Published online 20 April 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 11* ***e61***
doi: 10.1093/nar/gkaa257

Harnessing intrinsic fluorescence for typing of secondary structures of DNA

Michela Zuffo [1,2,*], Aurélie Gandolfini [1,2], Brahim Heddi [3] and Anton Granzhan [1,2,*]

[1] CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France, [2] CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France and [3] Laboratoire de Biologie et de Pharmacologie Appliquée, CNRS UMR8113, École Normale Supérieure Paris-Saclay, F-94235 Cachan, France

Received January 16, 2020; Revised March 28, 2020; Editorial Decision March 30, 2020; Accepted April 03, 2020

ABSTRACT

High-throughput investigation of structural diversity of nucleic acids is hampered by the lack of suitable label-free methods, combining fast and cheap experimental workflow with high information content. Here, we explore the use of intrinsic fluorescence emitted by nucleic acids for this scope. After a preliminary assessment of suitability of this phenomenon for tracking conformational changes of DNA, we examined steady-state emission spectra of an 89-membered set of oligonucleotides with reported conformation (G-quadruplexes (G4s), i-motifs, single- and double-strands) by means of multivariate analysis. Principal component analysis of emission spectra resulted in successful clustering of oligonucleotides into three corresponding conformational groups, without discrimination between single- and double-stranded structures. Linear discriminant analysis was exploited for the assessment of novel sequences, allowing the evaluation of their G4-forming propensity. Our method does not require any labeling agent or dye, avoiding the related bias, and can be utilized to screen novel sequences of interest in a high-throughput and cost-effective manner. In addition, we observed that left-handed (Z-) G4 structures were systematically more fluorescent than most other G4 structures, almost reaching the quantum yield of 5′-d[(G_3T)$_3$$G_3$]-3′ (*$G_3$T*, the most fluorescent G4 structure reported to date).

INTRODUCTION

Since the elucidation of the double-helical structure of DNA, a great deal of effort has been devoted to understand whether the genomic material could also adopt other conformations. It is now firmly established that DNA can fold into a wealth of secondary structures that are intertwined in delicate equilibria. These range from single-, double- and triple-stranded conformations (i.e., random-coil, A-, B- and Z-DNA, and triplexes, respectively) to three- and four-way junctions as well as tetra-stranded structures (chiefly, G-quadruplexes, or G4s, and i-motifs, or iMs). Specifically, G4 structures are formed by stacks of guanine quartets, stabilized by monovalent cations (mostly K^+ and Na^+). Despite this common scaffold, G4 structures are themselves extremely polymorphic in terms of strand topology (parallel, anti-parallel, hybrid), loop geometry, groove size etc. (1). On the other hand, iMs consist of two interpenetrated duplexes formed by hemi-protonated C:CH^+ base pairs. These structures are most stable in moderately acidic conditions, although certain factors, such as molecular crowding and negative supercoiling, were reported to stabilize them at near-physiological pH (2,3). Thus, the conformational status of any given sequence clearly depends on the specific arrangement of nucleotides, but also on environmental conditions (e.g. pH, solvent, salinity, molecular crowding agents) and responds to external stimuli. Notably, these phenomena are not limited to the *in vitro* space. In fact, the existence of non-canonical secondary structures *in vivo* is supported by a large body of immunochemical, biochemical and biophysical evidence, which has been growing over the last decade. Although a full understanding of these phenomena is still missing, it is generally accepted that such structures act as regulators of genetic and epigenetic transactions (2,4–7). Numerous enzymatic partners interacting with these structures and involved in their homeostasis, either as chaperons or by promoting their unfolding, have been identified (8,9). In fact, imbalances in the delicate equilibria among the different DNA conformations seem to trigger the development of different pathologies such as cancer, infectious, and neurodegenerative diseases (10–13).

Despite the clear relevance of the structural polymorphism of nucleic acids, its high-throughput investigation and the identification of novel secondary structure-forming sequences are hampered by current methodological limita-

*To whom correspondence should be addressed. Tel: +33 169 86 30 89; Email: anton.granzhan@curie.fr
Correspondence may also be addressed to Michela Zuffo. Email: michela.zuffo@gmail.com
Present address: Michela Zuffo, Minakem Recherche, 59310 Beuvry-la-Forêt, France.

tions. On the one hand, high-resolution methods such as solution NMR (14–17) and X-ray crystallography (18–20) can provide structural details at atomic resolution but are expensive, labor- and time-consuming, and require significant amounts of material. Moreover, in order to obtain high-resolution structures, these techniques need either a well-defined spectrum with single species (for NMR) or a well-diffracting crystal (for crystallography), which are not always straightforward. In addition, the structural information obtained by these methods may be affected by the formation of higher-order structures (oligomerization) at high concentrations required for NMR spectroscopy (21,22), or artefacts related to crystal packing effects (23,24). On the other hand, low-resolution biophysical techniques have reduced costs and operate at physiologically relevant concentration of DNA and buffer compositions but, also, lack the possibility of a high-throughput implementation. This is perfectly illustrated by the example of circular dichroism (CD), the benchmark technique for this scope. Despite providing a good compromise between information content (25,26), especially when coupled to multivariate (i.e. chemometric) analysis (27,28), and experimental cost, no CD-based high-throughput screening has been reported, to date. In this sense, fluorescence-based methods constitute a valuable alternative. The combination of their high-throughput potential with the decoding power of chemometrics constitutes an important step forward in the quest for a rapid screening method for DNA structures. Recently, we reported a sensor array of fluorescent dyes designed for this scope, providing a proof-of-concept of its applicability for the typing of secondary structures of DNA oligonucleotides (29). However, the use of non-covalent (external) reporter dyes can introduce an intrinsic bias, since their non-covalent interactions with nucleic acids may lead to modification, or induction, of secondary structures of the latter (30,31), engendering a skewing of the screening results. An alternative to external reporter dyes lies in the use of fluorescent nucleoside analogues, which minimally perturb the secondary structure of oligonucleotides and can be used as structural reporters in a variety of conditions (32–34), including those mimicking intracellular environment (35). However, custom synthesis procedures required for the incorporation of fluorescent nucleoside analogues preclude the use of this method for high-throughput investigations, and limit it to specialized applications. Finally, chemical probing methods allow the mapping of various structural motifs in nucleic acids (RNA in particular), allowing the analysis of long sequences *in vitro* and even *in vivo* (36–38). However, most these methods are tedious and require radioactive or fluorescent labeling of nucleic acids. For these reasons, the implementation of a high-throughput, *label-free* method for structural assessment of nucleic acids would be an invaluable asset.

In this context, we reasoned that the intrinsic fluorescence emitted by nucleic acids might be exploited for the scope. Similar to isolated nucleotides, single- and double-strands are known to emit in the near-UV spectral range with a low quantum yield (10^{-5} to $\sim 10^{-4}$); however, certain secondary structures display strongly enhanced fluorescence properties. Thus, the remarkable fluorescent properties of G4 structures were documented about ten years ago (39–41). Upon folding in highly saline buffers, they display a broad emission band, typically peaking between 330 and 420 nm (λ_{ex} = 255–270 nm) (41–43). Most importantly, the quantum yield of G4s (Φ = 2.9–3.5 × 10^{-4}) is at least 3-fold higher than that of the corresponding single strands (40,41), although these values vary greatly depending on the sequence and other factors. It was proposed that this behavior arises from the interactions of guanine residues in the excited state, including the formation of guanine excimers (44,45). An even stronger emission is observed in some particular cases (e.g. 5′-d[$(G_3T)_3G_3$]-3′ sequence, hereafter referred to as *G_3T*, and its variants, displaying Φ values of up to 2 × 10^{-3}, i.e., almost 6-fold more fluorescent than other G4s) (43,46), although the reason of such behavior is still a matter of debate. The long-lived red emitting state (λ_{em} = 380–390 nm) has been ascribed to the stacking of guanine bases in defined orientations in the inner part (core) of the G4 structure (43,46,47), or to the formation of excimers of external G-tetrads in 5′–5′ stacked dimeric structures (42,48). Over the years, the effects of guanine base orientation, size and nature of loops, presence and nature of bulges have been investigated, contributing to establish some correlations between the G4 conformation and the emissive properties (47,49,50). Very recently, the investigation on intrinsic fluorescence has been extended to iM structures (51): upon excitation at 267 or 300 nm, iMs display a long-lived, broad emission band centered ~410–420 nm (Φ = 3.4–14 × 10^{-4}, depending on the specific iM sequence, pH, and λ_{ex}). The emission is postulated to arise from the stacking of $C{:}CH^+$ base pairs from the two interpenetrated duplexes (51).

Based on these premises, we hypothesized that the features of steady-state emission spectra (e.g. intensity, maxima and bandshape) could be exploited to assess the secondary structures adopted by DNA in various conditions. Towards this end, we report herein a large-scale fluorescence analysis of DNA sequences, encompassing several types of secondary structures. Using a systematic assessment and multivariate analysis of emission spectra of a panel of synthetic oligonucleotides with established conformations (G4, iM, single and double strands), we were able to confidently discriminate between different structural motifs. The results of this work can be exploited for the implementation of a label-free and high-throughput test that could be further extended to genome-wide analyses.

MATERIALS AND METHODS

Oligonucleotides and buffer solutions

All chemicals were obtained from Sigma–Aldrich and used as supplied, without further purification. Experiments were performed in aqueous buffers containing 0.01 M lithium cacodylate and 0.1 M of the relevant chloride salt (KCl for buffers A and B, NaCl for buffer C, Table 1), unless stated otherwise. The pH was adjusted by addition of 0.1 M LiOH solution. Oligonucleotides (sequences: Supplementary Table S1) were purchased from Eurogentec (RP-Gold Cartridge purification grade) and used without further purification. Stock solutions of oligonucleotides with strand concentration of 100 μM (except for *46AG*: 50 μM and *DDD*: 200 μM) were prepared in deionized water and stored at

Table 1. Buffer solutions used in this work

Buffer	Li^+	K^+ or Na^+	pH
A	≈ 10 mM	K^+, 100 mM	7.2
B	≈ 10 mM	K^+, 100 mM	5.5
C	≈ 10 mM	Na^+, 100 mM	7.2

−20 °C. Samples for CD and fluorescence experiments (referred to as 'working solutions') were prepared by diluting the stock solutions with relevant buffers to a concentration of 5.7 μM (except for *46AG*: 2.85 μM and *DDD*: 11.4 μM). Heteroduplexes (Supplementary Table S1) were prepared by mixing equal volumes of working solutions of the corresponding single strands, to give heteroduplex concentration of 2.85 μM, accounting for the doubled number of nucleotides. Calf thymus DNA (*ct DNA*, Invitrogen, 10 mg ml^{-1}) was diluted with deionized water to $c \approx 4.2$ mM (nucleotides), and further diluted with the relevant buffer to 125 μM so as to obtain a working solution with a comparable nucleotide concentration as in oligonucleotide samples (considering 22 as the average length of oligonucleotides in Supplementary Table S1). Working solutions were subsequently annealed (5 min at 95 °C), let equilibrate to 20 °C overnight, and stored at 4 °C. For experiments involving mixtures of buffers, DNA samples were annealed separately according to the same protocol and, after equilibration, mixed in the relevant quantities. For experiments carried out at increasing KCl concentration, a 5.7 μM sample of *22AG* in 0.01 M lithium cacodylate buffer (pH 7.2) was gradually supplemented with KCl by addition of aliquots of concentrated (0.01, 0.1 or 1 M) KCl solutions. For experiments carried out at increasing pH, a 5.7 μM sample of *EPBC* in buffer B was titrated with aliquots of 1 M LiOH solution. The resulting solutions were left to equilibrate at 20 °C for 10 min before analysis. The spectra obtained in the last two cases were corrected by multiplication by the dilution factor, to account for the dilution effect.

CD spectroscopy

CD spectra were recorded with a Jasco J-1500 spectropolarimeter. Spectra were recorded using working solutions of DNA in pure buffers (A, B or C) or in mixtures of buffers A and C (1:99, 2:98, 5:95, 10:90, 15:85, 20:80, 40:60, 70:30), unless otherwise stated, in quartz cuvettes with rectangular cross-section (path length 1 × 0.4 cm), with the beam passing through a path length of 0.4 cm. Parameters used for spectra acquisition: wavelength range, 210–330 nm; scan speed, 50 nm min^{-1}; number of averaged scans, 3; data pitch, 0.5 nm; bandwidth, 2 nm; integration time, 1 s; temperature, 22 °C. Spectra were subsequently corrected for the blank. Finally, spectra were converted to molar dichroic absorption $\Delta\varepsilon$ [M^{-1} cm^{-1}] = θ / (32980 × c × ℓ), where θ is the CD ellipticity in millidegrees (mdeg), c is DNA concentration in M, and ℓ is the path length in cm.

