Effects of sequence and base composition on the CD and TDS profiles of i-DNA

Nunzia Iaccarino, Mingpan Cheng, Dehui Qiu, Bruno Pagano, Jussara Amato, Anna Di Porzio, Jun Zhou, Antonio Randazzo, Jean-louis Mergny

RESEARCH ARTICLE

Effects of sequence and base composition on the CD and TDS profiles of i-DNA

Nunzia Iaccarino, [+][a] Mingpan Cheng, [+][b,c] Dehui Qiu, [b] Bruno Pagano, [a] Jussara Amato, [a] Anna Di Porzio, [a] Jun Zhou, [b] Antonio Randazzo, *[a] and Jean-Louis Mergny [b,c,d]

[a] Dr. N. Iaccarino, Prof. B. Pagano, Prof. J. Amato, A. Di Porzio, Prof. A. Randazzo
Department of Pharmacy
University of Naples Federico II
Via D. Montesano 49, 80131 Naples, Italy
E-mail: antonio.randazzo@unina.it

[b] Dr. M. Cheng, D. Qiu, 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.

[c] Dr. M. Cheng, Dr. J.-L. Mergny
ARNA Laboratory
Université de Bordeaux, Inserm U 1212, CNRS UMR5320, IECB
Pessac 33607, France.

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

[+] These authors contributed equally to this work

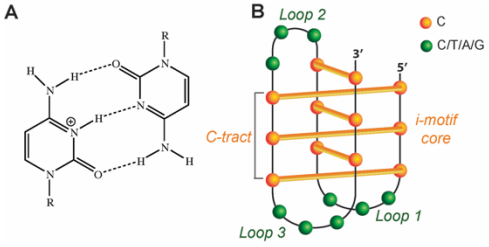
Supporting information for this article is given via a link at the end of the document.

Abstract: The i-motif DNA, also known as i-DNA, is a non-canonical DNA secondary structure formed by cytosine-rich sequences, consisting of two intercalated parallel-stranded duplexes held together by hemi-protonated cytosine-cytosine$^+$ (C:C$^+$) base pairs. The growing interest for the i-DNA structure as a target in anticancer therapy increases the need for tools allowing a rapid and meaningful interpretation of the spectroscopic data of i-DNA samples. Herein, we analyzed the circular dichroism (CD) and thermal difference UV-absorbance spectra (TDS) of 255 DNA sequences by means of multivariate data analysis, aiming at unveiling peculiar spectral regions that could be used as diagnostic features during the analysis of i-DNA-forming sequences.

Introduction

In the last decades, a variety of DNA secondary structures other than the canonical Watson-Crick duplex, have been documented. Such structural polymorphism depends on sequence, hydration, ions and/or ligands and superhelical stress; it occurs during biological processes such as replication and transcription, thus having an impact on genetic stability.[1] Non-canonical DNA structures include hairpins, cruciforms, triplexes, G-quadruplexes (G4s), and i-motifs (i-DNAs).[2–7] i-DNA was first observed in 1993 for the hexamer sequence d(TCCCCC) under acidic conditions.[8] It consists of two intercalated parallel- stranded duplexes held together by hemi-protonated cytosine-cytosine$^+$ (C:C$^+$) base pairs (Figure 1A)[8,9]. i-DNA formation at physiological pH has been recently reported.[10,11] Moreover, the generation of an antibody able to detect i-DNA has proved its presence in the nucleus of human cells, arguing for regulatory roles in the genome, *e.g.*, at proto-oncogene promoters and telomeres,[12] making this DNA structure a potential target for anticancer therapy. Putative i-DNA-forming sequences occur in C-rich strands complementary to G-rich regions that may form G4s. However, if the G4 counterpart folding conditions have already been extensively studied, i-DNA's optimal features for its formation near physiological pH *in vitro* are still under investigation. In fact, several studies have been conducted recently to better understand the influence of external conditions, such as presence of metals[13,14] or ligands,[15-17] molecular crowding,[11,18-19] cation type and ionic strength,[20] on i-DNA

Figure 1. (**A**) Hemi-protonated cytosine-cytosine$^+$ (C:C$^+$) base pair. (**B**) Schematic representation of an i-motif structure.



formation. However, i-DNA stability also depends on sequence composition. A typical formula of an intramolecular i-DNA-forming sequence is $(C_nX_N)_3C_n$, where X can be either a C or non-C (T, A, G); the presence of four C tracts (C_n) allows the generation of a C-stem, while the three spacers (X_N) , connect the four cytosine tracts (C-tracts) and form the loops (Figure 1B).

Much attention has been paid to the influence of the loops' length and composition as well as to the length of the C-tracts. In general, it was found that thymines confer a higher i-DNA stability compared to other non-C deoxynucleotides.[21,22] Very recently, the Vorlickova's group reported a systematic investigation of sequence requirements for i-DNA formation. They found that the lower number of residues are present in the spacers, the more i-DNA is destabilized. This is due to the loss of $C:C^+$ base pairs as several Cs need to be incorporated into the loops to compensate for the short linkers.[23]

In the companion paper of this investigation, some of us have explored the simultaneous variation of both C-tract length and loop arrangements (see Cheng *et al.*). In particular, by analyzing a first set of 180 different DNA sequences (Table S1), the contribution of C-tracts and spacer length on i-DNA stability was explored. The general formula of the majority of i-DNA-forming oligos employed in that study is $C_{(3\text{-}6)}T_{(1\text{-}6)}C_{(3\text{-}6)}T_{(1\text{-}6)}C_{(3\text{-}6)}T_{(1\text{-}6)}C_{(3\text{-}6)}$. We investigated sequences with cytosine tracts of equal length made of 3 up to 6 Cs and three spacers made of 1 up to 6 thymines, in such a combination to have from 4 to 12 total thymines. The samples were named according to the following rationale: a "T" was used as prefix because the three spacers were composed of thymines only; three consecutive numbers were used to describe the lengths of the three spacers in the 5' to 3' direction; while the suffix referred to the length of the C-tracts ("-3", "-4", "-5" or "-6"' for C_3, C_4, C_5, and C_6, respectively). Thus, for example, the sample T124-3 corresponded to the following DNA sequence: 5'-CCCTCCCTTCCCTTTTCCC-3'. An additional set of 75 i-DNA-forming sequences was added to evaluate the effects of different spacer lengths and compositions, terminal nucleobases, and non-equal C-tracts. The nomenclature of these additional sequences, listed in Table S2, uses the same rationale employed for the previous 180 samples. In particular, a subset of 40 samples was designed to generate sequences characterized by five Cs in each C-tract and differing for: (i) the length of central spacer, ranging from seven up to fifteen Ts (4 samples); (ii) the presence of adenines in the different positions of the spacers (24 samples); (iii) the length of the first and third spacers (12 samples). A second subset of 35 samples (all variants based on the original sequence 'T252-5') included: (i) sequences having As, Ts or Gs as flanking nucleobases (9 samples); (ii) sequences with an increasing number of As or Gs in the spacers (20 samples); (iii) sequences with four non-equally sized C-tracts (6 samples). Thus, overall, a total of 255 samples was employed. In particular, ultraviolet (UV) and circular dichroism (CD) spectroscopies at different pH and temperature values were used to evaluate the thermal and pH stability of each i-DNA (see the companion paper by Cheng *et al.*).

In the present work, we analyze the CD spectral profiles and UV thermal difference spectra (TDS) of these 255 samples by means of multivariate data analysis to detect hidden but potentially informative bands in the spectra of the i-DNA-forming sequences. Indeed, to our knowledge, only the well-known i-DNA characteristic bands of the CD (positive at 288 and negative at 264 nm) and TDS (positive at 240 and negative at 295 nm) spectra have been considered so far, limiting the informative power of the CD and UV-absorbance spectroscopies.

Results and Discussion

Considering the large number of spectra to be compared and the very high number of variables (*i.e.*, the intensity at each sampled wavelength) they contain, a plain visual inspection of the TDS and CD spectra may not be sufficient to reveal hidden information in these spectra. However, this huge amount of data can be handled by multivariate data analysis. In particular, the Principal Component Analysis (PCA) is an unsupervised multivariate method that allows the reduction of the dimensionality of a data set, providing a visual representation of the major variance in the data.[24] Particularly, the original variables are transformed into a smaller set of new uncorrelated variables, called principal components (PCs), which are ordered according to the variance they explain (PC1 explains the greatest variance, PC2 contains the second greatest variance, and so on). The principal components are visualized in two plots, termed "scores plot" (where the samples appear close to each other when they are similar, and apart when they are dissimilar) and "loadings plot" (which highlights the variables responsible for the separation of the samples along each PC, in our case the spectral regions). Therefore, this method allows the clustering of the samples based on their similarities and the identification of the spectral regions of the spectra that are characteristic for each cluster. In the following two paragraphs the multivariate data analysis of the CD and TDS spectra is reported.

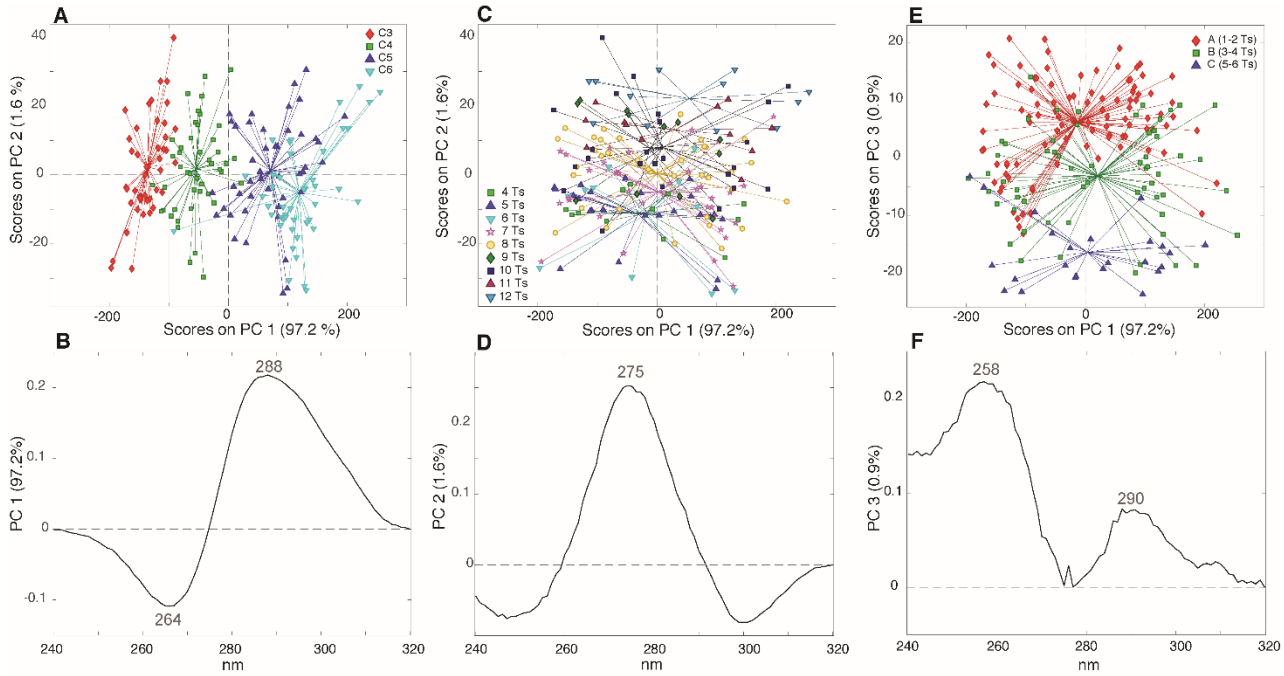


Figure 2. PC1/PC2 score plots of the PCA model calculated using the 180 CD spectra (acquired at pH 5.0) colored according to (**A**) C-tract length, and (**C**) total number of Ts. (**E**) PC1/PC3 score plot colored according to the length of the central spacer. Loading plots of (**B**) PC1, (**D**) PC2, and (**F**) PC3.

Circular Dichroism

We decided to first consider the 180 samples, listed in Table S1, characterized by cytosine tracts of equal length and thymine-based spacers. The characteristic CD profile of an i-DNA structure having a positive and a negative band, respectively, at 288 and 264 nm, is clearly distinguishable at acidic pH values (Figure S1). The intensity values of each data point of the 180 CD spectra acquired at pH 5 were used as variables in a PCA (Figure 2) that produced three meaningful principal components (PCs). Coloring the samples in the PC1/PC2 score plot (Figure 2A) according to the numbers of Cs in the C-tract reveals the first information of this analysis. In fact, the samples turn out to be separated along PC1, where those having a higher number of Cs lie on the right side of the plot, and those having shorter C-tracts lie on the left part. The PC1 loading plot (Figure 2B) reports the variables mainly responsible for the separation along the first principal component. Interestingly, this loading plot closely resembles the CD spectrum of an i-DNA structure. This result indicates that samples with longer C-tracts (that are in the positive region of the PC1) are characterized by more intense CD signals at 288 and 264 nm compared to those having shorter C-tracts. This is also evident from the superimposition of the 180 CD spectra colored according to C-tract length (Figure S2). This observation is not surprising since these bands are directly correlated to the number of C:C$^+$ base pairs in the i-DNA structure.[23,25] Indeed, calculating the Pearson's correlation coefficient *r*, a very good positive correlation is found between the CD signal intensity at 288 nm and the number of Cs in the C-tract (r = 0.90). Interestingly, if the samples are colored according to the total number of Ts (Figure 2C), they turn out to be distributed along PC2. Particularly, the samples characterized by higher number of Ts are placed in the top of the score plot, while samples having a lower number of Ts lie in the bottom of the plot. The PC2 loading plot (Figure 2D) shows that the more Ts are present, the more intense is the signal at 275 nm. This can be explained considering that the CD spectrum of a single-stranded poly(dT) shows a positive band at 275 nm (Figure S3), thus the more Ts are in the sequence, the higher is the signal intensity at 275 nm, independently from the presence of a secondary structure. Indeed, also in this case, a good correlation coefficient between the signal intensity at 275 nm and the total number of Ts in the spacers is found (r = 0.65). Valuable information can also be retrieved by analyzing the PC1/PC3 score plot.

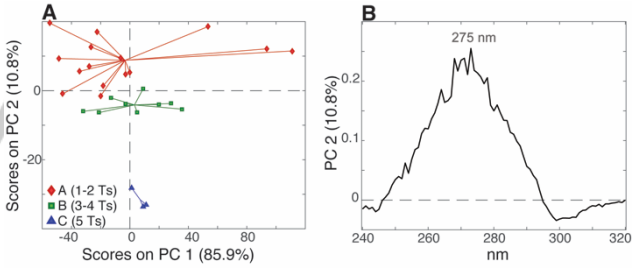


Figure 3. (**A**) PC1/PC2 score plot of the PCA model calculated using the 9 samples, having a total of twenty Cs and seven Ts, acquired at pH 5.00, 5.25, and 5.50 (27 CD spectra in total), colored according to the central spacer length, and (**B**) relative PC2 loading plot.

Indeed, an interesting trend along PC3 is observed when the samples are colored according to the length of the central spacer. This is particularly evident if the samples are grouped in three classes: 'class A' for samples containing 1 or 2 Ts in the central spacer; 'class B' for 3 or 4 Ts; and 'class C' for 5 or 6 Ts (Figure 2E). Unfortunately, the PC3 loading plot (Figure 2F) is difficult to be interpreted, as it is characterized by two positive bands at 258 and 290 nm, apparently not related to any i-DNA bands or base composition of the DNA. The profile reported in the PC3 loading plot is ascribable to a combination of effects, maybe related, to different extents, to sequence composition and conformational features.

