Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

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, and Jean-Louis Mergny*

anie_202016822_sm_miscellaneous_information.pdf

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

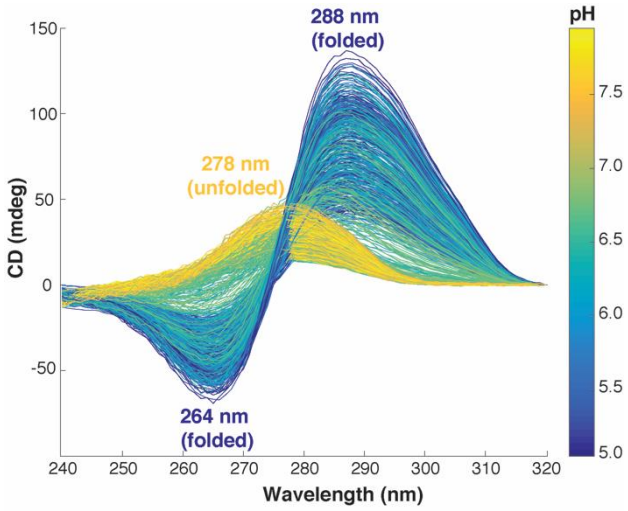

Figure S1. CD spectra of the 180 DNA samples acquired at 13 different pH values (2340 spectra). The color scale goes from yellow (pH 8.00) to dark blue (pH 5.00). Sequences are given in Table S1.

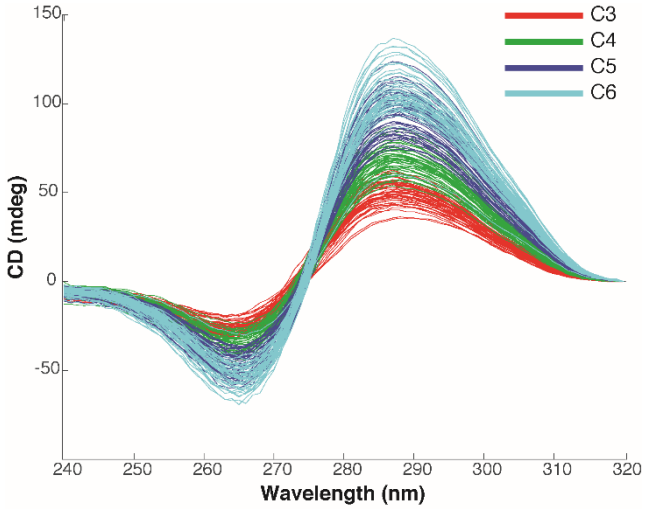

Figure S2. CD spectra of the 180 DNA samples acquired at pH 5.00 colored according to the C-tract length. Sequences are given in Table S1.

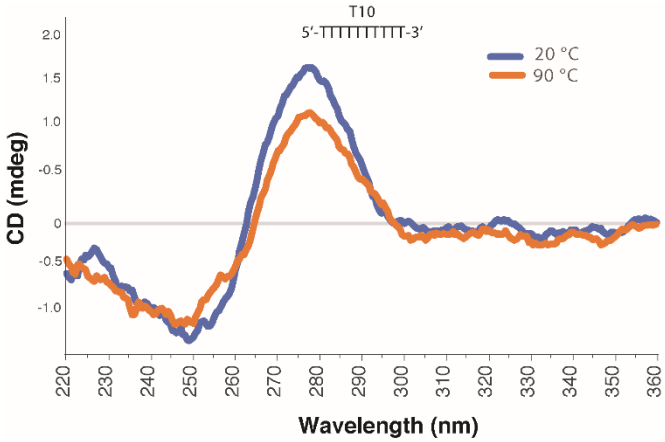

Figure S3. CD spectrum of a poly(dT) 'T10' in water, acquired at 20 °C (blue) and 90 °C (orange).

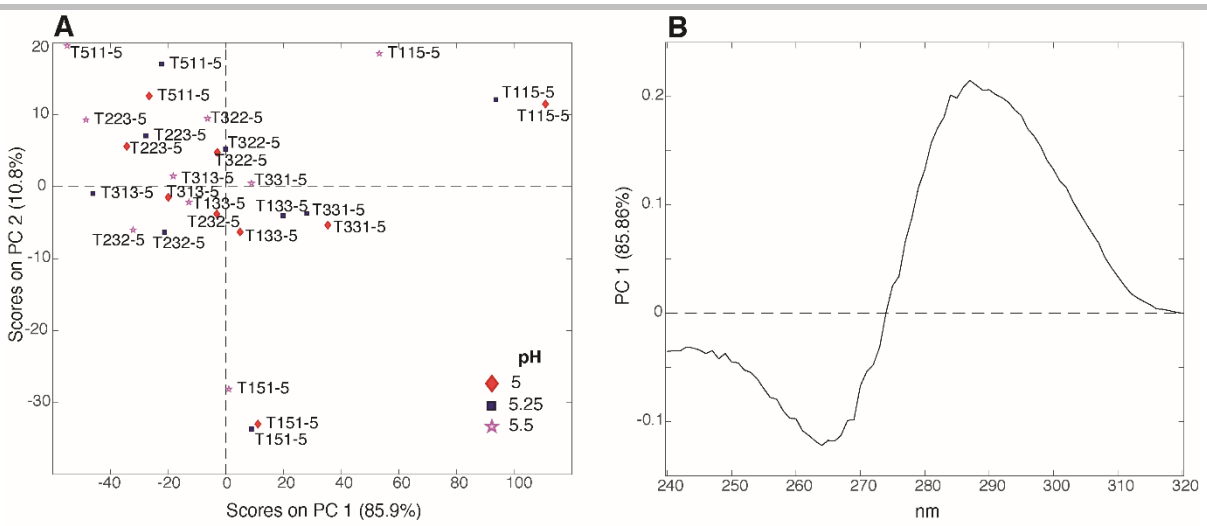

Figure S4. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 9 samples, having a total of twenty Cs and seven Ts, acquired at pH 5.00, 5.25, and 5.50 (27 CD spectra in total), colored according to the pH values, and (B) relative PC1 loading plot.

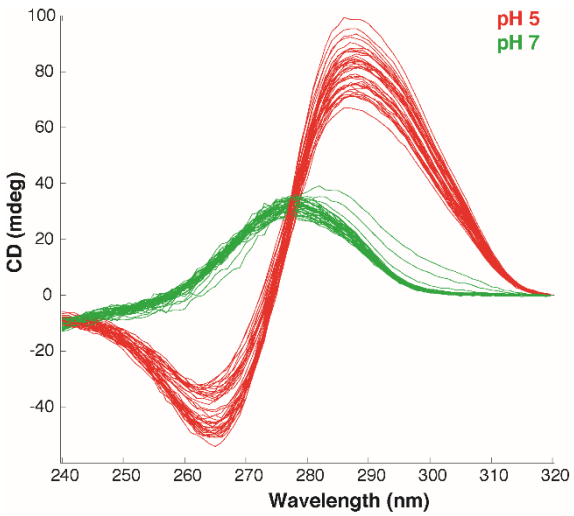

Figure S5. CD spectra of the additional 40 DNA samples acquired at pH 5.00 (red) and pH 7.00 (green). Sequences are given in Table S2.

