Machine Learning Aids Classification and Discrimination of Non-canonical DNA Folding Motifs by an Arrayed Host:guest Sensing System

Junyi Chen,[2] Adam D. Gill,[3] Briana L. Hickey,[1] Ziting Gao,[1] Xinping Cui,[4] Richard J. Hooley[1,3]* and Wenwan Zhong[1, 2]*

[1]Department of Chemistry; [2]Environmental Toxicology Graduate Program; [3]Department of Biochemistry; [4]Department of Statistics; University of California-Riverside, Riverside, CA 92521, U.S.A.

Supporting Information Placeholder

ABSTRACT: An arrayed host:guest fluorescence sensor system can discriminate and classify multiple different non-canonical DNA structures by exploiting selective molecular recognition. The sensor is highly selective, and can discriminate between folds as similar as native G-quadruplexes and those with bulges or vacancies. The host and guest can form heteroternary complexes with DNA strands, with the host acting as mediator between the DNA and dye, modulating the emission. By applying machine learning algorithms to the sensing data, prediction of the folding state of unknown DNA strands is possible with high fidelity.

INTRODUCTION

DNA strands can adopt a number of different secondary structures other than the classical double helix.[1] These non-canonical folding motifs influence DNA replication, gene transcription, and genome stability,[2] so are involved in diseases such as cancers, neurodegenerative diseases, and genetic disorders.[3] Examples of non-canonical folds include G-quadruplexes (G4s),[4] Hoogsteen triplexes,[5] hairpins and i-motifs,[6] among others.[7] While some of these motifs are quite structurally different from each other, each broad type of non-canonical fold has a variety of substructures. For example, G4s can exist in parallel, antiparallel or hybrid orientations (referring to the orientation of the phosphate backbone around the G4 stacks), as well as incorporating different numbers of G-quartets.[4] They can also incorporate bulges,[8] i.e. interruption of the consecutive guanine stretches by at least one non-guanine base, or vacancies,[9] where one of the G quartets has a missing G. Triplex DNA can display different orientations of the third strand, and are termed parallel or antiparallel.[10]

Understanding the formation and control mechanisms of non-canonical nucleic acid folding can help better interpret their biological roles and guide design of therapeutics targeting these structures.[11] However, the large number of non-canonical nucleic acid folded structures, some of which are highly similar, makes their identification challenging. [12] In addition, the structures can be transient, and controlled by various external factors such as oligonucleotide sequence, ion type and concentration, pH, or external effectors such as ligands or proteins,[13] which further complicates identification and mechanistic analysis. Complete structural analysis requires X-ray crystallography[14] and/or multidimensional NMR spectroscopy,[15] which, while powerful, require large amounts of sample and are too time-consuming for rapid analysis. Simple grouping into secondary structural types is possible with Circular Dichroism (CD) spectroscopy,[16] but this is not capable of differentiating small differences in structure.

Optical methods are potentially a simple, yet powerful method of detecting and analyzing non-canonical nucleotide structures. While there are examples of dyes and probes that can selectively target G4 structures,[3a] other motifs such as triplexes and i-motifs are much less easily detected, [17] and single fluorescent markers are rarely capable of distinguishing between substructures of a folding motif. Pattern recognition-based differential sensing[18] can be a powerful tool for creating fluorescent probes that selectively recognize and differentiate DNA folding. This has been used to identify folding patterns in fluorescently labeled RNAs,[19] and fluorescence displacement assays paired with multivariate analysis allow classification of DNA structure,[20] or identify ligands that can bind these structures.[21]

We recently described a host:guest fluorescence sensing system that was capable of sensing, discriminating and classifying different G4 types.[22] This technique does not require high selectivity of individual dyes for specific DNA folded structures, but rather relies on differential binding of multiple components. While pattern recognition-based sensing is extremely powerful, it can create large pools of data when used in complex systems, and this requires detailed statistical analysis.[23] To maximize the information gained from sensing arrays, machine learning can be employed,[24] which allows analysis of large datasets and prediction of unknown outcomes. Machine learning has been widely used in biomedical research,[25] including bioinformatics and drug discovery, and has more recently been used to solve chemical problems, such as reaction outcomes and mechanisms. [26] Machine learning is especially powerful for pattern recognition sensing, because it can detect hidden patterns in large, noisy or complex data set and prediction of unknown

groups is possible via data set training.[27] Here, we describe the use of a multicomponent host:guest sensing array to discriminate and classify a wide variety of different folded DNA structures, and apply a machine learning algorithm to optimize the array components and predict the conformation of a set of unknown DNA strands.

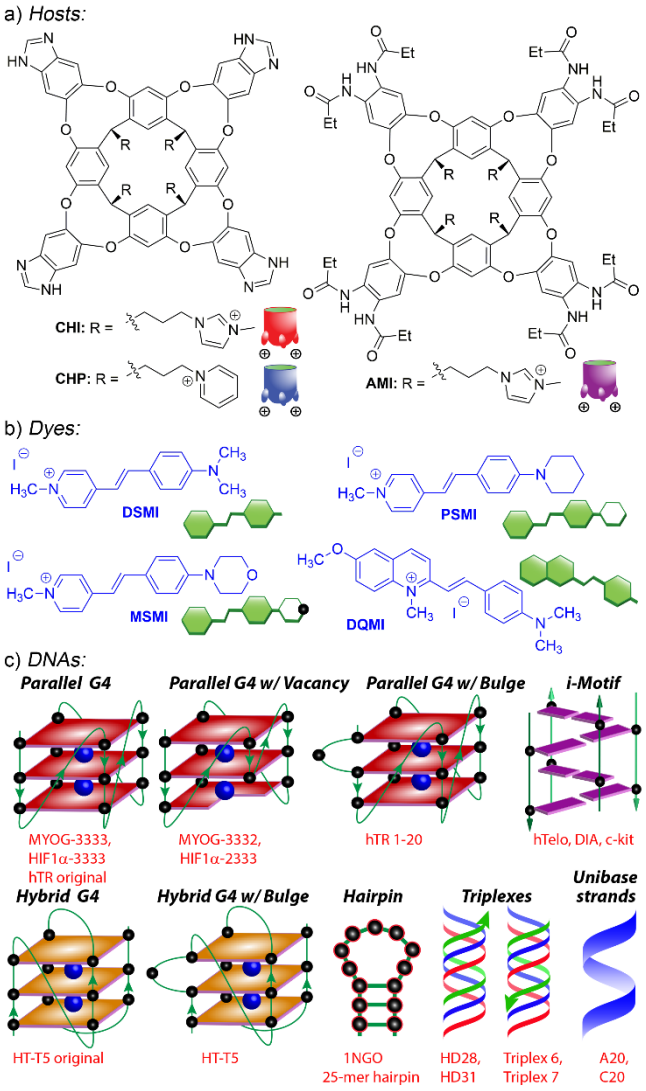


Figure 1. Host:Guest fluorescence sensing array for nucleotide structural discrimination. Structures of a) hosts and b) dyes in the screen; c) Pool of 19 DNA elements tested, of 10 different folding types. See Supporting Information for sequences.

RESULTS AND DISCUSSION

The components of the arrayed host:guest sensor and the DNA targets are shown in Figure 1. We targeted 10 types of DNA secondary structures (totaling 19 different strands, with lengths from 17 to 31 nt, Figure 1c). These targets range from entirely different folding motif structures, such as triplexes, to those with very small differences, such as bulges or vacancies in G4 structure. The intactness of each of the DNA strands and their folded structures were confirmed by gel electrophoresis and CD, respectively (see Supporting Information). The sensing array consists of a series of cationic, water-soluble deep cavitand hosts and a set of styrylpyridinium dyes that can variably bind both the DNA target and the hosts, with concomitant effects on the dye emission. In our previous work sensing G4s,[22] we used two dyes (**DSMI** and **PSMI**, Figure 1) with 5 cavitand hosts. This array functions at neutral pH and was optimized for G4 structures. For the larger set of targets described here, a wider range of sensor components that can function at lower pH is required. Three cationic cavitands **CHI**, **CHP** and **AMI** (Figure 1a) were used as host array components, and four dyes that showed large differences in fluorescence emission while mixed with the hosts or DNA during a brief initial screening were tested: **DSMI**, **PSMI**, the morpholine variant **MSMI** and quinoline dye **DQMI** (Figure 1b). Each of these dyes was synthesized from the corresponding aldehyde and methylated pyridinium salt.[28] As the i-motif and triplex motifs are only persistent in solution at low pH (~4-6),[29] we analyzed the interaction between the four dyes and the three cavitands at pH 5.5, using 20 mM KOAc buffer in the presence of 5 mM $MgCl_2$. While guest binding in deep cavitands is quite sensitive to pH,[30] each of the four dyes bound to the three cavitands under these conditions, and showed fluorescence increase upon binding (see Supporting Information). Similarly, the fluorescence of the four dyes increased with the addition of a representative folded DNA structure. From these spectra, the optimal λ_{ex} and λ_{em} for each dye were determined. The dye concentrations used were those displaying maximal F/F_0 upon titration into a DNA target at 0.1 μM, and optimal host concentration was determined by titration to the dye-DNA mixture (see Supporting Information). The F and F_0 values are defined in this case as dye fluorescence with or without the DNA target. From this, a 16-element array was created, consisting of the various combinations of the four dyes and three cavitands, as well as the dyes by themselves.

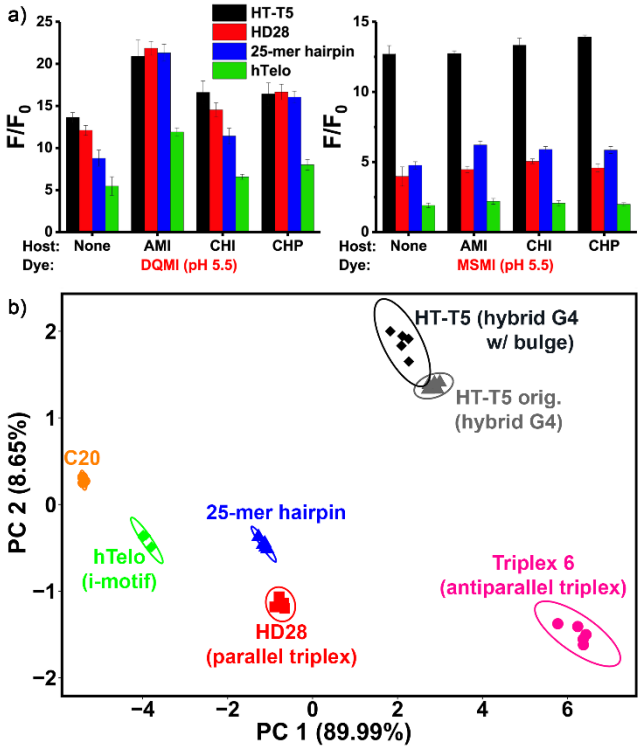


Figure 2. Selective array-based sensing of variable DNA structures. a) Selected fluorescence responses of upon addition of the four DNA strands to the host•dye components, (F_0:

emission at [DNA] = 0). b) PCA scores plot generated from the data using 7 DNAs and 16 array elements: **DSMI/PSMI/MSMI/DQMI** with **CHI/CHP/AMI/**No cavitand. [Dye] = 0.156 μM; [Host] = 0.125 μM; [DNA] = 0.1 μM; 20 mM KOAc, 5 mM $MgCl_2$, pH 5.5.

The initial tests were performed on a smaller, 7 DNA subset using the 16-element array. The seven DNA strands were HT-T5, HT-T5 original, HD28, Triplex 6, 25-mer hairpin, hTelo and the unibase oligonucleotide C20. The DNAs all exhibit varied fluorescence responses to the sensors, with **DQMI** and **DSMI** showing a wider range of F/F_0 (from ~ 5 to > 40) than **PSMI** and **MSMI**. A subset of the fluorescence emission bar plots is shown in Figure 2a (see Supporting Information Figure S-13 for full fluorescence response plots), and they illustrate the differential sensing nicely. The changes in emission are dependent on both dye structure and cavitand type. The variances in emission are complex, and not easily explained, but some notable trends can be seen. The dyes themselves show some selectivity for different folds, but not enough for robust discrimination in the absence of host. The greatest increases in emission for the various dyes is seen with Hoogsteen triplexes (e.g. Triplex 6, Fig. S13) and G4s (e.g. HT-T5 original), and the lowest response changes are seen with the unibase oligonucleotide C20. The effect of the three different cavitands was greatest when paired with **DSMI** and **DQMI**, and the morpholinyl dye **MSMI** appeared least affected by cavitand. The imidazole-footed hosts **AMI** and **CHI** also showed greater variability in emission signature than the pyridyl-footed **CHP**. However, these observations are merely qualitative, so we performed more detailed analysis of the fluorescence responses using Principal Component Analysis (PCA).[31] Subjecting the fluorescence profiles to PCA (Figure 2b) showed high reproducibility in DNA structure detection, and clear separation of most of the seven DNA tested. The least separation was observed between the two strands that are closest in structure, HT-T5 and HT-T5 original. The only difference in structure between these two hybrid G4s is the presence of one thymine residue, which forms a "bulge" in the middle of the G4 stacks of HT-T5: the other sequence elements are completely conserved. Even so, the array is capable of distinguishing these highly similar structures with only minimal overlap.

The array was next applied to the full 18-element DNA pool (the DNAs shown in Figure 1c, not including C20, which gave low fluorescence responses to the dyes, so wasn't a useful control test). Ten repeated measurements were conducted for each DNA strand, and the F/F_0 values were subjected to multivariate analysis. First, PCA was used to confirm the classification ability of the array, and this PCA plot is shown in Figure 3. As expected, the array shows excellent discrimination and classification of G4s. All parallel G4s are well-separated from hybrid G4s, and differentiation of different structures within the same folding type was possible. As was seen in Figure 2b with HT-T5 and HT-T5 original, the highly structurally similar hTR 1-20 original (a parallel G4) and hTR 1-20 (parallel G4 with a bulge) are closely located on the scores plot. Again, the structural differences are minimal: in this case, a C base interrupts one of the G_3 sequences, but otherwise the sequences are identical. The array was far more capable of distinguishing parallel G4s from their counterparts that show a *vacancy*, i.e. HIF1α-3333 vs HIF1α-2333 and MYOG-3333 vs MYOG-3332. This is remarkable, as the differences are only a single base: one G is missing in each vacancy G4, but otherwise the sequences are identical.

Other types of fold are easily distinguished from the various G4s, but the intra-class differentiation is more variable. The two hairpins (1NGO and 25-mer hairpin) are fully differentiated from each other and the other folding motifs, but the three i-motifs (DIA, c-kit and hTelo) are closely grouped. *Classification* is excellent - all i-motif strands show highly similar responses, but the differentiation between the three is minimal. This might be expected based on sequence, as all three i-motifs are 21-22 nt long, and vary only in the spacer bases between the C_3 regions (see Table S-1). Still, smaller changes in sequence between G4 structures are discriminated. Selective classification of triplexes was also successful: the antiparallel and parallel triplexes are highly separated on the scores plot, and Triplex 6 is well-distinguished from Triplex 7, despite their highly similar structures.