In order to verify this hypothesis, we computed another PCA on samples having the same nucleotide composition, but different residue sequences. By way of example, we show the results

obtained by selecting the nine samples having twenty cytosines (five Cs in each C-tract) and seven Ts (differently distributed in the three spacers) namely T115-5/T151-5/T511-5, T223-5/T232-5/T322-5, T331-5/T313-5/T133-5. In order to improve the reliability of the multivariate analysis, we decided to increase the size of this data set by also including the spectra of the selected samples acquired at pH 5.25 and 5.50, after having carefully checked the irrelevance of the slight pH variation on CD spectra (Figure S4A). Thus, a total of 27 spectra (9 samples, whose CD spectra have been acquired at 3 different pH values) were used to compute a new PCA, and the resulting PC1/PC2 score and loading plot are showed in Figures 3A and 3B, respectively. The variance on PC1 was again explained by the bands at 288 nm and 264 nm (Figure S4B), however, in this case, the distribution of the samples is obviously not related to the content of Cs, since all the samples have the same base composition. Most probably, the variance observed in PC1 has to be ascribed simply to an intrinsic uncertainty of the DNA concentration and extinction coefficient values of the samples. On the other hand, the samples turn out to be distributed according to the length of the central spacer along PC2. In this case, the PC2 loading plot (Figure 3B) appears different from the one obtained in the previous analysis (Figure 2F). Indeed, the signal intensity at 275 nm seems to carry the information about the central spacer length; in particular, samples with shorter central spacer show more intense CD signal at 275 nm compared to those with longer central spacer. This result confirms that the PC3 loading plot of the analysis performed on all the samples (Figure 2F) is probably polluted by the information related to the total number of thymines in the sequences that also affects the signal around 275 nm. To shed light on this point, we decided to calculate the correlation coefficients between the CD signal intensity at 275 nm and the central spacer length (r = -0.17). The expected low correlation improves to a value of r = -0.61 when the CD signal intensity at 275 nm is divided by the total number of Ts of each sample. This confirms that the 275 nm wavelength hides both information. Furthermore, the negative sign of the correlation agrees with the PCA outcome, indicating that the longer is the central spacer the less intense is the band at 275 nm.

To better understand, in practice, how the chemical composition of the samples and the length of the central spacer affect the CD profile, it is useful to look at the superimposition of CD spectra of some samples having the same C-tract length and a different number of Ts and samples with the same number of Cs and Ts, but different length of the central spacer. By way of example, the comparison between the CD spectra of T112-4 and T336-4, which have the same number of Cs and different number of Ts, is reported in Figure 4A. The increased intensity around 275 nm, expected for the sample having a higher number of Ts (T336-4), turns into an overall shift of the spectrum towards shorter wavelengths, while the relative intensities of the two bands at 264 and 288 nm remain basically unchanged. In contrast, the comparison of CD spectra of samples that have the same number of cytosines and thymines, but different length of the central spacers (T116-5 and T161-5) (Figure 4B), shows that the decrement of the intensity of the signal around 275 nm for the sample having the longer central spacer (T161-5) turns into a general shift of the spectrum towards longer wavelengths.

Interestingly, in this case, a change in the relative intensities of the bands at 264 and 288 nm is also observable. In order to better visualize this, we decided to normalize each data point of the CD spectra by the intensity of the signal at 264 nm (Figure 5).

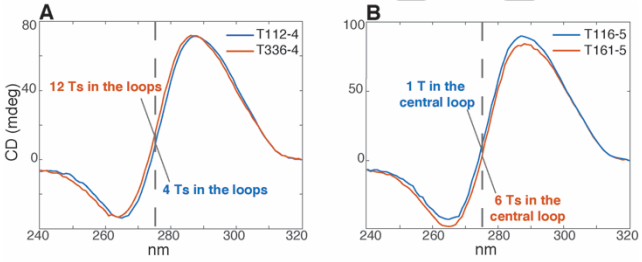


Figure 4. Superimposition of CD spectra of (A) samples having the same C-tract length with different number of Ts (T112-4 vs T336-4), (B) samples having the same chemical composition but different central spacer lengths (T116-4 vs T161-4). The dashed line is centered at 275.

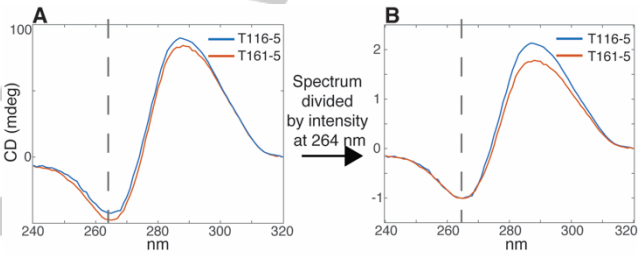


Figure 5. Superimposition of CD spectra of samples having the same chemical composition but with different central spacer length (T116-5 vs T161-5). (**A**) before and (**B**) after dividing the spectra by their intensity value at 264 nm. The dashed line is centered at 264 nm.

This basically provides normalized CD spectra whose bands' intensities are no more related to the number of Cs or Ts. Such normalization was applied to the 180 CD spectra and the new data set was submitted to a PCA (Figure 6).

As expected, the major variance (PC1) is no more related to the length of the C-tract, but it is related to the length of the central spacer (Figure 6A): indeed, samples having 5-6 nucleobases in the central spacer are mainly located in the left part of the plot, while samples having 1 or 2 Ts in their central spacer are placed on the right.

Interestingly, some samples having 1-2 Ts in the central spacer and three Cs in the C-tracts (T211-3, T112-3, T113-3, T311-3, T411-3, T114-3, T611-3, T116-3, T511-3, T115-3, T122-3, T121-3, T212-3) do not follow the general trend observed along PC1 (Figure 6A, inset). The reason of this apparent anomaly may be ascribed to the formation of bimolecular i-DNA structures, as proposed by Vorlickova's group.[23] Indeed, they observed that reducing the length both of the first and third, or the second and the third spacers, bimolecular i-DNAs are preferentially formed, and this only happens when the C-tract is made of three Cs.

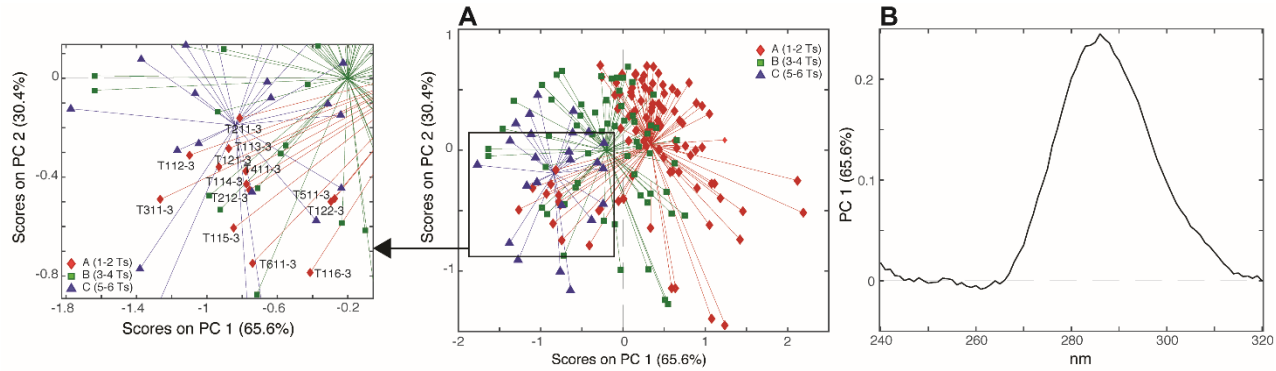


Figure 6. (**A**) PC1/PC2 score plot and relative inset of the PCA model calculated using the 180 CD spectra (normalized by the signal intensity at 264 nm) colored according to the length of the central spacer; (**B**) PC1 loading plot.

As suggested by the PC1 loading plot (Figure 6B), samples having longer central spacer are characterized by a low intensity ratio between the bands at 288 and 264 nm. The reason of this can be explained taking into account the structural requisites for the formation of an i-DNA structure. In particular, the central spacer is often responsible for the formation of a loop that spans the major groove of the i-DNA (see companion paper, Cheng *et al.*) and that this loop requires at least 3 residues.[23] If the central spacer contains a lower number of residues, the i-DNA structure can be formed in any case using some Cs of the adjacent C-tracts. In this case, a lower number of C:C$^+$ pairs is formed and the structure will contain some unpaired Cs. Obviously, the CD spectrum of the sample will proportionally contain information about both paired and unpaired Cs in the structure. As already mentioned, the C:C$^+$ base pairs in the stem provide a CD spectrum having a negative and a positive band at 264 and 288 nm respectively, while cytosines not involved in the pairing are characterized by a CD spectrum having only a positive band centered around 275 nm.[26] As a matter of fact, when both type of Cs contribute to define the general appearance of the CD spectrum, the positive band of the unpaired Cs is summed to the negative band at 264 nm of the paired ones, reducing in this way the intensity of this latter band, while the intensity of the positive band at 288 nm is further increased. Therefore, overall, the relative intensities of the two bands changes proportionally to the number of unpaired Cs in the i-DNA structure. Hence, the higher is the number of unpaired Cs, the higher is the ratio between the intensity of the bands at 288 and 264 nm.

In order to verify the robustness of these findings, we calculated the ratio between the intensity of the positive band at 288 nm and the intensity of the negative band at 264 nm for all the CD spectra (not normalized) employed in the study. A graphical representation of the calculated values is reported in the bar graph in Figure 7. This bar graph confirms that the higher is the central spacer length, the lower is the intensity ratio regardless of the number of Cs in the stem. The exception to this general 'rule' is represented by sequences having very short spacers (T112, T121, T211) and samples having a spacer combination where two spacers are longer than the third one (T221, T212, T122, T331, T313, T133, T332, T323, T332).

In order to evaluate if the findings reported for the first 180 sequences, characterized by equally sized cytosine tracts and thymine-based spacers, were still valid for a more heterogeneous set of i-DNA-forming sequences, we decided to perform multivariate data analyses of the CD spectra of 75 additional sequences divided in two subsets of 40 and 35 samples (Table S2), respectively.

The PCA computed on the subset made of 40 samples (characterized by five Cs in each C-tract and by different lengths and nucleotide composition of the spacers) generated a score plot where the samples are distributed along PC1 according to the length of the central spacer. In particular, the samples having a very short central spacer are characterized by lower intensities i-DNA bands around 264 and 288 nm, compared to the samples having a longer spacer, suggesting that a short central spacer is detrimental to the formation of i-DNA structures (Figures S5 and S6).

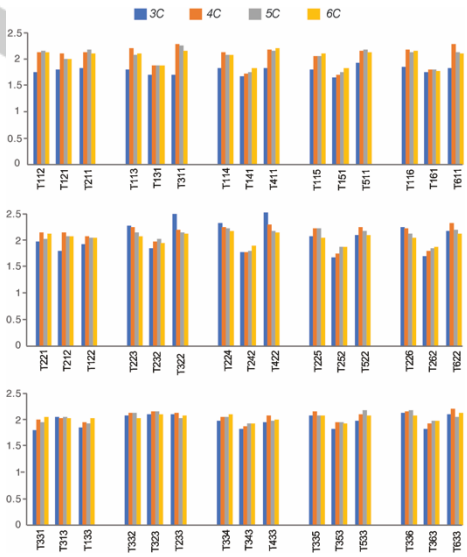


Figure 7. Graphic representation of the maximum (288 nm)/minimum (264 nm) ratio for the 180 CD spectra. Blue, orange, gray and yellow represent, respectively, C-tracts characterized by 3, 4, 5 and 6 Cs.

However, it should be noted that one sample ('T1151-5'), characterized by fifteen thymines in the central spacer, is clustered together with the samples having a single nucleotide in this central position. This observation suggests that a particularly long central spacer is also detrimental for the formation of i-DNA. This is in agreement with the pH and thermal stabilities reported in the companion paper (Cheng *et al.*).

Then, the CD spectra of these 40 samples were normalized by the signal intensity at 264 nm (Figure S7) and the resulting spectra were submitted to PCA. As observed for the previously

normalized 180 CD spectra, the samples having a longer central spacer are characterized by a lower ratio between the intensity of the bands at 288 and 264 nm. Interestingly, the T1151-5 sample is now clustered along those having a long central spacer, indicating that in this case, as expected, formation of the central loop involves this very long spacer and does not require the unpairing of C:C$^+$ base pairs (Figure S8). Thus, this additional data set allows to corroborate the fact that, independently from their composition, short and very long central spacers are detrimental for the formation of i-DNA structures and that, when the central spacer is shorter than three residues, the ratio between the signal intensity at 288 nm and 264 nm increases.

The same analysis has also been performed on the subset made of 35 samples (Figure S9) listed in Table S2. In particular these samples were analyzed in three subgroups to evaluate the impact on the i-DNA formation of three main sequence modifications: (i) addition of flanking bases, (ii) increasing number of purines in the spacers and (iii) non-equal C-tracts. In particular, the CD spectra of 9 samples having Ts, As, or Gs at the 5'- or 3'-ends (or both) were submitted to PCA. The PC1 loading plot indicates that the samples are distributed according to their propensities to form i-DNA. In particular, two samples (TT252-5 and TT252-5T), both characterized by a thymine at the 5'-end, turn out to be characterized by higher intensities of the i-DNA bands at 264 and 288 nm. We speculate that this is due to the formation of a T:T base pair between the thymine at the 5'-end and one of the thymines present in the central spacer (Figures S10A and S10B). The samples are also distributed along PC2 according to the Gs content (increasing number of Gs from the top to the bottom) and As content (increasing number of As from the bottom to the top) (Figures S10C and S10D). The PC2 loading plot indicates that the more Gs and, in turn, the less As, are in the sequence, the lower is the intensity of the band around 270-280 nm and the higher is the intensity of the band around 250-255 nm (Figure S10E). These spectral regions perfectly agree with the characteristic CD bands of a single-stranded poly(dG)[26] and a single-stranded poly(dA).[27] Thus, the PC2 basically explains the contributions of the Gs and As to the i-DNA CD spectrum. Therefore, the analysis of this first subset of samples suggests that the addition of flanking residues to the i-DNA forming sequence may have effects on the CD spectrum both for the change in chemical composition of the samples and also for the different propensity of the sample to form an i-DNA structure, especially for the samples having an additional T at the 5'-end.

The CD spectra of twenty samples belonging to the second subgroup of samples, which contains an increasing number of As or Gs in the spacers, were also submitted to PCA. From the PC1/PC2 score plot (Figure S11A), we first observed the presence of two outliers (A252-5 and G252-5) that bothered the interpretation of the data (see discussion in the Supporting Information, Figure S11). Thus, a new PCA without these two samples was computed. The resulting PC1/PC2 score plot (Figure S12A) shows that samples are distributed according to the number of Gs present in the spacers along PC1. The PC1 loading plot (Figure S12B) reveals that the more Gs are in the spacers, the less intense are the bands around 250-265 nm and 275-300 nm. These bands are wider than those observed for the samples having Gs as flanking sequences (respectively, 250-255 nm and 270-280 nm) and include the wavelength typical of the i-DNA structure (264 and 288 nm). Therefore, this observation could be explained in two ways: (i) the presence of Gs in the sequence may favor the formation of G:C base pairs at the expense of the number of C:C$^+$ pairs, so that the bands of i-DNA structure decrease in intensity; (ii) since the presence of Gs in the sequence naturally increases the positive band at around 255 nm and the negative band at 278 nm,[26] these bands are summed to the bands of the i-DNA structure generating a decrease of their intensities. Instead, we were not able observe a clear contribution of the As (Figure S13). Thus, the analysis of this second subset of samples suggests that the complete substitution of the thymines with purines changes the appearance of the i-DNA CD spectrum, suggesting the formation of additional DNA secondary structures in solution. Instead, when few thymines are substituted with guanines, the i-DNA general spectral profile is still observable even though its characteristic bands are less intense compared to the same sequence having only thymine-based spacers (Figure S14). Also, we found that peculiar bands (255 and 278 nm) are indicative of the content of Gs, in agreement with the observations made for the first subgroup of samples.