Fluorescence emission spectra

Fluorescence excitation and emission spectra were recorded with a HORIBA Jobin–Yvon FluoroMax-3 spectrofluorimeter, in asymmetric quartz cuvettes (path lengths of 1 × 0.4 cm for emission and excitation beams, respectively). For each sample, two spectra were acquired using $\lambda_{ex} = 260$ and 300 nm and emission range of 270–510 nm and 310–590 nm, respectively (slit widths: 5 nm for both excitation and emission beams). For both spectra, data pitch was fixed to 1 nm and integration time to 1 s. All spectra were corrected for the blank, recorded with the appropriate buffer, and truncated to the regions devoid of Raman or Rayleigh scattering bands (i.e. $\lambda_{ex} = 260$ nm: $\lambda_{em} = 305$–505 nm; $\lambda_{ex} = 300$ nm: $\lambda_{em} = 350$–585 nm). Of note, mathematical models allowing the suppression of scattering effects could also be applied for more precise corrections (52,53); however, due to the lack of significant emission and spectral variation in the regions where the scatters appear, we deemed this simple and straightforward process suitable for our scope. The obtained spectra were corrected for the inner filter effect, neglecting the re-absorption of the emitted light (Equation 1):

$$F_{corr}=F_{obs}\times 10^{A_{ex}/2}, \qquad (1)$$

where F_{corr} and F_{obs} are the corrected and recorded emission, respectively, and A_{ex} is sample absorption at the excitation wavelength. For multivariate bandshape analysis, the truncated spectra were normalized to the 0–1 interval.

Single-wavelength absorption measurements

Absorption measurements were performed for all samples at 260, 300 and 400 nm, on a HITACHI U-2900 UV-VIS spectrophotometer. After zeroing of the absorption on a blank sample, the absorption of the sample at the three wavelengths was recorded in a quartz cuvette (1 cm path length). The data at 260 and 300 nm were then used to correct the emission spectra obtained upon excitation at these wavelengths. The absorption at 400 nm was used as an internal control.

Multivariate analysis

Principal component analysis (PCA) and linear discriminant analysis (LDA) of normalized emission spectra was performed with Origin Pro 2018b (OriginLab, Northampton, MA). Data for LDA are presented as Canonical Variables 1 versus 2 plots, using 85% confidence ellipses. Leave-one-out test was used for internal validation of the LDA method.

Quantum yield measurements

An appropriate aliquot of the DNA working solution was diluted to 1 ml with buffer A (except for *EPBC* and *i-HRAS2* sequences which were tested in buffer B) so as to obtain a maximum absorbance value of 0.11 at 265 nm. Afterwards, the sample fluorescence was recorded using $\lambda_{ex} = 265$ nm and other parameters as described above. The resulting spectrum was integrated between 305 and 505 nm, after blank subtraction. 200 μl of the solution were removed from the cuvette and substituted with 200 μl of the appropriate buffer, and the absorption and emission were measured again. This procedure was performed four times, to obtain a total of five data points. The integrated areas were then plotted as a function of the absorbance at 265 nm and fitted to a linear model. The same protocol was applied to the reference (quinine sulfate in 0.5 M H_2SO_4, $\Phi = 0.546$) (54), except that its spectrum was acquired and integrated between 275 and 600 nm. The quantum yield (Φ) for each DNA was calculated according to Equation (2):

$$\Phi_{\text{sample}} = \Phi_{\text{ref}} \times \frac{\text{Grad}_{\text{sample}}}{\text{Grad}_{\text{ref}}} \times \frac{n^2_{\text{sample}}}{n^2_{\text{ref}}}, \qquad (2)$$

where Grad is the slope calculated from each (Integrated area) versus (Absorbance) plot, and n is the refractive index of the solvent (1.3325 and 1.346 for buffer and 0.5 M H_2SO_4, respectively); 'sample' and 'ref' denote the DNA sample and the quinine sulfate reference, respectively.

RESULTS

Intrinsic fluorescence reveals secondary structures of DNA oligonucleotides and their conformational changes

In the first instance, we sought to verify whether steady-state emission spectra could be used to monitor the folding of secondary structures. First, we focused on the sequence *22AG* (cf. Supplementary Table S1), a well-studied model G4-forming oligonucleotide (55,56), and proved that its transition from random coil to a G4 structure could be monitored by emission spectroscopy in as much detail as by CD spectroscopy. We recorded CD as well as fluorescence spectra using two excitation wavelengths (λ_{ex} = 260 and 300 nm, as previously used for observing iM fluorescence (51)), in K^+/Na^+-free and K^+-containing buffers (Figure 1). CD spectra of *22AG* show a typical signature (25) of hybrid G4 conformations in the K^+-containing buffer A, and a spectrum characteristic of a random coil in a K^+/Na^+-free buffer (Figure 1A). The emission spectra mirror this change: in the K^+-containing buffer, *22AG* displays a strong emission band peaking at 350 nm (similar features were observed with a related telomeric sequence (41)), while in K^+/Na^+-free conditions the emission drops to about a half of the intensity and the broad maximum red-shifts to 395 nm (Figure 1B). The spectra acquired upon 300-nm excitation also display some differences: in the K^+-containing buffer the emission spectrum peaks at 356 nm, with a large shoulder around 410 nm. In the K^+-free buffer, the maximum is red-shifted to 400 nm, and is slightly less intense (Figure 1C). Interestingly, comparable results were observed for other G4-forming sequences (Supplementary Figure S1).

The descriptive power of intrinsic emission is not limited to the characterization of end-point conditions. Thus, upon progressive increase of K^+ concentration, *22AG* undergoes a gradual folding (57,58), which is directly reflected in both emission and CD spectra (Figure 1D; cf. Supplementary Figure S2 for the full CD and emission spectra). Interestingly, the comparison of characteristic parameters for the two sets of spectra (i.e. the integral fluorescence for emission spectra recorded upon excitation at 260 or 300 nm, and the molar dichroic absorption at 294 nm for CD spectra) leads to overlapping transition profiles, demonstrating the sensitivity of intrinsic fluorescence to the conformation changes accompanying the folding of a G4 structure.

Next, we assessed whether intrinsic fluorescence could be used to monitor more subtle conformation changes. As reported in the literature, *22AG* sequence shifts from a major anti-parallel conformation in Na^+-rich conditions to a mixture of hybrid forms in K^+-containing buffers (56,57). This transition can be monitored by recording CD spectra of solutions containing a variable proportion of the two cations: upon increasing the K^+/Na^+ ratio, the CD spectrum gradually switches from the one typical of an anti-parallel G4 (positive maxima at 240 and 295 nm, negative maximum at 265 nm) to that of a hybrid G4 (positive maximum at 290 nm, with two positive shoulders at 265 and 250 nm) (Figure 2A). Interestingly, an analogous transition is observed in the corresponding fluorescence spectra (Figure 2B, C). Upon increasing K^+/Na^+ ratio, the emission decreases by about one third of its initial value and the emission maximum blue-shifts from 375 to 350 nm. The information provided by the two methods is perfectly aligned, as shown in Figure 2D (cf. Supplementary Figure S3 for the analysis of fluorescence spectra obtained with λ_{ex} = 300 nm).

Finally, we studied the transition of *EPBC* from random-coil to iM structure. In this case, the switch from neutral to acidic pH is accompanied by the formation of a strong positive CD band at 289 nm and a moderately intense negative band at 265 nm (Figure 3A). In the emission spectra obtained upon 260-nm excitation, the intensity of the 330-nm maximum increases about 3-fold and a second broad emission band appears, centered at ~420 nm (Figure 3B). The emission spectra obtained upon excitation at 300 nm vary accordingly: the emission maximum (around 410 nm) at pH 5.5 is about five times more intense than at neutral pH (Figure 3C). A similar behavior was observed for other iM-forming oligonucleotides (Supplementary Figure S4). As previously observed for the folding of *22AG*, the gradual conformational transition can also be monitored for *EPBC*. In fact, the unfolding of the iM structure upon increase of pH is directly reflected in both CD and emission spectra (Figure 3D; cf. Supplementary Figure S5A–B for the full spectra), and the comparison of the characteristic parameters for the two sets of spectra (i.e. integrated emission intensity and the molar dichroic absorption at a characteristic wavelength, from CD spectra) leads to overlapping transition profiles. Similar trends were observed when using the data from emission spectra obtained upon excitation at 300 nm (Supplementary Figure S5C, D).

Emission properties of various DNA sequences

Based on these promising data, we assessed the potential of intrinsic fluorescence as a reporter of oligonucleotide conformation on a larger scale. Towards this end, we screened an 89-membered set of DNA analytes with different and well-established conformations (Supplementary Table S1). Specifically, the selected sequences comprise 46 G4 structures (including 14 hybrid, 18 parallel, and 14 anti-parallel folds), 14 iM-forming sequences (iMFS) of varying stability, 14 duplexes (including three auto-complementary sequences, three hairpins, seven hetero-duplexes and one highly polymerized genomic DNA), and 12 single-strands with different base composition and purine versus pyrimidine content. For the sake of homogeneity of our assay,

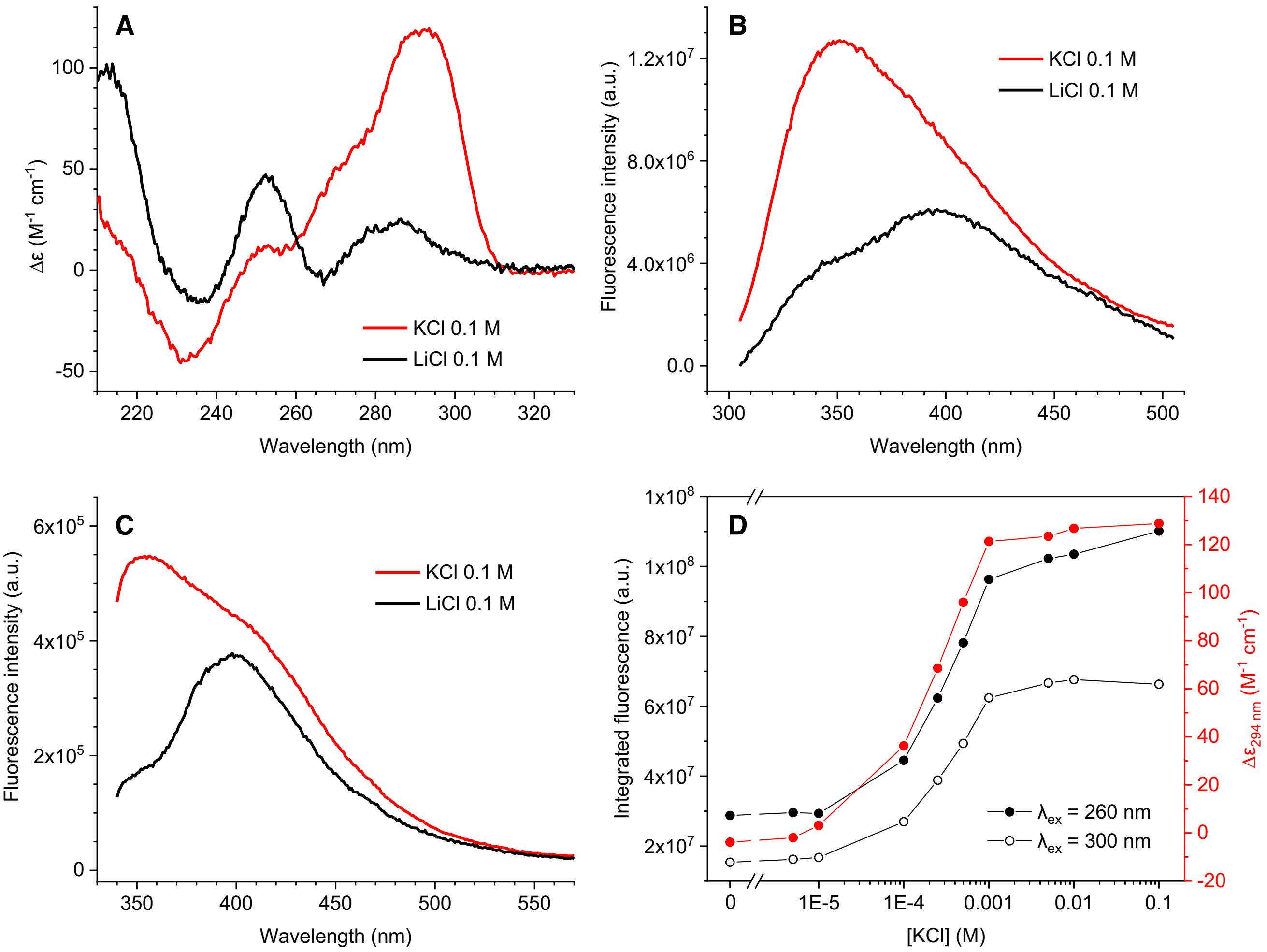

Figure 1. Following the folding of *22AG* by CD and intrinsic fluorescence. (**A**) CD spectra of *22AG* (c = 5.7 μM) in K^+-rich (buffer A) and K^+-free buffers (0.1 M LiCl, 0.01 M lithium cacodylate, pH 7.2). (**B, C**) Fluorescence spectra (B: λ_{ex} = 260 nm, C: λ_{ex} = 300 nm) of the samples presented above. (**D**) Comparison of (black) integrated fluorescence emission intensity and (red) molar dichroic absorption at 294 nm (from CD spectra) of 5.7 μM *22AG* solutions in 0.01 M lithium cacodylate buffer with variable KCl content (0, 5×10^{-6}, 1×10^{-5}, 1×10^{-4}, 2.5×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}, 1×10^{-2} and 1×10^{-1} M), pH 7.2. Fluorescence spectra were corrected for the inner filter effect.