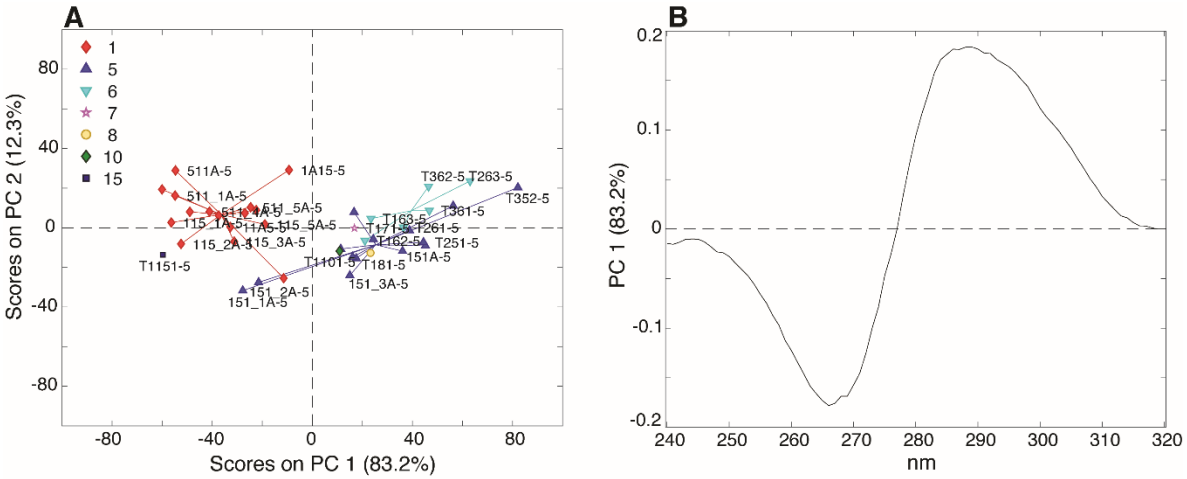

Figure S6. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 40 additional samples, acquired at pH 5.00, colored according to the length of the central spacer, and (B) relative PC1 loading plot.

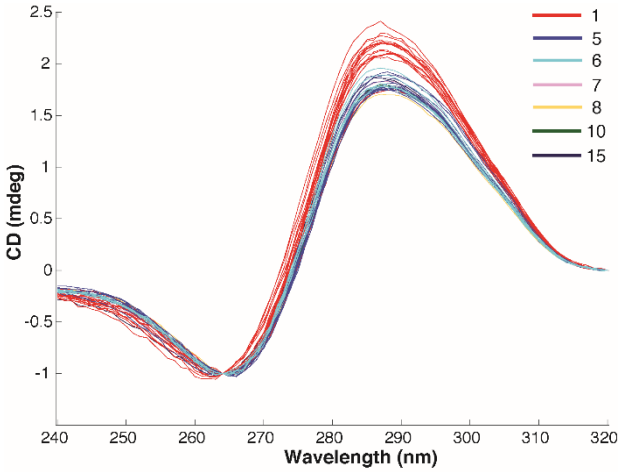

Figure S7. CD spectra of the additional 40 DNA samples, acquired at pH 5.00, and normalized by the intensity of the CD signal at 264 nm. The spectra are colored according to the length of the central spacer. Sequences are given in Table S2.

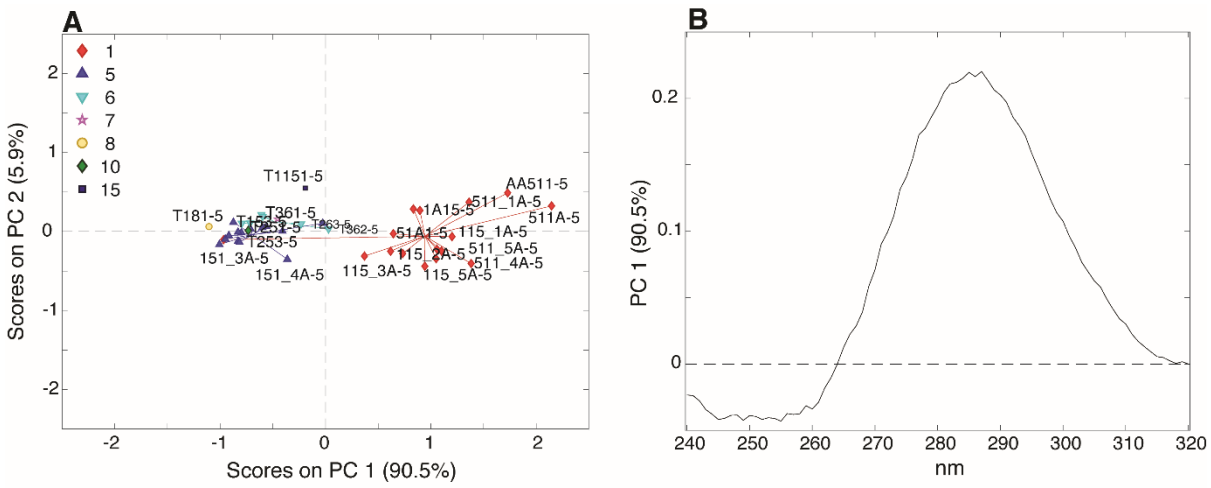

Figure S8. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 40 additional samples, acquired at pH 5.00, normalized by the intensity of the signal at 264 nm and colored according to the length of the central spacer, and (B) relative PC1 loading plot.

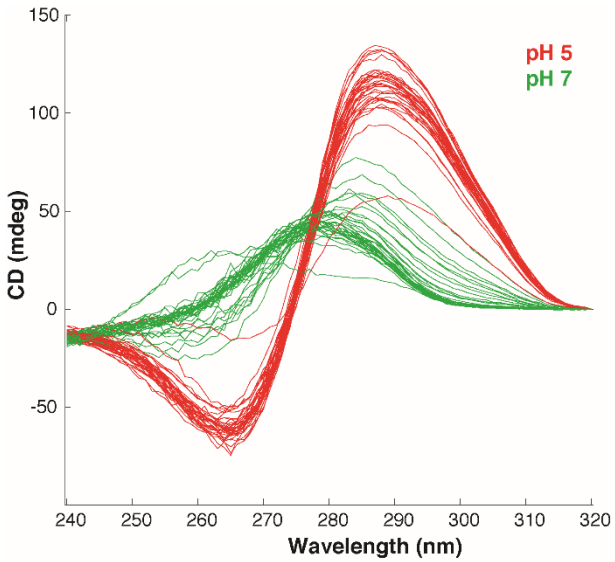

Figure S9. CD spectra of the additional 35 DNA samples acquired at pH 5.00 (red) and pH 7.00 (green). Sequences are given in Table S2.