The PCA scores plots are a useful illustration of the sensing power of the array: they show that strands with even small structural differences can be distinguished from each other. However, when the pool of data becomes large, it is not obvious how to determine specific regions where the individual motif types reside, as those regions intersect. This is where machine learning algorithms can be applied: by training the algorithm, precise boundaries can be determined, and unknown structures can be predicted with a greater level of confidence. To achieve this, the array data was treated with SVM-RFE (Figure 4), using the sklearn library in Python 3.9. SVM (support vector machine)[32] is a supervised machine learning algorithm, in which a hyperplane in the form of linear functions is used to separate different classes.[33] SVM-RFE (recursive feature elimination) can select the informative features for sample classification among all those used to generate the database, after recursively removing the non-important features based on their importance ranking. SVM-RFE is fast and is not prone to overfitting.[34]

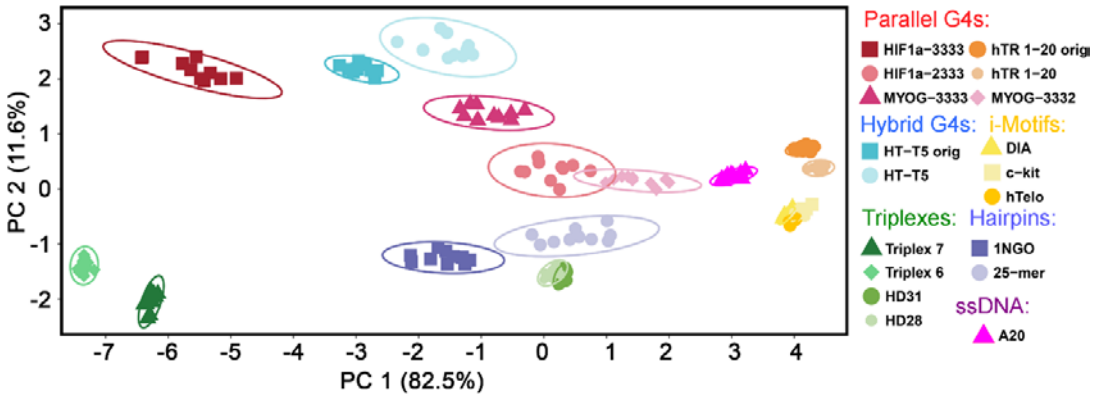


Figure 3. PCA scores plot generated from analysis of the 18-DNA pool using the 16-element array. Array elements: **DSMI/PSMI/MSMI/DQMI** with **CHI/CHP/AMI/**No cavitand. [Dye] = 0.156 μM; [Host] = 0.125 μM; [DNA] = 0.1 μM; 20 mM KOAc, 5 mM $MgCl_2$, pH 5.5.

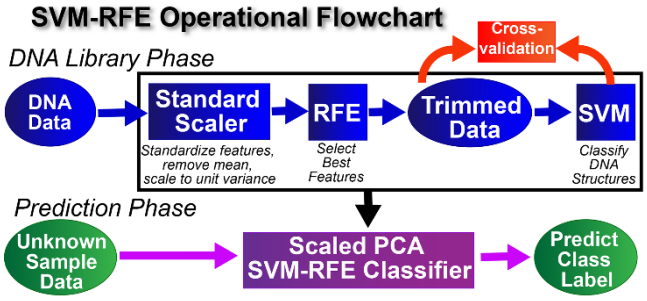


Figure 4. Operational flowchart of the SVM-based machine learning approach for DNA folding classification and prediction.

The analysis procedure is illustrated schematically in Figure 4: StandardScaler was initially applied for standardization of the F/F_0 data. To minimize data dimensions so that the folding classification can be visualized in a 2D plot, a PCA step was added to convert the scaled data into principal component (PC) values while retaining most of the information. Then, SVM-RFE determined the best two PC values for sample classification (PC 1 and PC 2 for this dataset), and determined a set of classifiers to build the model for folding classification. An SVM Decision Region Boundary plot was made using PC 1 and PC 2, with each region colored differently and dedicated to one DNA folding class (Figure 5). For example, each repeat of MYOG-3333, HIF1α-3333 and hTR 1-20 original were counted as a parallel G4, and located within the region colored in dark red, which was defined as the parallel G4 folding region. While the grouping effects for individual DNAs were comparable to that observed in Figure 3, the Decision Region Boundary plot clearly shows the regions where DNAs with the same folding motif can be found. The classification performance is excellent: the average ("macro") scores of accuracy, sensitivity, specificity, precision, and AUC from 3 repeated 8-fold cross validation tests were all > 0.96 (Table S-2, S-3).

The true potential of the SVM-RFE process is in structure *prediction*. While simple PCA scores plots can show effective discrimination between different folds, and can provide a qualitative grouping effect, it is not well-suited for assigning an unknown target into a specific group. As such, we can easily determine that two targets are different from each other, but accurately determining the structural motif of an unknown DNA target from its PCA placement is beyond the scope of the method. However, by training the SVM-RFE algorithm with the data from the known DNA pool, a classification model can be obtained that permits the use of the fluorescence responses from an "unknown" DNA to predict its folding motif. In our case, the 18-DNA dataset (180 samples data in total) can be viewed as the training dataset, and the classification model can then be used to predict the folded structure of new sequences, using the classifiers obtained from the training set (illustrated in Figure 4).

Four DNA strands with known folding motif were chosen as "unknown" targets to test the predictive abilities of the algorithm: c-myc 2345 and EAD4 (known to form a parallel G4 structure), APE 1-4 track (an i-motif), and Telo24 (a hybrid G4). The correct placement of these known strands will illustrate the accuracy of the prediction. Finally, we also tested unibase ssDNA G20, to ask a more complex question of the array: how does it handle complex DNAs that can adopt multiple different folded states? This "disordered" DNA is more complex than the other unibase equivalents (A20, etc.), because it can occupy multiple interconverting conformations in solution, including multiple G4 folds, and similar polyG_x strands have been reported to fold into parallel structures with guanine bulges, dependent on conditions.[35]

The five newly selected DNAs were exposed to the 16-element array as before, and the F/F_0 values acquired. These signals were exposed to the classifier resulting from running Scaled 2D PCA-coupled SVM-RFE on the training dataset to predict the folding, and the prediction results are shown as solid blocks in Figure 5. This clearly shows that all of the 10 repeats for c-myc 2345 were successfully projected into the "correct" folding region, i.e. parallel G4, despite the similarity in structures in the pool, including parallel vs hybrid G4s, and parallel G4s with either bulges or vacancies that differ in only one base in the sequence. Similarly, all of the repeats for APE 1-4, Telo24 and EAD4 were accurately predicted as i-motif, hybrid G4 and parallel G4, respectively.

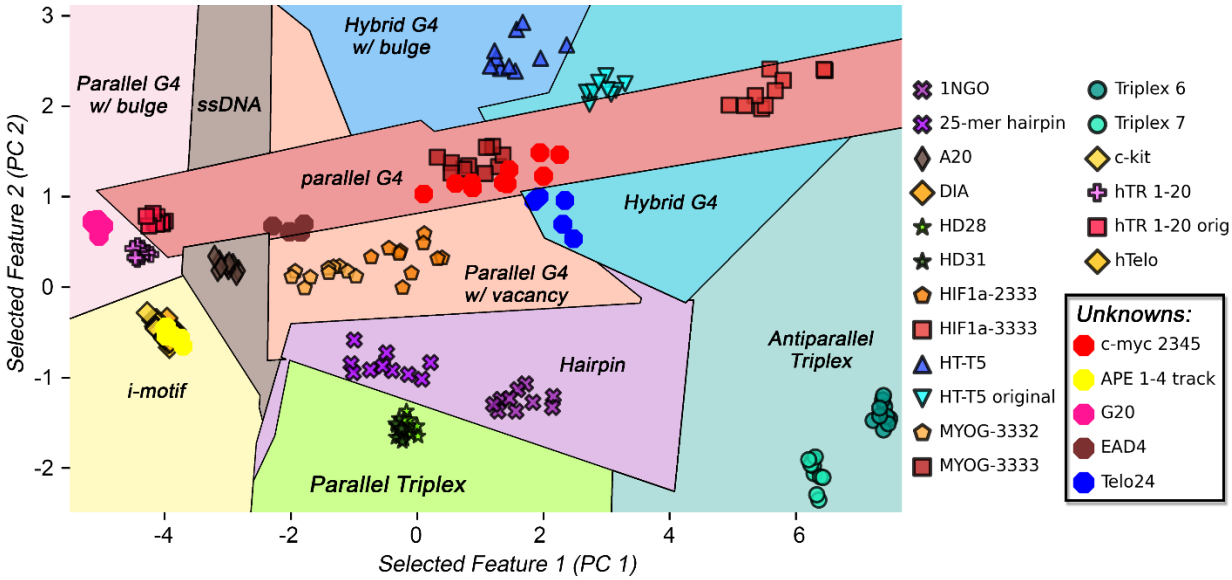


Figure 5. Decision Region Boundary plot using PC 1 and PC 2 obtained from subjecting the 16-element array data acquired from the 18-DNA pool by PCA-SVM-RFE. Five unknowns were projected to the regions representing the predicted folding structures.

Interestingly, the complex target G20, was placed in the parallel G4 with bulge region by the predictor. While this structure can display multiple G4 stacks in solution, gel electrophoresis (Figure S-2) also shows multiple bands with higher molecular weights than the monomeric strand, indicating the formation of intermolecular structures, as opposed to the truly unstructured single A20 strand. Despite this, it has been reported that G20 folds into stacks with hanging guanines (i.e. bulges).[35] Simply describing G20 as a G4 with a bulge is not truly "correct", but it is notable that the sensor and prediction module can distinguish between transiently folded structures (i.e. G20) and other unfolded unibase DNAs, and identify the presence of G4 motifs even when multiple states are present in solution.

As well as classification and prediction, the SVM-RFE process can be applied to determine the most important array elements by removing those that are dependent and linearly correlated. This can enable future array optimization and minimalization. For small datasets, the common approach is to manually subtract certain elements and re-run PCA to determine how well the classification performance is retained. For example, to determine whether the full 16-element array is necessary, the 7 DNA subset from Figure 2b was re-analyzed using fewer array components by manually removing one or more of the array elements and repeating the PCA. This iterative process showed that using only 4 of the 16 elements (four dyes + **CHI**) was sufficient to achieve a visually similar differentiation (Figure S-14). While this manual process did allow array optimization for the 7 DNA subset, it is subjective, labor-intensive and poorly suited to analyzing larger datasets, i.e. the full 18-DNA pool.

To explore the ability of machine learning in identifying the most important features in our array for DNA folding classification, we directly applied SVM-RFE to the 18-DNA dataset without PCA. By running the SVM-RFE cross-validation algorithm in Python, 7 features were chosen to be most important in determining folding classification, i.e. the subset of 7 features achieving the cross-validation score > 0.99: **DQMI** + **AMI** or **CHP**, **MSMI** + **CHI** or **CHP**, **DSMI** + **CHP**, and **PSMI** + **AMI** or alone. Compared to the minimal array needed to differentiate the smaller 7-DNA pool (i.e. the four dyes + **CHI** host), more hosts are required to clearly classify all 10 different folded structures in this much larger pool of 18 DNA. This illustrates the importance of the combination of all four dyes and three cavitands for completely successful folding classification. To test the efficacy of this minimized array, we subjected the array data collected from these 7 selected features to scaled PCA-SVM-RFE. The resultant Decision Region Boundary plot indeed showed a similarly effective classification effect as that shown in Figure 5 (Figure S-16b), albeit with less distinct separation of the hybrid and parallel G4. This is understandable, because we are trying to use only 7 sensor elements to distinguish 18 DNA strands. This suggests that a minimal effective array for classifying large DNA datasets should contain a sufficient number of elements that function orthogonally.

The sensing array requires both hosts and dyes for discrimination, and the SVM-RFE process can shed light on the most effective combinations, but the specific sensing mechanism is still not completely clear. The dyes are all of the correct size/shape and charge to interact with folded DNA, and indeed, the quinolinium dye **DQMI** is a turn-on sensor for DNA: whereas the pyridinium dyes all show strong emission that is enhanced by DNA, **DQMI** is a poor fluorophore in the absence of DNA, but turns on significantly when DNA is added. The cationic hosts provide a "second layer" of discrimination: we had previously postulated that dyes and cavitands form heteroternary complexes with the G4 DNAs,[22] and the sensing data shown here further supports that concept. Further evidence for the effect of the hosts on the dye-DNA system was gained by evaluating the UV/Vis absorbance behavior of two dyes (**PSMI** and **DQMI**) with the three hosts and two DNA targets, HT-T5 (a folded, hybrid G4 with bulge) and unibase DNA A20 (a representative unfolded structure). The absorbance of both **DQMI** and **PSMI** was minimally affected by A20, even in the presence of cavitand, which mirrors the small changes in fluorescence seen with unfolded structures. The folded HT-T5 was far more enlightening, however. Two sets of plots are shown in Figure 6 (with **PSMI**, see Supporting Information

for full spectral data, including that with **DQMI**), which illustrate the synergistic effects of *both* cavitand and DNA on the dye absorbance. While increasing DNA concentrations added to **PSMI** did not affect the UV absorption spectra (Figure 6b), significance differences were observed in the presence of the host **AMI** (Figure 6a). While the **AMI•PSMI** complex showed a 29 nm red shift and lowered absorbance, addition of HT-T5 increased both the absorbance and the red shift, dependent on DNA concentration: only 2.5 μM HT-T5 caused an additional 36 nm shift. This effect was also seen with **DQMI**, which showed large absorbance changes in the presence of both cavitand and HT-T5 (Figure S-21), but only small red shifts with HT-T5 alone (Figure S-19). In addition, the effect was not cavitand-specific: all three hosts **AMI**, **CHI** and **CHP** effected similar behavior when added to the dye•DNA complexes.

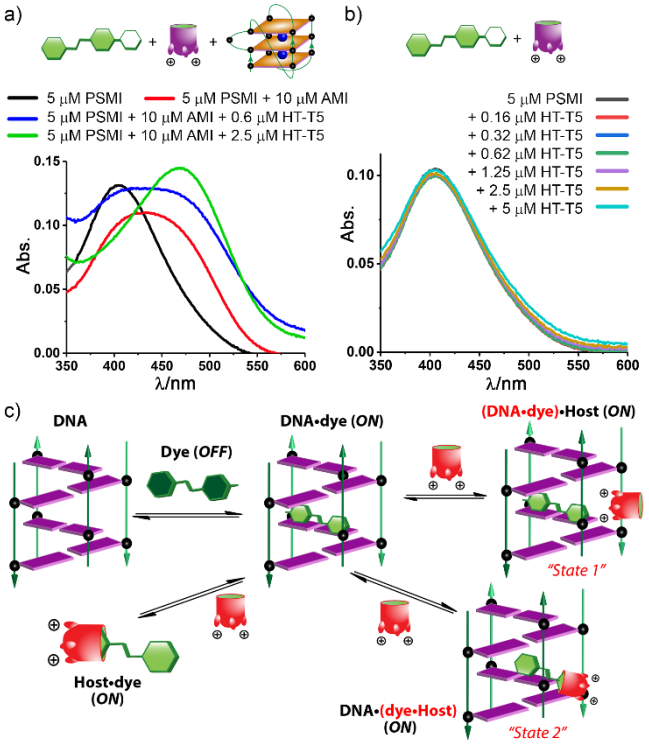


Figure 6. Host•Dye•DNA Interactions. UV-Vis spectra of **PSMI** dye (5 μM) and increasing concentrations of G4 DNA (HT-T5) with a) 10 μM cavitand (AMI) or b) no host, in 20 mM KOAc, 5 mM $MgCl_2$, pH 5.5. c) Schematic illustration of the multiple interaction equilibria between DNA, host, and dye leading to differential sensing of DNA folding motifs, using an i-motif as the example structure.

The emission and absorbance analysis allows a simplified discussion of the sensing mechanisms, albeit not a complete one. Obviously, multiple mechanisms contribute to the different emissions for each dye/host/DNA combination. An illustration of the equilibrium states that can contribute to the sensing mechanism is shown in Figure 6c: the dyes bind to both the hosts and the DNA, and show enhanced fluorescence emission in each case. The affinities for the dyes and different DNA folded structures can vary, as can the affinity for the dyes and the hosts,[22] and so competitive binding between DNA•dye and host•dye is an important contributor. The most interesting equilibrium states are those that involve both dye and host interacting with the DNA, i.e. heteroternary complexes. The UV/Vis absorbance data clearly shows the presence of heteroternary Dye•host•DNA complexes, although their exact structure is not obvious. Either (Dye•DNA)•host (i.e. state 1, Figure 6c) or (Dye•host)•DNA (i.e. state 2, Figure 6c) are possible: the dyes protrude from the cavitand when bound, so the cavitand•dye complexes could easily interact with the DNA, or the cationic cavitand could bind to the DNA•dye complex, altering the absorbance and emission. The requirement for folded DNA to effect maximal emission and absorbance changes on the dye suggests that the flat cationic dyes bind in an intercalative manner, and the large changes in emission and absorbance when cavitand is added suggest that state 2 is the most likely heteroternary complex, but this is only conjecture at this point.