Finally, the CD spectra of the last subgroup (six samples) characterized by samples having non-equally sized C-tracts (designed to obtain an odd number of C:C$^+$ pairs – Table S2) was also analyzed by PCAs. Interestingly all observations retrieved from the analysis of the previous 180 samples are perfectly applicable to this sample subset (Figure S15).

The entire data set of CD spectra acquired at pH 7 was also analyzed. As discussed in the Supporting Information (Figures S16-S18), the analysis confirmed the absence of i-DNA structure for the majority of the analyzed samples and revealed the contribution of Cs, Ts, As and Gs to the spectrum. Interestingly, the presence of flanking bases (in particular adenines) seems to induce the i-DNA structure at neutral pH.

Thermal Difference Spectra (TDS). The same approach used for CD was employed to study the TDS of the 255 samples investigated in this study. A TDS is calculated by subtracting the UV spectrum of the folded structure, at low temperature, from that of the unfolded structure, at high temperature. The resulting profile is unique and can be used to obtain a specific signature for DNA secondary structures.[28] At pH 5.0, the TDS profiles are in perfect agreement with the typical i-DNA signature (Figure 8) with a positive peak at 240 nm (used to normalize the spectra) and a negative one at 295 nm.[28] As for CD spectra, we first analyzed the initial set of 180 sequences, characterized by equally sized cytosine tracts and thymine-based spacers. The PCA computed on the such profiles revealed that the band around 250-265 nm is more intense when there are more Ts in the sequence (Figures 9A and 9B) (for a more detailed discussion see the SI and Figures S19 and S20). Moreover, the lower is the C content the higher is the band around 295-310 nm (Figures 9C and 9D). Then, as done for the CD analysis, we decided to get rid of the variability related to the different number of Ts. Thus, we generated a data set including only sequences with seven and eight Ts in the spacers, accounting for a total of 72 samples and a new PCA was computed. The resulting PC1/PC2 score plot shows that the sequences with the longest central spacers (5 or 6 Ts) are separated from the rest of the samples along PC2 and that they are characterized by low values of ΔA around 250–265 nm (Figures 10A and 10B).

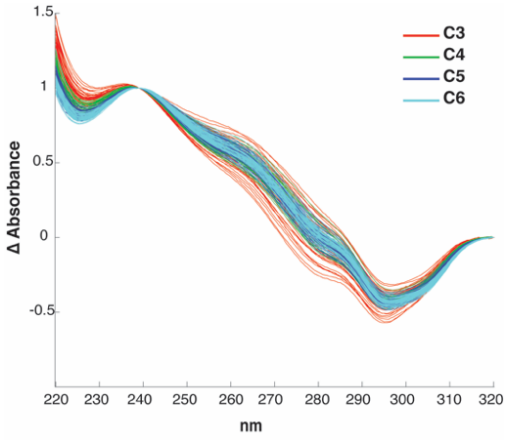


Figure 8. Thermal difference spectra (TDS) of the 180 samples acquired at pH 5.0. The different colors indicate distinct lengths of the C-tract.

Thus, once again, as observed in the CD data set, the information about the total number of Ts and central spacer seems hidden under the same wavelength. This observation was also mathematically verified (see Supporting Information). Then, by coloring the samples in the PC1/PC3 score plot according to the C-tract length (Figure 10C), it is possible to observe the same trend along PC3 observed considering all the sample (Figure 9C), thus confirming that samples having longer central spacers are characterized by lower ΔA around 295–310 nm (Figure 10D). These results can be easily observed comparing the TDS profiles of the samples as showed in Figure 11.

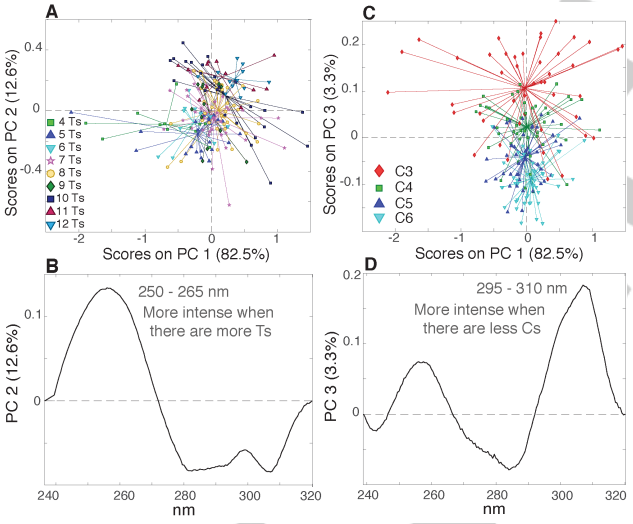


Figure 9. (**A**) PC1/PC2 score plot of the PCA model calculated using the 180 TDS acquired at pH 5.0, colored according to the total number of Ts; (**B**) PC2 loading plot. (**C**) PC1/PC3 score plot colored according to the C-tract length; (**D**) PC3 loading plot.

As in the case for CD, the TDS profiles of the 75 additional sequences were analyzed through PCA. As discussed in Supporting Information (Figures S21-S23), all the observations retrieved from the analysis of the 180 samples turned out to be perfectly applicable also to this additional subset of sequences.

Unfortunately, no peculiar band could be ascribable to the presence of As or Gs. However, some samples (G252-5, TT252-5T, TT252-5, and 252-5_A6) turned out to have unprecedented TDS profiles that, for the time being, we are not able to explain.

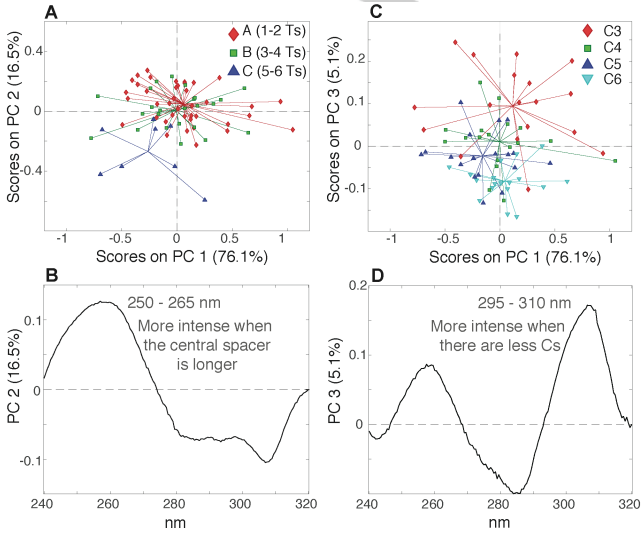


Figure 10. (**A**) PC1/PC2 score plot of the PCA model calculated using the 72 TDS of samples having 7 and 8 Ts (acquired at pH 5.0), colored according to the number of Ts in the central spacer; (**B**) PC2 loading plot. (**C**) PC1/PC3 score plot colored according to the C-tract length; (**D**) PC3 loading plot.

Conclusion

In this work, multivariate data analysis was used to study the TDS and CD spectral profiles of an unprecedented large selection of i-DNA forming sequences (255 in total). This analysis reveals the impact of several kinds of sequence modifications on i-DNA formation and on the entirety of its CD spectral profile. In particular, our findings confirm that, the higher is the number of cytosines in a sequence, the higher its propensity to form i-DNA. This result is true both for sequences having an even and an odd number of C:C$^+$ base pairs. Moreover, our results corroborate the importance of the length of the central spacer for the i-DNA structure. Indeed, we observe that both a particularly short (*i.e.*, one base) or long (*i.e.*, fifteen bases) central spacers are detrimental for i-DNA. Interestingly, the presence of terminal bases (at 5' and 3' ends) is not detrimental for the formation of i-DNA structure. Instead, the presence of a T at the 5' end seems to favor i-DNA formation. Moreover, the complete absence of thymines in the spacers (obtained by their substitution with purines) changes the appearance of the i-DNA CD spectrum, suggesting the formation of additional DNA secondary structures in solution. Instead, a partial substitution of some of the thymines with a few purines still allows i-DNA formation, even if to a lesser extent compared to the same sequence having only thymine-based spacers.

Interestingly, a principal component analysis allows to detect other CD bands than those strictly related to the i-DNA (264 and 288 nm). In particular, we find peculiar informative bands that have never been reported before, related to the presence of thymines, guanines and adenines in the sequences. The intensity of the CD signal intensity at 275 nm is positively correlated with the total number of thymines, while the bands around 250-255 nm

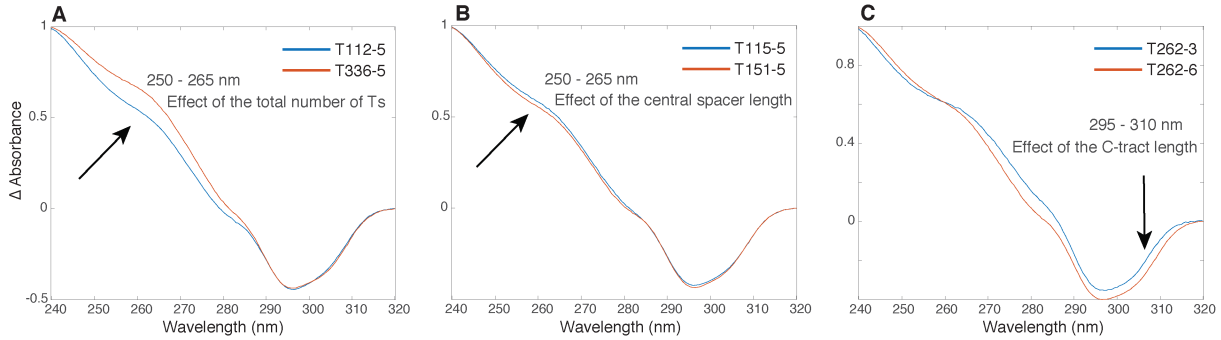


Figure 11. Superimposition of TDS of samples having (**A**) the same C-tract length and different number of total Ts, (**B**) the same chemical composition and different length of the central spacer, (**C**) the same spacers compositions and different C-tract length.

and 270-280 nm are indicative of the presence of adenines and guanines in the sequence. Moreover, the band around 275 nm is negatively correlated with the length of the central spacer. Moreover, the ratio of the intensities of the CD signals at 288 and 264 nm is negatively correlated with the length of the central spacer, when shorter than 3 residues.

The analysis of the CD spectra acquired at pH 7 confirmed the absence of i-DNA structure for the majority of the analyzed samples and revealed the contribution of Cs, Ts, As and Gs to spectrum. Interestingly, the presence of flanking bases (in particular adenines) seems to induce the i-DNA structure at neutral pH.

Furthermore, the multivariate analysis of the TDS data set reveals that the band around 250–265 nm is positively correlated with the total number of Ts, while it is negatively correlated with the length of the central spacer (if divided by the total number of Ts of the sequence). Moreover, the band around 295–310 nm deepens as the number of cytosines in the C-tracts increases.

Our results demonstrate that CD and TDS are much more informative for these structures than previously believed and that they can be used to retrieve interesting structural information on i-DNA.

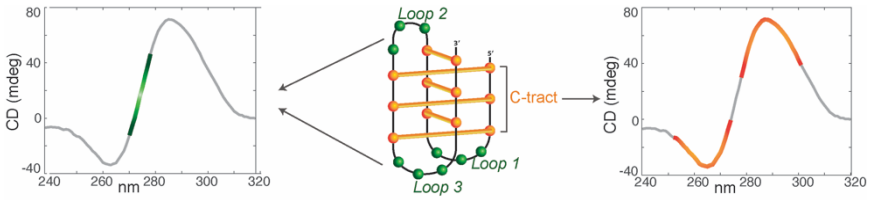
Keywords: Nucleic acids (DNA) • Multivariate data analysis • i-motif • Thermal difference spectra • Circular dichroism

Acknowledgements

This work was supported in part by the Italian Association for Cancer Research AIRC (IG 18695) and by Regione Campania-POR Campania FESR2014/2020 [B61G18000470007], National Natural Science Foundation of China (21977045), funding from Nanjing University (020514912216), and Fundamental Research Funds for the Central Universities (02051430210).

[1] Wang, K. M. Vasquez, *DNA Repair (Amst)*. **2014**, 19, 143–151.
[2] A. Bacolla, R. D. Wells, *Mol. Carcinog.* **2009**, 48, 273–285.
[3] J. van de Sande, N. Ramsing, M. Germann, W. Elhorst, B. Kalisch, E. von Kitzing, R. Pon, R. Clegg, T. Jovin, *Science* **1988**, 241, 551–557.
[4] M. Guéron, J.-L. Leroy, *Curr. Opin. Struct. Biol.* **2000**, 10, 326–331.
[5] M. Gajarský, M. L. Živković, P. Stadlbauer, B. Pagano, R. Fiala, J. Amato, L. Tomáška, J. Šponer, J. Plavec, L. Trantírek, *J. Am. Chem. Soc.* **2017**, 139, 3591–3594.
[6] L. Cerofolini, J. Amato, A. Giachetti, V. Limongelli, E. Novellino, M. Parrinello, M. Fragai, A. Randazzo, C. Luchinat, *Nucleic Acids Res.* **2014**, 42, 13393–13404.
[7] a) S. Neidle, *Nat. Rev. Chem.* **2017**, 1, 41; b) V. Sanchez-Martin, C. Lopez-Pujante, M. Soriano-Rodriguez, J. A. Garcia-Salcedo, *Int. J. Mol. Sci.* **2020**, 21, 8900.
[8] K. Gehring, J. L. Leroy, M. Guéron, *Nature* **1993**, 363, 561–565.
[9] J. Amato, N. Iaccarino, A. Randazzo, E. Novellino, B. Pagano, *ChemMedChem* **2014**, 9, 2026–2030.
[10] J. Zhou, C. Wei, G. Jia, X. Wang, Z. Feng, C. Li, *Mol. BioSyst.* **2010**, 6, 580–586.
[11] A. Rajendran, S. Nakano, N. Sugimoto, *Chem. Commun.* **2010**, 46, 1299.
[12] 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.
[13] H. A. Day, C. Huguin, Z. A. E. Waller, *Chem. Commun.* **2013**, 49, 7696.
[14] S. Saxena, S. Joshi, J. Shankaraswamy, S. Tyagi, S. Kukreti, *Biopolymers* **2017**, 107, e23018.
[15] H. A. Day, P. Pavlou, Z. A. E. Waller, *Bioorganic Med. Chem.* **2014**, 22, 4407–4418.
[16] S. Fernández, R. Eritja, A. Aviñó, J. Jaumot, R. Gargallo, *Int. J. Biol. Macromol.* **2011**, 49, 729–736.
[17] A. Pagano, N. Iaccarino, M. A. S. Abdelhamid, D. Brancaccio, E. U. Garzarella, A. Di Porzio, E. Novellino, Z. A. E. Waller, B. Pagano, J. Amato, A. Randazzo, *Front. Chem.* **2018**, 6, 1–13.
[18] Y. P. Bhavsar-Jog, E. Van Dornshuld, T. A. Brooks, G. S. Tschumper, R. M. Wadkins, *Biochemistry* **2014**, 53, 1586–1594.
[19] J. Zhou, G. Jia, Z. Feng, C. Li, *Chin. J. Chem. U.* **2010**, *31*, 309–311.
[20] N. Iaccarino, A. Di Porzio, J. Amato, B. Pagano, D. Brancaccio, E. Novellino, R. Leardi, A. Randazzo, *Anal. Bioanal. Chem.* **2019**, 411, 7473–7479.
[21] M. McKim, A. Buxton, C. Johnson, A. Metz, R. D. Sheardy, *J. Phys. Chem. B* **2016**, 120, 7652–7661.
[22] A. M. Fleming, K. M. Stewart, G. M. Eyring, T. E. Ball, C. J. Burrows, *Org. Biomol. Chem.* **2018**, 16, 4537–4546.
[23] P. Školáková, D. Renčiuk, J. Palacký, D. Krafčík, Z. Dvořáková, I. Kejnovská, K. Bednářová, M. Vorlíčková, *Nucleic Acids Res.* **2019**, 9, 2177–2189.
[24] a) H. Hotelling, *J. Educ. Psychol.* **1933**, 24, 417–441; b) J. Jaumot, R. Eritja, S. Navea, R. Gargallo, *Anal. Chim. Acta* **2009**, 642, 117–126.
[25] J. L. Mergny, L. Lacroix, C. Hélène, X. Han, J. L. Leroy, *J. Am. Chem. Soc.* **1995**, 117, 8887–8898.
[26] D. M. Gray, F. J. Bollum, *Biopolymers* **1974**, 13, 2087–2102.
[27] J. Greve, M. F. Maestre, A. Levin, *Biopolymers* **1977**, 16, 1489–1504.
[28] J. L. Mergny, J. Li, L. Lacroix, S. Amrane, J. B. Chaires, *Nucleic Acids Res.* **2005**, 33, e138.

