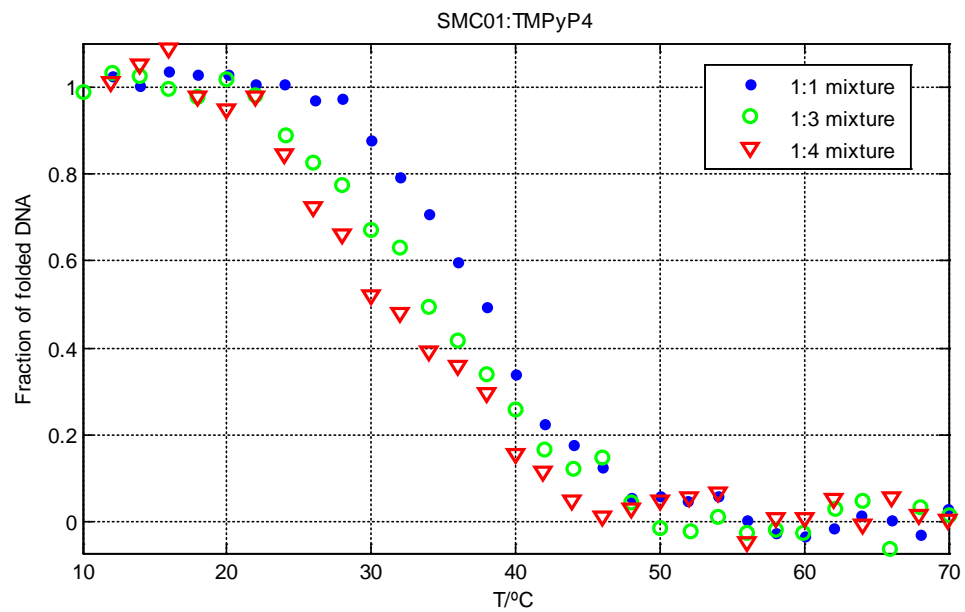




Figure S14. Melting experiment of mixtures of TMPyP4 with SMC01 at different DNA:ligand ratios, pH 5.2.

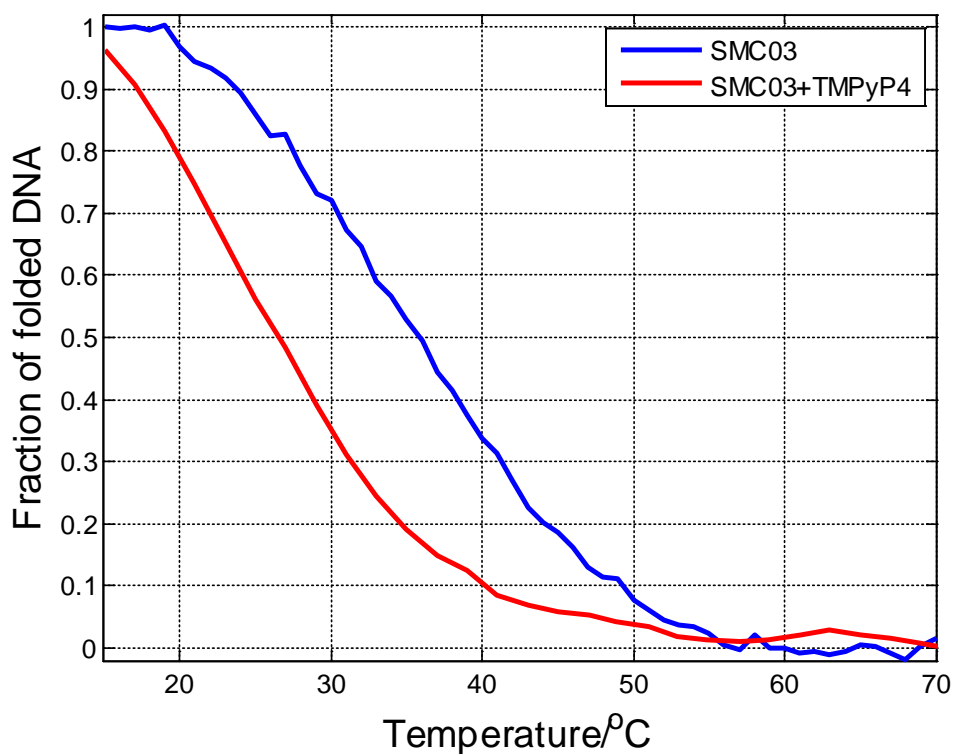


**Table S4.**  $T_m$  values determined for the unfolding of DNA in absence and in presence of TMPyP<sub>4</sub> at pH 5.2 (1:2 DNA:ligand ratio).