Importantly, though, the exact nature of the ternary complexes is not important for the differential sensing concept: the combination of multiple different hosts and dyes, all of which can interact synergistically, is the driving force for the sensitivity and selectivity of the recognition. As DNA sequences are highly diverse, the combination of multiple hosts and dyes provides a greater diversity in signal than merely using a single fluorescent probe molecule, and allows application of the sensing array to detect and discriminate many of those diverse folds with small differences, not just target a single folding type.

CONCLUSIONS

In conclusion, we have shown that an arrayed suite of synthetic hosts and dyes is capable of sensing different oligonucleotide secondary structures, including G-quadruplexes, hairpins, triplexes, and i-motifs. Multiple recognition mechanisms can be exploited to create a unique sensing fingerprint consisting of variable fluorescence enhancements in the presence of different DNA folded structures. Discrimination between DNA strands with highly similar structures, such as G-quadruplex strands with bulges and vacancies, as well as triplexes with parallel and antiparallel orientations can be achieved. By applying machine learning algorithms, a classification model can be established from the training set, and this model can provide accurate prediction of the folding state of unknown sequences.

The design of highly specific fluorescent probes for different non-canonical folding patterns of DNA is very challenging, and this method overcomes this by introducing synthetic hosts to tune the fluorophore-DNA interaction, introducing multiple recognition equilibria that modulate the fluorescence signal depending on the small difference in the folded target structures. Machine learning allows rapid analysis of complex datasets and confirms the classification and prediction power of the synthetic array. This strategy can easily be expanded to a broad scope of DNA-interacting dyes and synthetic hosts to sense more diverse nucleic acid structures. Compared to existing characterization methods such as CD, NMR, and X-ray crystallography, pattern-recognition-based fluorescence sensing is far quicker, more straightforward, more compatible with high-throughput screening, and more sensitive.

EXPERIMENTAL

General Information. Cavitands **CHI**,[36] **CHP**,[36] **AMI**[37] and fluorophore **PSMI**[37] were synthesized according to literature

procedures. ^{1}H and ^{13}C spectra were recorded on Bruker Avance NEO 400 MHz or Bruker Avance 600 MHz NMR spectrometer. The spectrometers were automatically tuned and matched to the correct operating frequencies. Proton (^{1}H) and carbon (^{13}C) chemical shifts are reported in parts per million (δ) with respect to tetramethylsilane (TMS, δ=0), and referenced internally. Deuterated NMR solvents were obtained from Cambridge Isotope Laboratories, Inc., Andover, MA, and used without further purification. All other materials, including *trans*-4-[4-(dimethylamino)-styryl]-1-methyl-pyridinium iodide were obtained from Aldrich Chemical Company (St. Louis, MO), or Fisher Scientific (Fairlawn, NJ), and were used as received. Solvents were dried through a commercial solvent purification system (Pure Process Technologies, Inc.). Oligonucleotides were purchased from Integrated DNA Technologies (IDT) with standard desalting and no further purification, the sequence and structural information of which are given in Table S-1. The concentrations of DNA stock solutions were determined by NanoDrop 2000 (Thermo Fisher Scientific) using the corresponding molar extinction coefficients provided by IDT after background subtraction. Before the experiments, the DNA stock solutions were diluted with 20 mM KOAc and 5 mM $MgCl_2$ at pH 5.5 and re-annealed to form the most stable folding topology, in which the DNA solutions were heated at 95 °C for 5 min, cooled on ice for 10 min and then held at room temperature for 30 min. Fluorescence measurements were performed with a BioTek™ Synergy™ H1 Hybrid Multi-Mode Microplate Reader at Fluorescence Endpoint or Spectral scanning read mode with the Ex/Em wavelengths at 520/600nm (**DSMI**), 500/600nm (**PSMI**), 480/600nm (**MSMI**), 560/640nm (**DQMI**), Gain=100. UV-Vis absorbance measurements were performed with an Agilent Technologies Cary 60 UV/Vis spectrophotometer using the disposable, methacrylate semi-micro cuvettes (path length = 10 mm). Principal Component Analysis (PCA) and confidence ellipses were performed with RStudio (Version 1.2.5019), an integrated development environment (IDE) for R (version 3.6.1). Classification and prediction were performed with Python 3.9 (64-bit), using StandardScaler for data standardization, PCA for orthogonal linear transformation and dimensionality reduction, Recursive Feature Elimination (RFE) for feature selection, Support Vector Machine (SVM) (kernel='linear') as the supervised classification model, and RepeatedStratifiedKFold (n_splits=8, n_repeats=3) for cross validation.

Fluorescence measurements. 1) Array constituents. The fluorescence assay was carried out by mixing 10 μL of the fluorescent dye (1.5625 μM **DSMI, PSMI, MSMI, DQMI** in water), 10 μL of the cavitand (1.25 μM **CHI, CHP, AMI** in water) or water, 70 μL of the incubation buffer, and 10 μL of 1 μM DNA in the 96-well plate, resulting in a final total volume around 100 μL in 20 mM KOAc and 5 mM $MgCl_2$ at pH 5.5. The mixture was incubated with mild shaking for 15 min at room temperature, before the fluorescence signal (F) was recorded. **2) Titrations. Dye-DNA:** The fluorescence titration curves were obtained by using 0-20 μM Dye and 0.1 μM DNA (HT-T5/HD28/25-mer hairpin/hTelo or no DNA). **Host Addition to Dye•DNA Complexes:** Fluorescence response curves of dye•DNA complexes upon titration of hosts were obtained by using 0.15625 μM dye, 0.1 μM DNA HT-T5/HD28/25-mer hairpin/hTelo or no DNA, 0-16 μM Host. **3) Fluorescence Spectra.** The emission and excitation fluorescence spectra were obtained from mixtures of the solution of dye (0.625 μM), host (4 μM), HT-T5 (0.2 μM).

UV-Vis Absorbance Spectra. All spectra were obtained with using 5 μM dye **(PSMI** or **DQMI)** and 0-5 μM HT-T5/A20, with or without the three hosts: **CHI, CHP** or **AMI**, at 10 μM, in 20 mM KOAc and 5 mM $MgCl_2$ at pH 5.5. The spectra were presented with baseline-correction in which the background signal from the buffer was subtracted.

Circular Dichroism (CD). CD spectra were recorded on a Jasco J-815 CD spectrophotometer over a wavelength range of 200 nm–350 nm at room temperature, with a band width of 1 nm and a data pitch of 1 nm. The instrument scanning speed was set at 100 nm/min, with a response time of 1 s. 10 μM of 200 μL oligonucleotide solution prepared in the 20 mM KOAc and 5 mM $MgCl_2$ at pH 5.5 buffer then was pipetted into a quartz cell with a path length of 0.1 cm. The CD spectra were presented with baseline-correction in which the background signal from the buffer was subtracted.

Gel Electrophoresis. The quality of the DNA solution was inspected by native gel electrophoresis using a gradient native polyacrylamide gel electrophoresis (PAGE) gel (4%-20%). 5 μL (or 10 μl for hTelo and C20) of a 2 μM DNA solution was loaded to the gel, after being denatured at 95°C for 5 min, cooled on ice for 10 min and then at room temperature for 30 min. The gel was run at 120 V for 60 min at room temperature in 1×TBE buffer, and stained with SYBR Gold (1.5:10000 dilution) before imaged using the UV transilluminator (SPECTROLINE).

Synthesis of New Compounds.

MSMI. 1,4-Dimethylpyridinium iodide (235 mg, 1.0 mmol) and 4-morpholinobenzaldehyde (191 mg, 1.0 mmol) were dissolved in ethanol (5 mL) inside a round bottom flask. While stirring, one drop of piperidine was added and the resulting solution was refluxed for 12 hours. The reaction was cooled, then diluted with water (10 mL). The resulting precipitate was filtered, rinsed with water and cold ethanol, then dried under vacuum to yield (E)-1-methyl-4-(4-morpholinostyryl)pyridin-1-ium iodide (388 mg, 95% yield) as a bright red powder. ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.75 (d, J = 6.6 Hz, 1H), 8.10 (d, J = 6.6 Hz, 1H), 7.93 (d, J = 16.2 Hz, 1H), 7.63 (d, J = 8.7 Hz, 1H), 7.28 (d, J = 16.2 Hz, 1H), 7.03 (d, J = 8.7 Hz, 1H), 4.20 (s, 3H), 3.74 (t, J = 4.5 Hz, 4H), 3.27 (t, J = 4.8 Hz, 4H). ^{13}C NMR (100 MHz, DMSO-d_6) δ 153.60, 152.90, 145.05, 141.65, 130.29, 125.63, 123.02, 119.34, 114.72, 66.36, 47.54, 47.05. ESI-MS: *m/z* $C_{18}H_{21}N_2O^+$ (M^+) calculated: 281.1617, found: 281.1654. UV/Vis: Exc. λ_{max} = 395 nm, Em. λ_{max} = 600 nm.

DQMI. 6-methoxy-2-methylquinoline (400mg, 2.3 mmol) was dissolved in iodomethane (3 mL) and refluxed for 12 hours. The solution was diluted into diethyl ether (10 mL) and the resulting precipitate was filtered, rinsed with diethyl ether, then dried under vacuum to yield 6-methoxy-1,2-dimethylquinolin-1-ium iodide (713 mg, 98%) as a bright yellow solid. ^{1}H NMR (600 MHz, DMSO-d_6) δ 8.95 (d, J = 8.6 Hz, 1H), 8.52 (d, J = 9.4 Hz, 1H), 8.06 (d, J = 8.6 Hz, 1H), 7.85 (dd, J = 9.4, 2.9 Hz, 1H), 7.84 (d, J = 2.9 Hz, 1H), 4.43 (s, 3H), 4.00 (s, 3H), 3.03 (s, 3H). 6-methoxy-1,2-dimethylquinolin-1-ium iodide (315 mg, 1.0 mmol) and 4-(dimethylamino)benzaldehyde (149 mg, 1.0 mmol) were dissolved in ethanol (5 mL) inside a round bottom flask. While stirring, one drop of piperidine was added and the resulting solution was refluxed for 12 hours. The reaction was cooled, then diluted with water (10 mL). The resulting precipitate was filtered, rinsed with water and cold ethanol, then dried under vacuum to yield (E)-2-(4-(dimethylamino)styryl)-6-methoxy-1-methylquinolin-1-ium iodide (375 mg, 84% yield) as a dark purple powder. ^{1}H NMR (600 MHz, DMSO-d_6) δ 8.71 (d, J = 9.1 Hz, 1H), 8.47 (d, J = 9.2 Hz, 1H), 8.38 (d, J = 9.3 Hz, 1H), 8.14 (d, J = 15.5 Hz, 1H), 7.82 (d, J = 8.7 Hz, 2H), 7.73 (d, J = 2.7 Hz, 1H), 7.72 (dd, J = 9.2, 2.7 Hz, 1H), 7.53 (d, J = 15.5 Hz, 1H), 6.83 (d, J = 8.8 Hz, 2H), 4.44 (s, 3H), 3.97 (s, 3H), 3.07 (s, 6H). ^{13}C NMR (100 MHz, DMSO-d_6) δ 158.62,

154.44, 153.00, 147.79, 141.36, 134.86, 131.99, 131.97, 129.12, 125.33, 122.90, 121.12, 112.44, 112.25, 109.23, 56.59, 40.20, 39.65. ESI-MS: *m/z* $C_{21}H_{23}N_2O^+$ (M^+) calculated: 319.1777, found: 319.1810. UV/Vis: Ex. λ_{max} = 490 nm, Em. λ_{max} = 595 nm.

ASSOCIATED CONTENT

Supporting Information

DNA sequences and characterization, NMR spectra, fluorescence spectra, UV-Vis absorption spectra and full array data, including the fluorescence bar plots and PCA results, for analysis of the 7-DNA and 18-DNA pools. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

* E-mail: richard.hooley@ucr.edu; wenwan.zhong@ucr.edu

ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation (CHE-1707347 to W.Z. and R.J.H.) for support, and are grateful to Prof. Min Xue and his group for the help with using the BioTek™ Synergy™ H1 Hybrid Multi-Mode Microplate Reader. J. C. would also like to thank Liyi Chen for discussion of Python usage.

REFERENCES

(1) Belmont, P.; Constant, J.-F.; Demeunynck, M. Nucleic acid conformation diversity: from structure to function and regulation. *Chem. Soc. Rev.* **2001**, *30*, 70–81.

(2) (a) Du, X.; Wojtowicz, D.; Bowers, A. A.; Levens, D.; Benham, C. J.; Przytycka, T. M. The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in Escherichia coli. *Nucleic Acids Res.* **2013**, *41*, 5965–5977. (b) del Mundo, I. M. A.; Zewail-Foote, M.; Kerwin, S. M.; Vasquez, K. M. Alternative DNA structure formation in the mutagenic human c-MYC promoter. *Nucleic Acids Res.* **2017**, *45*, 4929–4943. (c) Jansson, L. I.; Hentschel, J.; Parks, J. W.; Chang, T. R.; Lu, C.; Baral, R.; Bagshaw, C. R.; Stone, M. D. Telomere DNA G-quadruplex folding within actively extending human telomerase. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 9350–9359. (d) Wang, G.; Vasquez, K. M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair* **2014**, *19*, 143–151.

(3) (a) Di Antonio, M.; Ponjavic, A.; Radzevičius, A.; Ranasinghe, R. T.; Catalano, M.; Zhang, X.; Shen, J.; Needham, L.-M.; Lee, S. F.; Klenerman, D.; Balasubramanian, S. Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat. Chem.* **2020**, *12*, 832–837. (b) Balasubramanian, S.; Hurley, L. H.; Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.* **2020**, *10*, 261–275. (c) Kamat, M. A.; Bacolla, A.; Cooper, D. N.; Chuzhanova, N. A role for non-B DNA forming sequences in mediating microlesions causing human inherited disease. *Hum. Mutat.* **2016**, *37*, 65–73. (d) Zain, R.; Smith, C. I. E. Targeted oligonucleotides for treating neurodegenerative tandem repeat diseases. *Neurotherapeutics* **2019**, *16*, 248–262.

(4) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **2006**, *34*, 5402–5415.

(5) Sklenář, V.; Felgon, J. Formation of a stable triplex from a single DNA strand. *Nature* **1990**, *345*, 836–838.

(6) Choi, J.; Kim, S.; Tachikawa, T.; Fujitsuka, M.; Majima, T. pH-Induced Intramolecular Folding Dynamics of i-Motif DNA. *J. Am. Chem. Soc.* **2011**, *133*, 16146–16153.

(7) Choi, J.; Majima, T. Conformational changes of non-B DNA. *Chem. Soc. Rev.* **2011**, *40*, 5893–5909.

(8) Mukundan, V. T.; Phan, A. T. Bulges in G-quadruplexes: Broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.* **2013**, *135*, 5017–5028.

(9) (a) Li, X.-M.; Zheng, K.-W.; Zhang, J.-Y.; Liu, H.-H.; He, Y.-D.; Yuan, B.-F.; Hao, Y.-H.; Tan, Z. Guanine-vacancy–bearing G-quadruplexes responsive to guanine derivatives. *Proc. Natl. Acad. Sci. U.S.A.* **2015,** *112*, 14581–14586. (b) He, Y.-D.; Zheng, K.-W.; Wen, C.-J.; Li, X.-M.; Gong, J.-Y.; Hao, Y.-H.; Zhao, Y.; Tan, Z. Selective targeting of guanine-vacancy-bearing G-quadruplexes by G-quartet complementation and stabilization with a guanine–peptide conjugate. *J. Am. Chem. Soc.* **2020,** *142*, 11394-11403.

(10) Goñi, J. R.; de la Cruz, X.; Orozco, M. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.* **2004**, *32*, 354–360.

(11) (a) Debnath, M.; Fatma, K.; Dash, J. Chemical regulation of DNA i-motifs for nanobiotechnology and therapeutics. *Angew. Chem. Int. Ed.* **2019**, *58*, 2942–2957. (b) Ruggiero, E.; Richter, S. N. G-quadruplexes and G-quadruplex ligands: Targets and tools in antiviral therapy. *Nucleic Acids Res.* **2018**, *46*, 3270–3283.