Entry for the Table of Contents



i-DNA is an emerging non-canonical DNA secondary structure that represents a suitable target in anticancer therapy. A multivariate data analysis approach unveiled peculiar TDS and CD spectral regions that could be used for the structural determination of i-DNA-forming sequences.

Supporting Information

69451 Weinheim, Germany

SUPPORTING INFORMATION

Table of Contents

SUPPORTING INFORMATION

SUPPORTING INFORMATION

Experimental Procedures

Preparation of oligonucleotides and reagents

The 255 DNA sequences were purchased from Sangon Biotech (Shanghai, China), or Sigma-Aldrich (USA) while all the remaining chemicals were purchased from Sigma-Aldrich. All the details about TDS and CD spectra acquisition are reported in the companion paper.[1]

Data analysis

Circular Dichroism (CD). Before multivariate data analysis, the CD data matrix consisting of 510 rows (255 DNA sequences acquired both at pH 5.00 and pH 7.0) and 101 columns (variables: CD intensities in the 220-320 nm range) was imported in Matlab (R2015b). Spectral variables included in the region between 220 and 240 nm were discarded because they were affected by buffer-related noise. Then, each spectrum was zeroed at 320 nm meaning that the intensity registered at that wavelength was subtracted from each point of the spectrum. The obtained matrix (having 81 variables) was imported in the PLS Toolbox 8.6.1 that works in the Matlab environment. Then the variables were mean centered to perform the Principal Component Analyses on the various blocks of data.

Thermal Difference Spectra (TDS). The TDS data matrix, consisting of 255 rows (DNA sequences acquired at pH 5.00) and 201 columns (variables: intensities in the 220-320 nm range) was imported in PLS Toolbox 8.6.1 that works in the Matlab environment. Spectral variables included in the region between 220 and 240 nm were discarded because affected by buffer-related noise. Next, the variables were mean centered and the Principal Component Analysis was performed.

Principal Component Analysis (PCA). It is a variable reduction technique able to identify patterns in data. PCA aims to detect the correlation between variables; if a strong correlation between variables exists, PCA uses this information to reduce the dimensionality of the dataset into a smaller number of 'principal components' ('PC') that account for most of the variance of the observed variables. Two plots are generated from this analysis: a score plot, where the samples (DNA sequences, in our case) are displayed, and a loading plot that reports the spectral regions that strongly influence each PC.

Pearson's correlation coefficients. This correlation coefficient, indicated as r, was calculated between signal intensities (at a given wavelengths) and features under investigations such as the total number of Cs, total number of Ts, number of Ts in the central spacer and so on. This parameter can range from −1 for a perfect negative linear relationship to +1 for a perfect positive linear relationship between the two variables. A value around 0 indicates no relationship.

Results and Discussion

Circular Dichroism (CD)

Samples with purines in the spacers (20 sequences). The CD spectra of twenty samples belonging to the second subgroup of samples, which contains an increasing number of As or Gs in the spacers, were also submitted to PCA. From the PC1/PC2 score plot (Figure S11A), we first observed the presence of two outliers (A252-5 and G252-5)The PC1 and PC2 loading plots (Figure S11B and S11C) revealed the unusual spectral bands that characterized these two samples. Interestingly, A252-5 and G252-5 are the only sequences characterized by having only purines in the three spacers. This suggests that the complete substitution of thymines with purines in the spacers is detrimental for the i-DNA formation and probably favors the formation of additional secondary structures in solution. This is in agreement with the lower thermal stability of these two samples (compared to the original sequence 'T252-5' that has only thymines in the spacers) reported in the companion paper.[1] Since the interpretation of the PCA is clearly affected by the presence of these outliers, we decided to remove those samples from the dataset and proceed with the analysis.

Analysis at pH 7. The entire data set of CD spectra acquired at pH 7 was also analyzed. Also in this case, different principal component analyses were applied to the three subsets of samples (180, 40 and 35 sequences). Thus, we first considered the CD spectra of the 180 sequences characterized by equally sized cytosine tracts and thymine-based spacers. A PCA was performed and the PC1/PC2 score plot revealed a distribution of the samples, along PC1, according to the number of Cs present in the C-tracts (Figure S16A). In particular, as indicated by the PC1 loading plot (Figure S16B), the more Cs are in the sequence the higher is the intensity of the band at 278 nm. This is in agreement with the CD spectrum of a single-stranded poly(dC) which is characterized by a positive band around 278 nm.[2] Interestingly, along PC2 (Figure S16C), a distribution of the samples due to their thymines' content is visible. The band around 255 nm is responsible for this separation (Figure S16D), in fact the more thymines are in the spacers, the more the band around 255 nm deepens; this agrees with the CD spectrum of a single-stranded poly(dT) (Figure S3). In conclusion, the CD spectra at pH 7 of the 180 i-DNA forming sequences, do not show any evidence of structured i-DNA, while they contain information only about the chemical composition of the samples. Analogously, the PCAs of the CD spectra acquired at pH 7 of the two additional subsets composed of 40 and 35 samples (Figures S17 and S18), mainly revealed information about the chemical composition of the samples. In fact, we could just observe the contribution of the 'free' Cs, Ts, As and Gs to the CD spectrum of the unstructured i-DNA.

However, nine samples that contains terminal nucleobases at the 5'- and 3'-ends (belonging to the subset of 35 samples) revealed the presence of i-DNA. In particular, the PCA revealed that the presence of adenines at both ends (sample 'AT252-5A') tend to induce the formation of the i-DNA structure at neutral pH (Figures S18A and S18B). This agrees also with the thermal stability analysis performed at pH 7 in the companion paper.[1]

SUPPORTING INFORMATION

Thermal Difference Spectra (TDS)

TDS of the 180 sequences. The same approach used to analyze the CD spectra has been employed to study the TDS of the 255 samples investigated in this study. In analogy to the study performed on the CD spectra, we first analyzed the initial set of 180 sequences, characterized by equally sized cytosine tracts and thymine-based spacers. As shown in Figure 8, the samples having three Cs in the C-tracts are characterized by a significative variability in the 260–285 nm region. The PCA has been computed after removing the spectral region between 220-240 nm, since it may be affected by the buffer-related noise. PC1/PC2 and PC1/PC3 score plots with the relative loading plots are showed in Figure 9 and in Figure S19. As expected, the main variation, found by PC1 (82.5% of explained variance), was due to the sequences characterized by C-tracts with three Cs. In particular, six oligos (T311-3, T211-3, T121-3, T112-3, T131-3, T113-3) were rather separated from the rest of the samples (Figure S19A), as indicated in the PC1 loading plot by a small absorbance variation (ΔA) in the 260–285 nm region (Figure S19B). As observed during the analysis of the CD spectra, these samples may form bimolecular structures[3] and therefore could deviate from the regular TDS profiles of the other samples. Interestingly, coloring the samples according to the total number of Ts, it is possible to observe a distribution of the samples along PC2 (Figure 9A), from the samples having a high number of Ts in the top of the score plot to those having a low number of Ts in the bottom. As indicated by the loading plot (Figure 9B), the sequences located at the top of the score plot (characterized by a higher number of Ts) have a higher ΔA around the region 250–265 nm. It is also interesting to analyze the PC1/PC3 score plot (Figure 9C). In fact, by coloring the samples according to the C-tract length, a clear trend along PC3 could be observed. As indicated in the PC3 loading plot, the shorter is the C-tract, the higher is the ΔA between 295 and 310 nm (Figure 9D).

In order to retrieve potential information also on the bands related to the central spacer length, we decided to perform a second PCA on a more 'simplified' data set in which we got rid of the variability generated by a different number of Ts in the samples, as done for CD analysis. Indeed, the new dataset was generated by including only sequences with seven or eight Ts, accounting for a total of 72 samples. The PC1/PC2 score plot shows that PC1 (76.1% of explained variance) was still dominated by the variation detectable around 260–285 nm (Figure S20A and S20B). However, by coloring the 72 samples according to the central spacer length, we found that the samples with the longest central spacers (5 or 6 Ts) are separated from the rest of the samples along PC2 and that they are characterized by low values of ΔA around 250–265 nm (Figure 10A and 10B). Thus, once again, as observed in the CD dataset, the information about the total number of Ts and central spacer seems hidden under the same wavelength. To verify this point, we decided to calculate the Pearson's correlation coefficient (employing the 180 CD spectra acquired at pH 5.00) between the ΔA at 258 nm (representative of the region between 250 and 265 nm) and the length of the central spacer; a poor correlation (r = 0.05) was detected. Interestingly, after dividing the original intensity at 258 nm by the number of Ts in the sequence, the correlation coefficient improves significantly (r = -0.55) and its negative value corroborates the fact that the higher is the number of Ts in the central spacer, the lower is the signal intensity at 258 nm. Then, coloring the samples in the PC1/PC3 score plot according to the C-tract length (Figure 10C), it is possible to observe the same trend along PC3 observed considering all the sample (Figure 9C), thus confirming that samples having longer C-tracts are characterized by lower ΔA around 295–310 nm. These results can be easily observed comparing the TDS spectra of the samples. By way of example, the superimposition of the TDS of T112-5 and T336-5 shows the effect of a different composition in Ts of the analyzed sequences (Figure 11A). Figure 11B, comparing the samples T115-5 and T151-5, instead shows the effect of the central spacer length on the region between 250 and 265 nm. Finally, the superimposition of TDS of samples T262-3 and T262-6 (Figure 11C) shows the influence of the different number of Cs on the spectral region 295–310 nm.

TDS of the additional 75 sequences. As in the case for CD, the TDS profiles of the 75 additional sequences were analyzed through PCA. In particular, the PCA model computed employing the 40 additional sequences (Table S2) shows a slight distribution along PC2 according to the number of Ts in the spacers (Figure S21A). The main band responsible for this distribution is around 255-265 nm, as indicated by the PC2 loading plot (Figure S21B). In order to confirm this observation, we decided to remove the 24 samples containing one or two adenines that bothered the interpretation of the data, and we recomputed the PCA model employing the remaining 16 sequences. The PC1/PC2 score plot, reported in Figure S21C, shows a clearer distribution of the samples, along PC2, according to the number of Ts in the spacers. Also in this case we can observe that the higher is the number of Ts in the sequence, the more intense is the band around 255-265 nm (Figure S21D). This is in line with the observation made from the PCA model obtained employing the 180 sequences.

Then, we performed PCAs employing the TDS of the additional set composed of 35 samples. By plotting the 35 TDS (Figure S22) we could observe unusual spectral profiles generated from four samples ('G252-5', 'TT252-5T', 'TT252-5', and '252-5_A6') but, for the time being, we are not able to explain such behavior. Then, we computed three different PCAs, one for each subgroup of sequences. The first PCA included the 9 samples characterized by the presence of flanking bases and it confirmed the previous observation concerning the correlation between the number of Ts present in the sequence and the band around 250-260 nm (Figures S23A and S23B). Interestingly, the PCA obtained employing the 6 samples characterized by non-equal C-tracts, not only confirmed the observations concerning the contribution of the Ts to the TDS profile but it also confirmed that increasing the number of Cs in the C-tracts reduces the intensity of the band around 295-310 nm, as observed in the 180 sequences (Figures S23C-S23F)

Unfortunately, the PCA model generated from the TDS profiles of characterized by purines in the spacers (20 samples) did not reveal any useful information related to the contribution of adenines and guanines to the spectral profile (data not shown).

Table S1. List of the 180 DNA samples grouped by spacer permutation.

C3 - tract	C4 - tract	C5 - tract	C6 - tract	Total spacer length
Group 1	*Group 16*	*Group 31*	*Group 46*	
T112-3	T112-4	T112-5	T112-6	
T121-3	T121-4	T121-5	T121-6	4
T211-3	T211-4	T211-5	T211-6	
Group 2	*Group 17*	*Group 32*	*Group 47*	
T113-3	T113-4	T113-5	T113-6	
T131-3	T131-4	T131-5	T131-6	5
T311-3	T311-4	T311-5	T311-6	
Group 3	*Group 18*	*Group 33*	*Group 48*	
T114-3	T114-4	T114-5	T114-6	
T141-3	T141-4	T141-5	T141-6	6
T411-3	T411-4	T411-5	T411-6	
Group 4	*Group 19*	*Group 34*	*Group 49*	
T115-3	T115-4	T115-5	T115-6	
T151-3	T151-4	T151-5	T151-6	7
T511-3	T511-4	T511-5	T511-6	
Group 5	*Group 20*	*Group 35*	*Group 50*	
T116-3	T116-4	T116-5	T116-6	
T161-3	T161-4	T161-5	T161-6	8
T611-3	T611-4	T611-5	T611-6	
Group 6	*Group 21*	*Group 36*	*Group 51*	
T221-3	T221-4	T221-5	T221-6	
T212-3	T212-4	T212-5	T212-6	5
T122-3	T122-4	T122-5	T122-6	
Group 7	*Group 22*	*Group 37*	*Group 52*	
T223-3	T223-4	T223-5	T223-6	
T232-3	T232-4	T232-5	T232-6	7
T322-3	T322-4	T322-5	T322-6	
Group 8	*Group 23*	*Group 38*	*Group 53*	
T224-3	T224-4	T224-5	T224-6	
T242-3	T242-4	T242-5	T242-6	8
T422-3	T422-4	T422-5	T422-6	
Group 9	*Group 24*	*Group 39*	*Group 54*	
T225-3	T225-4	T225-5	T225-6	
T252-3	T252-4	T252-5	T252-6	9
T522-3	T522-4	T522-5	T522-6	
Group 10	*Group 25*	*Group 40*	*Group 55*	
T226-3	T226-4	T226-5	T226-6	
T262-3	T262-4	T262-5	T262-6	10
T622-3	T622-4	T622-5	T622-6	
Group 11	*Group 26*	*Group 41*	*Group 56*	
T331-3	T331-4	T331-5	T331-6	
T313-3	T313-4	T313-5	T313-6	7
T133-3	T133-4	T133-5	T133-6	
Group 12	*Group 27*	*Group 42*	*Group 57*	
T332-3	T332-4	T332-5	T332-6	
T323-3	T323-4	T323-5	T323-6	8
T233-3	T233-4	T233-5	T233-6	
Group 13	*Group 28*	*Group 43*	*Group 58*	
T334-3	T334-4	T334-5	T334-6	
T343-3	T343-4	T343-5	T343-6	10
T433-3	T433-4	T433-5	T433-6	
Group 14	*Group 29*	*Group 44*	*Group 59*	
T335-3	T335-4	T335-5	T335-6	
T353-3	T353-4	T353-5	T353-6	11
T533-3	T533-4	T533-5	T533-6	
Group 15	*Group 30*	*Group 45*	*Group 60*	
T336-3	T336-4	T336-5	T336-6	
T363-3	T363-4	T363-5	T363-6	12
T633-3	T633-4	T633-5	T633-6	

[a] The first 'T' letter means that all the spacers are composed of thymine bases only; three consecutive numbers refer to lengths of the three spacers in the 5' to 3' direction; '-3, -4, -5 and -6' refer to sequences with four C3, C4, C5, and C6 tracts (all of equal length), respectively. For example, the T112-3 sequence is 5'-CCCTCCCTCCCTTCCC-3' (four repeats of 3 cytosines separated by one, one, and two thymines).