we deliberately omitted bi- and tetramolecular G4 structures (such as d[TG_4T]$_4$ and d[$G_3T_4G_3$]$_2$), as well as unimolecular G4 structures with a known propensity to form dimers or higher aggregates, such as *N-myc* (59), *93del* (60), *G$_3$T* and its analogues (21). The conformations adopted by all sequences were initially assessed by CD spectroscopy in three different buffers (A, B and C). The results (Figures S6 and S7, summarized in Supplementary Table S2) confirmed that most putative G4-forming sequences adopted the expected conformation in buffer A (Supplementary Figure S6, A/D/G), with the exception of *LWDLN-1*, *19wt* and *SP-PGQ3* whose CD spectra did not agree with the reported conformations, most likely due to the differences in experimental conditions employed for their structural characterization (cf. Supplementary Table S2 footnote). Subtle or no changes were observed when these sequences were reassessed in K^+-containing acidic conditions (buffer B). All putative G4-forming sequences appear to be folded under these conditions, and most maintained the same conformation as in buffer A, except for *Bcl2Mid* which underwent a conformational change from hybrid to parallel form, and *hras-1* whose CD spectrum gave evidence of a partial conformational change (Supplementary Figure S6, B/E/H, and Supplementary Table S2). The use of the Na^+-containing buffer C was, instead, more problematic, since 12 out of 46 sequences were completely or partially unfolded under these conditions). In addition, 16 out of 46 sequences underwent considerable conformational changes, most typically a change from hybrid (*22AG* as described above, *46AG*, *26TTA*, *23TAG*, *24TTA*, *chl1*, *UpsB-Q-3*) or parallel (*c-kit2-T12T21*, *VEGF*, *Myc1245*) to anti-parallel forms (Supplementary Figure S6, C/F/I, and Supplementary Supplementary Table S2). Therefore, we decided to discard buffer C for the following emission studies. With regard to iMFS, these appeared mostly folded in buffer B, as expected in slightly acidic conditions. In buffer A, all iMFS display CD spectra compatible with a random-coil arrangement, likely due to insufficient protonation, although the presence of a small fraction of folded structure cannot be ruled out on the basis

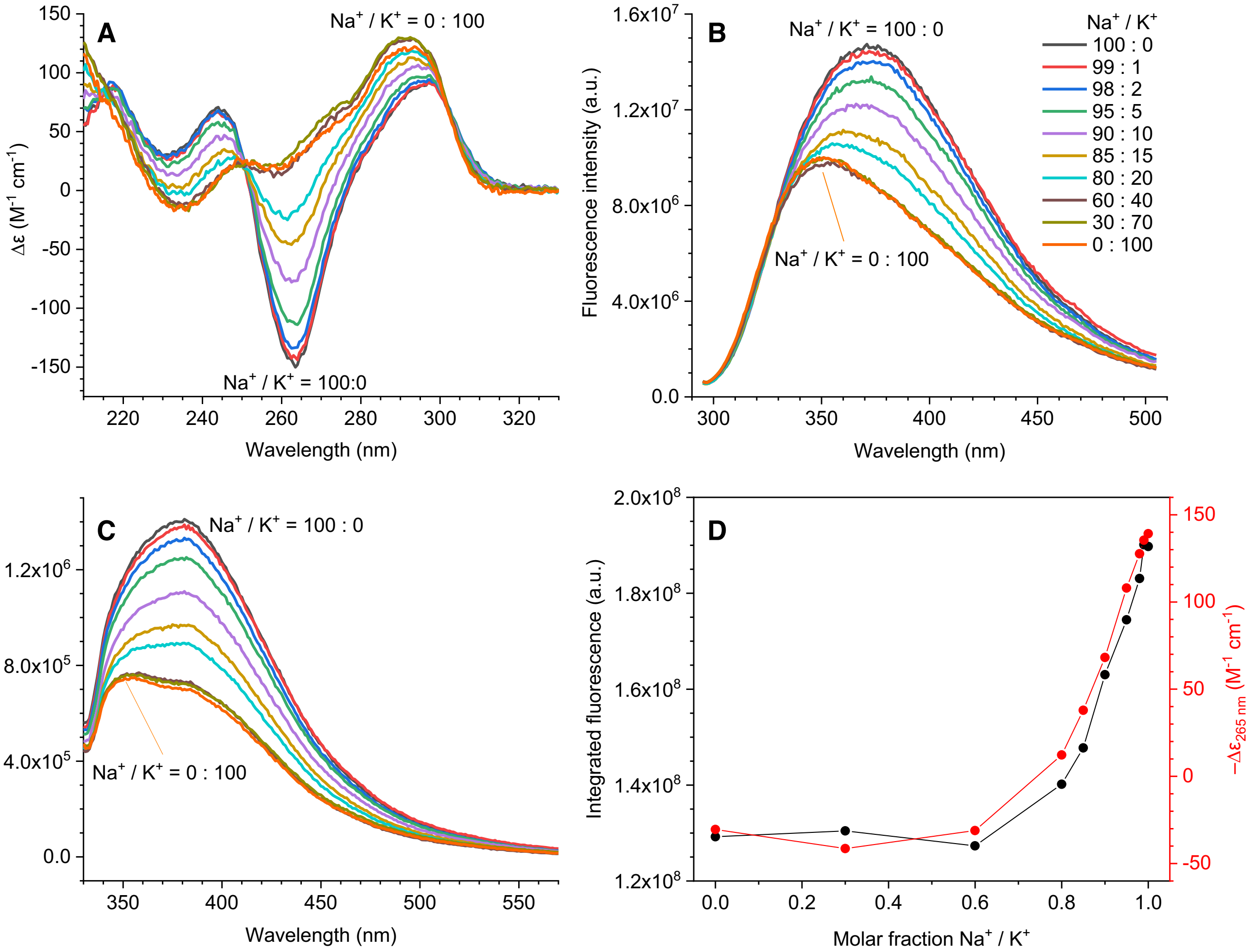

Figure 2. Conformational transition of *22AG* from anti-parallel to hybrid G4 monitored by CD and intrinsic fluorescence. (**A**) CD spectra of *22AG* solutions ($c = 5.7$ μM) containing variable proportions of Na^+ and K^+ (total NaCl + KCl concentration of 0.1 M in all solutions, lithium cacodylate buffer 0.01 M, pH 7.2). (**B, C**) Corresponding emission spectra (B: $\lambda_{ex} = 260$ nm, C: $\lambda_{ex} = 300$ nm). (**D**) Comparison of the integrated fluorescence intensity ($\lambda_{ex} = 260$ nm) and molar dichroic absorption at 265 nm (from CD spectra) at the various KCl/NaCl ratios.

of CD spectra (Supplementary Figure S7A, B). Finally, according to their CD spectra, double-stranded and single-stranded sequences adopted the expected conformation in all buffers, regardless of the specific pH (Supplementary Figure S7C–F).

Next, we recorded emission spectra of all aforementioned sequences in buffers A and B, using excitation at 260 and 300 nm (Figures S8 and S9, respectively). The features of the emission spectra (i.e. band shape and intensity) were found to vary to large extents inside each group. However, several patterns could be identified upon naked-eye examination of the spectra. On the overall, the emission intensities were higher upon excitation at 260 nm than at 300 nm, due to the differences in sample absorbances at these wavelengths. As a general behavior, the intrinsic emission of G4 structures is significantly more intense (∼3.5-fold, based on the comparison of the median integrated fluorescence intensities) than that of other structures upon 260-nm excitation (Figure 4A, B and Supplementary Figure S8). This difference is more limited upon excitation at 300 nm (G4 emission is only 1.6-fold higher than that of other structures, Figure 4C, D). Interestingly, iM structures become the most fluorescent ones in buffer B upon 300-nm excitation, with a median intensity 1.4-fold higher than that of G4 structures under the same conditions (Figure 4D). This difference is not observed in the data obtained upon excitation at 260 nm, due to the considerable spectral shape changes. In fact, in this case, the appearance of the iM-characteristic, broad peak centered around 425 nm in buffer B is accompanied by a reduction of the 320-nm maximum by approximately a half (Supplementary Figure S8G, H), leading to the decrease of the integrated emission intensity (Figure 4B). On the contrary, the iM-characteristic red-shifted peak is predominant in the spectra obtained upon excitation at 300 nm, with an intensity increase of ∼3.5-fold upon moving from buffer A to B (Supplementary Figure S9G, H). This is likely due to the fact that a longer wavelength selectively excites the iM-characteristic band that otherwise appears only as a shoulder.

The spectral shape was also found to be informative of the oligonucleotide conformation, especially considering the spectra obtained upon 260-nm excitation, as demonstrated