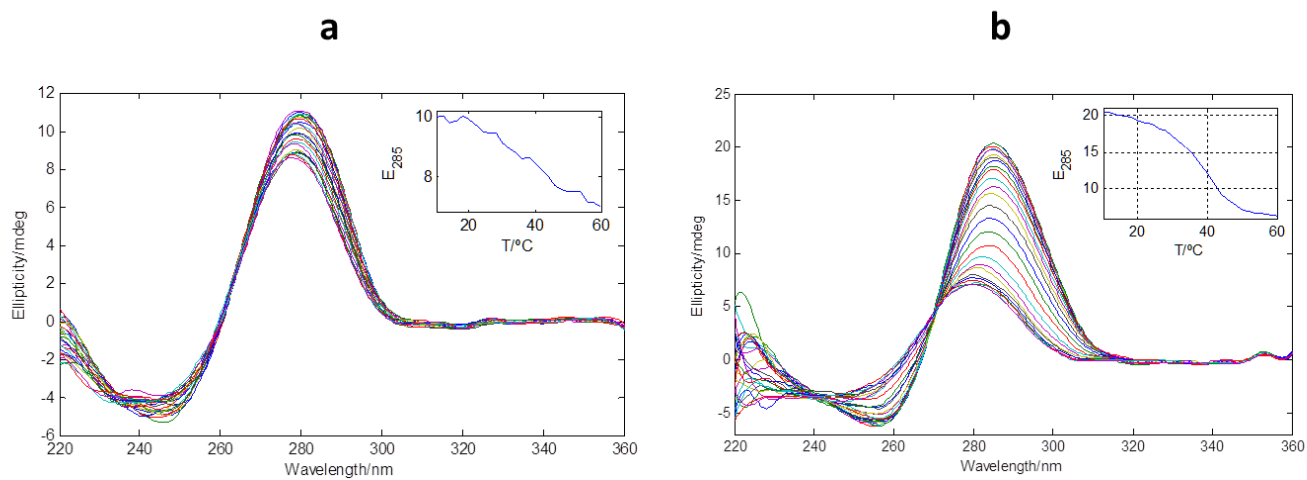
Sequence	$T_m$ (°C)	$T_m$ (°C)	$\Delta T_m$ (°C)
	DNA	DNA+TMPyP4	
SMC01	38	37	-1
SMC02	27 / 42	~38	
SMC03	36	~25	~ -11
SMC04	36	~25	~ -11
TT	55	52	-2

TT denotes the sequence 5'-T<sub>2</sub>C<sub>3</sub>TTTC<sub>3</sub>T<sub>3</sub>C<sub>3</sub>TTTC<sub>3</sub>T<sub>2</sub>-3' that forms an intramolecular i-motif structure stabilized by six C·C<sup>+</sup> base pairs (Benabou et al., *Physical Chemistry Chemical Physics* **2016**, 18, 7997-8004).