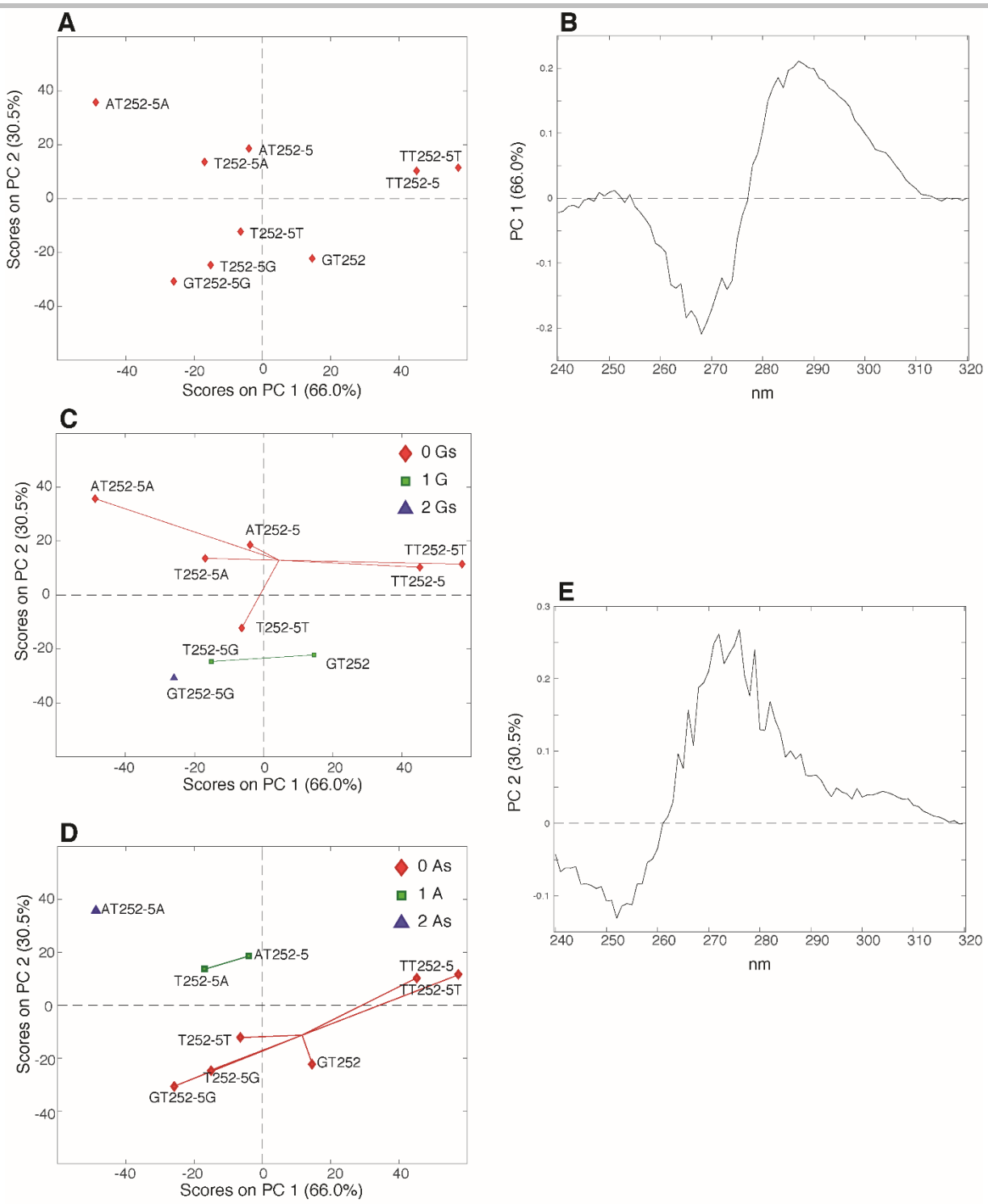

Figure S10. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 9 samples characterized by flanking bases (from the dataset made of 35 sequences) acquired at pH 5.00, colored according to the number of (C) guanines (Gs) and (D) adenines (As) and relative (B) PC1 and (E) PC2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

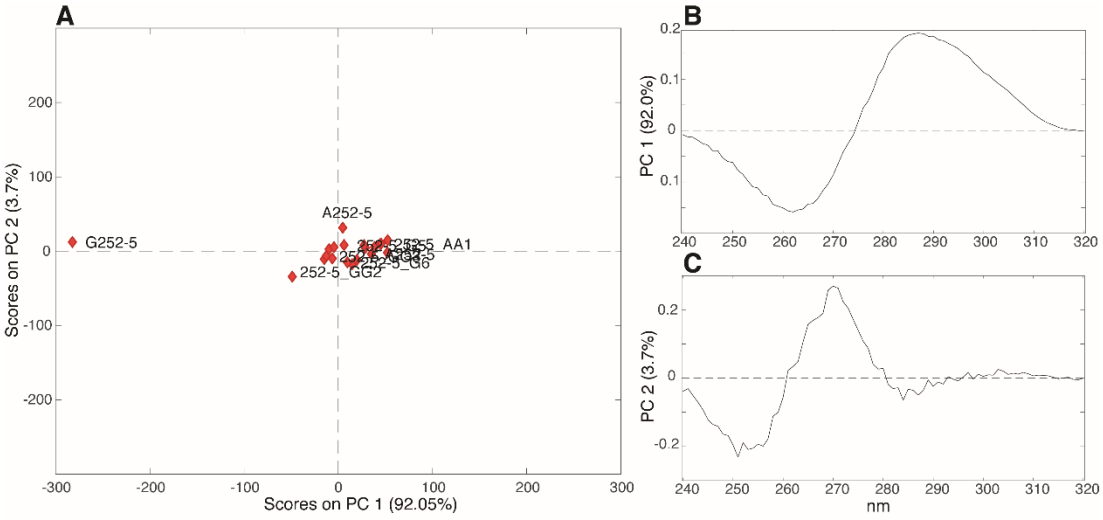

Figure S11. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 20 samples characterized by adenines and guanines in the spacers (from the dataset made of 35 sequences) acquired at pH 5.00 and relative (B) PC1 and (C) PC2 loading plots.