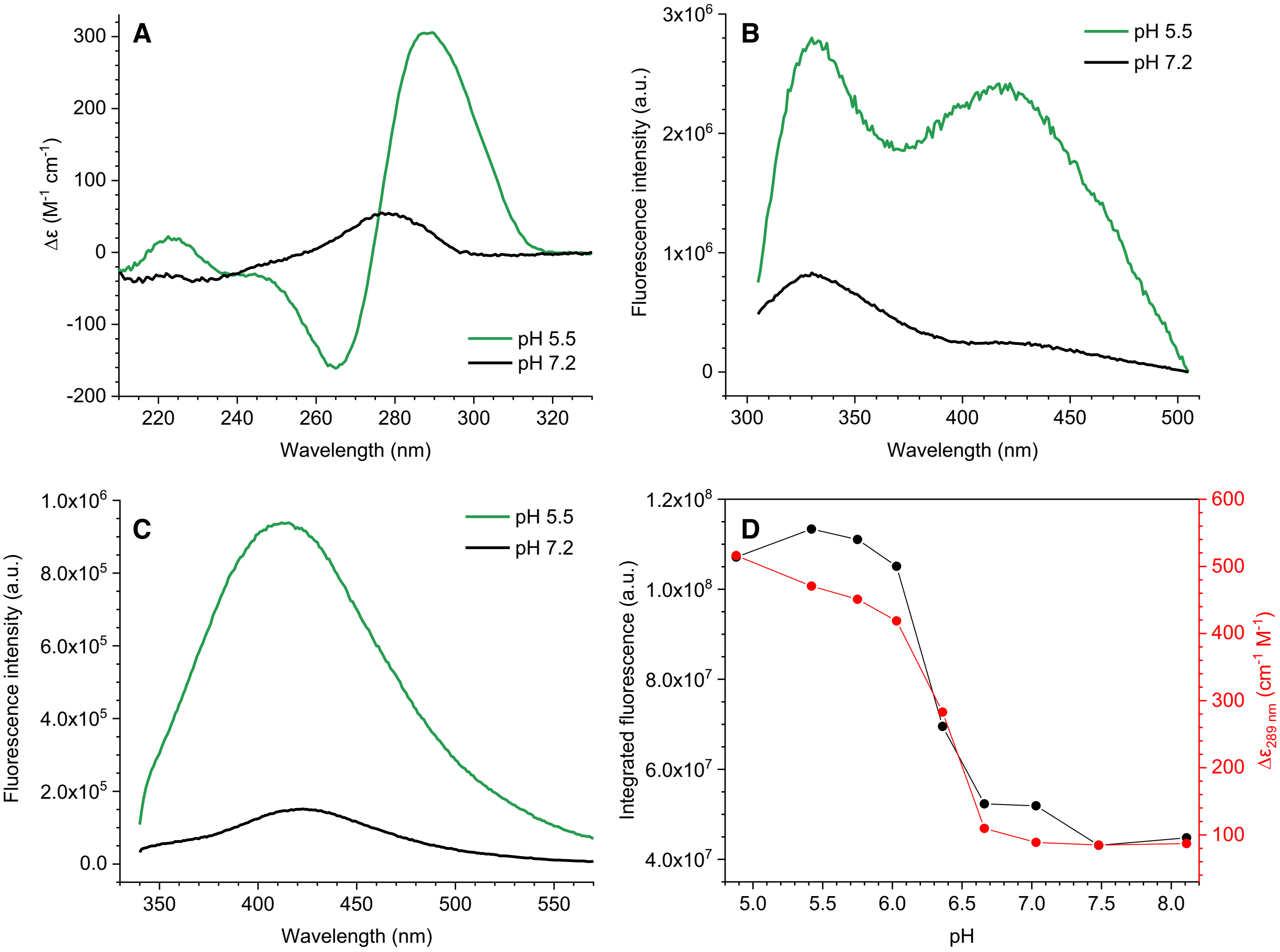

Figure 3. Following the folding of *EPBC* to iM structure by CD and intrinsic fluorescence. (**A**) CD spectra of *EPBC* ($c = 5.7$ μM) in neutral and acidic conditions (buffers A and B, respectively). (**B, C**) Fluorescence spectra (B: $\lambda_{ex} = 260$ nm, C: $\lambda_{ex} = 300$ nm) of the samples presented above. (**D**) Comparison of (black) integrated fluorescence emission ($\lambda_{ex} = 260$ nm) and (red) molar dichroic absorption at 289 nm (from CD spectra) of 5.7 μM *EPBC* solutions at variable pH (4.9–8.1, 0.01 M lithium cacodylate buffer, 0.1 M KCl). Fluorescence spectra were corrected for the inner filter effect.

by the comparison of group-averaged normalized emission spectra (Figure 5). In addition, the use of normalized emission spectra avoids the differences in intensity arising from the differences in the length of oligonucleotides (and thus in their absorbance at λ_{ex}). As shown in Figure 5A–C, and Supplementary Figure S10A–C, G4 structures give broad emission spectra in both buffers A and B, with maxima centered between 330 and 350 nm and gradually decreasing intensity at longer wavelengths. No general trends are observed that could enable the discrimination between different G4 topologies. In contrast, single- and double-stranded structures display significantly sharper spectra in both conditions, peaking around 350 and 330 nm, respectively (Figure 5D, E and Supplementary Figure S10D, E). In all cases, a more or less pronounced shoulder is observed between 400 and 500 nm. The same is true for iM-forming sequences in buffer A (Figure 5F), in which they are mostly unfolded. However, at lower pH (buffer B), a characteristic shoulder appears between 375 and 475 nm, whereas the maximum remains fixed at around 330 nm (Figure 5G). The spectra obtained using $\lambda_{ex} = 300$ nm do not display the same degree of shape variability (Supplementary Figure S11). In this case, most sequences display a maximum between 310 and 330 nm and no shoulders in any of the buffers, regardless of the adopted conformation. The only remarkable difference is observed for some G4 structures, displaying a blue-shifted shoulder of varying intensity.

Principal component analysis of the emission spectral dataset

To extract more information, the emission spectral dataset was subjected to principal component analysis (PCA). PCA is an unsupervised multivariate method, which describes the variance of the examined data matrix through a reduced number of variables and highlights the similarities between the analytes displaying similar response patterns. Thus, four emission spectra generated for each analyte (corresponding to two excitation wavelengths and two buffers, A and B) would account for 894 variables if the information obtained at each wavelength (1 nm data pitch) were to be used. These would be reduced to 447 variables for the spectra obtained in just one of the buffers, or to 201 (or 246) variables for a single spectrum (corresponding to the emission spectral windows, cf. Materials and Methods). PCA analysis enables

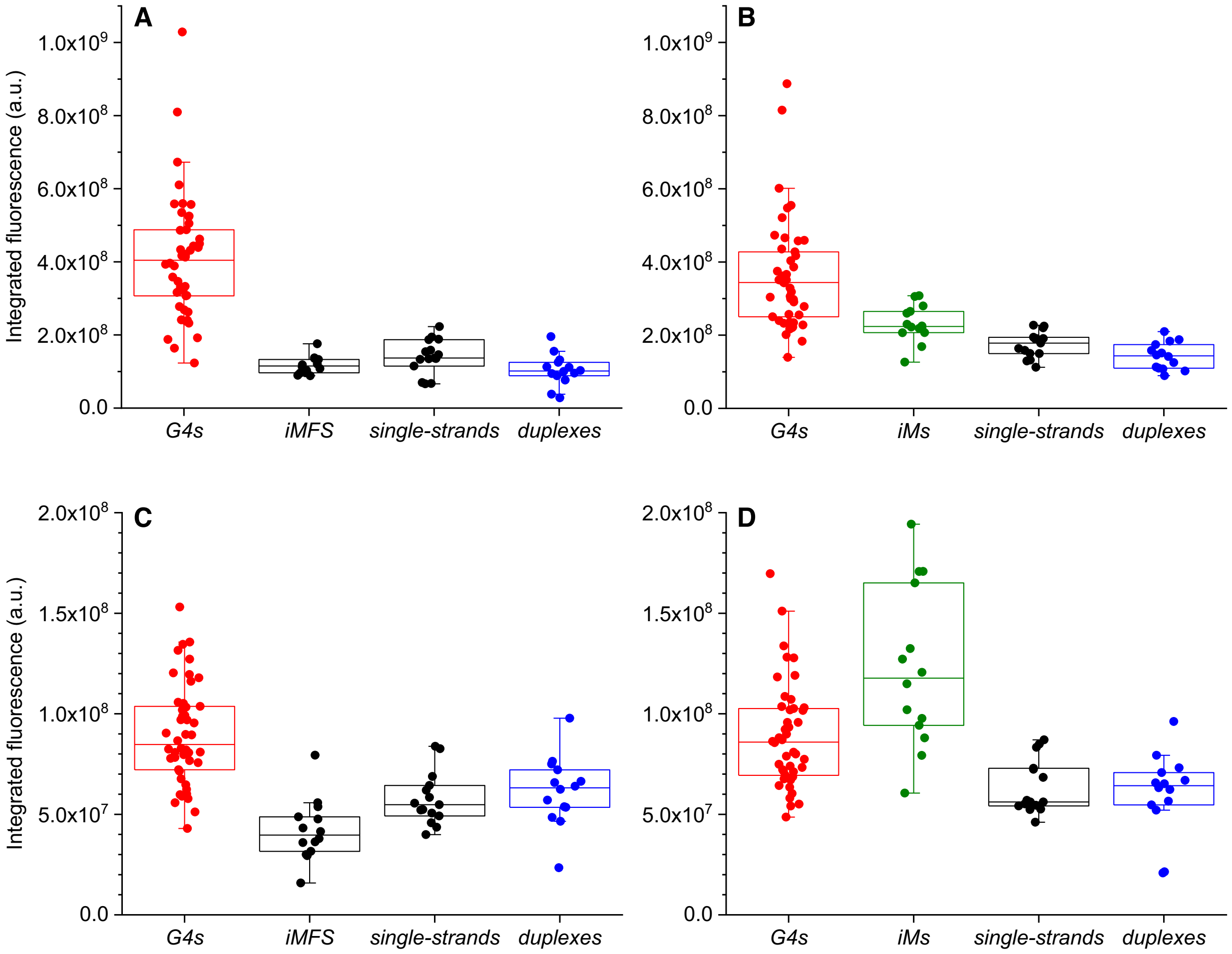

Figure 4. Box plots of the integrated fluorescence intensity of DNA samples grouped according to their conformation. (**A, B**) $\lambda_{ex} = 260$ nm in (A) buffer A and (B) buffer B, integration limits: 305–505 nm; (**C, D**) $\lambda_{ex} = 300$ nm in (C) buffer A and (D) buffer B, integration limits: 350–585 nm. Fluorescence spectra were corrected for the inner filter effect prior to integration.

to combine these data into a limited number of principal components (PCs) and facilitates their interpretation.

In order to understand which spectra were more suitable for the analysis, we first performed a preliminary PCA on each separate set of spectra acquired with different excitation and buffer conditions. Remarkably, upon using normalized spectra obtained in buffer A upon 260-nm excitation, we readily observed a good separation between G4 structures, mostly occupying the upper part of the PCA plot, and all other conformations (Figure 6A). In this case, the putative iM-forming sequences (iMFS) were combined with the group of single-stranded oligonucleotides, since they are mostly unfolded under these conditions. No clear difference could be observed between single- and double strands, in agreement with what could be qualitatively inferred from inspection of the spectra. When the same analysis was run on the data obtained in buffer B under the same excitation wavelength conditions, a partial separation could be inferred for i-motif structures: these sequences form a broad cluster overlapping with those of G4s and single- and double-strands (Figure 6B). In both buffers, the spectra obtained upon excitation at 300 nm were less suitable for the analysis (Supplementary Figure S12). In fact, in both cases the separation between the groups was less efficient due to the little amount of variance described by PC2 (18.4% and 16.4% for buffers A and B, respectively). PCA of raw emission spectra (Figures S8 and S9) was also attempted, but resulted in less efficient clustering (Supplementary Figure S13), presumably due to strong intragroup variance of emission spectra G4 structures.

In order to improve the clustering of iM structures and thus be able to detect the three separate groups (i.e. G4s, iMs and ‘others’, from now on used to designate the group of single-and double-strands) in a single plot, we attempted PCA on a combined dataset, including the spectra obtained upon 260-nm excitation in both buffers A and B. We speculated that, since the data obtained in each buffer alone were not sufficient to reveal the cluster of iM structures, a change in fluorescence properties between the buffers A and B, specific for iM-forming sequences, could provide an additional information element. Indeed, the visualization and separation of the three groups was significantly better in this case, especially when using the first three principal components (Figure 7A). The three principal components presented in the plot account for 82.3% of the dataset variance. Inspection of the scree plot (Supplementary Figure S14) shows