Table S2. List of the 75 additional DNA samples, and relative sequences, included in the study.

Name	Sequence (5' → 3')
Subset of 40 sequences	
Sequences with longer (7-15) central spacer (4 samples)	
T171-5	CCCCC T CCCCC TTTTTTT CCCCC T CCCCC
T181-5	CCCCC T CCCCC TTTTTTTT CCCCC T CCCCC
T1101-5	CCCCC T CCCCC T_{10} CCCCC T CCCCC
T1151-5	CCCCC T CCCCC T_{15} CCCCC T CCCCC
Sequences with 1 or 2 adenines in the spacers (24 samples)	
AA115-5	CCCCC A CCCCC A CCCCC TTTTT CCCCC
AA151-5	CCCCC A CCCCC TTTTT CCCCC A CCCCC
AA511-5	CCCCC TTTTT CCCCC A CCCCCACCCCC
1A15-5	CCCCC A CCCCCT CCCCC TTTTT CCCCC
11A5-5	CCCCC T CCCCC A CCCCC TTTTT CCCCC
1A51-5	CCCCC A CCCCC TTTTT CCCCC T CCCCC
151A-5	CCCCC T CCCCC TTTTT CCCCCACCCCC
51A1-5	CCCCC TTTTT CCCCC A CCCCC T CCCCC
511A-5	CCCCC TTTTT CCCCC T CCCCC A CCCCC
115_1A-5	CCCCC T CCCCC T CCCCC ATTTT CCCCC
151_1A-5	CCCCC T CCCCC ATTTT CCCCC T CCCCC
511_1A-5	CCCCC ATTTT CCCCC T CCCCC T CCCCC
115_2A-5	CCCCC T CCCCC T CCCCC TATTT CCCCC
151_2A-5	CCCCC T CCCCC TATTT CCCCC T CCCCC
511_2A-5	CCCCC TATTT CCCCC T CCCCC T CCCCC
115_3A-5	CCCCC T CCCCC T CCCCC TTATT CCCCC
151_3A-5	CCCCC T CCCCC TTATT CCCCC T CCCCC
511_3A-5	CCCCC TTATT CCCCC T CCCCC T CCCCC
115_4A-5	CCCCC T CCCCC T CCCCC TTTAT CCCCC
151_4A-5	CCCCC T CCCCC TTTAT CCCCC T CCCCC
511_4A-5	CCCCC TTTAT CCCCC T CCCCC T CCCCC
115_5A-5	CCCCC T CCCCC T CCCCC TTTTA CCCCC
151_5A-5	CCCCC T CCCCC TTTTA CCCCC T CCCCC
511_5A-5	CCCCC TTTTA CCCCC T CCCCC T CCCCC
Sequences with two short spacers of different length (12 samples)	
T152-5	CCCCC T CCCCC TTTTT CCCCC TT CCCCC
T251-5	CCCCC TT CCCCC TTTTT CCCCC T CCCCC
T153-5	CCCCC T CCCCC TTTTT CCCCC TTT CCCCC
T351-5	CCCCC TTT CCCCC TTTTT CCCCC T CCCCC
T253-5	CCCCC TT CCCCC TTTTT CCCCC TTT CCCCC
T352-5	CCCCC TTT CCCCC TTTTT CCCCC TT CCCCC
T162-5	CCCCC T CCCCC TTTTTT CCCCC TT CCCCC
T261-5	CCCCC TT CCCCC TTTTTT CCCCC T CCCCC
T163-5	CCCCC T CCCCC TTTTTT CCCCC TTT CCCCC
T361-5	CCCCC TTT CCCCC TTTTTT CCCCC T CCCCC
T263-5	CCCCC TT CCCCC TTTTTT CCCCC TTT CCCCC
T362-5	CCCCC TTT CCCCC TTTTTT CCCCC TT CCCCC
Subset of 35 sequences (T252-5 based variants)	
T252-5	CCCCC TT CCCCC TTTTT CCCCC TT CCCCC
Sequences with flanking bases (9 samples)	
TT252-5	T CCCCC TT CCCCC TTTTT CCCCC TT CCCCC
T252-5T	CCCCC TT CCCCC TTTTT CCCCC TT CCCCC T
TT252-5T	T CCCCC TT CCCCC TTTTT CCCCC TT CCCCC T
AT252-5	A CCCCC TT CCCCC TTTTT CCCCC TT CCCCC
T252-5A	CCCCC TT CCCCC TTTTT CCCCC TT CCCCC A
AT252-5A	A CCCCC TT CCCCC TTTTT CCCCC TT CCCCC A
GT252-5	G CCCCC TT CCCCC TTTTT CCCCC TT CCCCC
T252-5G	CCCCC TT CCCCC TTTTT CCCCC TT CCCCC G
GT252-5G	G CCCCC TT CCCCC TTTTT CCCCC TT CCCCC G

Sequences with purines in the spacers (20 samples)	
252-5_A1	CCCCC AT CCCCC TTTTT CCCCC TT CCCCC
252-5_A2	CCCCC TA CCCCC TTTTT CCCCC TT CCCCC
252-5_A3	CCCCC TT CCCCC ATTTT CCCCC TT CCCCC
252-5_A4	CCCCC TT CCCCC TTTTA CCCCC TT CCCCC
252-5_A5	CCCCC TT CCCCC TTTTT CCCCC AT CCCCC
252-5_A6	CCCCC TT CCCCC TTTTT CCCCC TA CCCCC
252-5_AA1	CCCCC AA CCCCC TTTTT CCCCC TT CCCCC
252-5_AA2	CCCCC TT CCCCC AAAAA CCCCC TT CCCCC
252-5_AA3	CCCCC TT CCCCC TTTTT CCCCC AA CCCCC
A252-5	CCCCC AA CCCCC AAAAA CCCCC AA CCCCC
252-5_G1	CCCCC GT CCCCC TTTTT CCCCC TT CCCCC
252-5_G2	CCCCC TG CCCCC TTTTT CCCCC TT CCCCC
252-5_G3	CCCCC TT CCCCC GTTTT CCCCC TT CCCCC
252-5_G4	CCCCC TT CCCCC TTTTG CCCCC TT CCCCC
252-5_G5	CCCCC TT CCCCC TTTTT CCCCC GT CCCCC
252-5_G6	CCCCC TT CCCCC TTTTT CCCCC TG CCCCC
252-5_GG1	CCCCC GG CCCCC TTTTT CCCCC TT CCCCC
252-5_GG2	CCCCC TT CCCCC GGGGG CCCCC TT CCCCC
252-5_GG3	CCCCC TT CCCCC TTTTT CCCCC GG CCCCC
G252-5	CCCCC GG CCCCC GGGGG CCCCC GG CCCCC
Sequences with non-equal C-tracts (6 samples)	
T225-45	CCCC TT CCCCC TT CCCC TTTTT CCCCC
T252-45	CCCC TT CCCCC TTTTT CCCC TT CCCCC
T522-45	CCCC TTTTT CCCCC TT CCCC TT CCCCC
T225-56	CCCCC TT CCCCCC TT CCCCC TTTTT CCCCCC
T252-56	CCCCC TT CCCCCC TTTTT CCCCC TT CCCCCC
T522-56	CCCCC TTTTT CCCCCC TT CCCCC TT CCCCCC

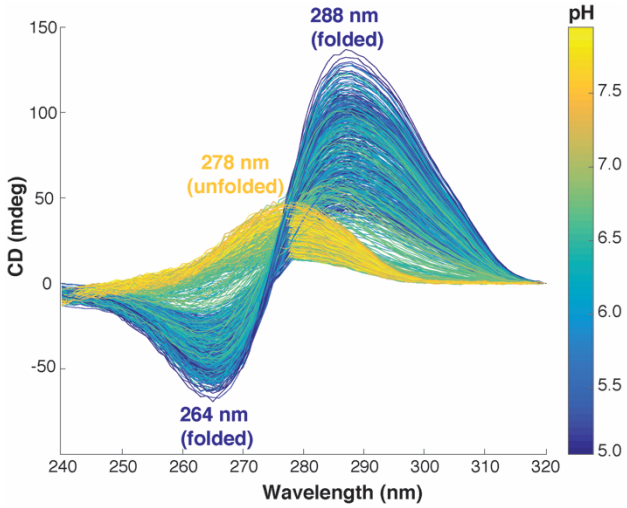


Figure S1. CD spectra of the 180 DNA samples acquired at 13 different pH values (2340 spectra). The color scale goes from yellow (pH 8.00) to dark blue (pH 5.00). Sequences are given in Table S1.

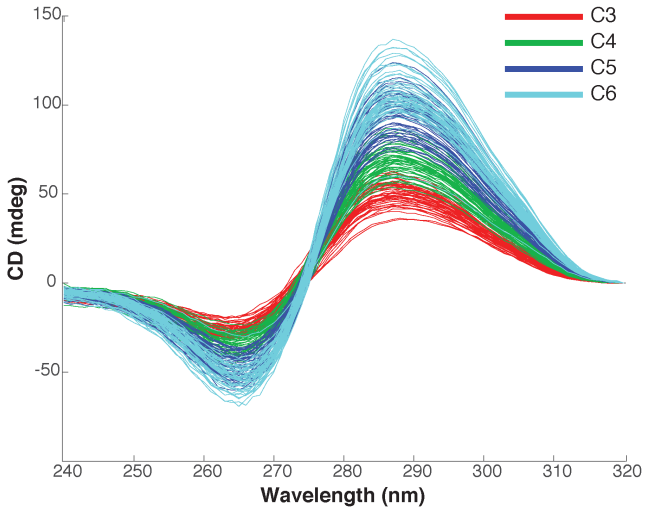


Figure S2. CD spectra of the 180 DNA samples acquired at pH 5.00 colored according to the C-tract length. Sequences are given in Table S1.

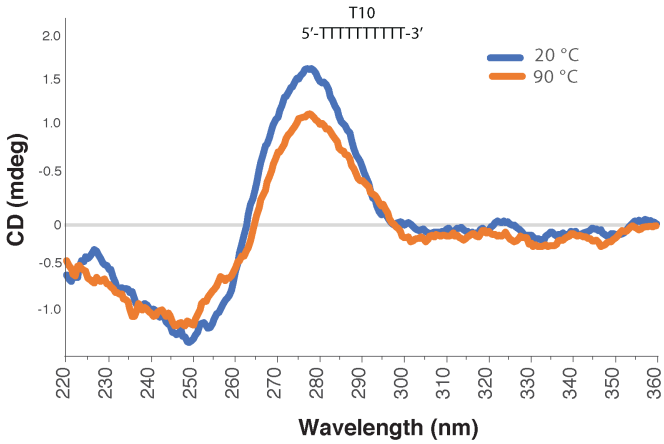


Figure S3. CD spectrum of a poly(dT) 'T10' in water, acquired at 20 °C (blue) and 90 °C (orange).

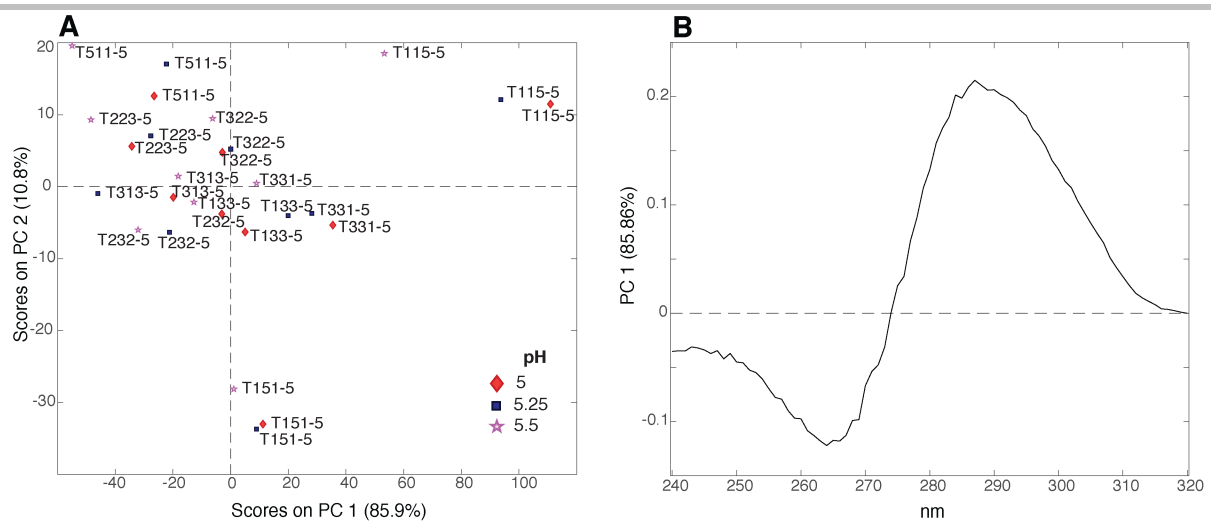


Figure S4. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 9 samples, having a total of twenty Cs and seven Ts, acquired at pH 5.00, 5.25, and 5.50 (27 CD spectra in total), colored according to the pH values, and (B) relative PC1 loading plot.

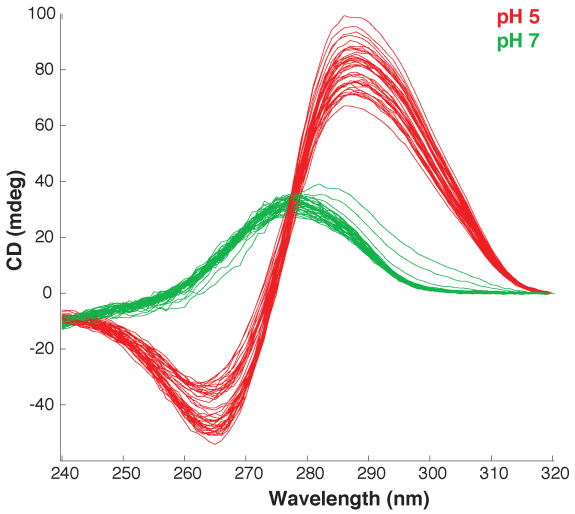


Figure S5. CD spectra of the additional 40 DNA samples acquired at pH 5.00 (red) and pH 7.00 (green). Sequences are given in Table S2.

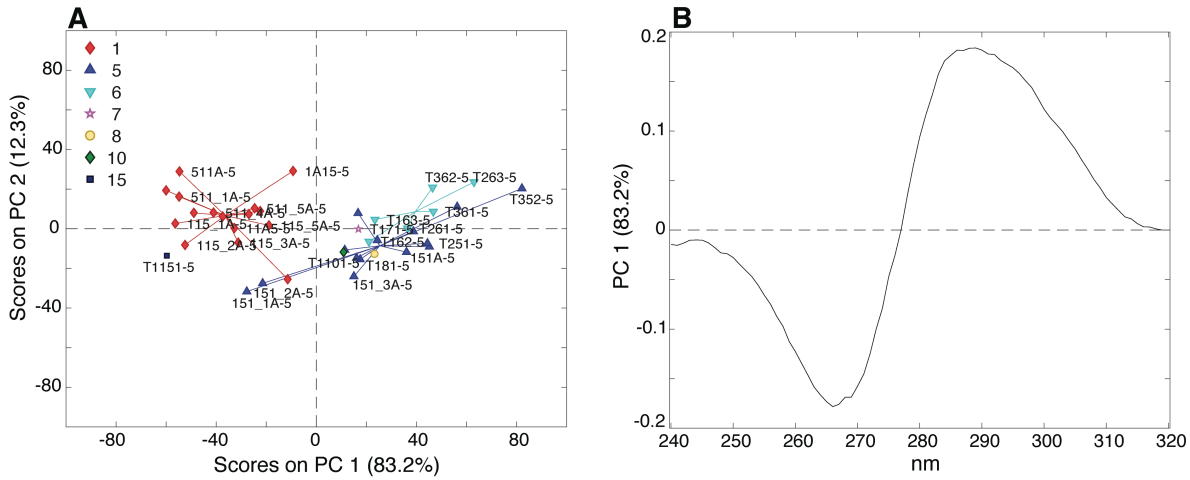


Figure S6. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 40 additional samples, acquired at pH 5.00, colored according to the length of the central spacer, and (B) relative PC1 loading plot.