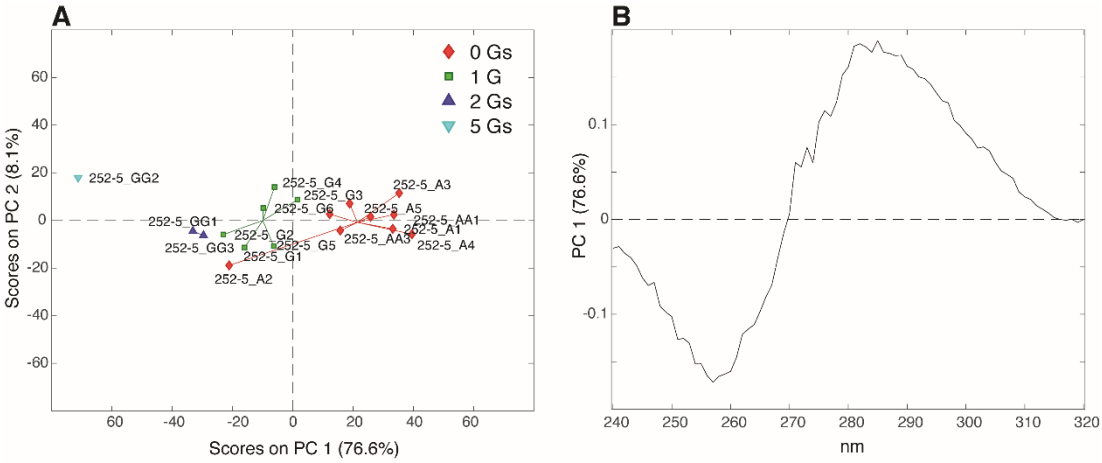

Figure S12. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 18 samples characterized by adenines and guanines in the spacers (samples 'A252-5' and 'G252-5' have been excluded) acquired at pH 5.00, colored according to the number of Gs, and relative (B) PC1 loading plot**.**

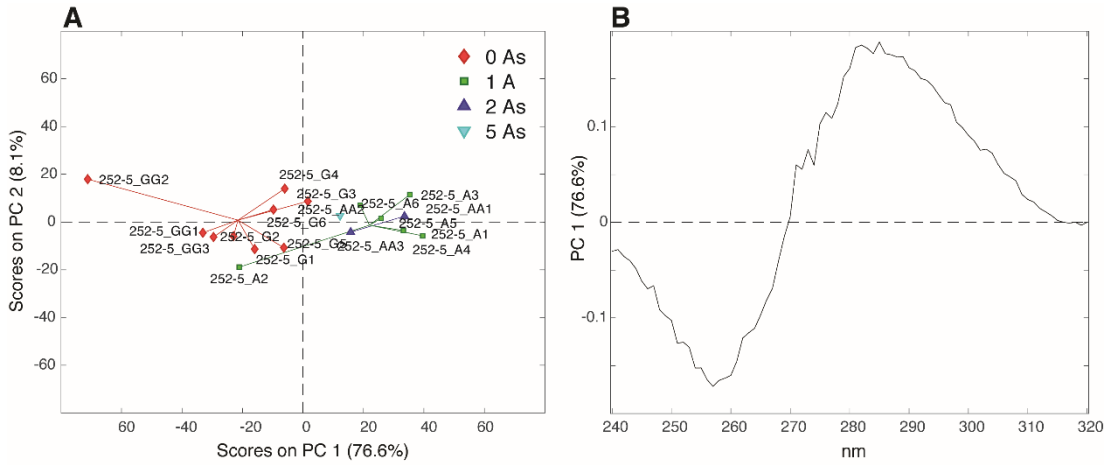

Figure S13. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 18 samples characterized by adenines and guanines in the spacers (samples 'A252-5' and 'G252-5' have been excluded) acquired at pH 5.00, colored according to the number of As, and relative (B) PC1 loading plot.

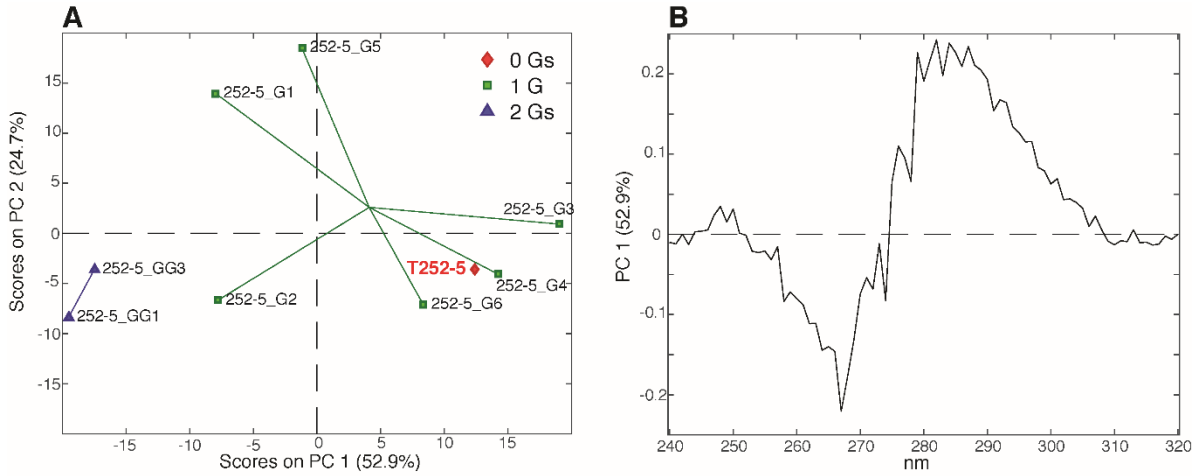

Figure S14. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra (pH 5.00) of the 8 samples characterized by guanines in the spacers ('G252-5' and '252-5_GG2' have been excluded because they would drive the separation of the samples for their peculiar spectral profiles) together with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