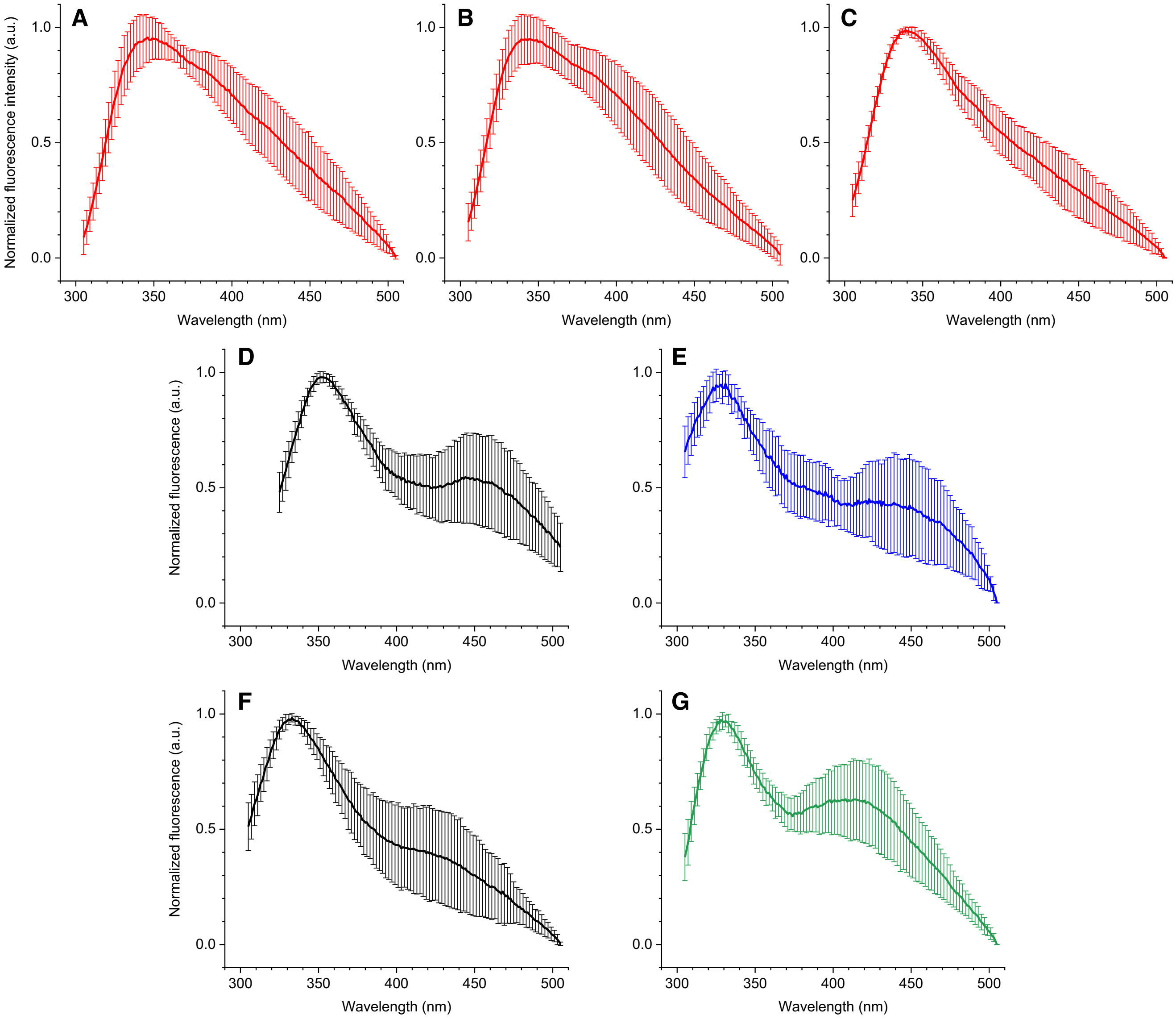

Figure 5. Group-averaged, normalized emission spectra ($\lambda_{ex} = 260$ nm, buffer A, unless stated otherwise) of the tested oligonucleotides, grouped according to their conformation (as confirmed by CD spectra). (**A**) hybrid G4s; (**B**) parallel G4s, (**C**) anti-parallel G4s, (**D**) single strands, (**E**) duplexes, (**F**) iM-forming sequences, (**G**) the same as F in buffer **B**. Error bars represent the standard deviation of the emission at a given wavelength inside each conformational group. *LWDLN-1*, *19wt* and *SP-PGQ3* were excluded from the analysis (see text).

that five components out of twenty that had been calculated are actually significant and describe up to 92.3% of the dataset variance. In order to better understand the concrete meaning of each of the PCs, we examined the various 2D plots (PC1 versus PC2, PC1 versus PC3 and PC2 versus PC3, Figure 7B–D). For PC2 and PC3, we could infer a clear correlation with the propensity of samples to adopt a G4 or iM structure, respectively. In fact, G4 structures have generally positive PC2 scores (‘*G4-likeness*’), whereas iMs, single- and double-stranded sequences have negative ones (Figure 7B and D). At the same time, iMs are the only structures with high negative value scores for PC3 (‘*iM-likeness*’), whereas all the other analytes display values around zero (Figure 7C and D). With respect to PC1, data points are distributed quite evenly along the axis, suggesting that the correlation is not purely conformational.

Dataset reduction and linear discriminant analysis of emission spectra

As a next step, we attempted to reduce the number of variables (and thus the dataset size) by selecting spectral regions that would be most relevant for sample grouping. Ideally, these should correspond to the wavelengths at which the analytes belonging to the same conformational group share similar spectral shapes, whereas the analytes from different groups show significant differences. To identify such zones, we calculated the intra-group variance for the normalized values obtained at each wavelength and, from the resulting three sets of values, the inter-group variance at each wavelength. These were plotted as a function of the wavelength (Supplementary Figure S15). From this graph, four zones (324–333 nm and 374–393 nm for the spectra obtained in

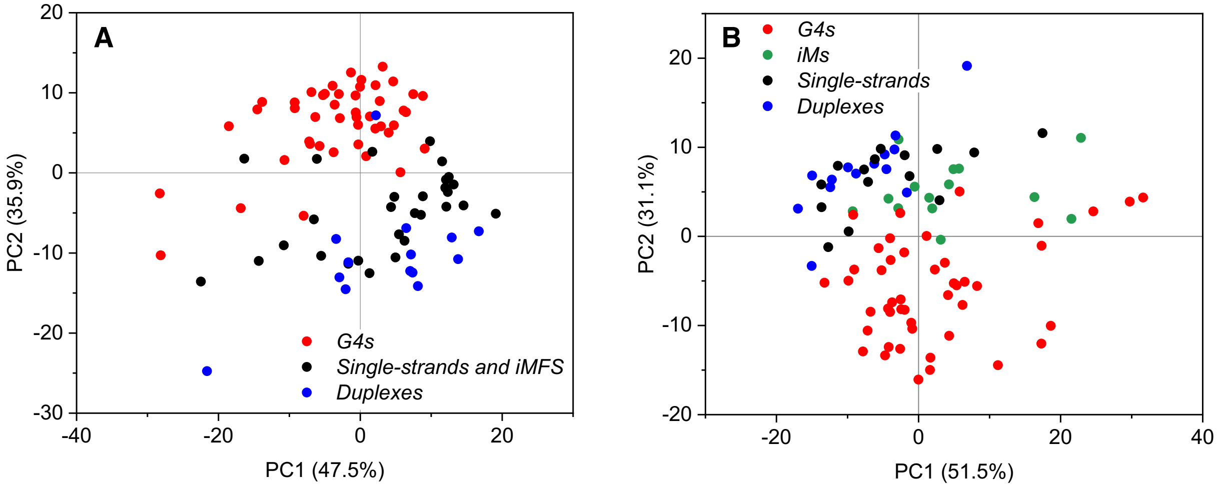

Figure 6. PC1 versus PC2 plots resulting from the PCA of normalized emission spectra recorded (**A**) in buffer A (λ_{ex} = 260 nm); (**B**) in buffer B (λ_{ex} = 260 nm). In (A), the putative iM-forming sequences (iMFS) sequences are grouped together with single-strands (in black), as they are mostly not folded in this buffer. In (B), they are shown in green, since they adopt a distinct conformation.

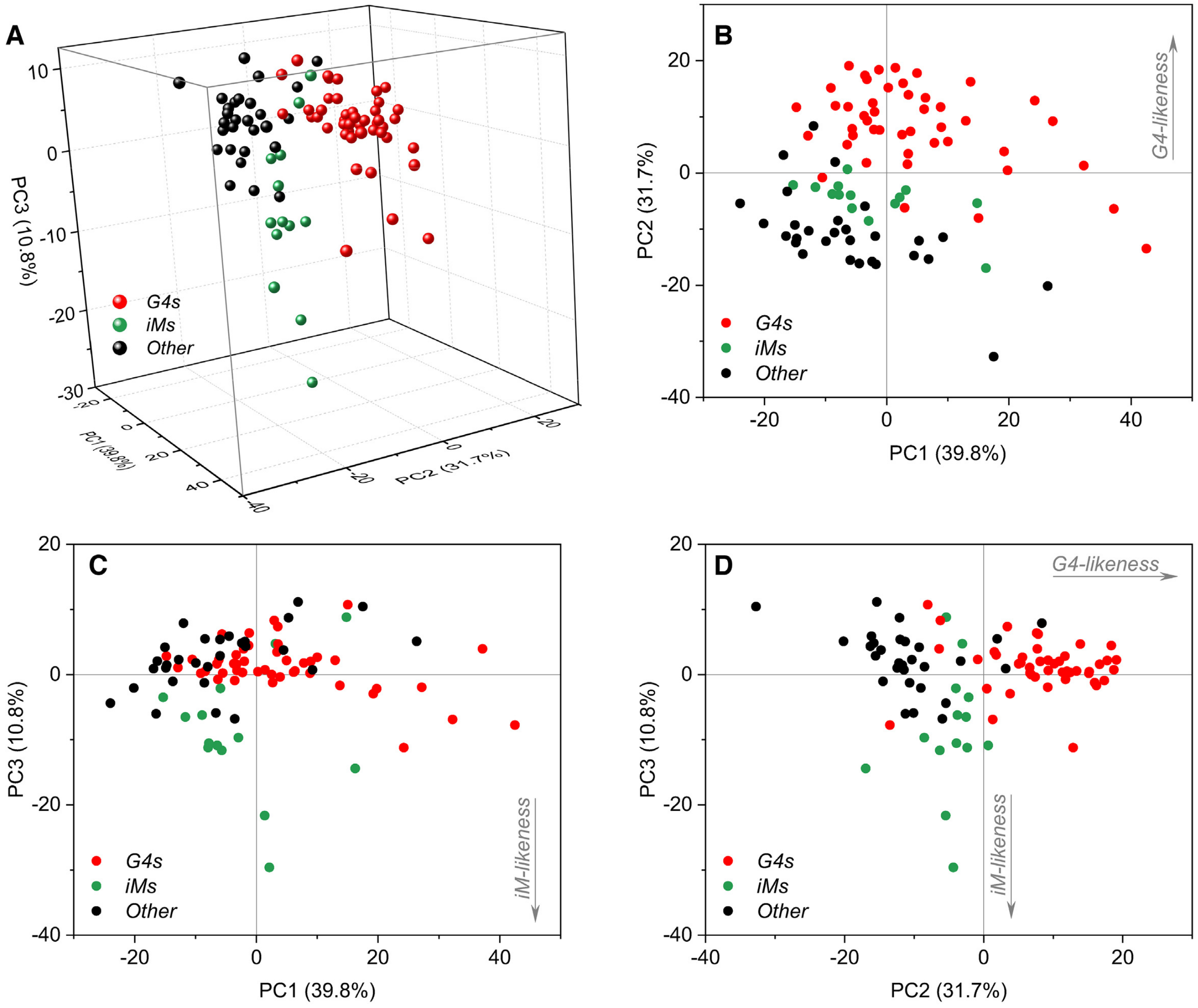

Figure 7. PCA of the combination of normalized emission spectra recorded in buffers A and B (λ_{ex} = 260 nm). (**A**) 3D PCA plot (PC1 versus PC2 versus PC3); (**B–D**) corresponding 2D plots.

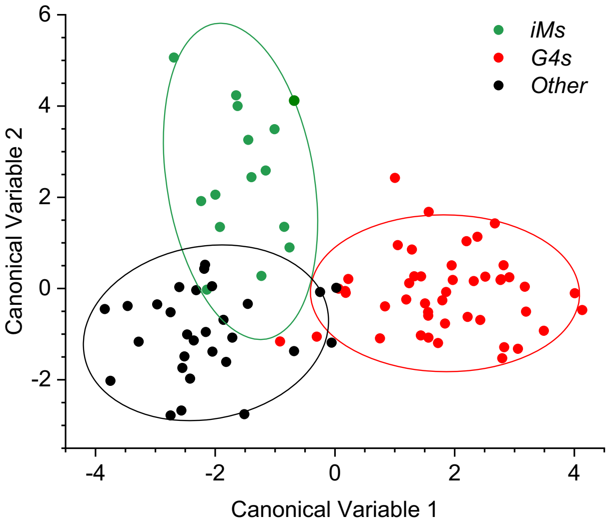

Figure 8. Canonical variable 1 versus Canonical variable 2 plot obtained from LDA of the reduced emission dataset (normalized emission data recorded at 335, 338, 341, 380, 383 and 386 nm in buffer A, and 329, 332, 376, 379, 382, 385, 388 and 391 nm in buffer B, $\lambda_{ex} = 260$ nm) on 89 characterized sequences. Ellipses represent 85% confidence zones for each group.

buffer B; 333–343 nm and 378–388 nm for buffer A) in which the inter-group variance was relatively high were selected for further examination. In this choice, we tried to favor the spectral zones in which the intra-group variance of iM structures was low, since this group showed the poorest clustering. The selected wavelengths account for 52 variables out of the 402 arising from the two combined spectra, corresponding to an 87.5% reduction of the dataset size. PCA was ran on this reduced dataset, to verify the quality of the clustering under these conditions (Supplementary Figure S16A), and gave satisfactory results, with the three groups still appearing as relatively well separated.

The resulting dataset was further reduced in order to subject to it to linear discriminant analysis (LDA). This is a supervised multivariate method, enabling to assign a test analyte to one of the *a priori* defined groups on the basis of the similarity of, in our case, emission spectrum profile, with those of the elements of the training set. The limiting factor in LDA is constituted by the group sizes, since the number of variables cannot exceed the number of training samples in each group. In our case, the group of iMs only contains 14 oligonucleotides, thus restricting the number of exploitable variables to 14. We thus evenly selected 14 wavelengths out of the 52 identified ones and ran PCA again to ascertain that the information content and the groups separation were retained. Satisfyingly, this was the case (Supplementary Figure S16B), and we proceeded to perform LDA on the established dataset (Figure 8).