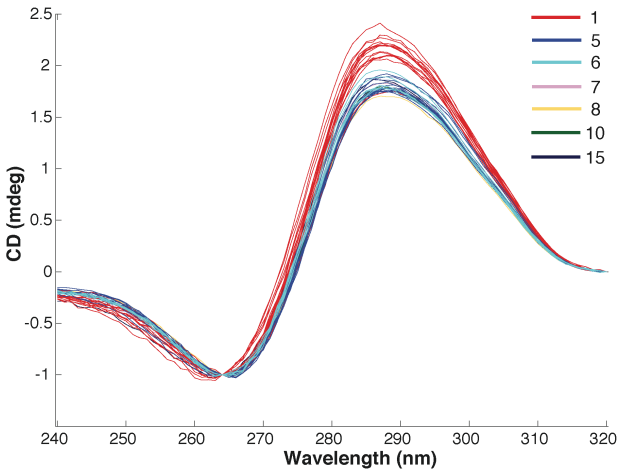


Figure S7. CD spectra of the additional 40 DNA samples, acquired at pH 5.00, and normalized by the intensity of the CD signal at 264 nm. The spectra are colored according to the length of the central spacer. Sequences are given in Table S2.

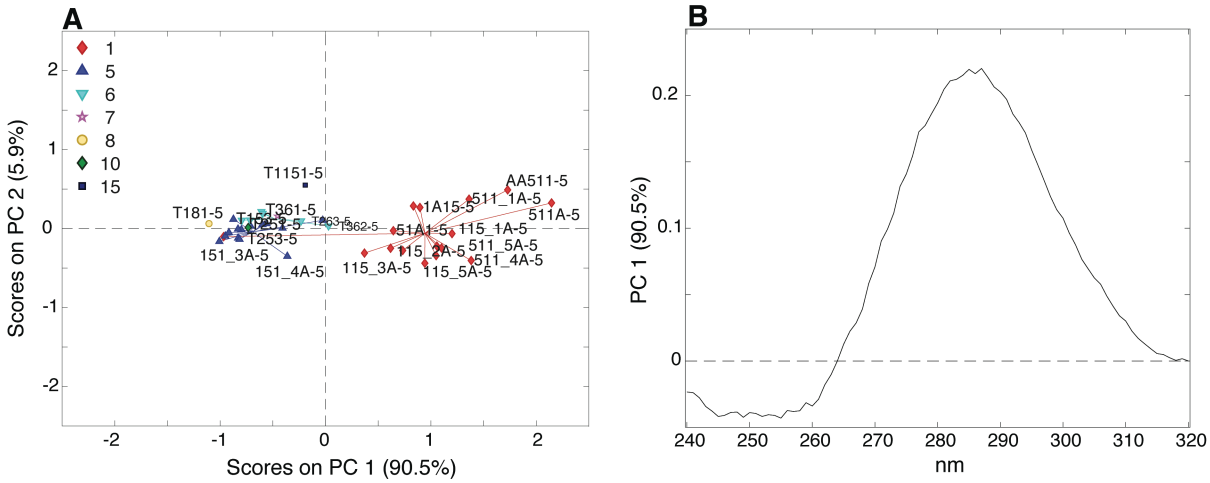


Figure S8. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 40 additional samples, acquired at pH 5.00, normalized by the intensity of the signal at 264 nm and colored according to the length of the central spacer, and (B) relative PC1 loading plot.

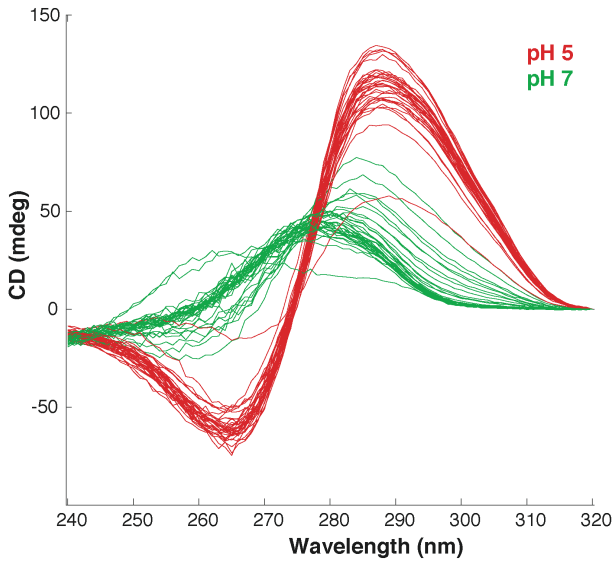


Figure S9. CD spectra of the additional 35 DNA samples acquired at pH 5.00 (red) and pH 7.00 (green). Sequences are given in Table S2.

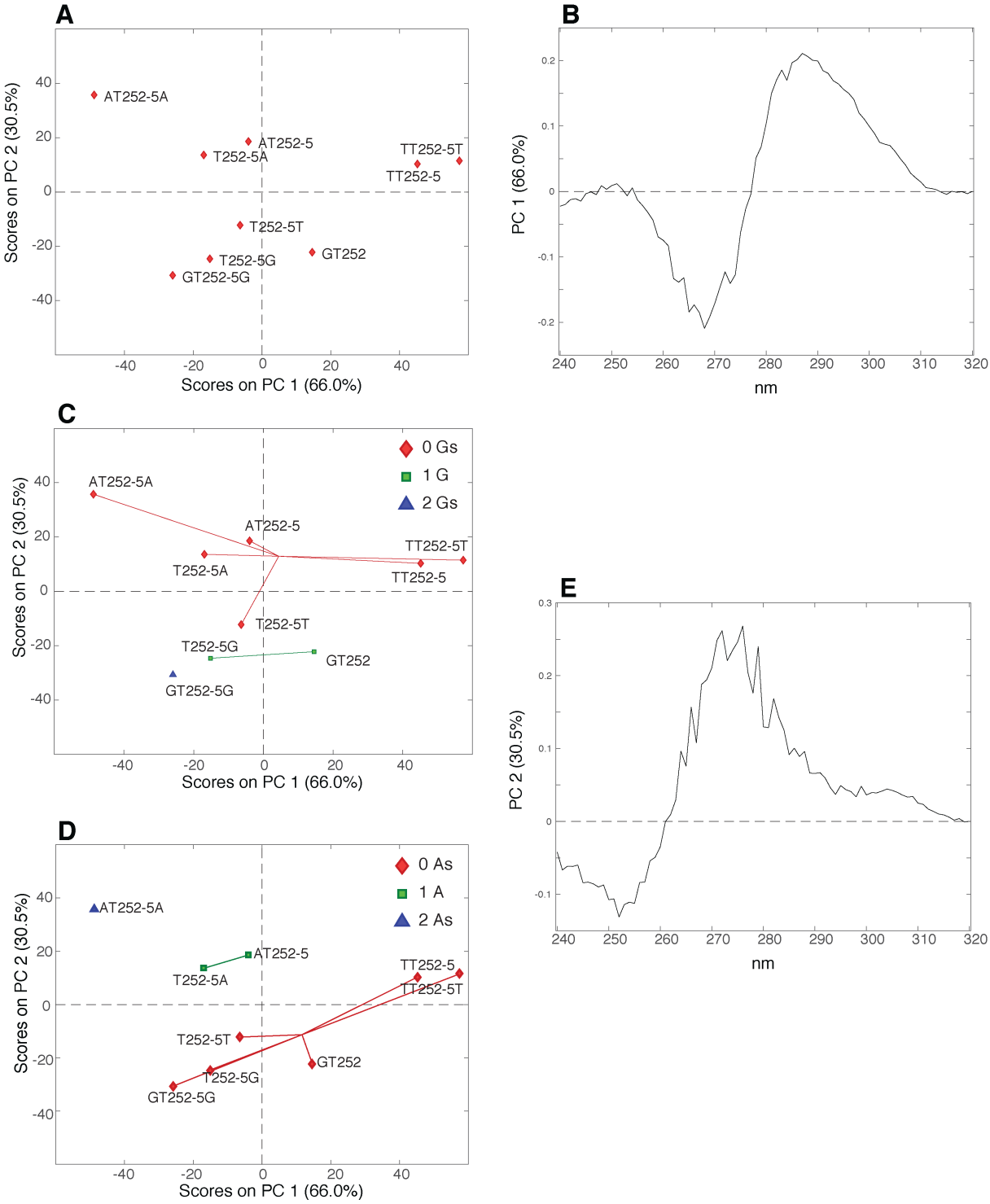


Figure S10. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 9 samples characterized by flanking bases (from the dataset made of 35 sequences) acquired at pH 5.00, colored according to the number of (C) guanines (Gs) and (D) adenines (As) and relative (B) PC1 and (E) PC2 loading plots.

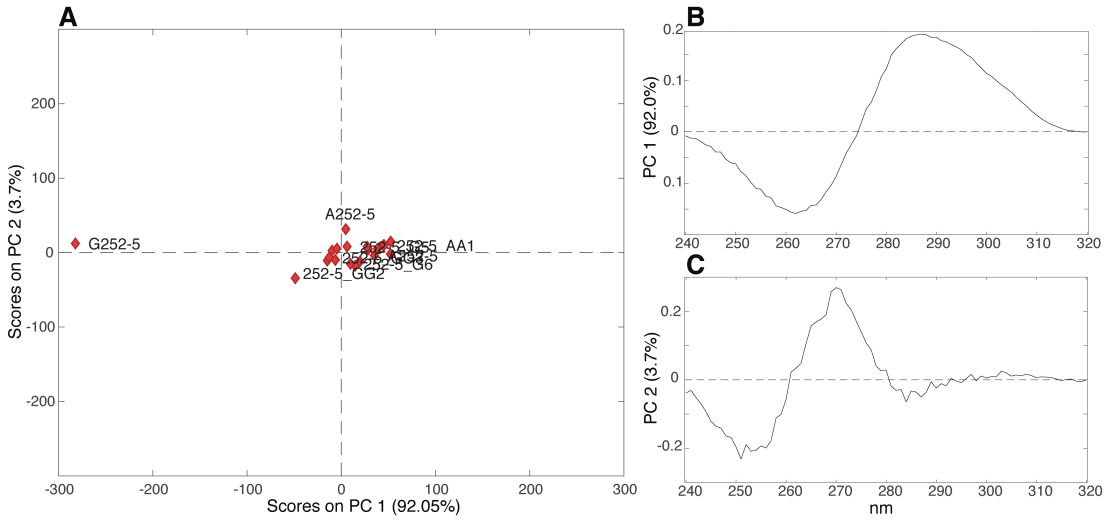


Figure S11. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 20 samples characterized by adenines and guanines in the spacers (from the dataset made of 35 sequences) acquired at pH 5.00 and relative (B) PC1 and (C) PC2 loading plots.

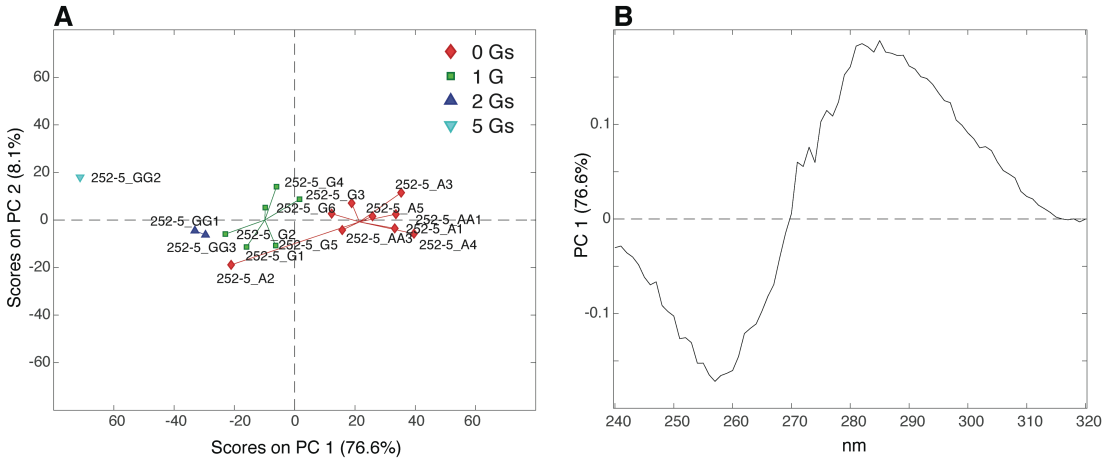


Figure S12. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 18 samples characterized by adenines and guanines in the spacers (samples 'A252-5' and 'G252-5' have been excluded) acquired at pH 5.00, colored according to the number of Gs, and relative (B) PC1 loading plot.

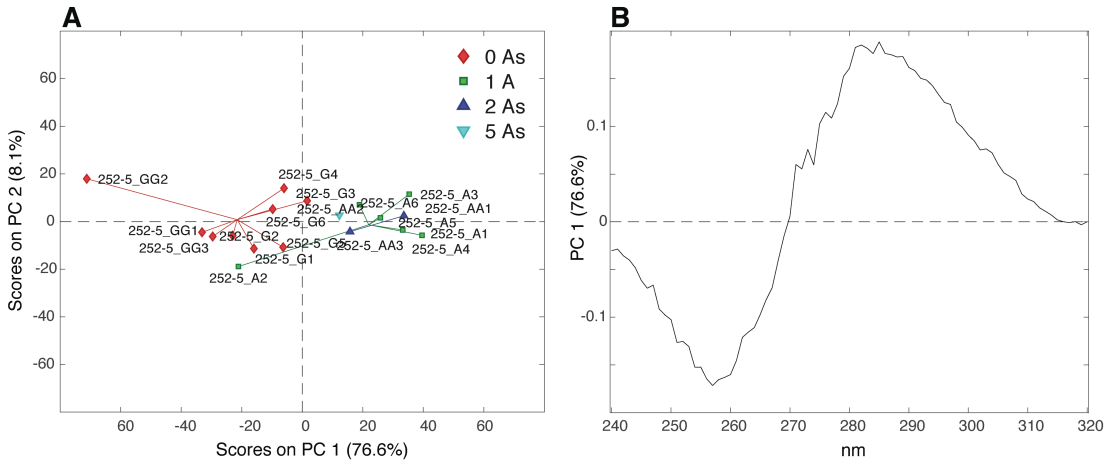


Figure S13. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 18 samples characterized by adenines and guanines in the spacers (samples 'A252-5' and 'G252-5' have been excluded) acquired at pH 5.00, colored according to the number of As, and relative (B) PC1 loading plot.