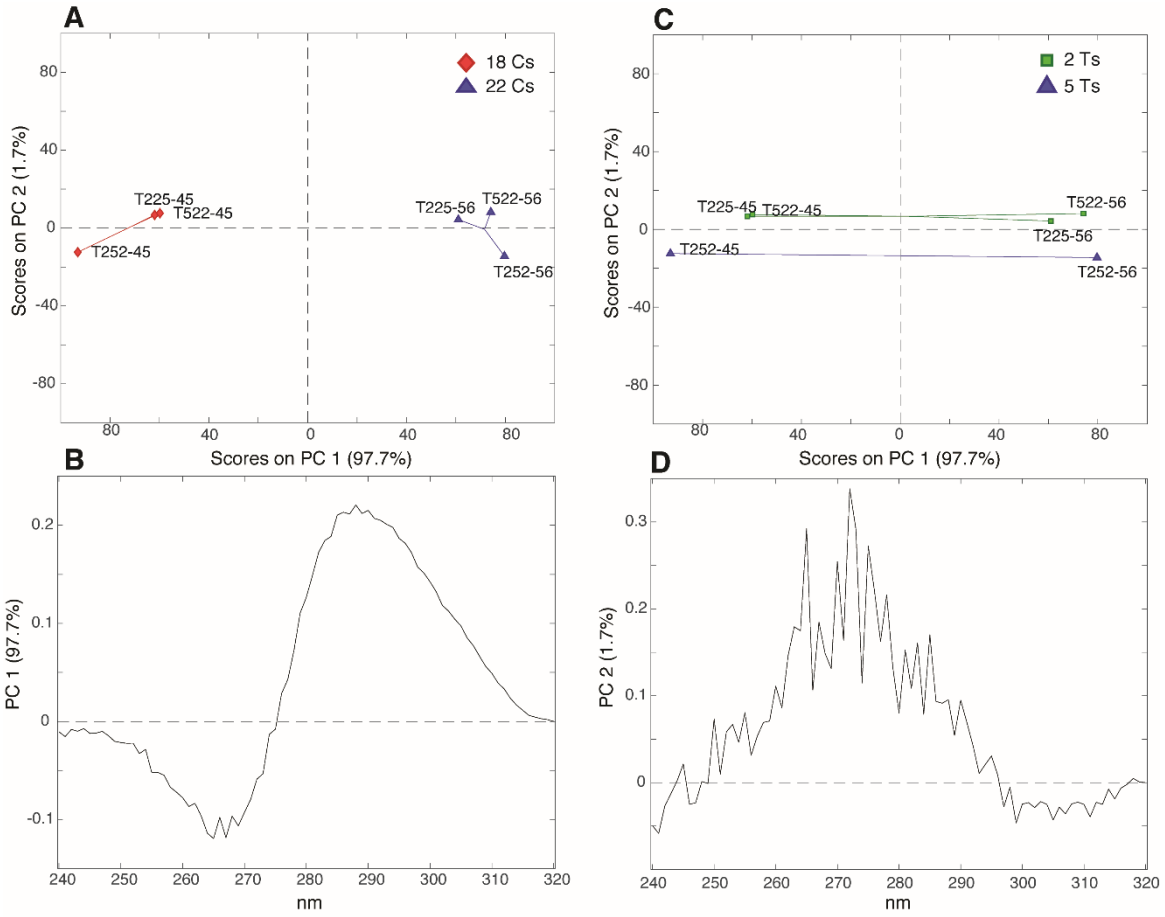

Figure S15. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 6 samples characterized by non-equally sized C-tracts (from the dataset made of 35 sequences) acquired at pH 5.00, colored according to the number of cytosines and (C) according to the length of the central spacer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

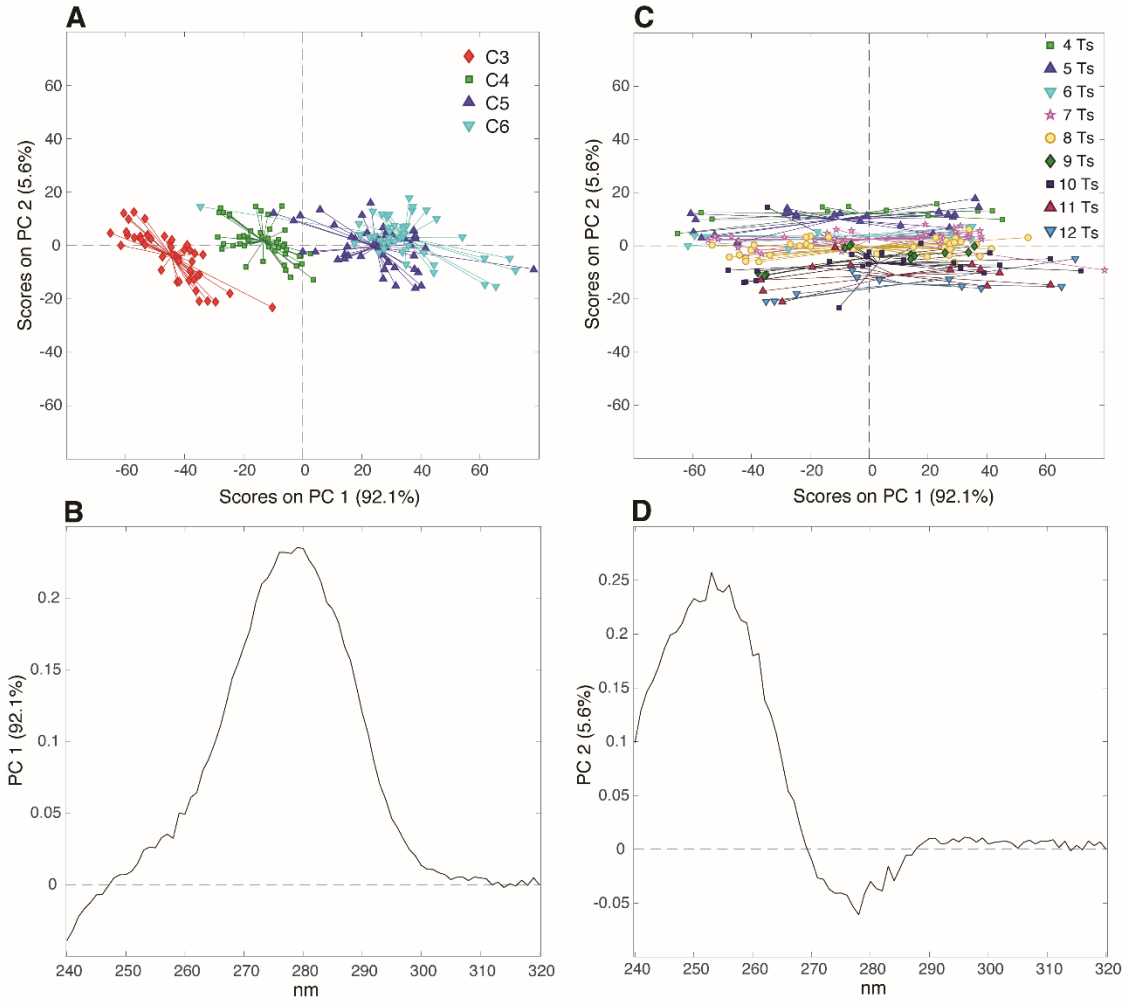

Figure S16. PC1/PC2 score plot of the PCA model calculated using the CD spectra of the 180 samples acquired at pH 7.00, colored according to the number of (A) cytosines, and (C) thymines, and relative (B) PC1 and (D) PC2 loading plots.