(12) Tateishi-Karimata, H.; Sugimoto, N. Chemical biology of non-canonical structures of nucleic acids for therapeutic applications. *Chem. Commun.* **2020**, *56*, 2379–2390.

(13) Nakano, S.-i.; Miyoshi, D.; Sugimoto, N. Effects of molecular crowding on the structures, interactions, and functions of Nucleic Acids. *Chem. Rev.* **2014**, *114*, 2733-2758.

(14) Winnerdy, F. R.; Bakalar, B.; Maity, A.; Vandana, J. J.; Mechulam, Y.; Schmitt, E.; Phan, A. T. NMR solution and X-ray crystal structures of a DNA containing both right-and left-handed parallel-stranded G-quadruplexes *Nucleic Acids Res.* **2019**, *47*, 8272–8281.

(15) Salgado, G. F.; Cazenave, C.; Kerkour, A.; Mergny, J.-L. G-quadruplex DNA and ligand interaction in living cells using NMR spectroscopy. *Chem. Sci.* **2015**, *6*, 3314–3320.

(16) del Villar-Guerra, R.; Trent, J. O.; Chaires, J. B. G-Quadruplex secondary structure obtained from Circular Dichroism spectroscopy. *Angew. Chem. Int. Ed.* **2018**, *57*, 7171–7175.

(17) Lubitz, I.; Zikich, D.; Kotlyar, A. Specific high-affinity binding of thiazole orange to triplex and G-quadruplex DNA. *Biochemistry* **2010**, *49*, 3567–3574.

(18) (a) You, L.; Zha, D.; Anslyn, E. V. Recent advances in supramolecular analytical chemistry using optical sensing. *Chem. Rev.* **2015**, *115*, 7840–7892. (b) Stewart, S.; Ivy, M. A.; Anslyn, E. V. The use of principal component analysis and discriminant analysis in differential sensing routines. *Chem. Soc. Rev.* **2014**, *43*, 70–84.

(19) (a) Eubanks, C. S.; Forte, J. E.; Kapral, G. J.; Hargrove, A. E. Small molecule-based pattern recognition to classify RNA structure. *J. Am. Chem. Soc.* **2017**,*139*, 409–416. (b) Eubanks, C. S.; Hargrove, A. E. RNA structural differentiation: opportunities with pattern recognition. *Biochemistry* **2019**, *58*, 199–213. (c) Eubanks, C. S.; Zhao, B.; Patwardhan, N. N.; Thompson, R. D.; Zhang, Q.; Hargrove, A. E. Visualizing RNA conformational changes via pattern recognition of RNA by small molecules. *J. Am. Chem. Soc.* **2019**, *141*, 5692–5698.

(20) Zuffo, M.; Xie. X.; Granzhan, A. Strength in numbers: development of a fluorescence sensor array for secondary structures of DNA. *Chem.-Eur. J.* **2019**, *25*, 1821–1818.

(21) del Villar-Guerra, R.; Gray, R. D.; Trent, J. O.; Chaires, J. B. A rapid fluorescent indicator displacement assay and principal component/cluster data analysis for determination of ligand–nucleic acid structural selectivity. *Nucleic Acids Res.* **2018**, *46*, e41.

(22) Chen, J.; Hickey, B. L.; Wang, L.; Lee, J.; Gill, A. D.; Favero, A.; Pinalli, R.; Dalcanale, E.; Hooley, R. J.; Zhong, W. Selective discrimination and classification of G-quadruplex structures with a host:guest sensing array *Nat. Chem.* **2021**, *13*, 488–495.

(23) Dreher, S. D.; Krska, S. W. Chemistry informer libraries: conception, early experience, and role in the future of cheminformatics. *Acc. Chem. Res.* **2021**, *54*, 1586–1596.

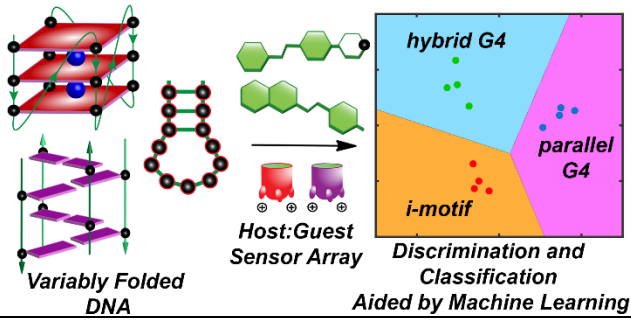
(24) Shi, Y.; Prieto, P. L.; Zepel, T.; Grunert, S.; Hein, J. E. Automated experimentation powers data science in chemistry. *Acc. Chem. Res.* **2021**, *54*, 546–555.

(25) Goecks, J.; Jalili, V.; Heiser, L. M.; Gray, J. W. How machine learning will transform biomedicine. *Cell* **2020**, *181*, 92–101.

(26) (a) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Ryan P. Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*

2021, *590*, 89–96. (b) Żurański, A. M.; Alvarado, J. I. M.; Shields, B. J.; Doyle, A. G. Predicting reaction yields via supervised learning. *Acc. Chem. Res.* **2021**, *54*, 1856–1865.
(27) Bradley, P.; Gordon, N. C.; Walker, T. M.; Dunn, L.; Heys, S.; Huang, B.; Earle, S.; Pankhurst, L. J.; Anson, L.; de Cesare, M.; Piazza, P.; Votintseva, A. A.; Golubchik, T.; Wilson, D. J.; Wyllie, D. H.; Diel, R.; Niemann, S.; Feuerriegel, S.; Kohl, T. A.; Ismail, N.; Omar, S. V.; Smith, E. G.; Buck, D.; McVean, G.; Walker, A. S.; Peto, T. E. A.; Crook, D. W.; Iqbal, Z. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **2015**, *6*, 10063.
(28) (a) Rosania, G. R.; Lee, J. W.; Ding, L.; Yoon, H.-S.; Chang, Y.-T. Combinatorial approach to organelle-targeted fluorescent library based on the styryl scaffold. *J. Am. Chem. Soc.* **2003**, *125*, 1130–1131. (b) Beatty, M. A.; Selinger, A. J.; Li, Y.; Hof, F. Parallel synthesis and screening of supramolecular chemosensors that achieve fluorescent turn-on detection of drugs in saliva. *J. Am. Chem. Soc.* **2019**, *141*, 16763–16771.
(29) (a) Phan, A. T.; Guéron, M.; Leroy, J.-L. The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *J. Mol. Biol.* **2000**, *299*, 123–144. (b) Day, H. A.; Pavlou, P.; Waller, Z. A. E. i-Motif DNA: Structure, stability and targeting with ligands. *Bioorg. Med. Chem.* **2014**, *22*, 4407–4418.
(30) Liu, Y.; Perez, L.; Mettry, M.; Gill, A. D.; Byers, S. R.; Easley, C. J.; Bardeen, C. J.; Zhong, W.; Hooley, R. J. Site selective reading of epigenetic markers by a dual-mode synthetic receptor array. *Chem. Sci.* **2017**, *8*, 3960 – 3970.
(31) Jurs, P. C.; Bakken, G. A.; McClelland, H. E. Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chem. Rev.* **2000**, *100*, 2649–2678.
(32) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
(33) Ivanciuc, O. Applications of Support Vector Machines in chemistry. *Rev. Comput. Chem.* **2007**, *23*, 291–400.
(34) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389–422.
(35) Joly, L.; Rosu, F.; Gabelica, V. d(TG_nT) DNA sequences do not necessarily form tetramolecular G-quadruplexes. *Chem. Commun.* **2012**, *48*, 8386–8388.
(36) Mosca, S.; Yu, Y.; Rebek, J., Jr. Preparative scale and convenient synthesis of a water-soluble, deep cavitand. *Nat. Protoc.* **2016**, *11*, 1371-1387.
(37) Gill, A. D.; Hickey, B. L.; Wang, S.; Xue, M.; Zhong, W.; Hooley, R. J. Sensing of citrulline modifications in histone peptides by deep cavitand hosts. *Chem. Commun.* **2019**, *55*, 13259–13262.

TOC Graphic:


hybrid G4
parallel G4
i-motif
Variably Folded DNA
Host:Guest Sensor Array
Discrimination and Classification Aided by Machine Learning

Machine Learning Aids Classification and Discrimination of Non-canonical DNA Folding Motifs by an Arrayed Host:guest Sensing System

Junyi Chen,[2] Adam D. Gill,[3] Briana L. Hickey,[1] Ziting Gao,[1] Xinping Cui,[4] Richard J. Hooley[1,3]* and Wenwan Zhong[1, 2]*

[1]*Department of Chemistry;* [2]*Environmental Toxicology Graduate Program;* [3]*Department of Biochemistry;* [4]*Department of Statistics; University of California-Riverside, Riverside, CA 92521, U.S.A.*

E-mail: richard.hooley@ucr.edu; wenwan.zhong@ucr.edu

Supporting Information

Table of Contents

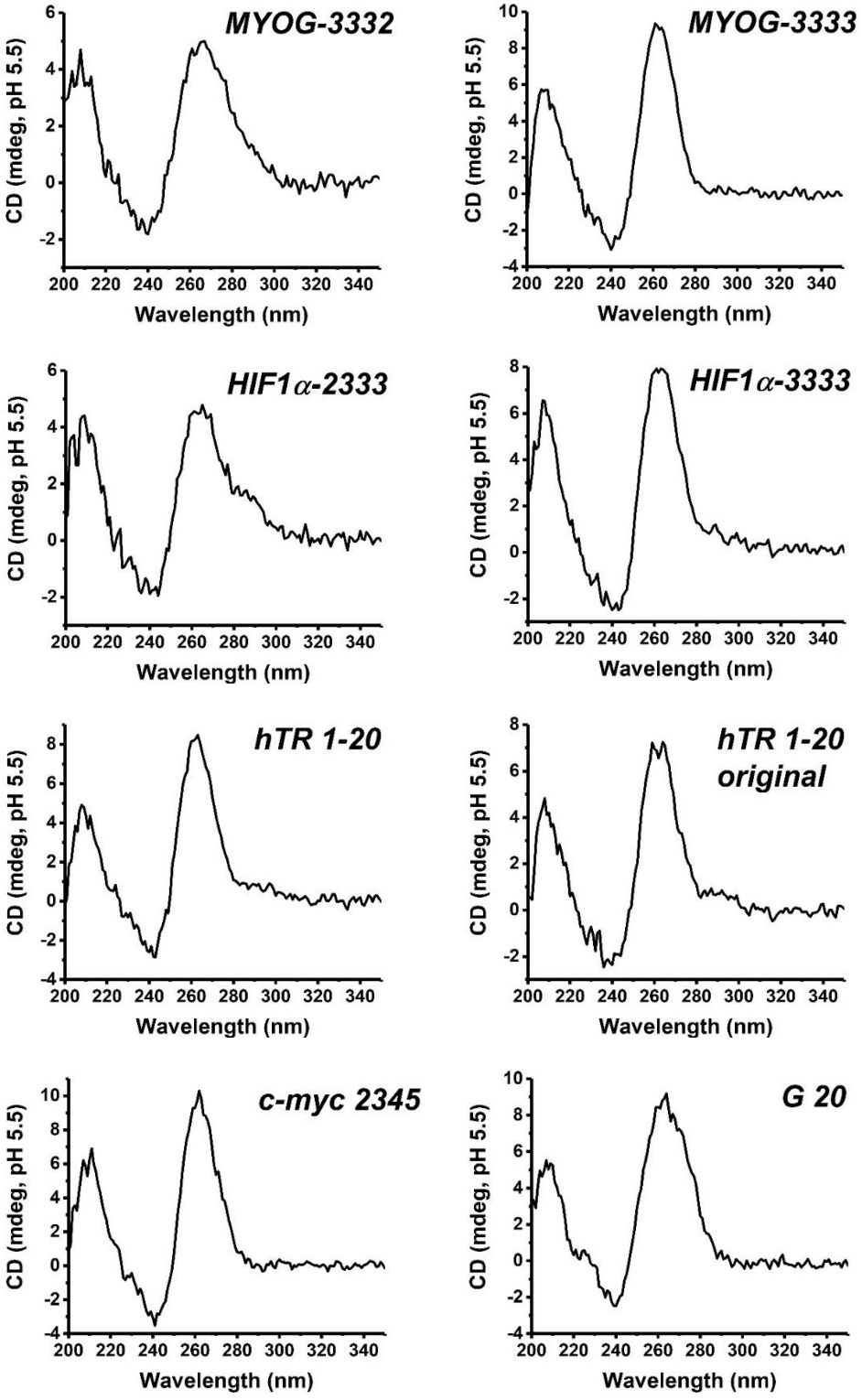
1. DNA Sequences and Characterization

1.1 DNA Sequences

Table *S*-1. DNA sequences used in this project.

Group	Name	Sequence	Motif	Bases	Ref.
DNA G4 with/ without a vacancy	MYOG-3332	AGGGTGGGCTGGGAGGT	Parallel G4 with a vacancy	17	1
	MYOG-3333	AGGGTGGGCTGGGAGGGT	Parallel G4	18	
	HIF1α-2333	AGGTGGGCGGGCTTGCGGGA	Parallel G4 with a vacancy	20	2
	HIF1α-3333	AGGGTGGGCGGGCTTGCGGGA	Parallel G4	21	
DNA G4 with/ without a bulge	hTR 1-20	GGGTTGCGGAGGGT GGG CCT	c1: parallel + antiparallel G4 c2: parallel dimeric G4	20	3
	hTR 1-20 original	GGGTTGGGAGGGTGGG CCT	Parallel G4	19	
	HT-T5	TTGGGTTAGGGTTAGTGGTTAGGGA	Hybrid G4 with a bulge	25	4
	HT-T5 original	TTGGGTTAGGGTTAGGGTTAGGGA	Hybrid G4	24	
i-motif	hTelo	CCCTAACCCTAACCCTAACCCT	i-motif (pH 4-7)	22	5
	c-kit	CCCTCCTCCCAGCGCCCACCCT	i-motif (pH 5-6.8)	22	6
	DIA	CCCAATCCCAATCCCAATCCC	i-motif (pH=4.8) ssDNA (pH=7.6)	21	7
Triplex	HD28	GAGAGAACCCCTTCTCTCTTTCTCTCTT	Parallel triplex (pH 5.5)	28	8
	HD31	AGAGAGAACCCCTTCTCTCTTTTTCTCTCTT	Parallel triplex (pH 5.5)	31	9
	Triplex 6	TCCCTCCTTTTTGGAGGGATTTTTTGGGTGG	Antiparallel triplex (pH 5.6)	31	10
	Triplex 7	TCCCTCCTTTTTGGAGGGATTTTTAGGGAGG	Antiparallel triplex (pH 5.6)	31	
Hairpin	25-mer	CCCCTTAGTAGTTCCTCACAAGGGG	hairpin	25	11
	1NGO	CTCTTTTTGTAAGAAATACAAGGAGAG	hairpin	27	12
Unibase strand	A20	AAAAAAAAAAAAAAAAAAAA	Unibase	20	13-14
	C20	CCCCCCCCCCCCCCCCCCCC	Unibase	20	15
Predict	APE1-4 track	TACCCACCCCCACCCTGCCCTG	i-motif (pH 5)	22	16
	c-myc 2345	TGAGGGTGGGGAGGGTGGGGAA	Parallel G4	22	17-19
	G20	GGGGGGGGGGGGGGGGGGGG	Parallel G4 or w/ bulges	20	20-22
	EAD4	CTGGGTTGGGTTGGGTTGGGA	Parallel G4	21	23
	Telo24	TTAGGGTTAGGGTTAGGGTTAGGG	Hybrid G4	24	24

1.2 CD Spectra for DNA Folding Confirmation



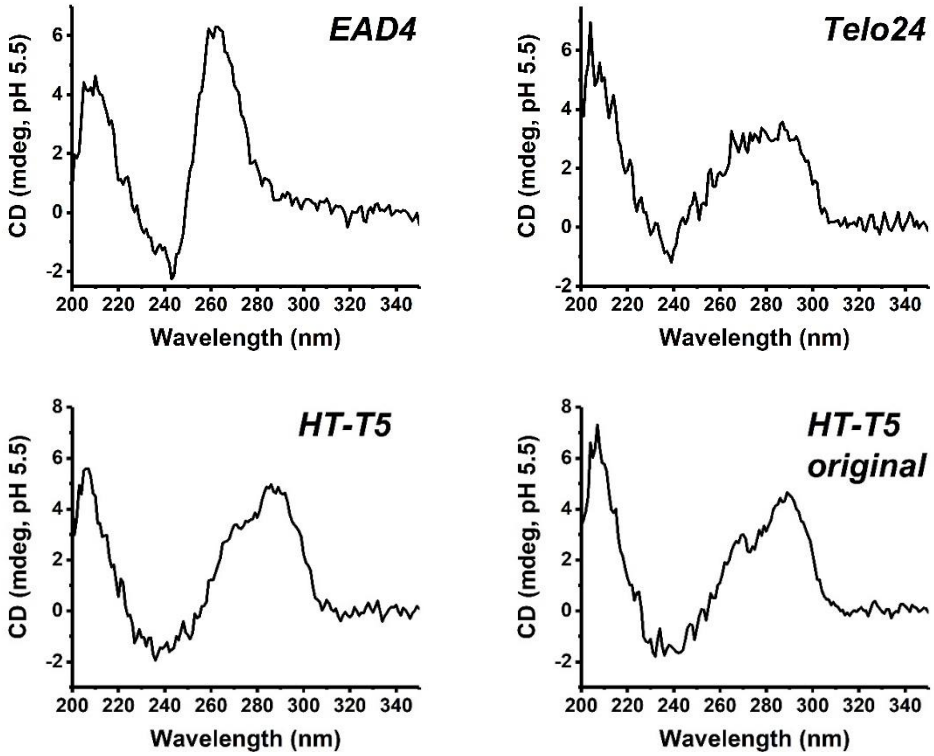


Figure *S*-1a. CD spectra with baseline correction of 10 μM G4 DNA in 20mM CH_3COOK 5mM $MgCl_2$ pH 5.5. DNA was denatured at 95 °C for 5 min, cooled on ice for 10 min and then held at room temperature for 30 min before the experiment.