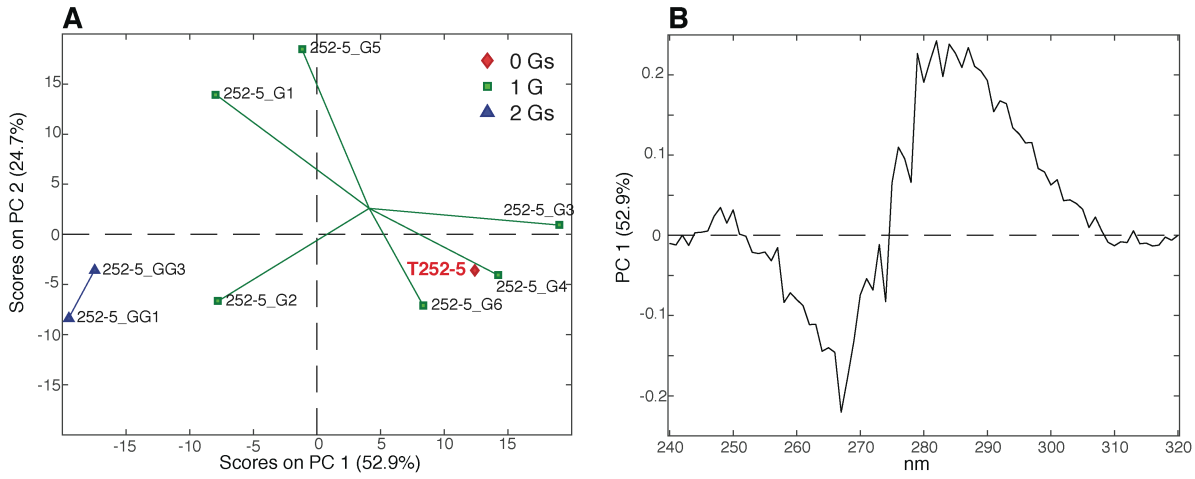


Figure S14. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra (pH 5.00) of the 8 samples characterized by guanines in the spacers ('G252-5' and '252-5_GG2' have been excluded because they would drive the separation of the samples for their peculiar spectral profiles) together with the unmodified sequence 'T252-5' characterized by only thymine-based spacers. Samples are colored according to the number of Gs; (B) PC1 loading plot.

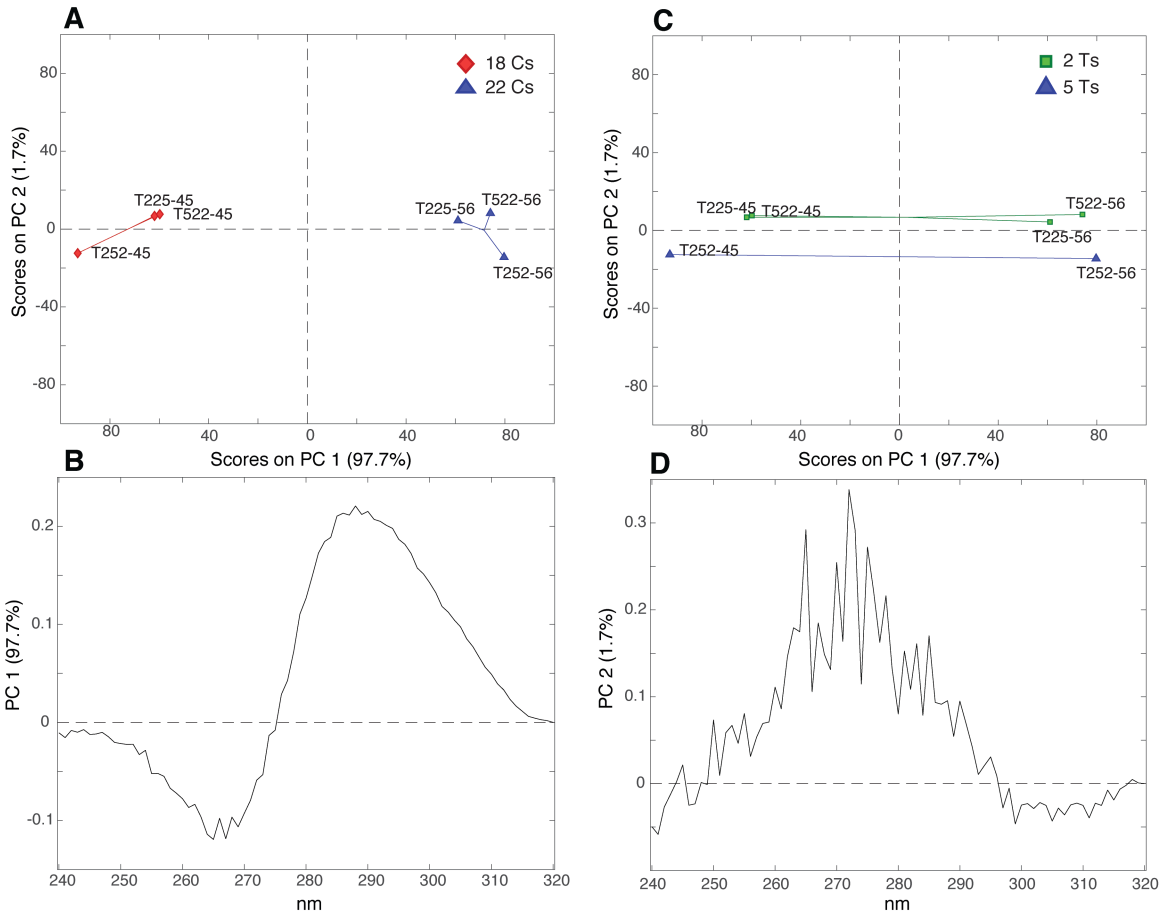


Figure S15. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 6 samples characterized by non-equally sized C-tracts (from the dataset made of 35 sequences) acquired at pH 5.00, colored according to the number of cytosines and (C) according to the length of the central spacer and relative (B) PC1 and (D) PC2 loading plots.

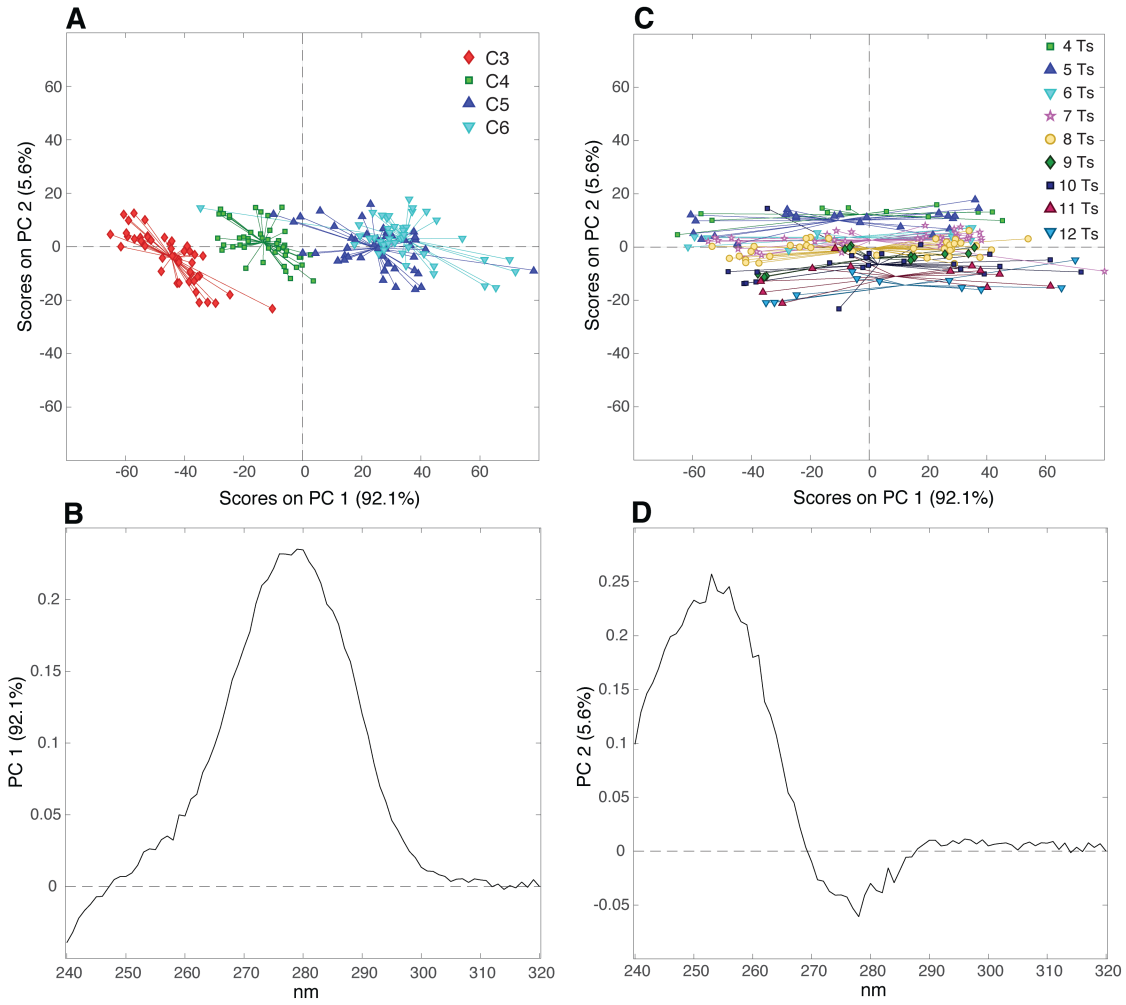


Figure S16. PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 180 samples acquired at pH 7.00, colored according to the number of (A) cytosines, and (C) thymines, and relative (B) PC1 and (D) PC2 loading plots.

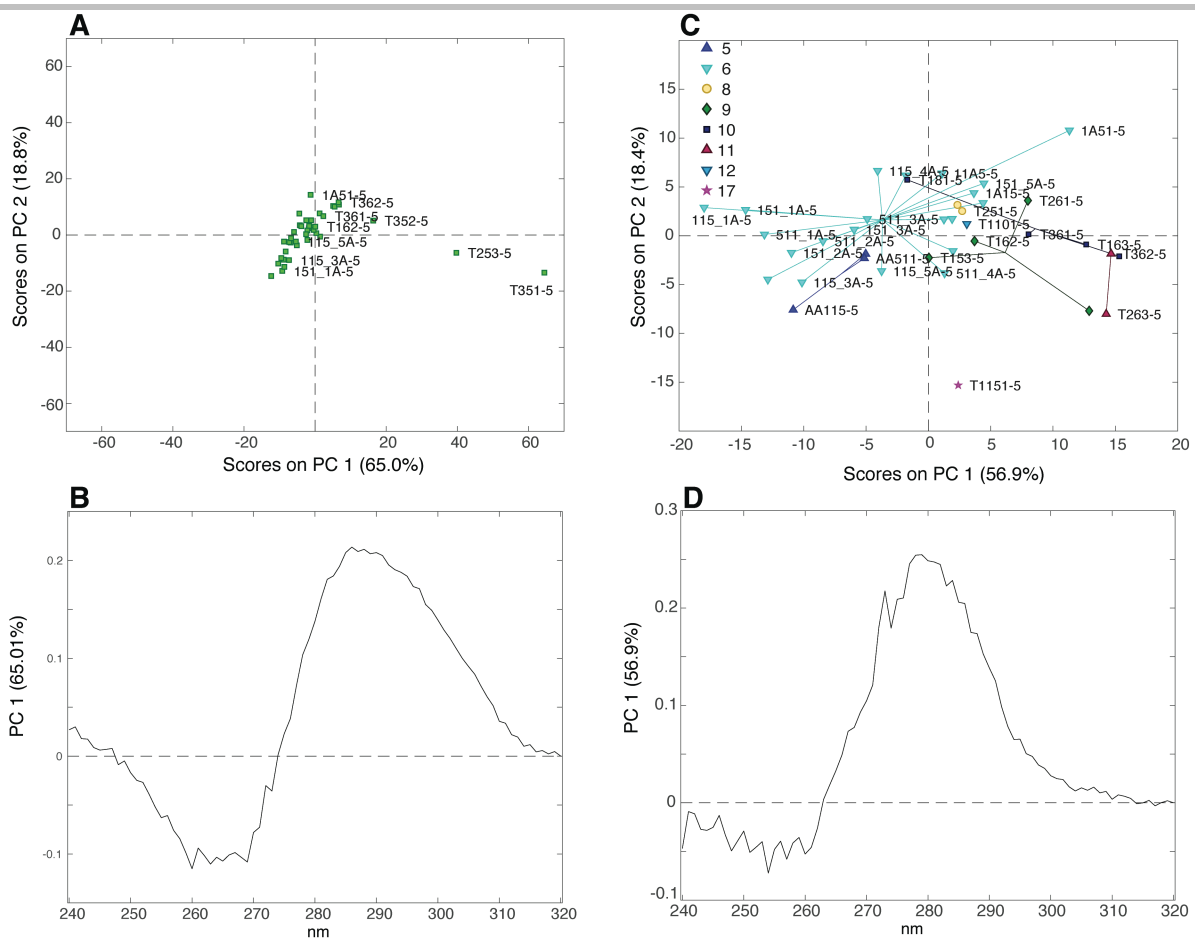


Figure S17. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the additional 40 samples acquired at pH 7.00 and (B) relative PC1 loading plot. (C) PC1/PC2 score plot of the PCA model calculated using 38 out of 40 samples ('T253-5' and 'T351-5' are excluded) acquired at pH 7.00 colored according to the number of thymines and (D) relative PC1 loading plot.

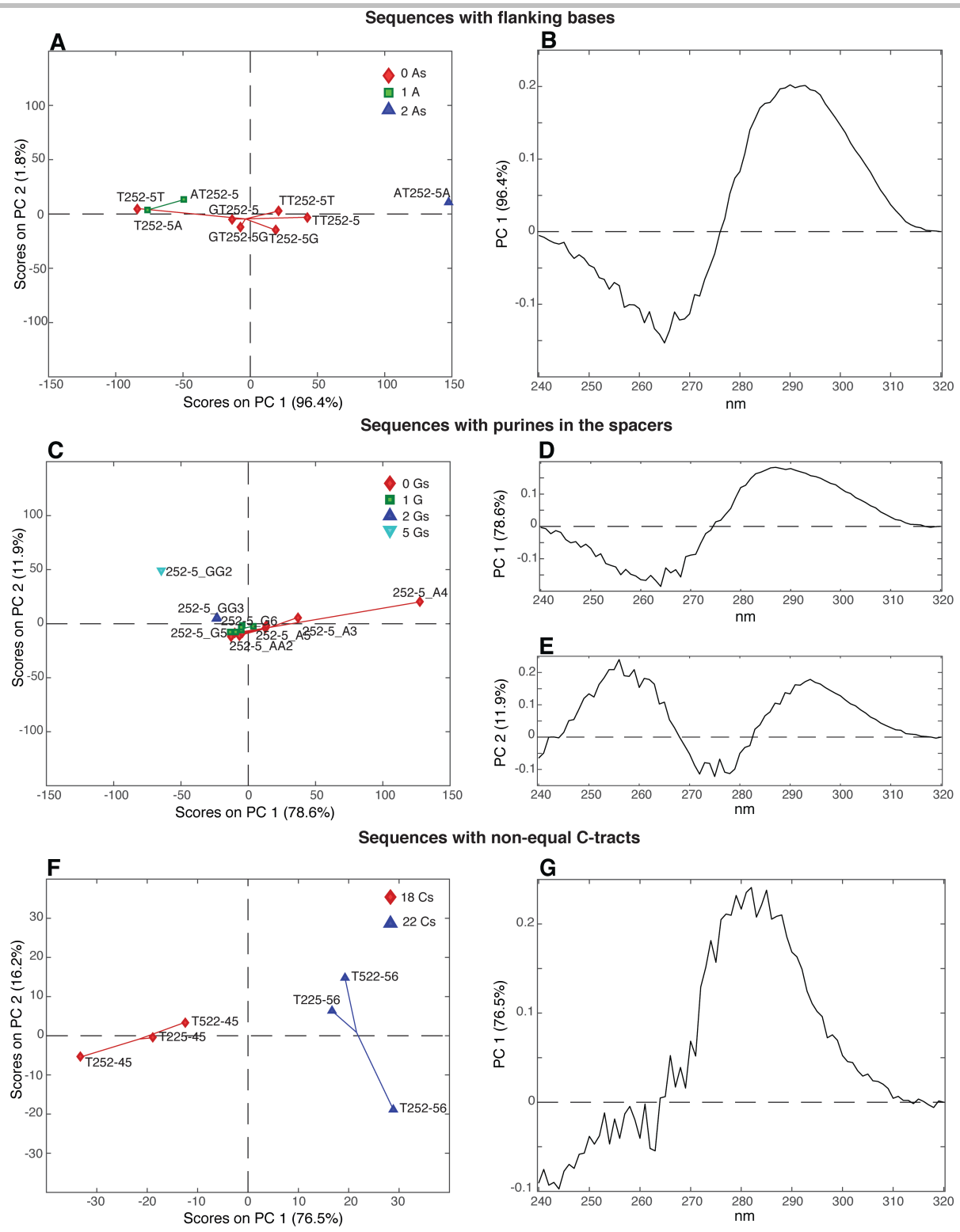


Figure S18. PC1/PC2 score plots of the PCA model calculated using the CD spectra acquired at pH 7.00 of the three subgroups of the additional 35 samples and relative loading plots. (A) Score plot of the PCA model calculated using the 9 samples having flanking bases and (B) relative PC1 loading plot; (C) Score plot of the PCA model calculated using the 18 samples ('A252-5' and 'G252-5' are excluded) having purines in the spacers and relative (D) PC1 and (E) PC2 loading plots; (F) Score plot of the PCA model calculated using the 6 samples having non-equal C-tracts and (G) relative PC1 loading plot.