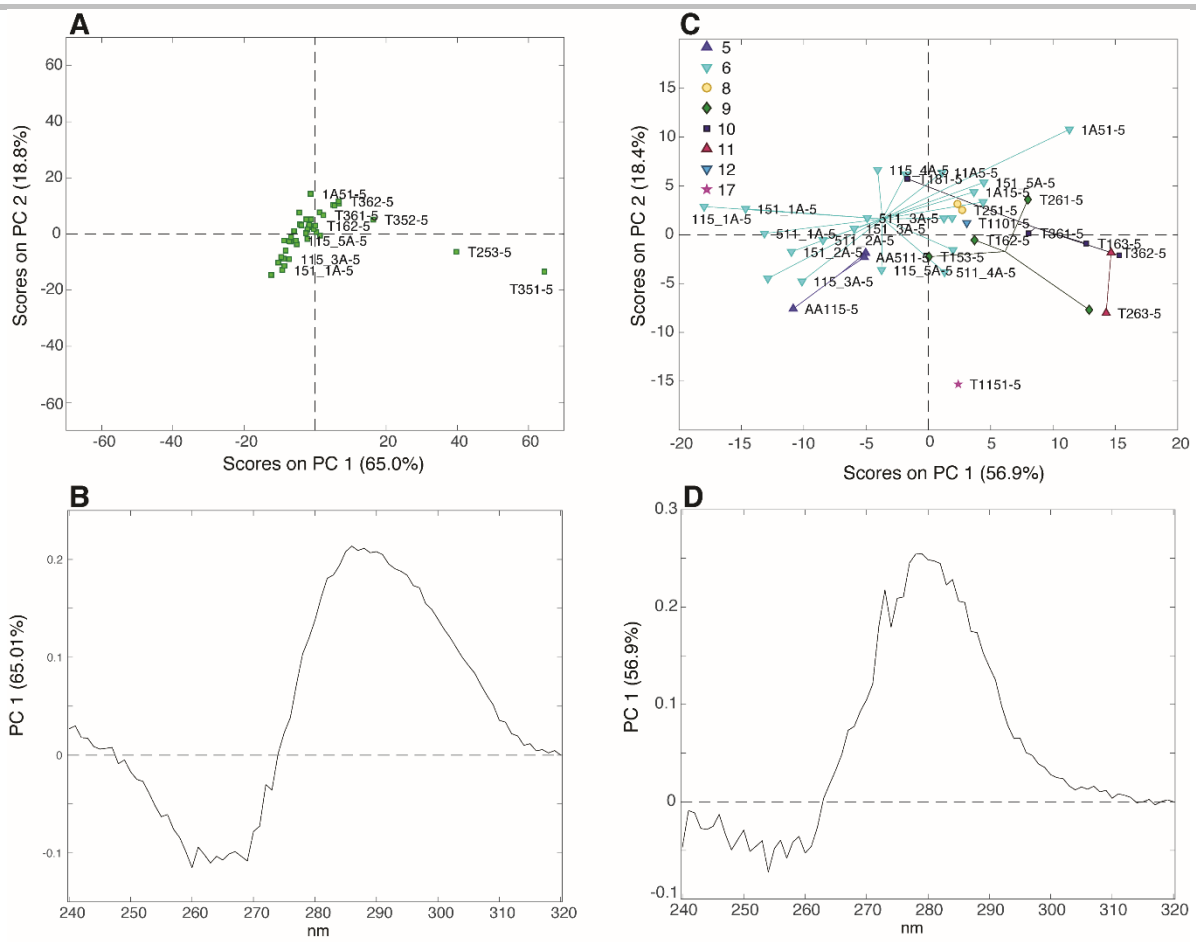

Figure S17. (A) PC1/PC2 score plot of the PCA model calculated using the CD spectra of the additional 40 samples acquired at pH 7.00 and (B) relative PC1 loading plot. (C) PC1/PC2 score plot of the PCA model calculated using 38 out of 40 samples ('T253-5' and 'T351-5' are excluded) acquired at pH 7.00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

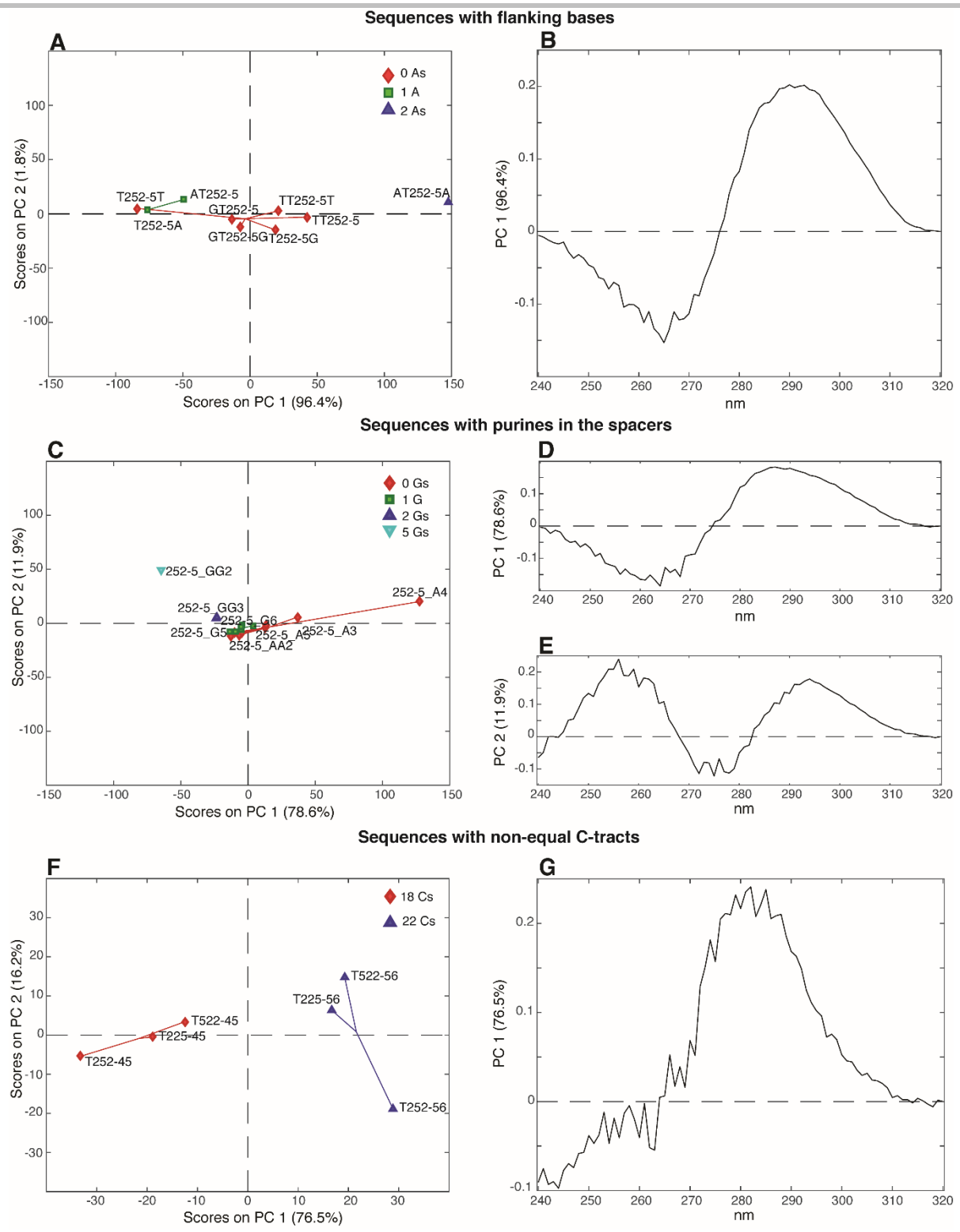

Figure S18. PC1/PC2 score plots of the PCA model calculated using the CD spectra acquired at pH 7.00 of the three subgroups of the additional 35 samples and relative loading plots. (A) Score plot of the PCA model calculated using the 9 samples having flanking bases and (B) relative PC1 loading plot; (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.