Remarkably, DNA sequences were found to cluster relatively well, according to their conformation. Although the partial overlap of the clusters could not be avoided (in particular, those of iMs and 'other' sequences), the validation by the leave-one-out method confirmed a 90.8% of correct identifications. Considering the complexity of our task with respect to usual chemometrics applications, this error rate is considered quite satisfactory.

Chemometric assessment and fluorescence properties of novel sequences

In order to assess the behavior of the established multivariate methods upon analysis of sequences with unknown conformation, we recorded the emission spectra of a number of test sequences and analyzed them by PCA and LDA. In the first instance, we analyzed four randomly generated, 24- to 25-mer sequences with moderately high *G4Hunter* scores (*RND-HS1* to *RND-HS4*, Supplementary Table S3). PCA analysis of their emission spectra revealed that in the PCA plot, *RND-HS1* fell definitely closer to the group of single- and double-strands ('other'), *RND-HS2* was located in between the two groups, and *RND-HS3* and *RND-HS4* fell closer to the center of the G4 group (Supplementary Figure S17, A). Subsequent LDA testing confirmed this interpretation (Supplementary Figure S17, B), assigning *RND-HS3* and *RND-HS4* to the G4 group (with probabilities $P = 0.98$ and 0.80, respectively) and *RND-HS1* and *RND-HS2* to the group of single- and double-strands (Supplementary Table S3); interestingly, the probability of *RND-HS2* to belong to the G4 group was not negligible ($P = 0.095$). To test the validity of the LDA prediction, we recorded ^{1}H NMR, CD and TDS spectra of the four test oligonucleotides. NMR data (Supplementary Figure S18) showed that all four sequences showed the presence of imino proton peaks characteristic of the formation of secondary structures (16). In the case of *RND-HS1* sharp peaks were observed between 12.5 and 14 ppm, indicative of formation of Watson–Crick base pairs and a duplex-type structure, despite the presence of a small hump at 10–11.5 ppm. *RND-HS2* displayed sharp peaks at 12.8–13.2 ppm indicative of a Watson–Crick base pair related structure and a broad peak centered at 10.7 ppm characteristic of Hoogsteen base pairings; it is noted that the formation of intermolecular structures could be favored by the high concentration of oligonucleotides required for NMR experiments (130 μM, i.e. 23-fold higher with respect to fluorescence and CD studies). In the case of *RND-HS3*, NMR spectrum in the imino region showed several sharp peaks in 10–12 ppm region and a sharp peak at 13 ppm, characteristic of G4-duplex hybrid structures; indeed, the formation of a hairpin-type structure is possible, considering the presence of two 6-nt complementary runs of GC base pairs in this sequence (Supplementary Table S3). Finally, *RND-HS4* exclusively displayed broad signals in the Hoogsteen base-pair region giving evidence of formation of multiple G4 structures, in agreement with the LDA prediction based on the fluorescence data. The inspection of CD spectra of these sequences in buffer A suggested that *RND-HS1* is unlikely to fold into a G4 structure, whereas the other three sequences might adopt such a conformation (Supplementary Figure S19A). Furthermore, the results of thermal difference spectra (61) were in agreement with this assignment: *RND-HS3* and, particularly, *RND-HS4* showed negative bands in the 295−300 nm region, giving evidence of at least partial formation of G4 structures, whereas *RND-HS1* and *RND-HS2* were devoid of this band (Supplementary Figure S19B), supporting our NMR data and fluorescence-based LDA typing.

Subsequently, we tested a number of sequences adopting secondary structures other than those represented

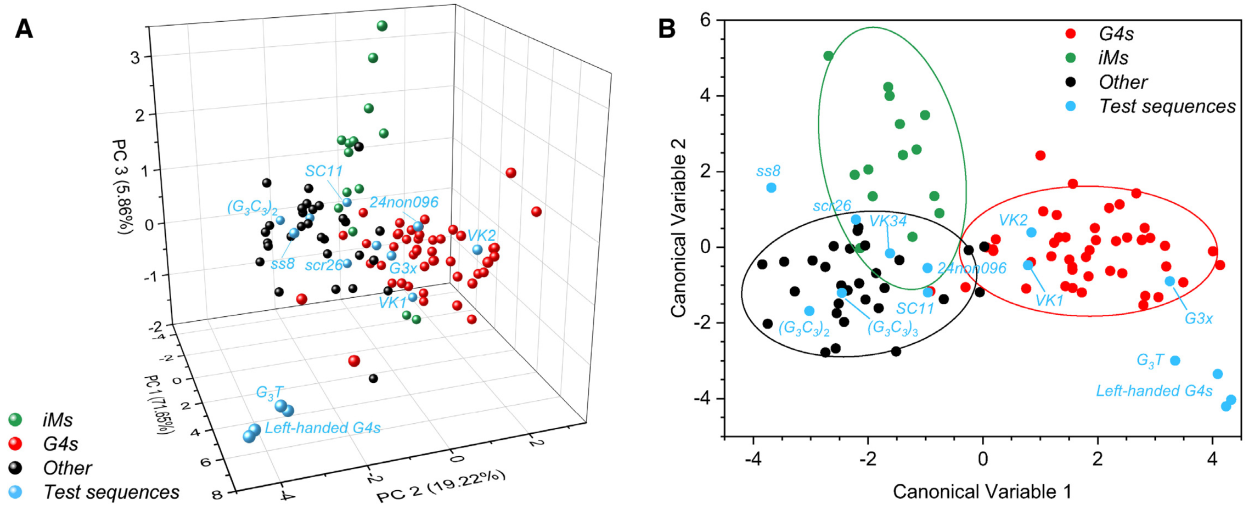

Figure 9. (**A**) PC1 versus PC2 versus PC3 plot obtained from the analysis of the training set (reduced dataset), supplemented with the emission data obtained for test sequences with peculiar conformations (*G_3T*, *ZG4*, *Block2Δ*, *2xBlock2*, *VK1*, *VK2*, *VK34*, *$(G_3C_3)_3$*, *$(G_3C_3)_2$*), *SC11*, *G3x*, *ss8*, *24non096* and *scr26*). (**B**) LDA plot for the same dataset.

in the training set. These include the stacked dimeric G4 structure known to be the most fluorescent unmodified oligonucleotide described to date (*G_3T*) (21,43,46), three left-handed G4 structures (*ZG4*, *Block2Δ*, *2xBlock2*) (62,63), three peculiar non-G4 quadruple helices (*VK1*, *VK2*, *VK34*) (64,65), two A-type duplexes (*$(G_3C_3)_3$* and *$(G_3C_3)_2$*) (66,67), a G-hairpin (*SC11*) (68), a G-triplex (*G3x*) (69,70), and three supposed single strands displaying peculiar CD signatures, namely *ss8*, *24non096* and *scr26* (Supplementary Table S4). The CD spectra of these sequences in buffers A and B are presented in Supplementary Figure S20. In the PCA plot based on the fluorescence data (Figure 9A), most non-G4-forming sequences were distributed in the zone of single- and double-strands, except for tetraplexes *VK1* and *VK2,* and G-triplex *G3x*, which fell in the G4 zone. In agreement with this result, all sequences were assigned to the corresponding groups upon LDA analysis (Figure 9B, and Supplementary Table S4), except for *ss8*, which was mis-assigned to the iM group ($P = 0.87$) and *scr26* which was assigned to iM and 'other' groups with almost equal probabilities ($P = 0.48$ and 0.51, respectively). This result indicates that the method yields a relatively low rate of false positive results with respect to iM and G4 groups. At the same time, the inconsistency between the group assignment and the expected structure (as can be inferred from the sequence analysis) points out to some peculiarities, for example in the case of *ss8* and *scr26*, as already evidenced by their CD spectra and, in the case of *ss8*, the anomalous results of the fluorescent-probe analysis (71). Such anomalies clearly call for further in-depth studies.

Another peculiar observation following from the results of the multivariate analysis is represented by the left-handed G4 structures which, along with *G_3T*, occupy an otherwise empty region of the PCA plot (Figure 9A). Accordingly, they spot in the bottom right corner of the LDA plot, relatively far away from all other clusters (Figure 9B); nonetheless, LDA unambiguously assigned these structures to the G4 group (Supplementary Table S4). This behavior is likely due to the peculiar emission spectra of these structures, which are characterized by a sharp emission band peaking around 385 nm in both buffers A and B (Supplementary Figure S21). This value is 35–55 nm red-shifted with respect to the values observed for other G4 structures and, more in general, for all the examined oligonucleotides. Of note, the emission completely fades off upon switching from a K^+ to a Na^+-rich buffer, as a result of the structure unfolding evidenced by the corresponding CD spectra (Figure 10). The peculiarities of left-handed G4s are not limited to the shape and position of their fluorescence bands, but also extend to the emission intensities. In fact, upon comparison of the raw emission data in the maxima, left-handed G4s appear to be on average 2.2-fold more fluorescent than other representative G4 sequences. A quantitative description of this phenomenon is offered by the values of fluorescence quantum yield (Φ, Table 2; cf. Supplementary Figure S22 for the determination of these values). Indeed, left-handed G4s prove in general more fluorescent ($\Phi = 2.0 \times 10^{-3}$ to 3.1×10^{-3}) than other G4 structures ($\Phi = 4.0 \times 10^{-4}$ to 2.4×10^{-3}) and other sequences ($\Phi = 2.8 \times 10^{-4}$ to 1.4×10^{-3}). Interestingly, the quantum yields of left-handed G4 structures are almost as high as that of the *G_3T* sequence ($\Phi = 3.7 \times 10^{-3}$ in our conditions, i.e. almost 2-fold higher than 2×10^{-3} as reported in (46)). The reason for this similarly enhanced fluorescence might reside in shared structural features. In fact, all reported left-handed G4 structures feature an interface of G4 blocks with a '5/6-ring' stacking mode of guanine residues (62,63). Notably, the same stacking mode is observed at the 5′–5′ interface of the dimeric structure formed by a close analogue of *G_3T* (*J19*, PDB: 2LE6, Supplementary Figure S23), and was proposed to be at the origin of the exceptional fluorescence properties of this sequence, according to quantum chemical calculations (42,48). Since the 5′–5′ stacking interface with a '5/6-ring' stacking mode is energetically favorable in G4 structures (48), we may expect that other sequences with this structural feature also display enhanced emission properties.