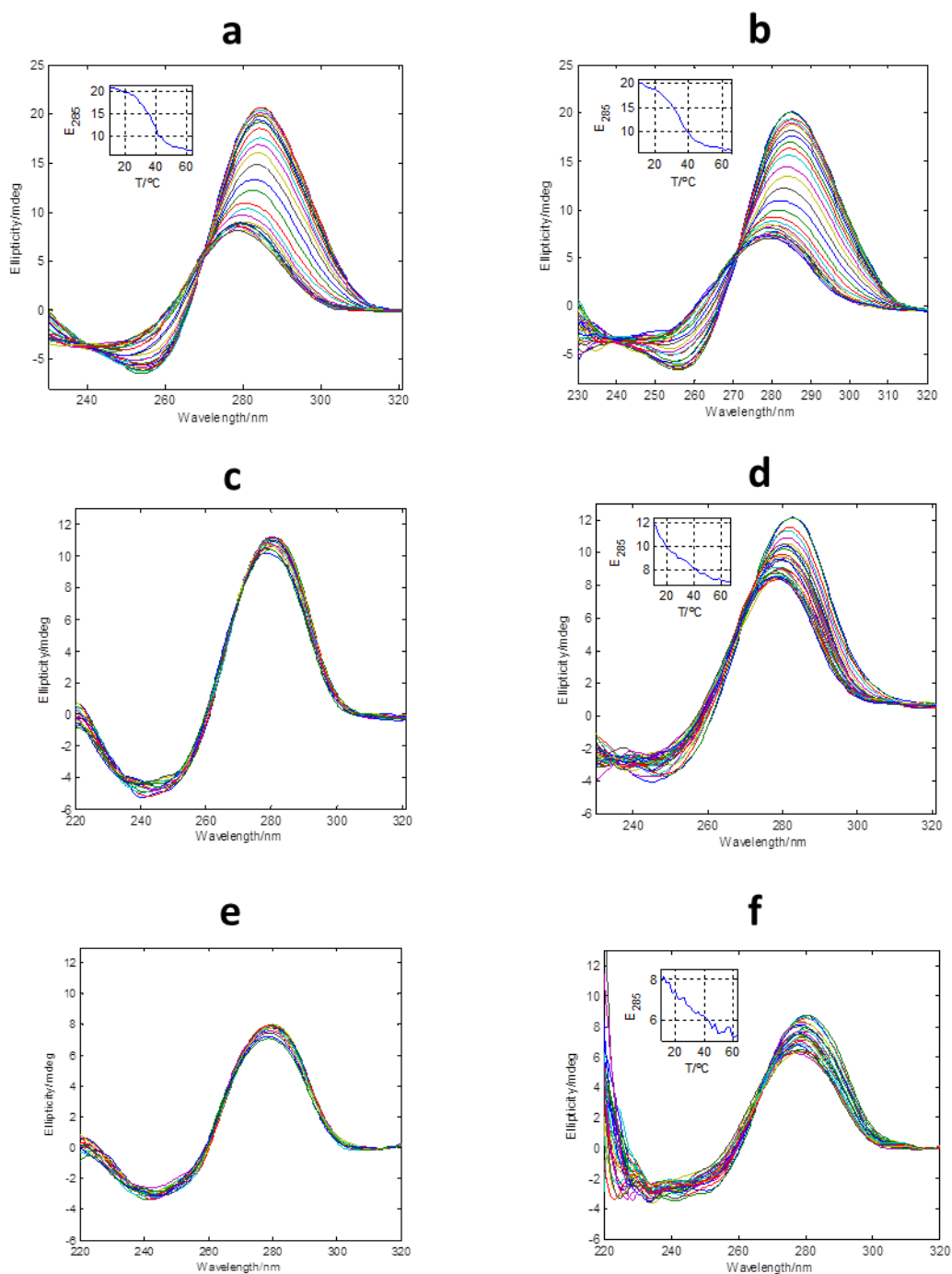
**Figure S15.** Fraction of folded SMC03 in absence and in presence of ligand. The low stability of the complex makes unreliable the determination of thermodynamic parameters.



**Figure S16.** Thermally-induced unfolding of SMC01 in absence and presence of PEG200 (20 % w/v) and 20 mM buffer. (a) pH 6.5, 20 mM phosphate buffer, 150 mM KCl. (b) pH 6.7, 20 mM phosphate buffer, 150 mM KCl, PEG200 (20 % w/v).  $C_{\text{SMC01}} = 2 \mu\text{M}$ .



**Figure S17.** Thermally-induced unfolding of SMC01 in absence (a, c, e) and presence (b, d, f) of PEG200 (20 % w/v) and 100 mM buffer. (a) pH 5.2, 100 mM acetate buffer, 150 mM KCl. (b) pH 5.2, 100 mM acetate buffer, 150 mM KCl, PEG200 (20 % w/v). (c) pH 6.5, 100 mM phosphate buffer, 150 mM KCl. (d) pH 6.5, 100 mM phosphate buffer, 150 mM KCl, PEG200 (20 % w/v). (e) pH 7.2, 100 mM phosphate buffer, 150 mM KCl. (f) pH 7.2, 100 mM phosphate buffer, 150 mM KCl, PEG200 (20 % w/v).  $C_{SMC01} = 2 \mu\text{M}$ . Insets show the ellipticity values at 285 nm versus temperature.



**Figure S18.** pH diagrams of species distribution for 5'-TT CCC TTT CCC TTT CCC TT-3' (TT, left) and 5'-TT CCC TAT CCC TTT CCC TAT CCC TT-3' (AA, right). The titrations were carried out at 25 °C [S. Benabou *et al.*, Phys. Chem. Chem. Phys. 2016, 18, 7997-8004].