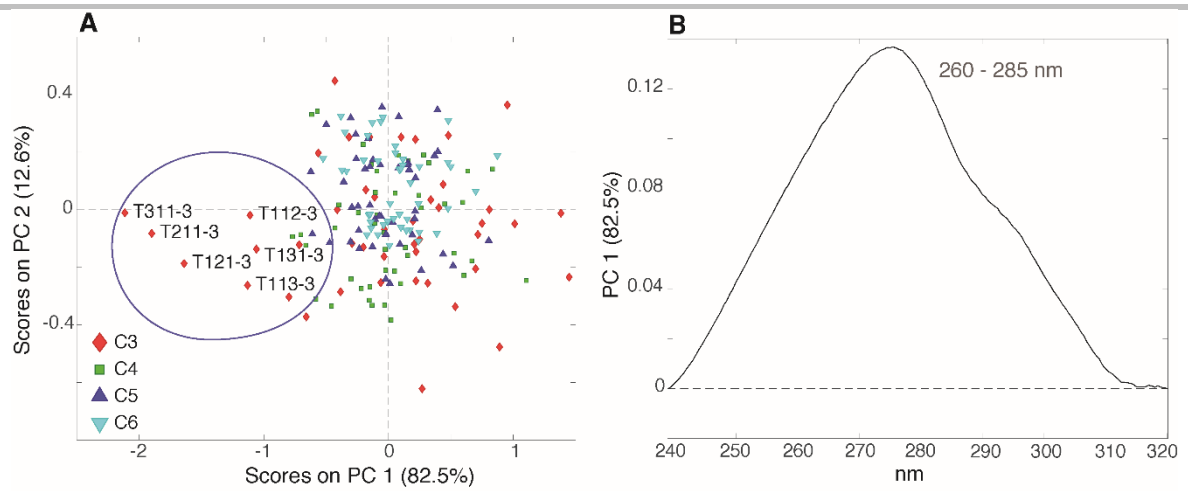

Figure S19. (A) PC1/PC2 score plot of the PCA model calculated using the 180 TDS acquired at pH 5.00, colored according to C-tract length and (B) PC1 loading plot.

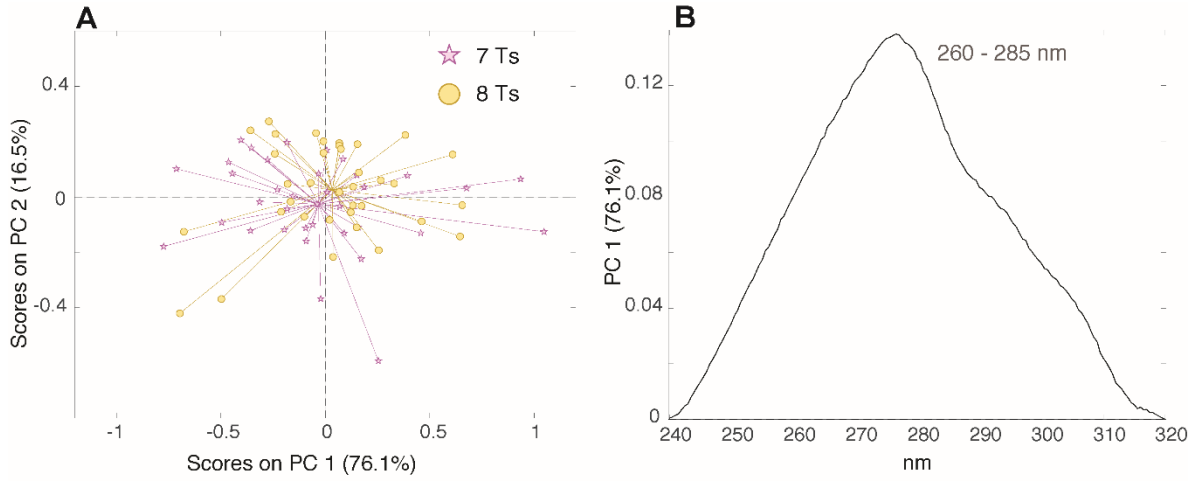

Figure S20. (A) PC1/PC2 score plot of the PCA model calculated using the 72 TDS of samples having 7 and 8 Ts (acquired at pH 5.00), colored according to the total number of Ts and (B) PC1 loading plot.