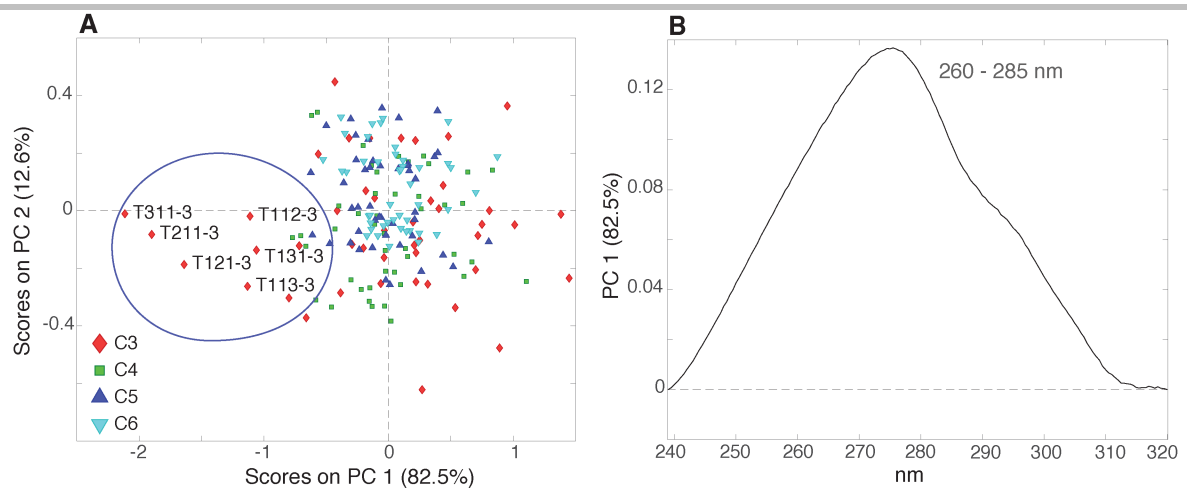


Figure S19. (A) PC1/PC2 score plot of the PCA model calculated using the 180 TDS acquired at pH 5.00, colored according to C-tract length and (B) PC1 loading plot.

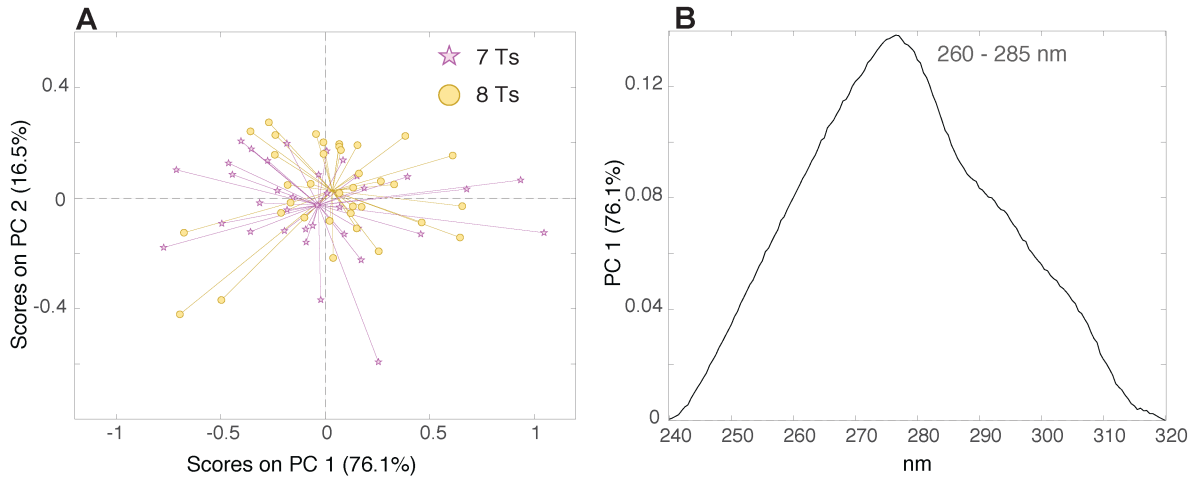


Figure S20. (A) PC1/PC2 score plot of the PCA model calculated using the 72 TDS of samples having 7 and 8 Ts (acquired at pH 5.00), colored according to the total number of Ts and (B) PC1 loading plot.

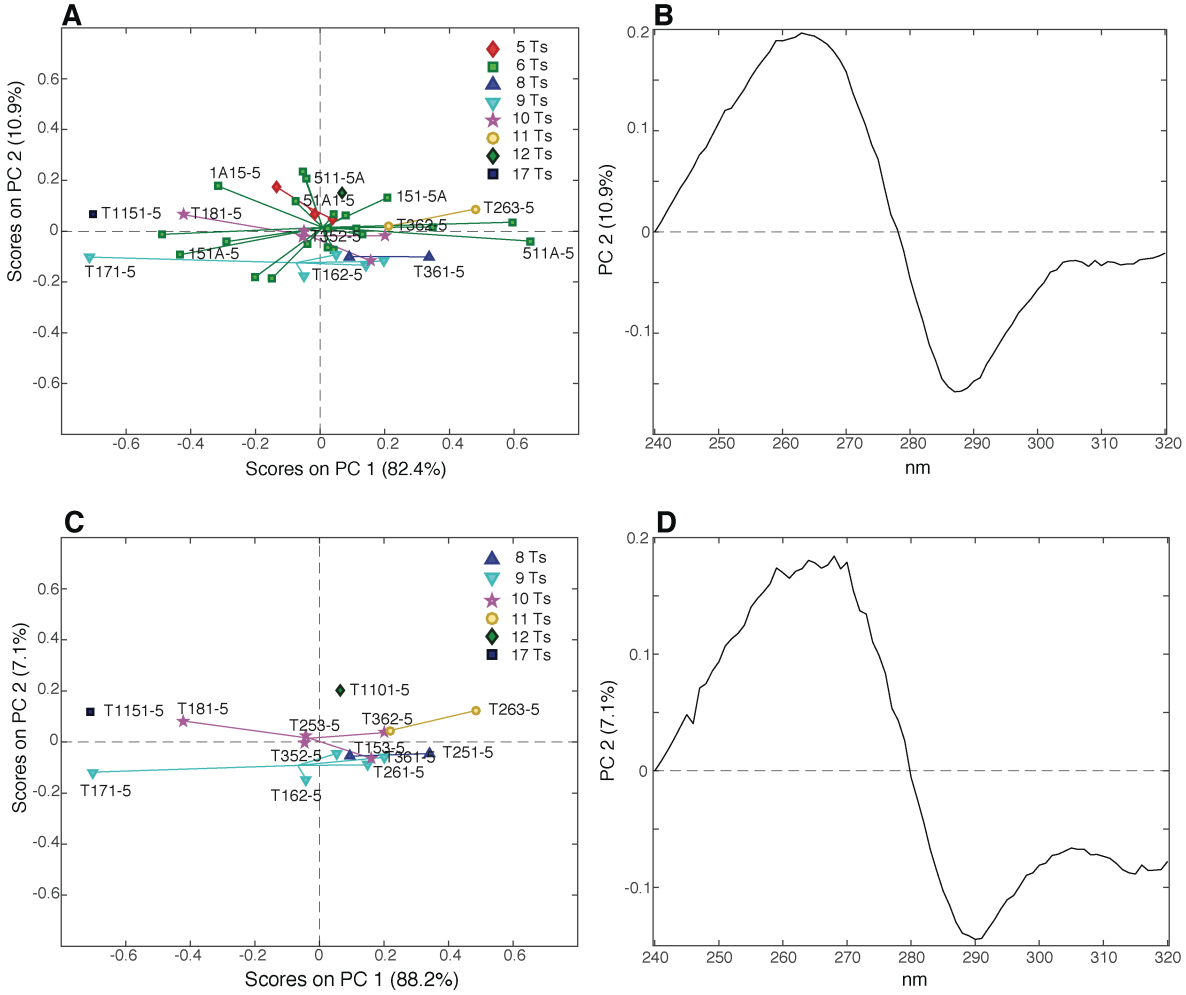


Figure S21. (A) PC1/PC2 score plot of the PCA model calculated using the TDS acquired at pH 5.00 of the additional subset of 40 samples, colored according to the number of Ts, and (B) relative PC2 loading plot. (C) PC1/PC2 score plot of the PCA model calculated using 16 out of 40 samples of the data set (all the samples containing adenines have been excluded) colored according to the number of Ts and (D) relative PC2 loading plot.

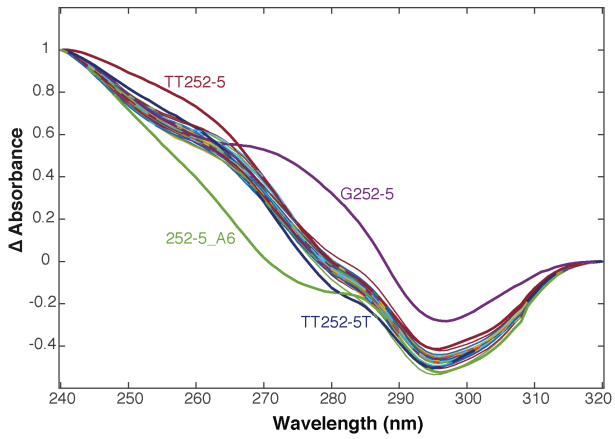


Figure S22. Superimposition of the TDS profiles of the additional 35 DNA samples acquired at pH 5.00. Unusual profiles of the samples 'G252-5', 'TT252-5T', 'TT252-5', and '252-5_A6' have been highlighted. Sequences are given in Table S2.

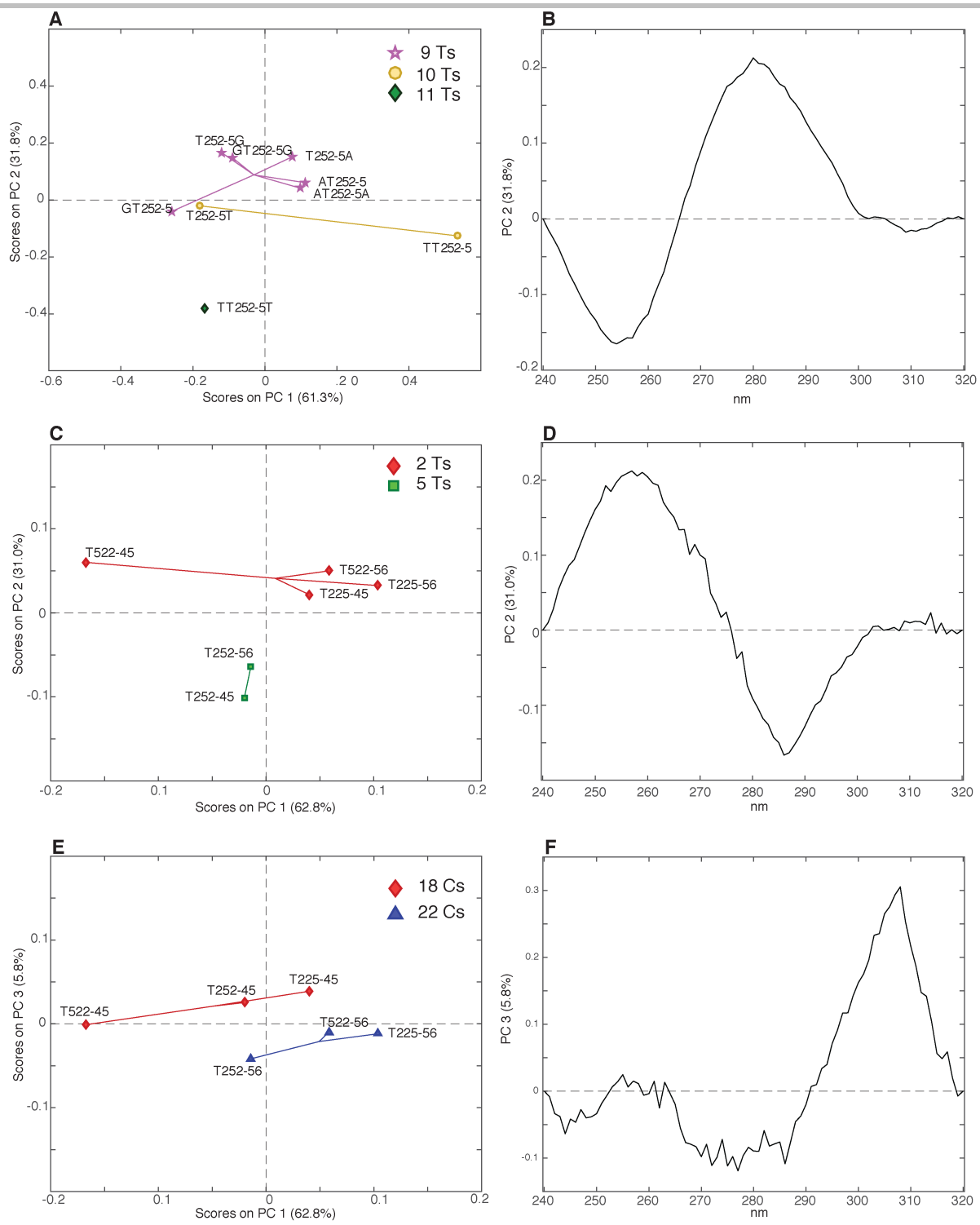


Figure S23. Score and loading plots of the PCA models calculated using the TDS acquired at pH 5.00 of two subgroups of the additional subset of 35 samples and relative loading plots. (A) PC1/PC2 score plot of the PCA model calculated using the 9 samples having flanking bases and (B) relative PC2 loading plot; PC1/PC2 score plot of the PCA model calculated using the 6 samples having non-equal C-tracts colored according to (C) the number of Ts in the central spacer and (E) the number of Cs in the C-tracts and relative (D) PC2 and (F) PC3 loading plots.

References

[1] M. Cheng, D. Qiu, L. Tamon, E. Maturová, P. Víšková, S. Amrane, A.Guédin, J. Chen, L. Lacroix, H. Ju, L. Trantírek, A. B. Sahakyan, J. Zhou, J. L. Mergny, *Angew. Chem., Int. Ed.* **2021**. DOI: 10.1002/anie.202016801

[2] D. M. Gray, F. J. Bollum, *Biopolymers* **1974**, 13, 2087–2102.

[3] P. Školáková, D. Renčiuk, J. Palacký, D. Krafčík, Z. Dvořáková, I. Kejnovská, K. Bednářová, M. Vorlíčková, *Nucleic Acids Res.* **2019**, 9, 2177–2189.