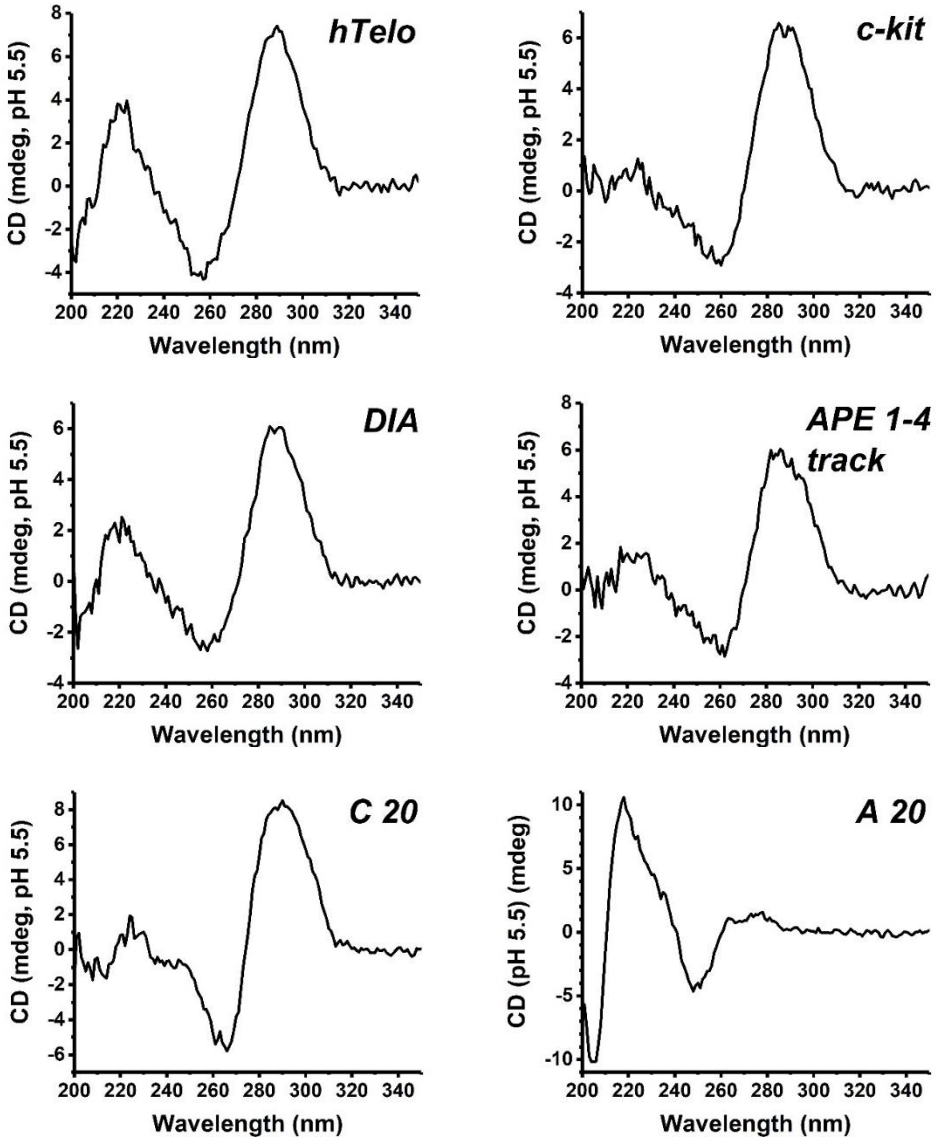


Figure *S*-1b. CD spectra with baseline correction of 10 μM i-motif and C20/A20 DNA. Other conditions are identical to those described in Figure *S*-1a.

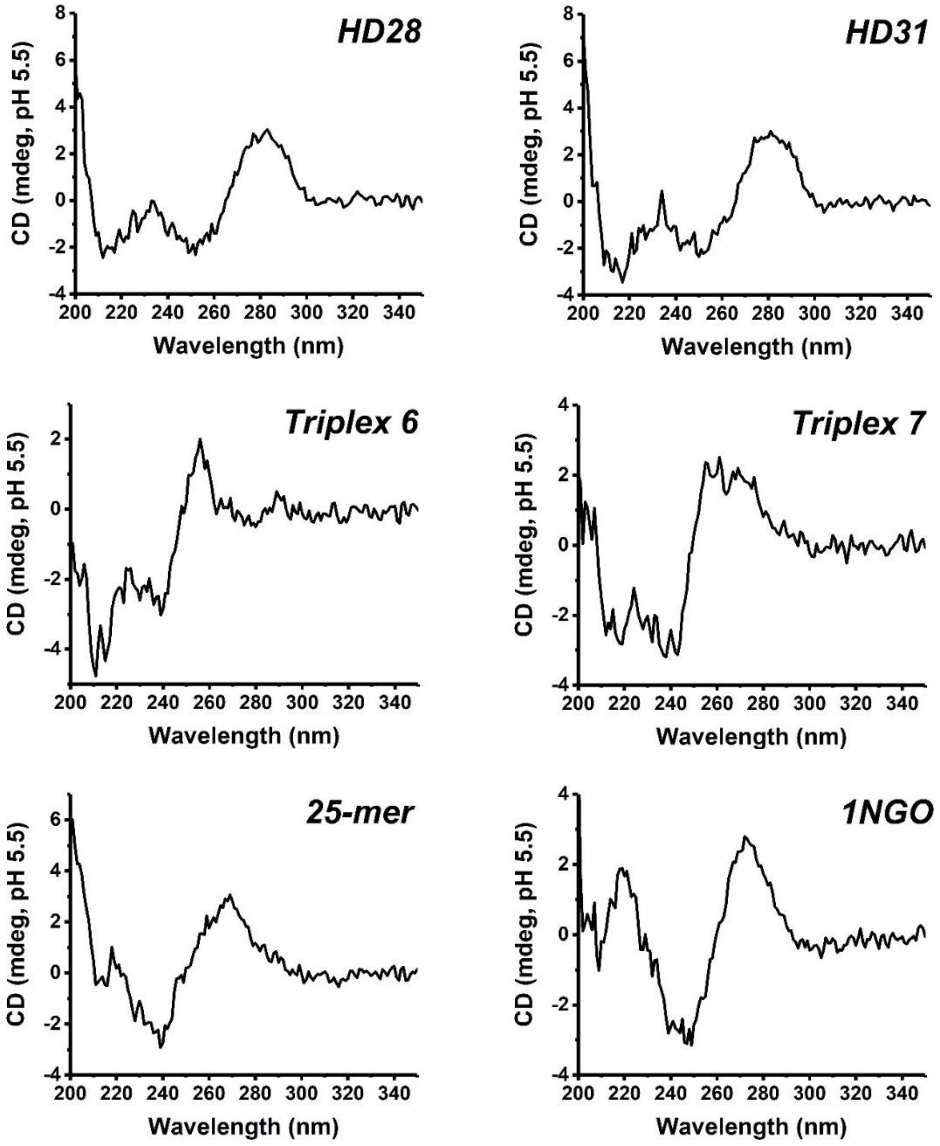


Figure *S*-1c. CD spectra with baseline correction of 10 μM triplex and hairpin DNA. Other conditions are identical to those described in Figure *S*-1a.

1.3 Gel Electrophoresis for DNA Quality Inspection

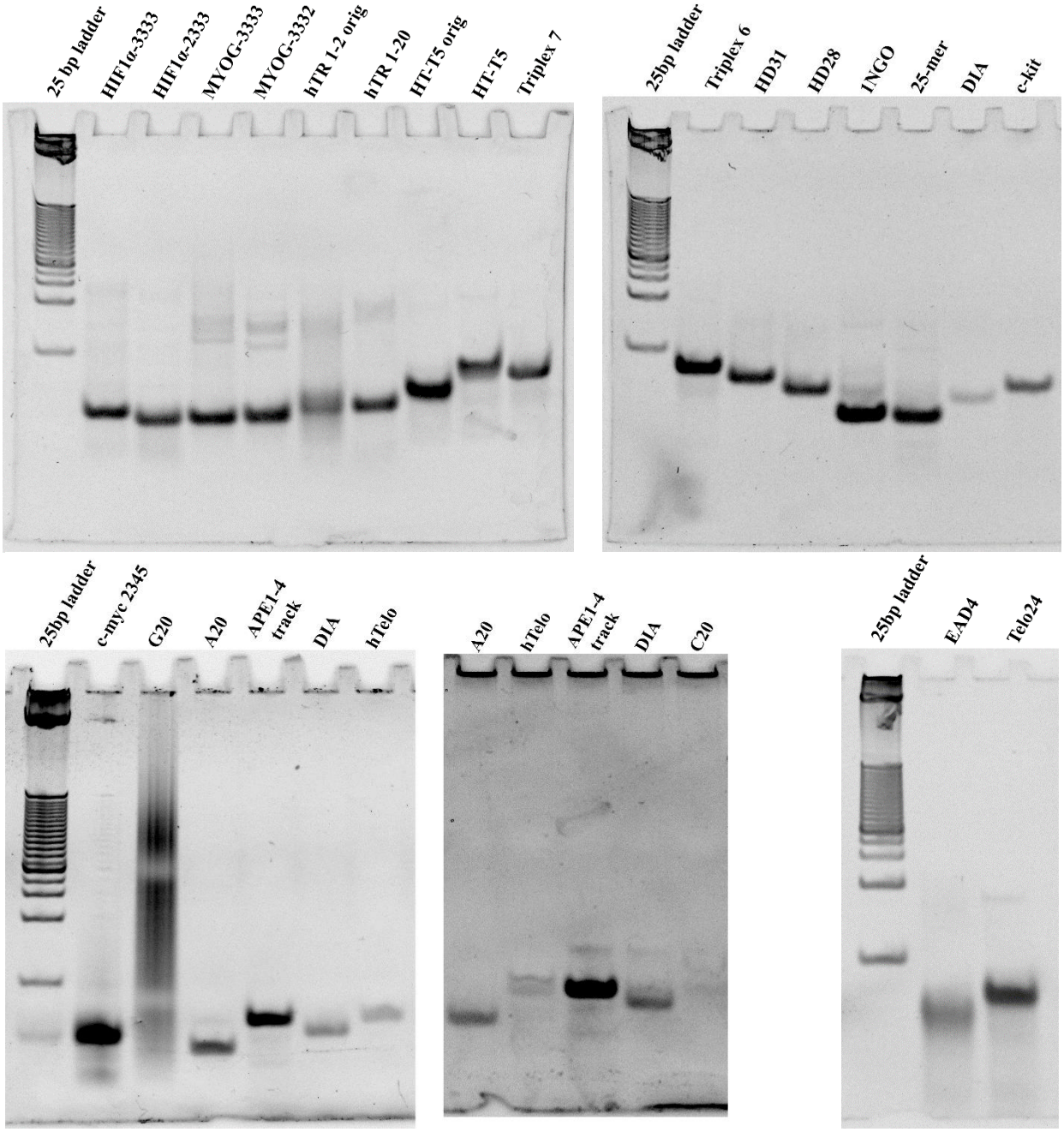


Figure *S*-2. The gradient native polyacrylamide gel electrophoresis (PAGE) gel (4%-20%) results of DNA sequences. The gel was loaded with 5 μl (10 μl for hTelo and C20) of a 2 μM DNA dissolved in 20mM CH_3COOK 5mM $MgCl_2$ pH 5.5 buffer, which had been denatured at 95 °C for 5 min, cooled on ice for 10 min and then held at room temperature for 30 min. The gel was run at 120 V for 60 min at room temperature in 1 × TBE buffer and stained with SYBR Gold (1.5:10000 dilution) before imaging using a UV transilluminator (SPECTROLINE).

2. NMR Spectra of Components Used

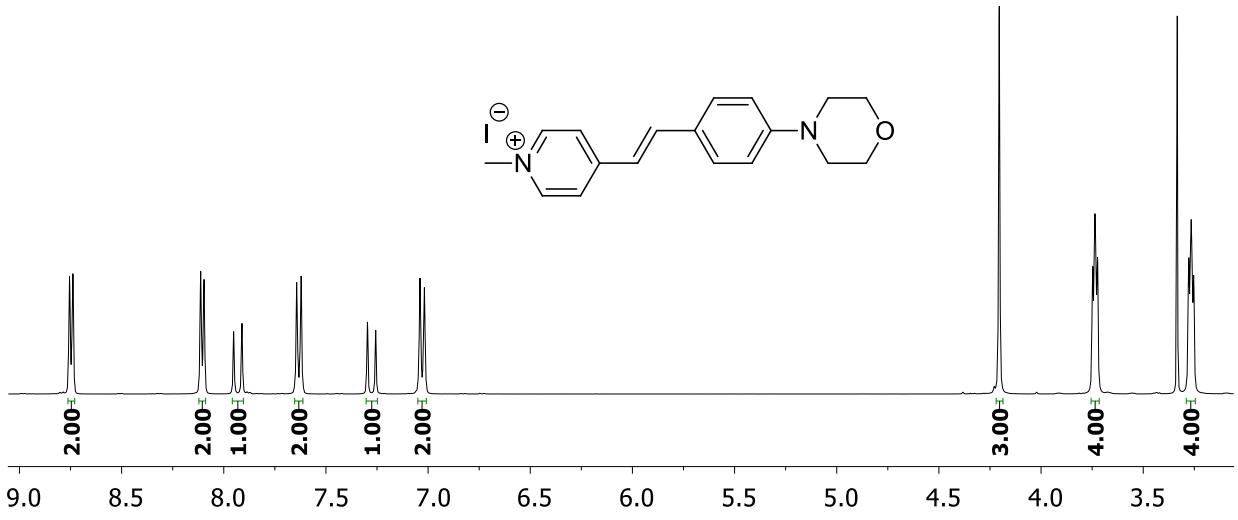


Figure *S*-3. ^{1}H NMR spectrum of **MSMI** (400 MHz, 298 K, DMSO-d_6).