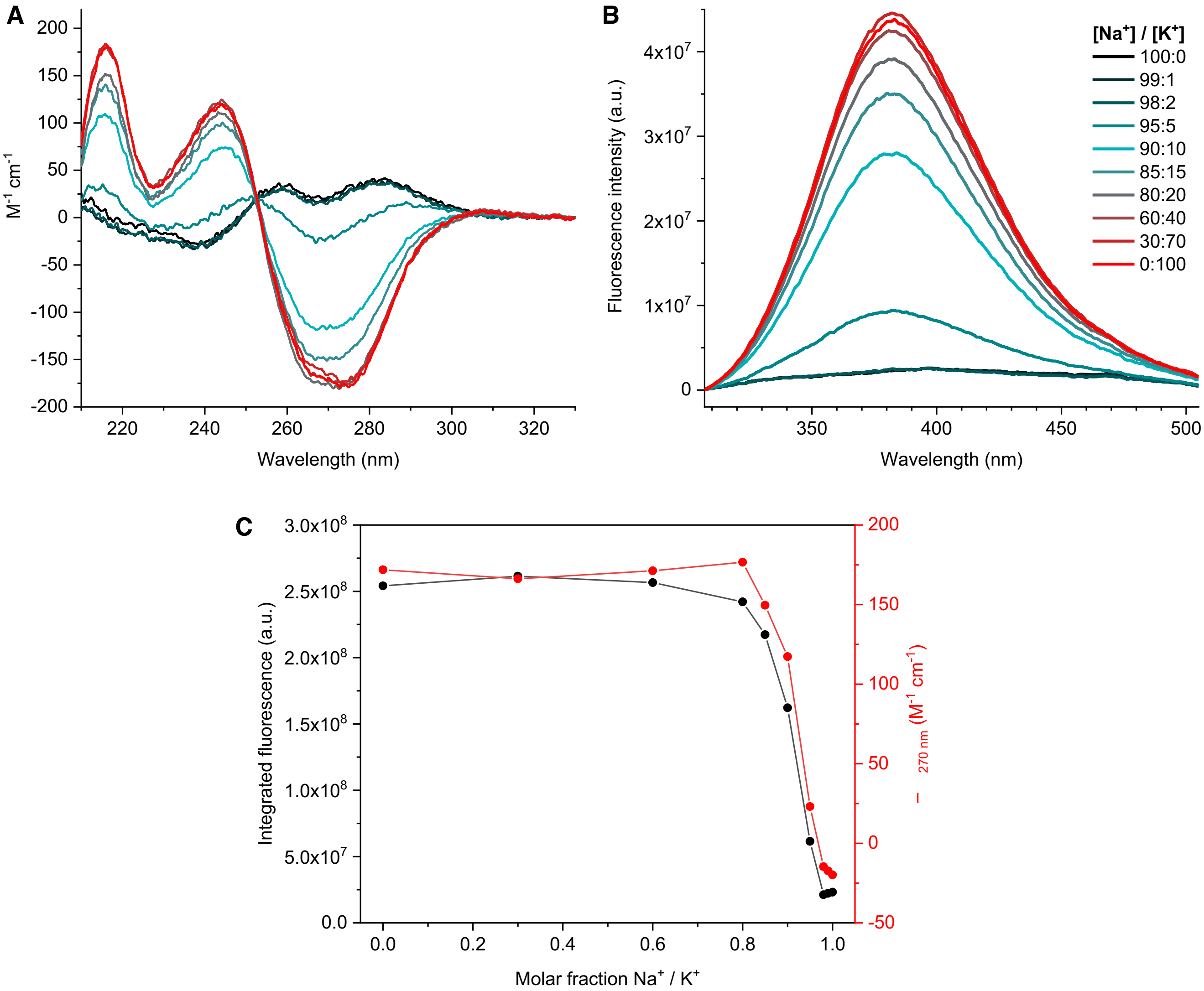

Figure 10. Following the conformational transition of *ZG4* from unfolded state to a left-handed G4 by CD and intrinsic fluorescence. (**A**) CD spectra of *ZG4* solutions ($c = 5.7$ μM) containing a variable proportion of Na^+ and K^+ (total NaCl + KCl concentration of 0.1 M, lithium cacodylate buffer 0.01 M, pH 7.2); (**B**) corresponding emission spectra ($\lambda_{ex} = 260$ nm); (**C**) comparison of the integrated fluorescence intensity ($\lambda_{ex} = 260$ nm, integration between 305 and 505 nm) and molar dichroic absorption at 270 nm (from CD spectra) at various Na^+/K^+ proportions.

DISCUSSION

While the fluorescence of DNA oligonucleotides has been documented since more than a decade, this phenomenon has barely been exploited for structural characterization of secondary structures (49). In this work, we attempted to systematically correlate the secondary structure of oligonucleotides with fluorescence properties using a relatively large training set 89 sequences, whose structures had been unambiguously established by independent methods. First, we confirmed the previous observations demonstrating that steady-state fluorescence spectra are sufficiently sensitive to monitor the conformation transitions of oligonucleotides, such as folding or isomerization of G4 structures (39,41), and extended this observation to the folding of iM structures. These observations corroborate the fact that intrinsic fluorescence can be used as a sensitive method for real-time, unbiased monitoring of conformational changes of oligonucleotides in various settings.

Next, we attempted to identify the characteristic spectral signatures of each conformational group of structures. We observed that intramolecular G4 structures were systematically more fluorescent (on average, 3.5-fold) than other types of structures, in spite of a strong intra-group variance, and display a characteristic shape of fluorescence spectra, featuring a broad, unstructured red edge devoid of additional shoulders (cf. Figure 5). On the other hand, iM structures are characterized by an enhanced fluorescence observed upon 300-nm excitation, which appears in the conditions appropriate for their folding. Further information could be obtained from the multivariate analysis of fluorescence spectra. The data obtained in a single buffer allowed a good discrimination of G4 structures, but were insufficient for the discrimination of iMs, single- and double-stranded

Table 2. Absolute and relative (with respect to G_3T) fluorescence quantum yields of selected oligonucleotides[a]

Sequence	Buffer	$\Phi \times 10^2$	Φ / Φ_{G3T}
22AG	A	0.069	0.19
26TTA	A	0.129	0.35
23TAG	A	0.076	0.20
Bom17	A	0.059	0.16
TBA	A	0.078	0.21
G4CT	A	0.150	0.40
Pu24T	A	0.214	0.57
26Ceb	A	0.040	0.11
KRAS-22RT	A	0.237	0.64
ds26	A	0.028	0.08
hairpin-3	A	0.063	0.17
ss6	A	0.045	0.12
RND3	A	0.039	0.10
EPBC	B	0.137	0.37
i-HRAS2	B	0.026	0.07
ZG4	A	0.195	0.52
Block2Δ	A	0.311	0.83
2xBlock2	A	0.270	0.73
G_3T	A	0.372	1

[a] Conditions: $\lambda_{ex} = 265$ nm, quantum yield standard: quinine sulfate in 0.5 M H_2SO_4 ($\Phi = 0.546$).

structures. However, a combination of the data obtained in two conditions (i.e., favorable and unfavorable for iM folding) allowed a separation of iM-forming sequences as a separate cluster in a PCA plot (cf. Figure 7). Finally, supervised LDA of the reduced dataset (consisting of most informative wavelength readings) allowed a relatively good discrimination (error rate of 9.1%) between the three distinct groups of G4, iM, and other (single- and double-stranded) oligonucleotide conformations (cf. Figure 8). Remarkably, the discrimination between the latter was not possible, neither by naked-eye inspection of the intensity or the shape of their fluorescence spectra (cf. Figures 4 and 5), nor by multivariate analysis of the latter. Notably, CD spectroscopy is not performing better in this case, since the CD spectra of these structural groups are relatively similar; in addition, single-stranded sequences demonstrate a significant intragroup variance in terms of CD spectra (Supplementary Figure S7E, F).

The power of supervised multivariate analysis lies in the use of the data obtained with the training set of substrates for the in-class assignment ('typing') of novel analytes. To test the practical applicability of this approach, we employed four newly generated G-rich sequences and demonstrated that structural predictions based on the LDA of their fluorescence spectra qualitatively agree with the information by two other biophysical techniques, CD and NMR spectroscopy. Furthermore, we assessed several 'problematic' sequences adopting unusual secondary structures not included into the training set. In this case, we observed that, in most cases, multivariate analysis of their fluorescent properties correctly assigned these analytes to the group of 'other' structures, with a notable exception of non-G4 tetraplexes (*VK1*, *VK2*) and G-triplex (*G3x*) structures, 'mis-assigned' to the G4 group. This fact implies that the fluorescence properties of this class of structures resemble those of G4s, most likely due to similar stacking of guanine residues (70), and also calls for a detailed investigation of this phenomenon. Notably, three left-handed G4 structures as well as the stacked G4 dimer G_3T clustered separately but were correctly assigned to the G4 group. This observation implies that the exceptional fluorescent properties of these structures are related to their shared structural feature, that is, the '5/6-ring' stacking of guanines at the 5′–5′ interface of G4 blocks.

Compared with CD spectroscopy, which is a well-established method for assessment of secondary structures of nucleic acids, the discriminatory power and utility of intrinsic fluorescence may seem more limited at a first glance. Indeed, we were not able to discriminate between different topological classes of G4 structures, which can be easily achieved through the multivariate (27,28) or naked-eye analysis of CD spectra (25). Of note, integrated fluorescence intensities showed no correlation with the amplitudes of the characteristic CD bands of G4 structures (data not shown), implying that fluorescence and CD spectroscopy provide independent information on secondary structures of nucleic acids. However, the discrimination between the classes of G4s, iMs and other structures was achieved with an acceptable error rate of 9.1%. We should stress out that discrimination of G4 and iM structures from single-stranded and/or duplex structures is not always trivial with CD spectroscopy (27), and often requires additional methods such as NMR spectroscopy, mass-spectrometry, and/or thermal methods (61). Thus, the intrinsic fluorescence may be exploited as a supplementary method to assess the folding preferences of oligonucleotides in different conditions. Clearly, this technique is limited to *in vitro* conditions devoid of strongly fluorescing components. However, the key advantage of fluorescence-based methods is their great potential of miniaturization and high-throughput screening, which are not accessible with other methods. In particular, recent advances in high-power, deep-ultraviolet light-emitting diodes hold promise for analysis of even small volumes of weakly emitting samples (72). Another advantage of the method presented here is the use of unmodified oligonucleotides, readily available in high-density formats. Thus, the implementation of this method in a DNA microarray format would allow the screening and chemometric analysis of up to hundreds of thousands of sequences in one experiment, paving a way to structural profiling of whole genomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Jean-Louis Mergny for helpful discussions.

FUNDING

French National Research Agency [ANR-17-CE07-0004-01 to A.G.]; Institut Curie (post-doctoral fellowship to M.Z.). Funding for open access charge: Institut Curie.