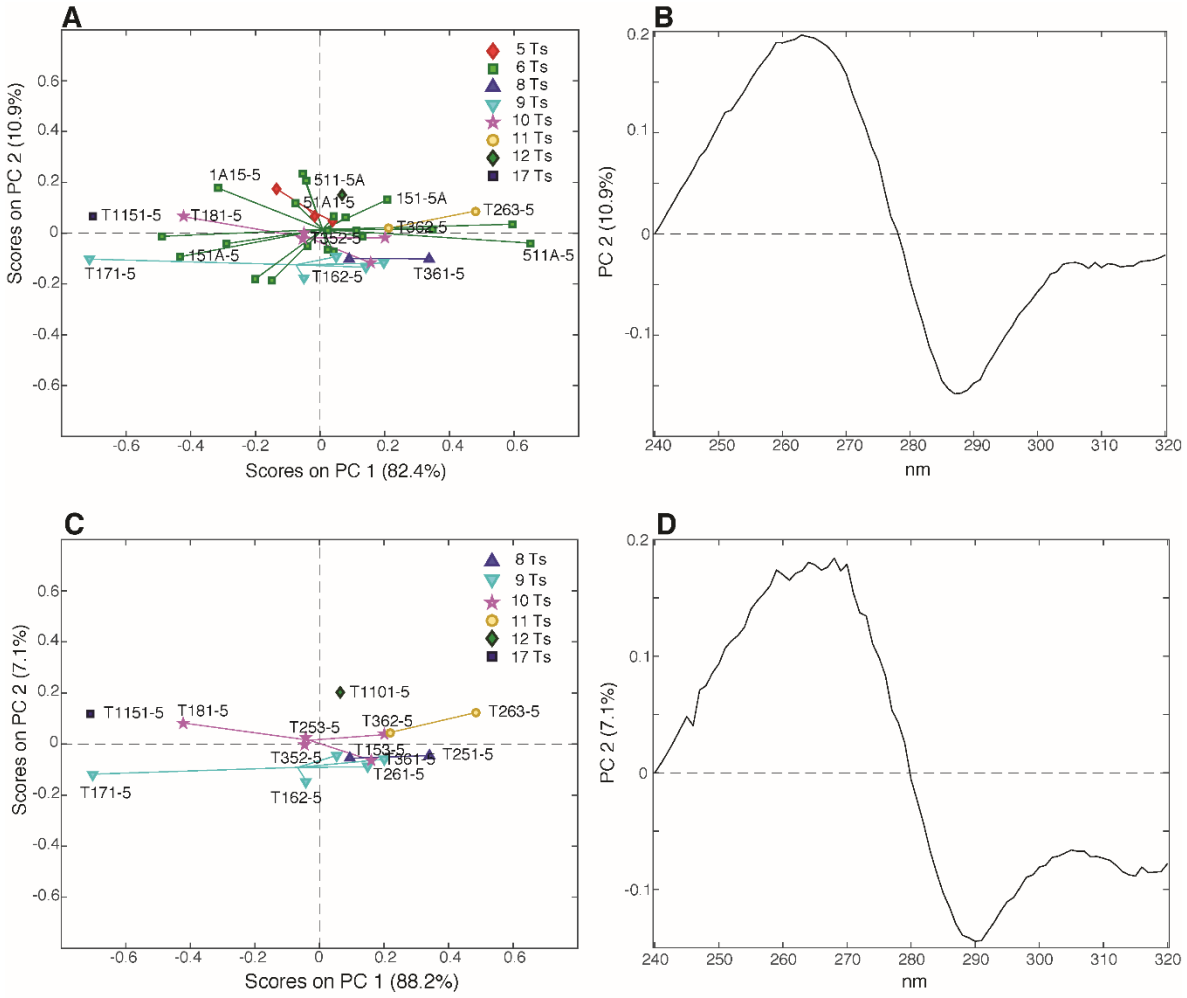

Figure S21. (A) PC1/PC2 score plot of the PCA model calculated using the TDS acquired at pH 5.00 of the additional subset of 40 samples, colored according to the number of Ts, and (B) relative PC2 loading plot. (C) PC1/PC2 score plot of the PCA model calculated using 16 out of 40 samples of the data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

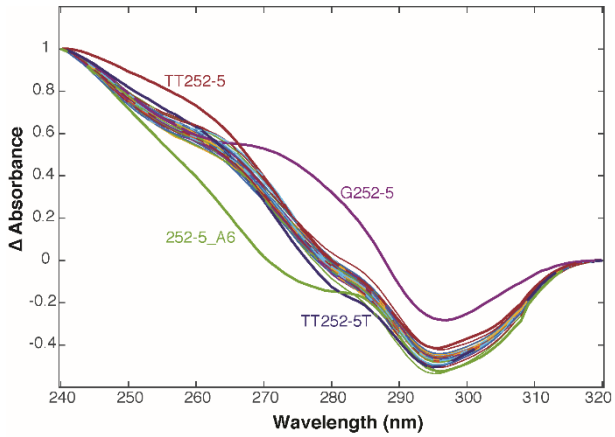

Figure S22. Superimposition of the TDS profiles of the additional 35 DNA samples acquired at pH 5.00. Unusual profiles of the samples 'G252-5', 'TT252-5T', 'TT252-5', and '252-5_A6' have been highlighted. Sequences are given in Table S2.

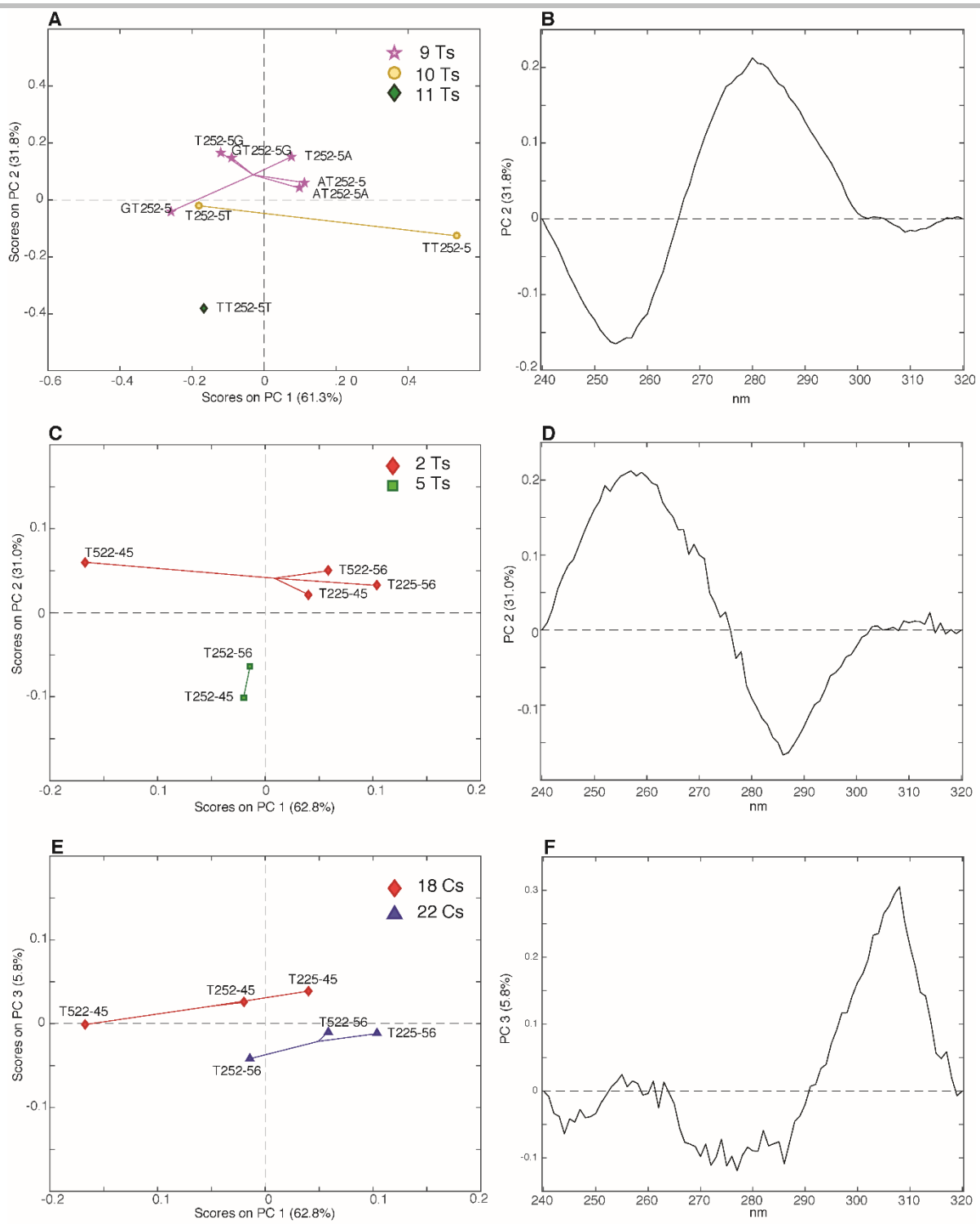

Figure S23. Score and loading plots of the PCA models calculated using the TDS acquired at pH 5.00 of two subgroups of the additional subset of 35 samples and relative loading plots. (A) PC1/PC2 score plot of the PCA model calculated using the 9 samples having flanking bases and (B) relative PC2 loading 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.* **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.