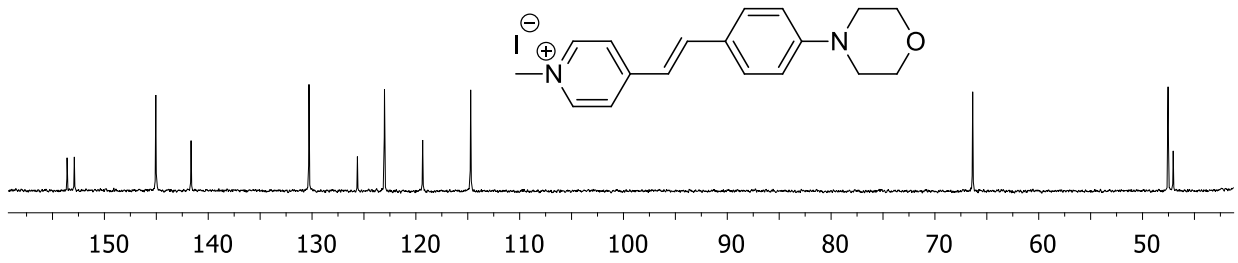


Figure *S*-4. ^{13}C NMR of **MSMI** (100 MHz, 298 K, DMSO-d_6).

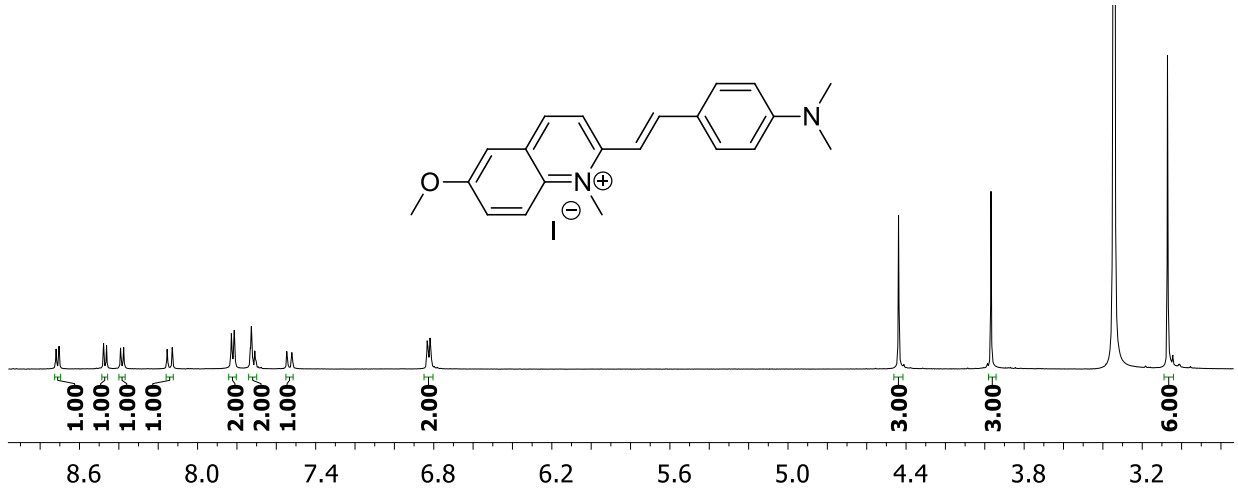


Figure *S*-5. ^{1}H NMR spectrum of **DQMI** (600 MHz, 298 K, DMSO-d_6).

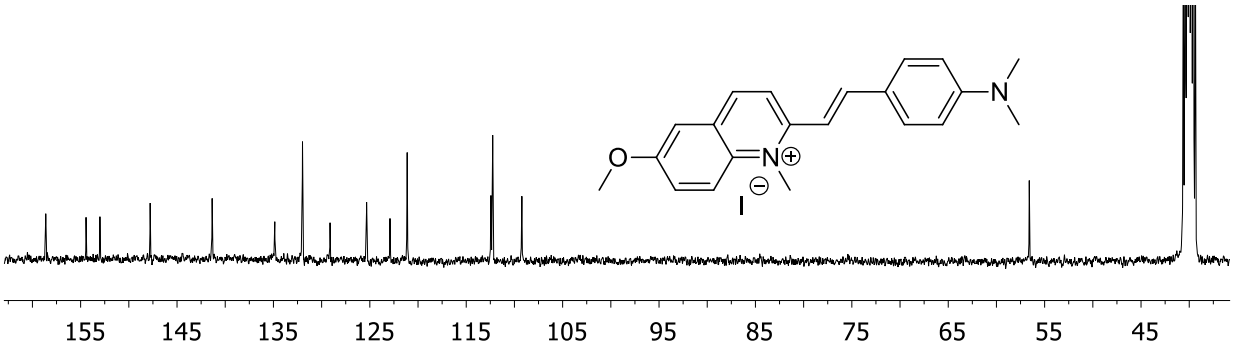


Figure *S*-6. ^{13}C NMR of **DQMI** (100 MHz, 298 K, DMSO-d_6).

3. Fluorescence Spectra and Titration Curves

3.1 Fluorescence Spectra of Fluorescent Guests with Hosts/DNA

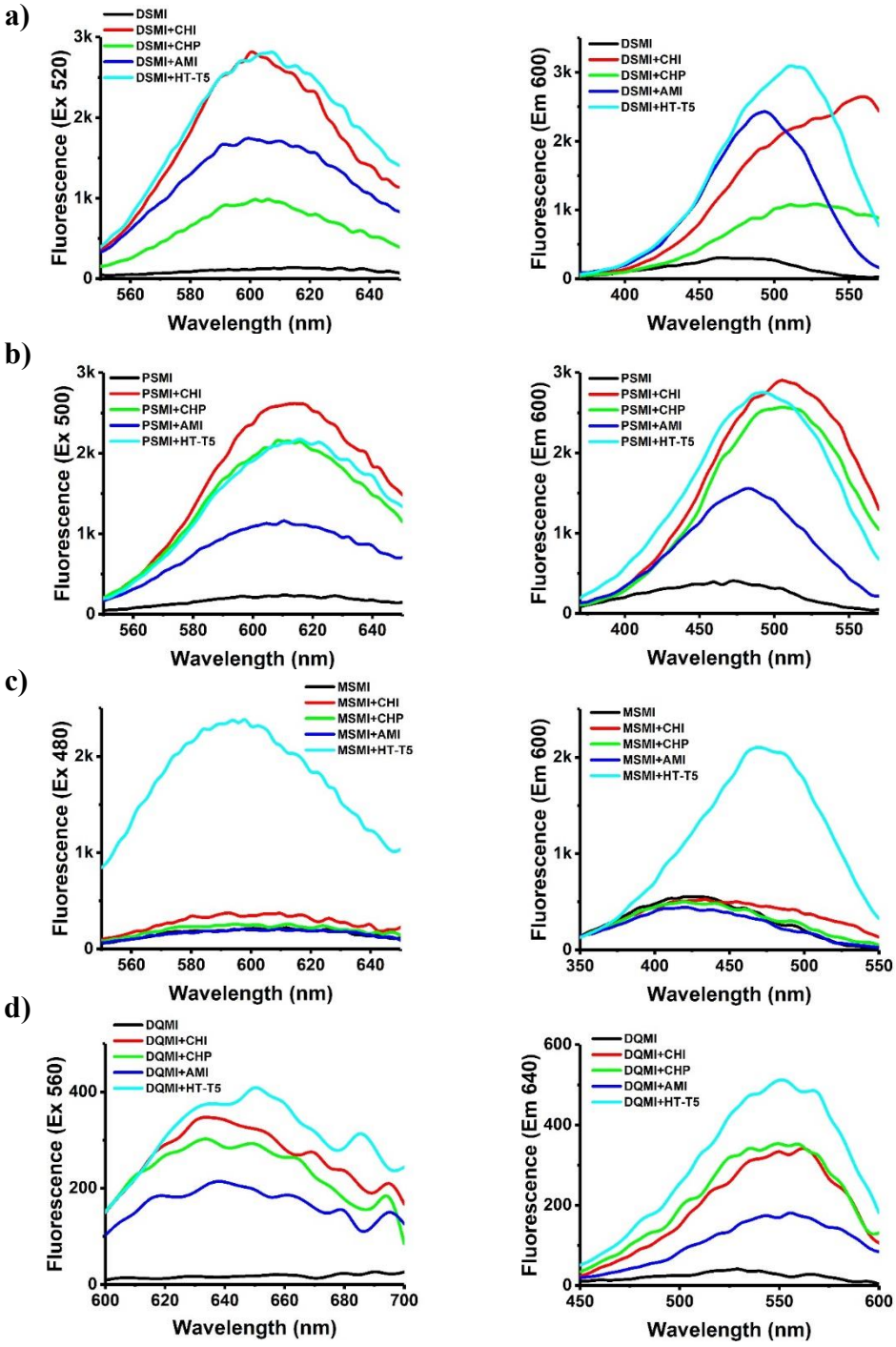


Figure *S*-7. Emission (left) and excitation (right) fluorescence spectra of Dye a) **DSMI**, b) **PSMI**, c) **MSMI**, and d) **DQMI** with **CHI/CHP/AMI**/HT-T5. Left: emission spectra; Right: excitation spectra. [Dye] = 0.625 μM, [**CHI/CHP/AMI**] = 4 μM, [HT-T5] = 0.2 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

3.2 Fluorescence Titration of Dye-DNA

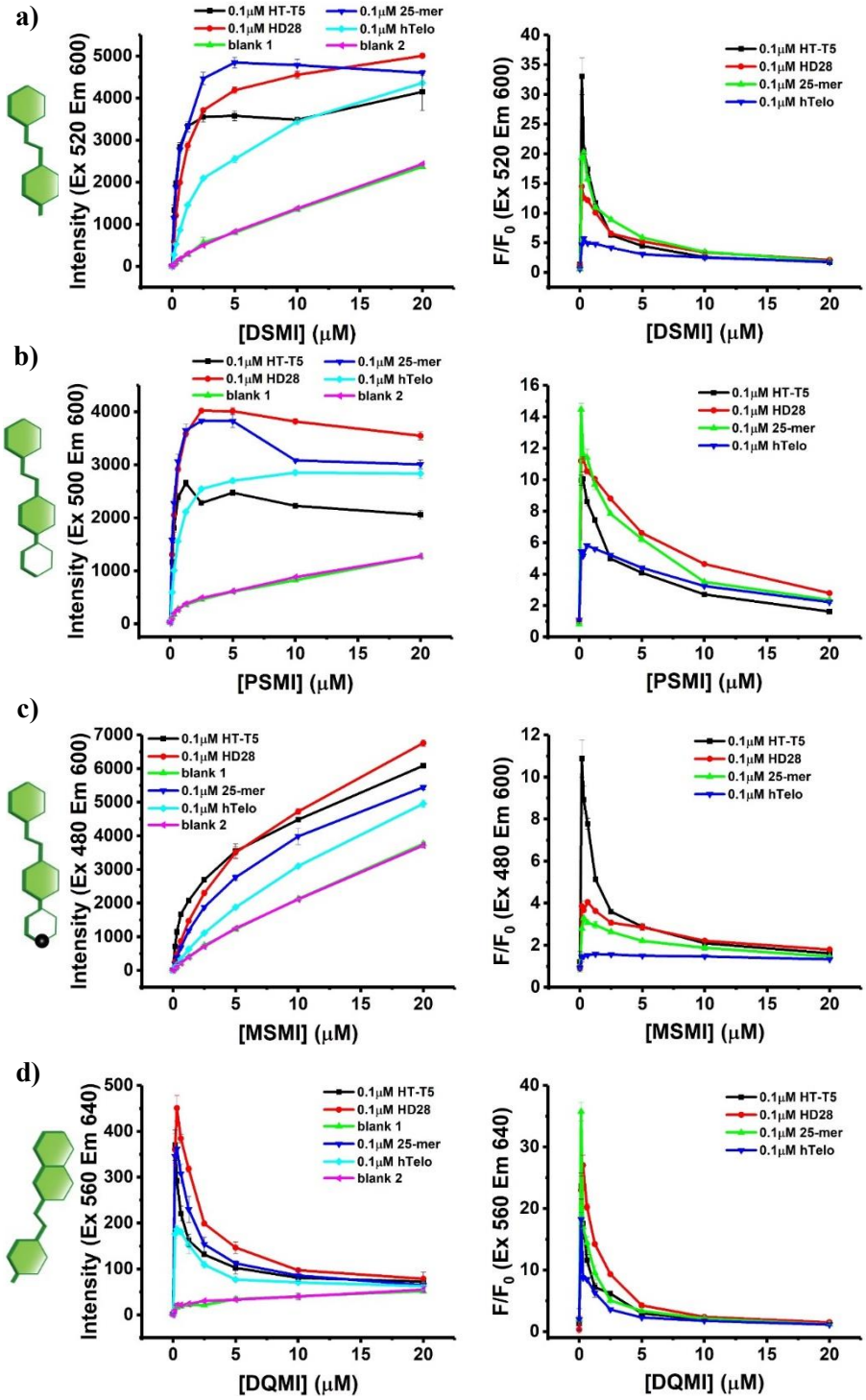


Figure *S*-8. Fluorescence response curves of HT-T5, HD28, 25-mer hairpin or hTelo with increasing concentration (0-20 μM) of Dye a) **DSMI**, b) **PSMI**, c) **MSMI**, and d) **DQMI**. Left: plots using the raw fluorescence counts; Right: plots using the fluorescence normalized against that of the dye (F_0 being the dye fluorescence in the absence of DNA). [Dye] = 0-20 μM, [DNA] = 0.1 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

3.3 Fluorescence Titration of Host Addition to Dye•DNA Complexes

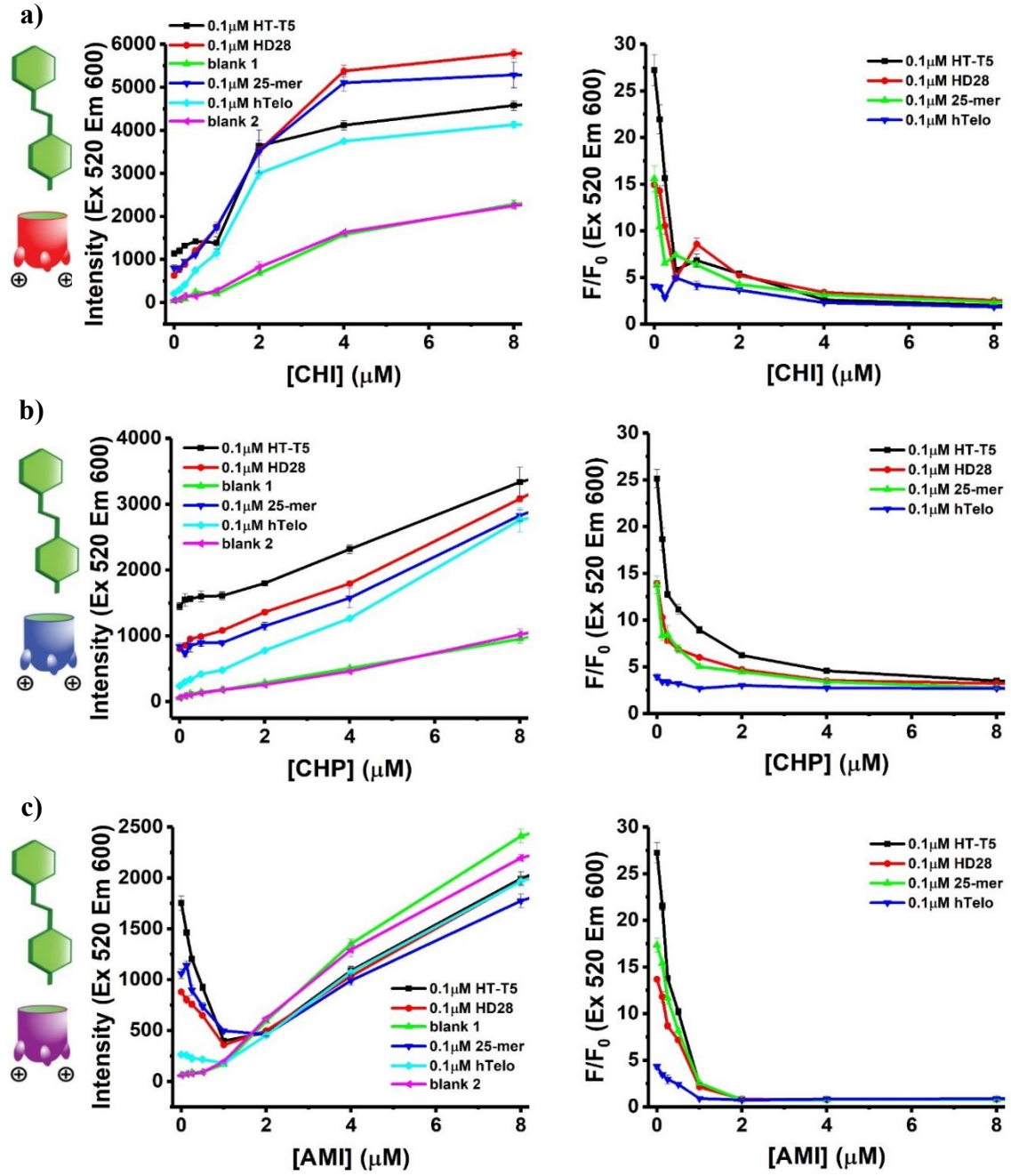


Figure *S*-9. Fluorescence response curves of **DSMI**•DNA complexes upon titration of hosts: a) **CHI**, b) **CHP**, and c) **AMI**. Left: the raw fluorescence counts (**DSMI** + Host); Right: plots normalized to the response of cavitand-**DSMI** in the absence of DNA (F_0). [**DSMI**] = 0.15625 μM, [DNA] = 0.1 μM, [Host] = 0-16 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