Conflict of interest statement. None declared

REFERENCES

1. Lightfoot,H.L., Hagen,T., Tatum,N.J. and Hall,J. (2019) The diverse structural landscape of quadruplexes. *FEBS Lett.*, **593**, 2083–2102.
2. Abou Assi,H., Garavís,M., González,C. and Damha,M.J. (2018) i-Motif DNA: structural features and significance to cell biology. *Nucleic Acids Res.*, **46**, 8038–8056.
3. Day,H.A., Pavlou,P. and Waller,Z.A.E. (2014) i-Motif DNA: structure, stability and targeting with ligands. *Bioorg. Med. Chem.*, **22**, 4407–4418.
4. Tian,T., Chen,Y.Q., Wang,S.R. and Zhou,X. (2018) G-Quadruplex: a regulator of gene expression and its chemical targeting. *Chem*, **4**, 1314–1344.
5. Mukherjee,A.K., Sharma,S. and Chowdhury,S. (2019) Non-duplex G-Quadruplex structures emerge as mediators of epigenetic modifications. *Trends Genet.*, **35**, 129–144.
6. Kouzine,F., Wojtowicz,D., Baranello,L., Yamane,A., Nelson,S., Resch,W., Kieffer-Kwon,K.-R., Benham,C.J., Casellas,R., Przytycka,T.M. *et al.* (2017) Permanganate/S1 nuclease footprinting reveals Non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.*, **4**, 344–356.
7. Belotserkovskii,B.P., Mirkin,S.M. and Hanawalt,P.C. (2013) DNA sequences that interfere with transcription: implications for genome function and stability. *Chem. Rev.*, **113**, 8620–8637.
8. Brázda,V., Hároníková,L., Liao,J. and Fojta,M. (2014) DNA and RNA quadruplex-binding proteins. *Int. J. Mol. Sci.*, **15**, 17493–17517.
9. Mendoza,O., Bourdoncle,A., Boulé,J.-B., Brosh,R.M. and Mergny,J.-L. (2016) G-quadruplexes and helicases. *Nucleic Acids Res.*, **44**, 1989–2006.
10. Rigo,R., Palumbo,M. and Sissi,C. (2017) G-quadruplexes in human promoters: a challenge for therapeutic applications. *Biochim. Biophys. Acta - Gen. Subj.*, **1861**, 1399–1413.
11. Maizels,N. (2015) G4-associated human diseases. *EMBO Rep.*, **16**, 910–922.
12. Wu,Y. and Brosh,R.M. (2010) G-quadruplex nucleic acids and human disease. *FEBS J.*, **277**, 3470–3488.
13. Lavezzo,E., Berselli,M., Frasson,I., Perrone,R., Palù,G., Brazzale,A.R., Richter,S.N. and Toppo,S. (2018) G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLoS Comput. Biol.*, **14**, e1006675.
14. Patel,D.J., Phan,A.T. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: Diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429–7455.
15. Webba da Silva,M. (2007) NMR methods for studying quadruplex nucleic acids. *Methods*, **43**, 264–277.
16. Adrian,M., Heddi,B. and Phan,A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, **57**, 11–24.
17. Lin,C., Dickerhoff,J. and Yang,D. (2019) NMR studies of G-quadruplex structures and G-quadruplex-interactive compounds. In: Yang,D and Lin,C (eds). *G-Quadruplex Nucleic Acids. Methods in Molecular Biology, vol. 2035*. Humana, NY, pp. 157–176.
18. Campbell,N.H. and Parkinson,G.N. (2007) Crystallographic studies of quadruplex nucleic acids. *Methods*, **43**, 252–263.
19. Campbell,N., Collie,G.W. and Neidle,S. (2012) Crystallography of DNA and RNA G-quadruplex nucleic acids and their ligand complexes. *Curr. Protoc. Nucleic Acid Chem.*, **50**, 17.6.1–17.6.22.
20. Parkinson,G.N. and Collie,G.W. (2019) X-ray crystallographic studies of G-quadruplex structures. In: Yang,D and Lin,C (eds). *G-Quadruplex Nucleic Acids. Methods in Molecular Biology, vol. 2035*. Humana, NY, pp. 131–155.
21. Do,N.Q. and Phan,A.T. (2012) Monomer-dimer equilibrium for the 5′-5′ stacking of propeller-type parallel-stranded G-quadruplexes: NMR structural study. *Chem. Eur. J.*, **18**, 14752–14759.
22. Kogut,M., Kleist,C. and Czub,J. (2019) Why do G-quadruplexes dimerize through the 5′-ends? Driving forces for G4 DNA dimerization examined in atomic detail. *PLoS Comput. Biol.*, **15**, e1007383.
23. Dickerson,R.E., Goodsell,D.S. and Neidle,S. (1994) '…the tyranny of the lattice…'. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 3579–3583.
24. Li,J., Correia,J.J., Wang,L., Trent,J.O. and Chaires,J.B. (2005) Not so crystal clear: The structure of the human telomere G-quadruplex in solution differs from that present in a crystal. *Nucleic Acids Res.*, **33**, 4649–4659.
25. Karsisiotis,A.I., Hessari,N.M., Novellino,E., Spada,G.P., Randazzo,A. and Webba da Silva,M. (2011) Topological characterization of nucleic acid G-Quadruplexes by UV absorption and circular dichroism. *Angew. Chem. Int. Ed.*, **50**, 10645–10648.
26. Randazzo,A., Spada,G.P. and da Silva,M.W. (2012) Circular dichroism of quadruplex structures. *Top. Curr. Chem.*, **330**, 67–86.
27. Jaumot,J., Eritja,R., Navea,S. and Gargallo,R. (2009) Classification of nucleic acids structures by means of the chemometric analysis of circular dichroism spectra. *Anal. Chim. Acta*, **642**, 117–126.
28. del Villar-Guerra,R., Trent,J.O. and Chaires,J.B. (2018) G-Quadruplex secondary structure obtained from circular dichroism spectroscopy. *Angew. Chem. Int. Ed.*, **57**, 7171–7175.
29. Zuffo,M., Xie,X. and Granzhan,A. (2019) Strength in numbers: development of a fluorescence sensor array for secondary structures of DNA. *Chem. Eur. J.*, **25**, 1812–1818.
30. Nicoludis,J.M., Barrett,S.P., Mergny,J.-L. and Yatsunyk,L.A. (2012) Interaction of human telomeric DNA with N-methyl mesoporphyrin IX. *Nucleic Acids Res.*, **40**, 5432–5447.
31. Xie,X., Reznichenko,O., Chaput,L., Martin,P., Teulade-Fichou,M.-P. and Granzhan,A. (2018) Topology-selective, fluorescent "Light-Up" probes for G-quadruplex DNA based on photoinduced electron transfer. *Chem. Eur. J.*, **24**, 12638–12651.
32. Tanpure,A.A. and Srivatsan,S.G. (2015) Conformation-sensitive nucleoside analogues as topology-specific fluorescence turn-on probes for DNA and RNA G-quadruplexes. *Nucleic Acids Res.*, **43**, e149.
33. Manna,S., Sarkar,D. and Srivatsan,S.G. (2018) A Dual-App nucleoside probe provides structural insights into the human telomeric overhang in live cells. *J. Am. Chem. Soc.*, **140**, 12622–12633.
34. Sabale,P.M., Tanpure,A.A. and Srivatsan,S.G. (2018) Probing the competition between duplex and G-quadruplex/i-motif structures using a conformation-sensitive fluorescent nucleoside probe. *Org. Biomol. Chem.*, **16**, 4141–4150.
35. Manna,S., Panse,C.H., Sontakke,V.A., Sangamesh,S. and Srivatsan,S.G. (2017) Probing human telomeric DNA and RNA topology and ligand binding in a cellular model by using responsive fluorescent nucleoside probes. *ChemBioChem*, **18**, 1604–1615.
36. Raguseo,F., Chowdhury,S., Minard,A. and Di Antonio,M. (2020) Chemical-biology approaches to probe DNA and RNA G-quadruplex structures in the genome. *Chem. Commun.*, **56**, 1317–1324.
37. Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.*, **20**, 295–304.
38. Bevilacqua,P.C., Ritchey,L.E., Su,Z. and Assmann,S.M. (2016) Genome-wide analysis of RNA secondary structure. *Annu. Rev. Genet.*, **50**, 235–266.
39. Mendez,M.A. and Szalai,V.A. (2009) Fluorescence of unmodified oligonucleotides: a tool to probe G-quadruplex DNA structure. *Biopolymers*, **91**, 841–850.
40. Miannay,F.A., Banyasz,A., Gustavsson,T. and Markovitsi,D. (2009) Excited states and energy transfer in G-quadruplexes. *J. Phys. Chem. C*, **113**, 11760–11765.
41. Dao,N.T., Haselsberger,R., Michel-Beyerle,M.E. and Phan,A.T. (2011) Following G-quadruplex formation by its intrinsic fluorescence. *FEBS Lett.*, **585**, 3969–3977.
42. Dao,N.T., Haselsberger,R., Michel-Beyerle,M.E. and Phan,A.T. (2013) Excimer formation by stacking G-quadruplex blocks. *ChemPhysChem*, **14**, 2667–2671.
43. Kwok,C.K., Sherlock,M.E. and Bevilacqua,P.C. (2013) Effect of loop sequence and loop length on the intrinsic fluorescence of G-Quadruplexes. *Biochemistry*, **52**, 3019–3021.
44. Improta,R. (2014) Quantum mechanical calculations unveil the structure and properties of the absorbing and emitting excited electronic states of guanine quadruplex. *Chem. Eur. J.*, **20**, 8106–8115.
45. Martinez-Fernandez,L., Changenet,P., Banyasz,A., Gustavsson,T., Markovitsi,D. and Improta,R. (2019) Comprehensive study of guanine excited state relaxation and photoreactivity in G-quadruplexes. *J. Phys. Chem. Lett.*, **10**, 6873–6877.
46. Sherlock,M.E., Rumble,C.A., Kwok,C.K., Breffke,J., Maroncelli,M. and Bevilacqua,P.C. (2016) Steady-state and time-resolved studies into the origin of the intrinsic fluorescence of G-quadruplexes. *J. Phys. Chem. B*, **120**, 5146–5158.
47. Chan,C.Y., Umar,M.I. and Kwok,C.K. (2019) Spectroscopic analysis reveals the effect of a single nucleotide bulge on G-quadruplex structures. *Chem. Commun.*, **55**, 2616–2619.

48. Lech,C.J., Phan,A.T., Michel-Beyerle,M.E. and Voityuk,A.A. (2015) Influence of base stacking geometry on the nature of excited states in G-quadruplexes: A time-dependent DFT study. *J. Phys. Chem. B*, **119**, 3697–3705.
49. Gao,S., Cao,Y., Yan,Y., Xiang,X. and Guo,X. (2016) Correlations between fluorescence emission and base stacks of nucleic acid G-quadruplexes. *RSC Adv.*, **6**, 94531–94538.
50. Majerová,T., Streckerová,T., Bednárová,L. and Curtis,E.A. (2018) Sequence requirements of intrinsically fluorescent G-quadruplexes. *Biochemistry*, **57**, 4052–4062.
51. Ma,C., Chan,R.C.T., Chan,C.T.L., Wong,A.K.W., Chung,B.P.Y. and Kwok,W.M. (2018) Fluorescence and ultrafast fluorescence unveil the formation, folding molecularity, and excitation dynamics of homo-oligomeric and human telomeric i-motifs at acidic and neutral pH. *Chem. Asian J.*, **13**, 3706–3717.
52. Larsson,T., Wedborg,M. and Turner,D. (2007) Correction of inner-filter effect in fluorescence excitation-emission matrix spectrometry using Raman scatter. *Anal. Chim. Acta*, **583**, 357–363.
53. Eilers,P.H.C. and Kroonenberg,P.M. (2014) Modeling and correction of Raman and Rayleigh scatter in fluorescence landscapes. *Chemom. Intell. Lab. Syst.*, **130**, 1–5.
54. Brouwer,A.M. (2011) Standards for photoluminescence quantum yield measurements in solution (IUPAC Technical Report). *Pure Appl. Chem.*, **83**, 2213–2228.
55. Wang,Y. and Patel,D.J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*, **1**, 263–282.
56. Phan,A.T. (2010) Human telomeric G-quadruplex: structures of DNA and RNA sequences. *FEBS J.*, **277**, 1107–1117.
57. Largy,E., Marchand,A., Amrane,S., Gabelica,V. and Mergny,J.-L. (2016) Quadruplex turncoats: cation-dependent folding and stability of quadruplex-DNA double switches. *J. Am. Chem. Soc.*, **138**, 2780–2792.
58. Zhang,M.-L., Xu,Y.-P., Kumar,A., Zhang,Y. and Wu,W.-Q. (2019) Studying the potassium-induced G-quadruplex DNA folding process using microscale thermophoresis. *Biochemistry*, **58**, 3955–3959.
59. Trajkovski,M., Webba da Silva,M. and Plavec,J. (2012) Unique structural features of interconverting monomeric and dimeric G-Quadruplexes adopted by a sequence from the intron of the N-myc gene. *J. Am. Chem. Soc.*, **134**, 4132–4141.
60. Phan,A.T., Kuryavyi,V., Ma,J.-B., Faure,A., Andreola,M.-L. and Patel,D.J. (2005) An interlocked dimeric parallel-stranded DNA quadruplex: a potent inhibitor of HIV-1 integrase. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 634–639.
61. Mergny,J.-L., Li,J., Lacroix,L., Amrane,S. and Chaires,J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
62. Chung,W.J., Heddi,B., Schmitt,E., Lim,K.W., Mechulam,Y. and Phan,A.T. (2015) Structure of a left-handed DNA G-quadruplex. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 2729–2733.
63. Bakalar,B., Heddi,B., Schmitt,E., Mechulam,Y. and Phan,A.T. (2019) A minimal sequence for left-handed G-quadruplex formation. *Angew. Chem. Int. Ed.*, **58**, 2331–2335.
64. Kocman,V. and Plavec,J. (2014) A tetrahelical DNA fold adopted by tandem repeats of alternating GGG and GCG tracts. *Nat. Commun.*, **5**, 5831.
65. Kocman,V. and Plavec,J. (2017) Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. *Nat. Commun.*, **8**, 15355.
66. Trantírek,L., Štefl,R., Vorlíčková,M., Koča,J., Sklenářář,V. and Kypr,J. (2000) An A-type double helix of DNA having B-type puckering of the deoxyribose rings. *J. Mol. Biol.*, **297**, 907–922.
67. Štefl,R., Trantírek,L., Vorlíčková,M., Koča,J., Sklenář,V. and Kypr,J. (2001) A-like guanine-guanine stacking in the aqueous DNA duplex of d(GGGGCCCC). *J. Mol. Biol.*, **307**, 513–524.
68. Gajarský,M., Živković,M.L., Stadlbauer,P., Pagano,B., Fiala,R., Amato,J., Tomáška,L., Šponer,J., Plavec,J. and Trantírek,L. (2017) Structure of a stable G-Hairpin. *J. Am. Chem. Soc.*, **139**, 3591–3594.
69. Limongelli,V., De Tito,S., Cerofolini,L., Fragai,M., Pagano,B., Trotta,R., Cosconati,S., Marinelli,L., Novellino,E., Bertini,I. *et al.* (2013) The G-triplex DNA. *Angew. Chem. Int. Ed.*, **52**, 2269–2273.
70. Cerofolini,L., Amato,J., Giachetti,A., Limongelli,V., Novellino,E., Parrinello,M., Fragai,M., Randazzo,A. and Luchinat,C. (2014) G-triplex structure and formation propensity. *Nucleic Acids Res.*, **42**, 13393–13404.
71. Xie,X., Renvoisé,A., Granzhan,A. and Teulade-Fichou,M.-P. (2015) Aggregating distyrylpyridinium dye as a bimodal structural probe for G-quadruplex DNA. *New J. Chem.*, **39**, 5931–5935.
72. Kneissl,M., Seong,T.-Y., Han,J. and Amano,H. (2019) The emergence and prospects of deep-ultraviolet light-emitting diode technologies. *Nat. Photonics.*, **13**, 233–244.