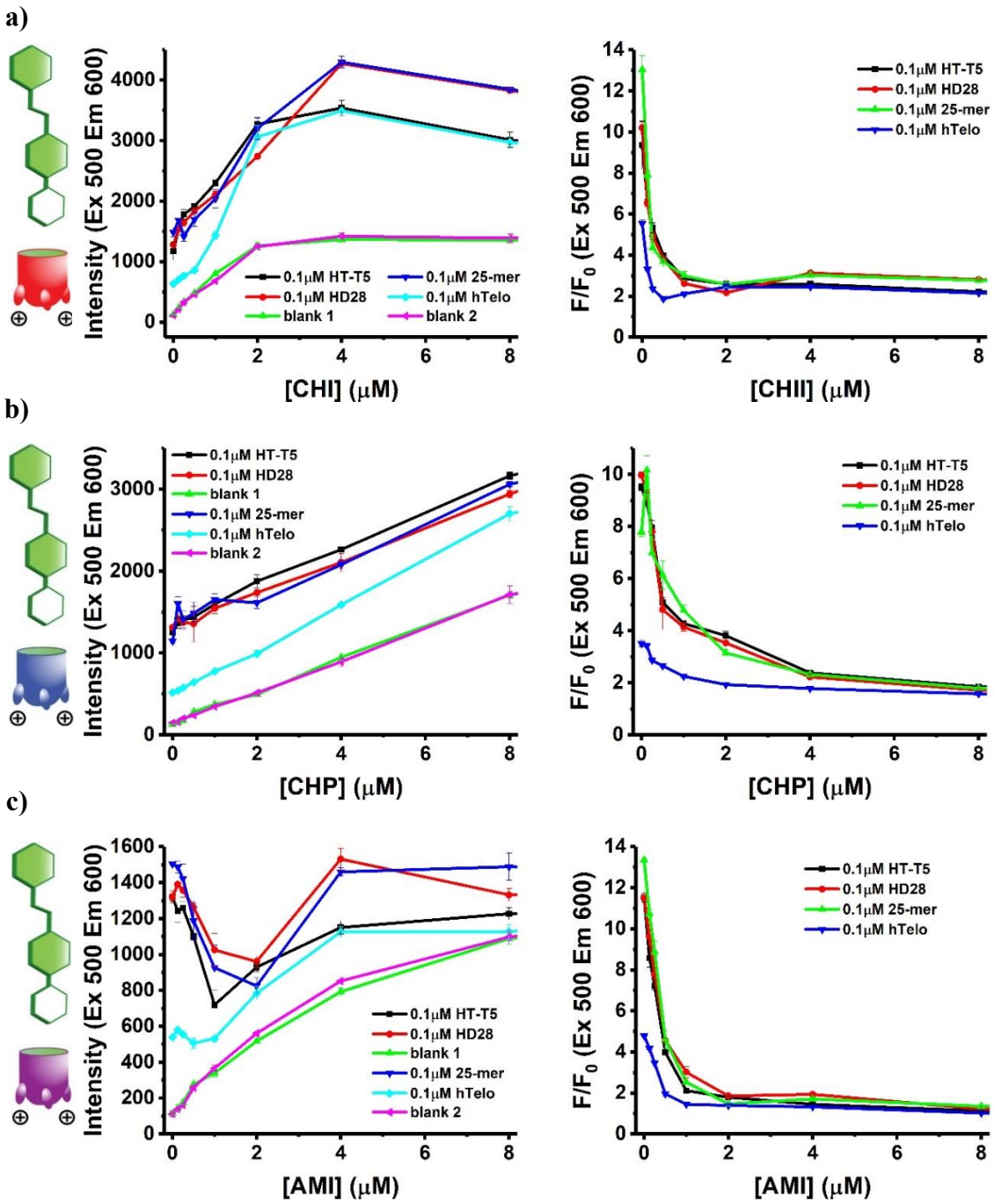


Figure *S*-10. Fluorescence response curves of **PSMI**•DNA complexes upon titration of hosts: a) **CHI**, b) **CHP**, and c) **AMI**. Left: the raw fluorescence counts (**PSMI** + Host); Right: plots normalized to the response of cavitand-**PSMI** in the absence of DNA (F_0). [**PSMI**] = 0.15625 μM, [DNA] = 0.1 μM, [Host] = 0-16 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

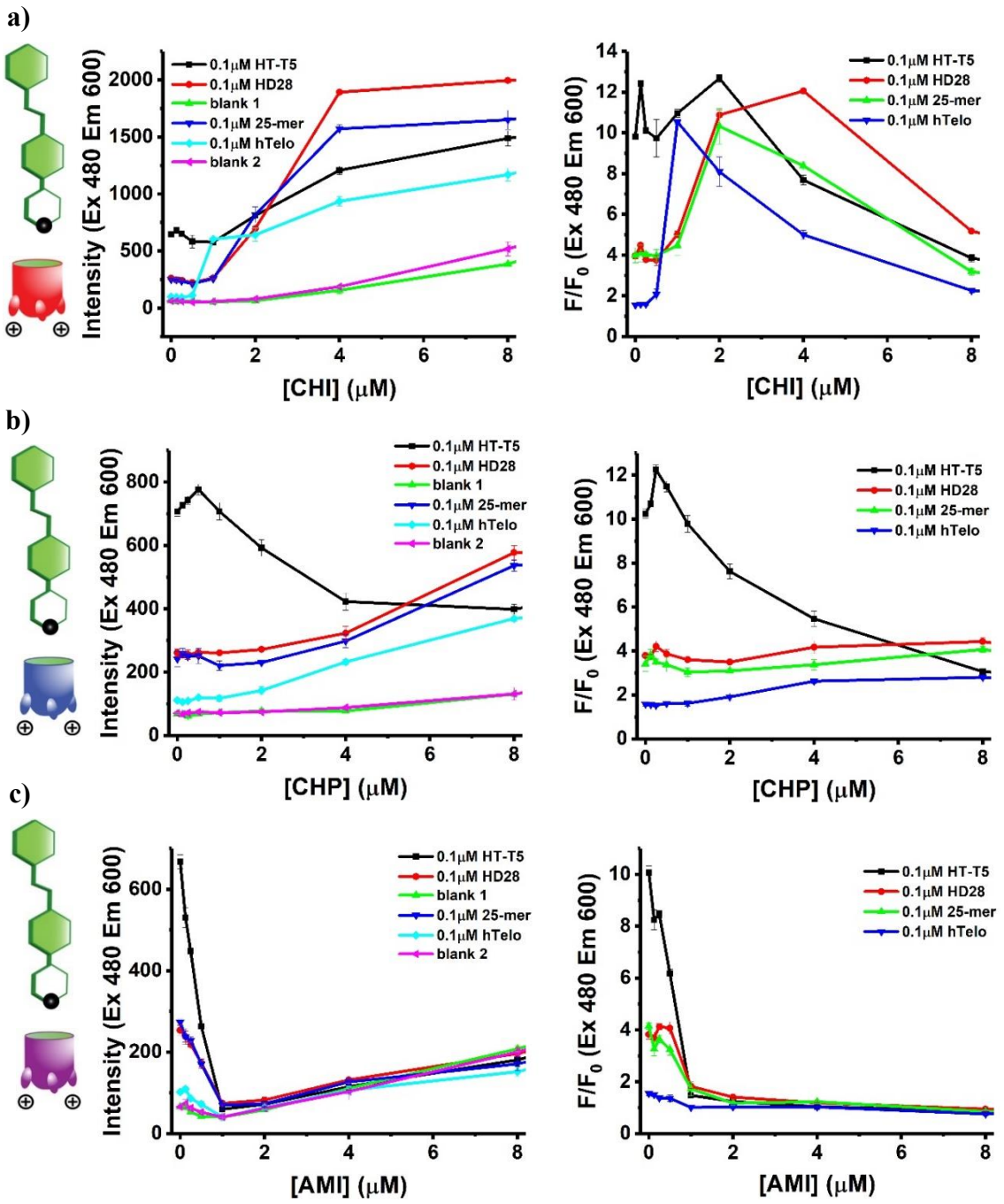


Figure *S*-11. Fluorescence response curves of **MSMI**•DNA complexes upon titration of hosts: a) **CHI**, b) **CHP**, and c) **AMI**. Left: the raw fluorescence counts (**MSMI** + Host); Right: plots normalized to the response of cavitand-**MSMI** in the absence of DNA (F_0). [**MSMI**] = 0.15625 μM, [DNA] = 0.1 μM, [Host] = 0-16 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

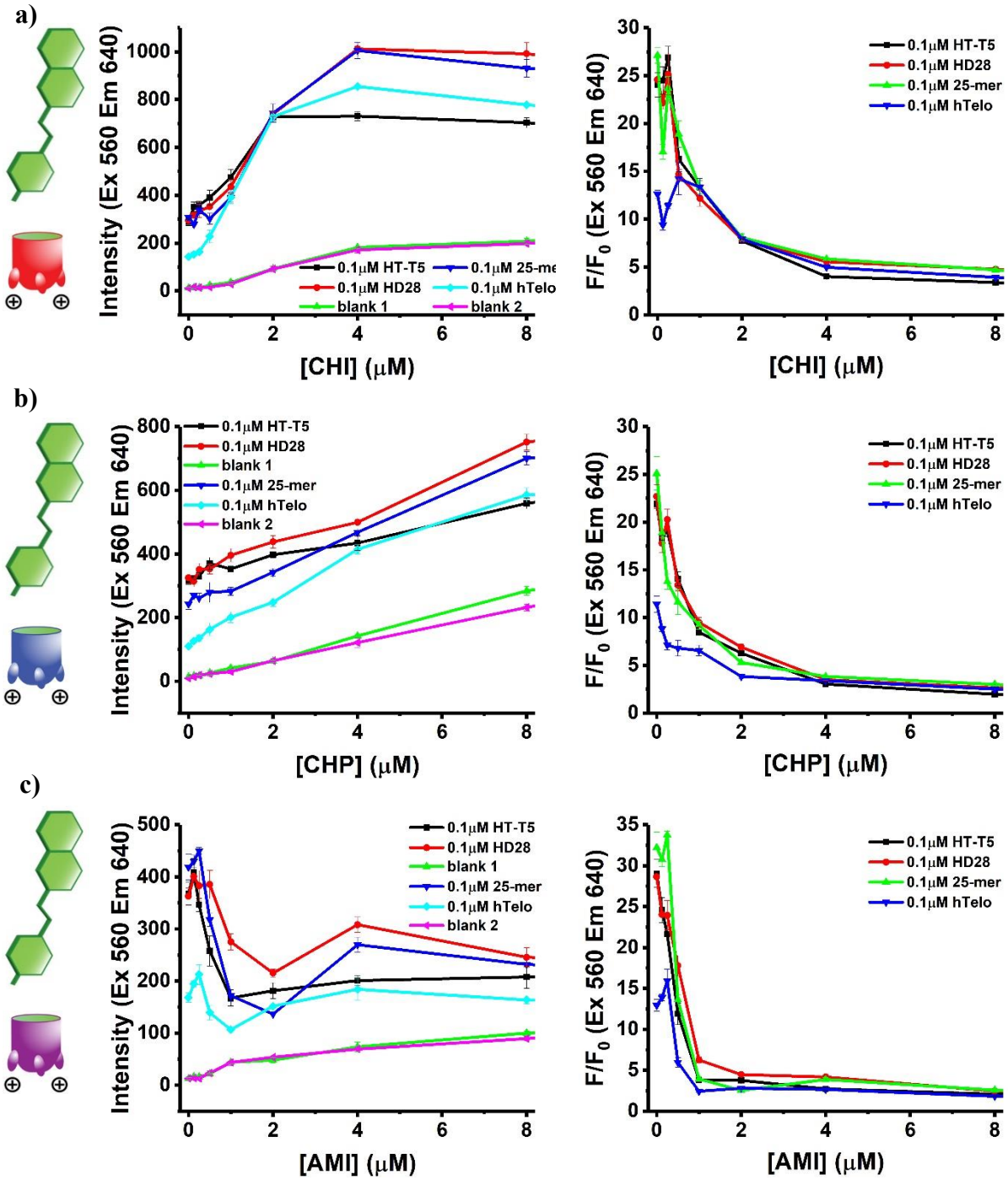


Figure *S*-12. Fluorescence response curves of **DQMI**•DNA complexes upon titration of hosts: a) **CHI**, b) **CHP**, and c) **AMI**. Left: the raw fluorescence counts (**DQMI** + Host); Right: plots normalized to the response of cavitand-**DQMI** in the absence of DNA (F_0). [**DQMI**] = 0.15625 μM, [DNA] = 0.1 μM, [Host] = 0-16 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

4. Array Analysis for Differentiation of 7 DNAs

4.1 Bar Plots for Array Signals from 7 DNAs

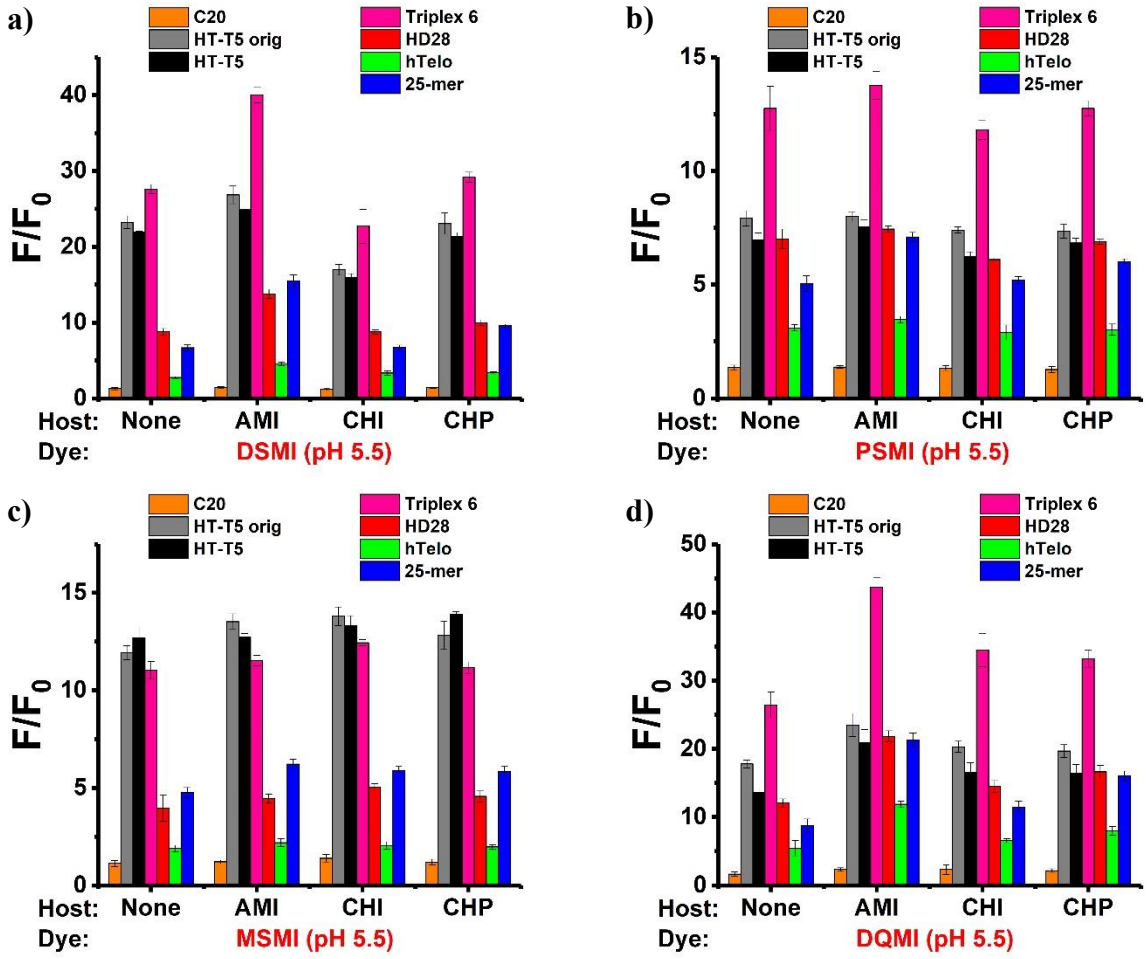


Figure *S*-13. Full fluorescence response plots of 7 DNA sequences, obtained with the full 16-element array: 4 dyes a) **DSMI**, b) **PSMI**, c) **MSMI**, and d) **DQMI** with **CHI/CHP/AMI**/No cavitand. [Dye] = 0.15625 μM, [Host] = 0.125 μM, [DNA] = 0.1 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer. **DSMI** Ex/Em = 520/600nm, **PSMI** Ex/Em=500/600nm, **MSMI** Ex/Em=480/600nm, **DQMI** Ex/Em=560/640nm.

4.2 PCA Plots with Different Host:Guest Array Elements Combinations

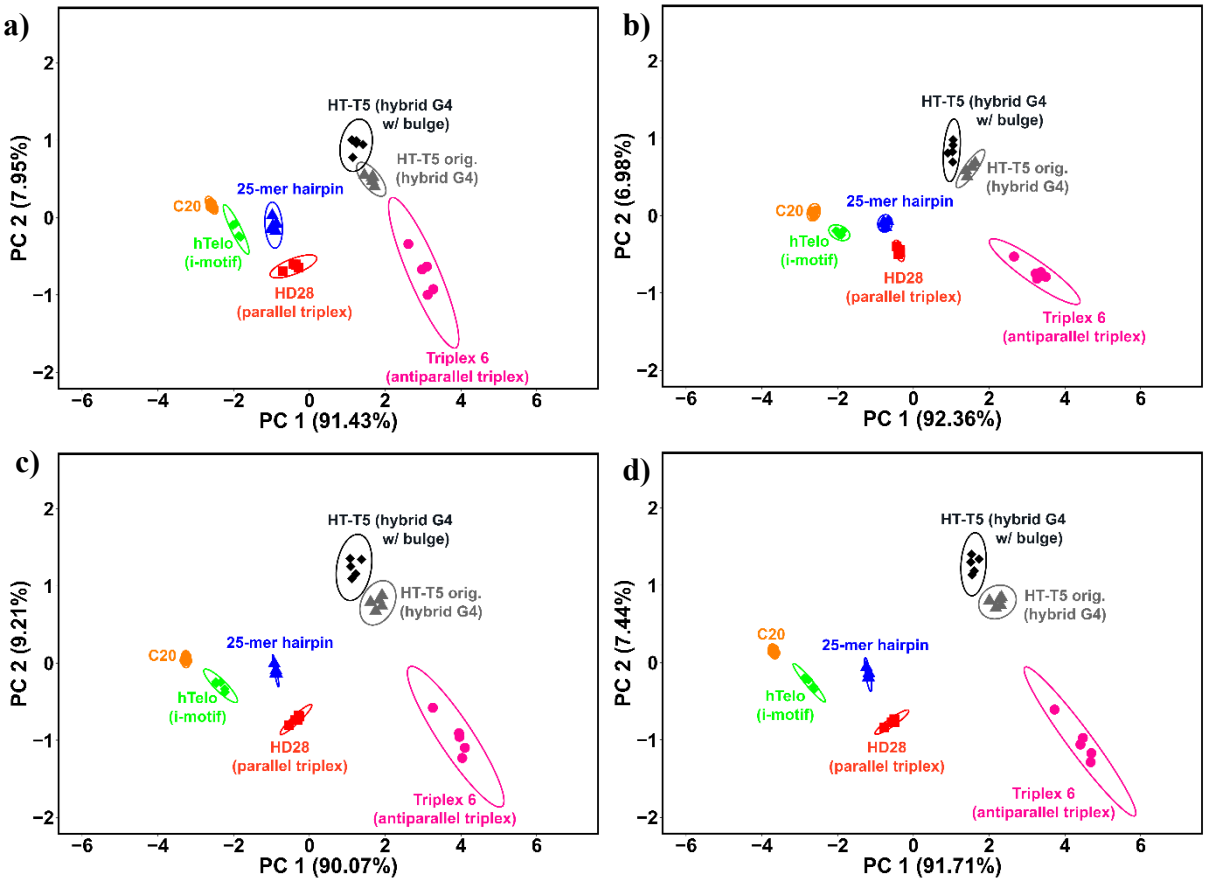


Figure *S*-14. PCA scores plot for selective sensing of 7 DNA sequences using a) four dyes only, b) four dye with **CHI** components, c) 6 elements: **DQMI/PSMI/MSMI** dyes only + **DQMI/PSMI/MSMI** with **CHI** components, d) 8 elements: four dyes only+four dyes with **CHI** components. [Dye] = 0.15625 μM, [Host] = 0.125 μM, [DNA] = 0.1 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer. Ellipses indicate 95% confidence.

5. Array Analysis for Sensing 18 DNAs

5.1 Bar Plots for Array Signals from 18 DNAs

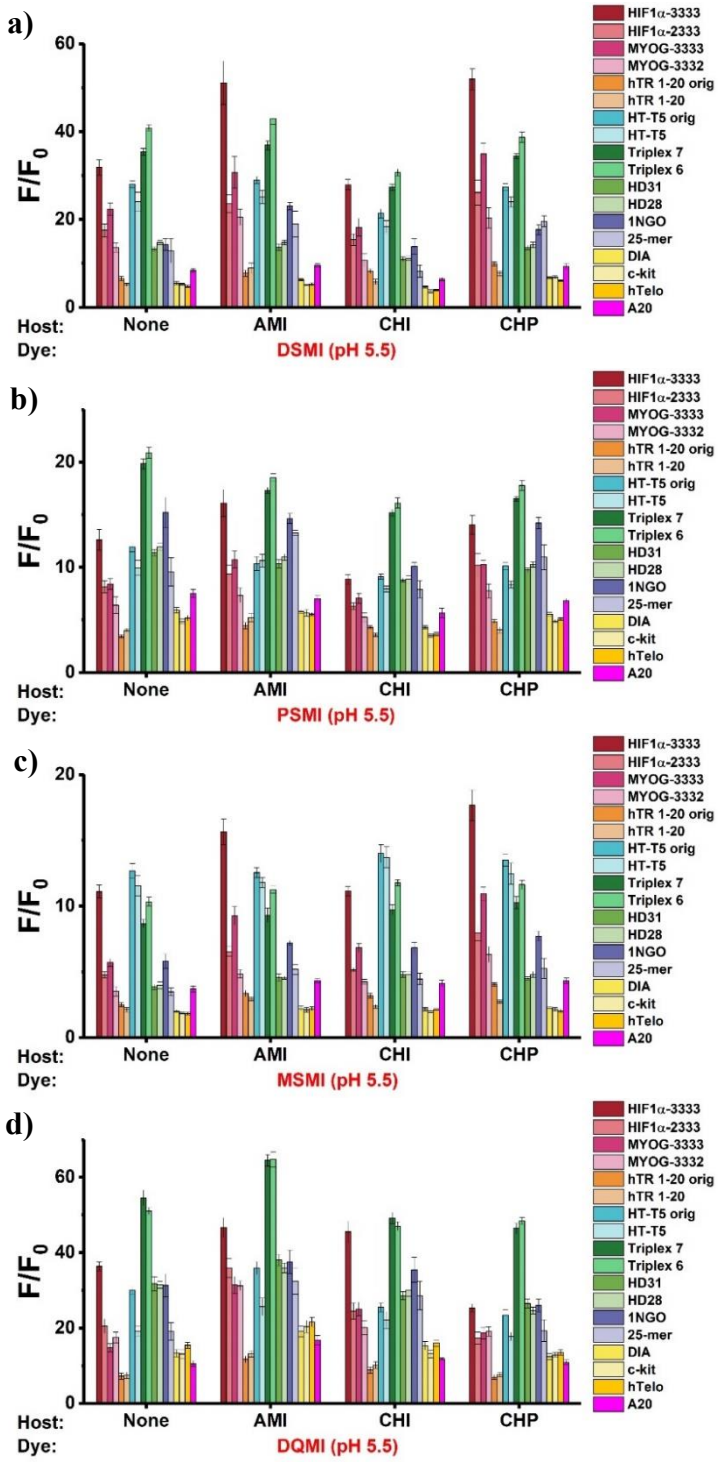


Figure *S*-15. Full fluorescence response plots of 18 DNA sequences, obtained with the full 16-element array: 4 dyes a) **DSMI**, b) **PSMI**, c) **MSMI**, and d) **DQMI** with **CHI/CHP/AMI**/No cavitand. Sensor conditions identical to those described in Figure *S*-13.

5.2 SVM plot of 18 DNA training set with 16-element Host:Guest Array

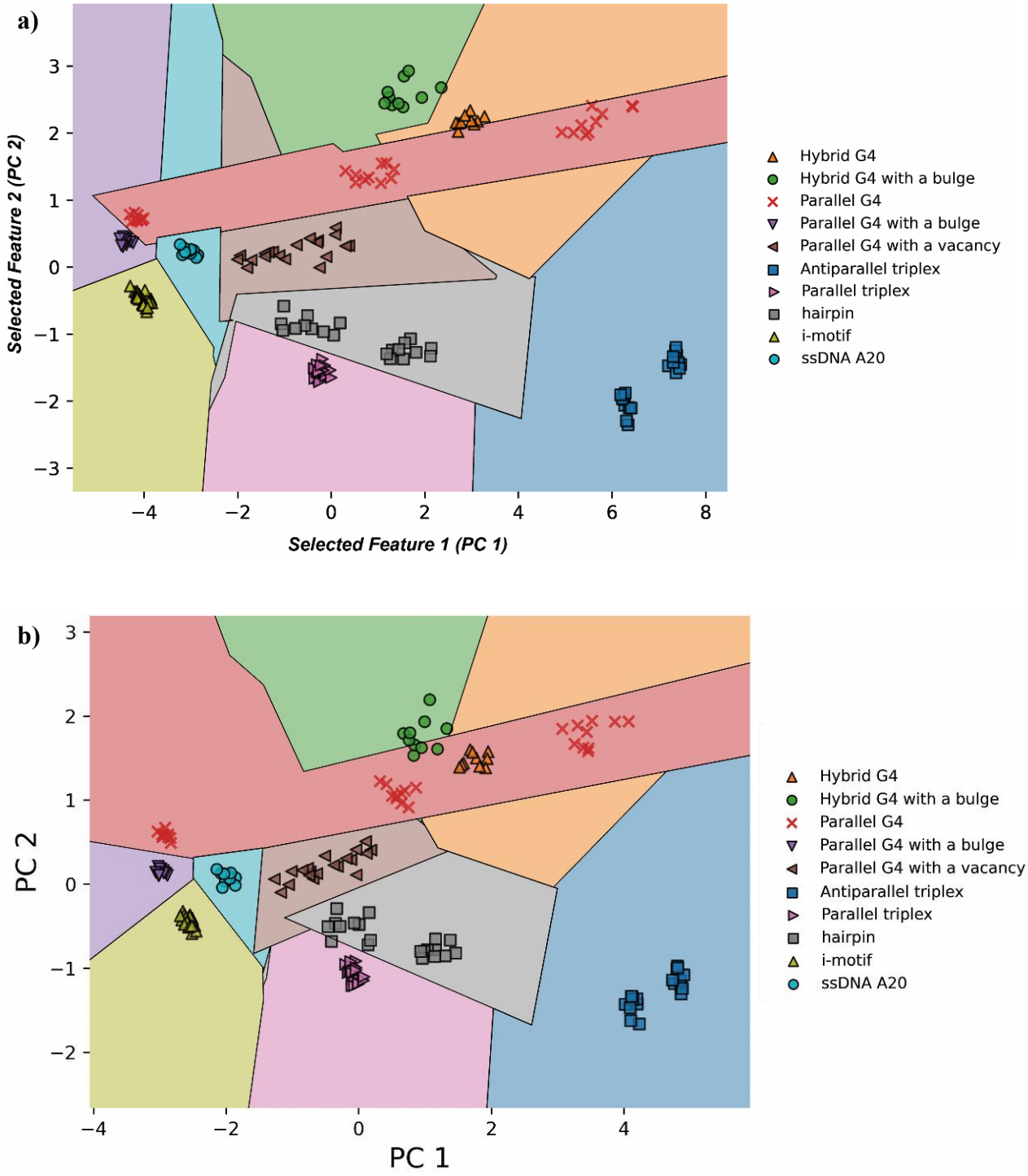


Figure *S*-16. SVM decision boundary plot of 2D PCA (using PC1 and PC 2) for classifying ten DNA classes of the training set using a) full 16-element array; or b) selected 7-element array. Sensor conditions identical to those described in Figure *S*-13.

5.3 Performance metrics of DNA Structures Classification

Table *S*-2. Performance metrics of 3 repeated 8-fold cross validation.

Evaluation Metrics	Score (standard deviation from 3 repeated running of the 8-fold cross validation)
Accuracy	0.9812 (0.0289)
Sensitivity	0.9688 (0.0475)
Specificity	0.9978 (0.0034)
Precision	0.9640 (0.0588)
F1 Score	0.9642 (0.0554)
AUC	0.9998 (0.0008)

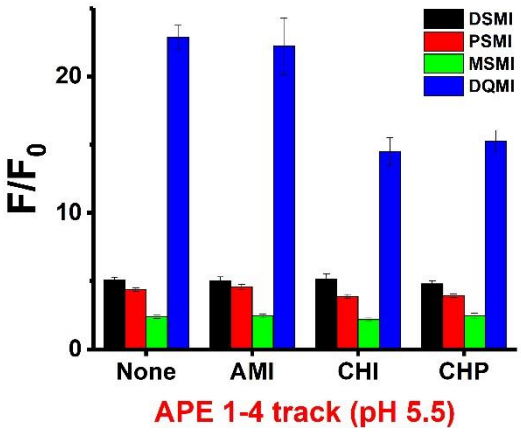
Table *S*-3. Performance metrics of each DNA class.

Class	Sensitivity	Specificity	Precision	Accuracy	AUC
Antiparallel triplex	1.0000	1.0000	1.0000	1.0000	1.0000
Hybrid G4	0.7000	0.9922	0.8380	0.9759	0.9986
Hybrid G4 with a bulge	0.9000	1.0000	1.0000	0.9944	1.0000
Parallel G4	0.9889	0.9800	0.9085	0.9815	0.9991
Parallel G4 with a bulge	1.0000	1.0000	1.0000	1.0000	1.0000
Parallel G4 with a vacancy	1.0000	1.0000	1.0000	1.0000	1.0000
Parallel triplex	1.0000	1.0000	1.0000	1.0000	1.0000
hairpin	1.0000	1.0000	1.0000	1.0000	1.0000
i-motif	1.0000	1.0000	1.0000	1.0000	1.0000
ssDNA	1.0000	1.0000	1.0000	1.0000	1.0000

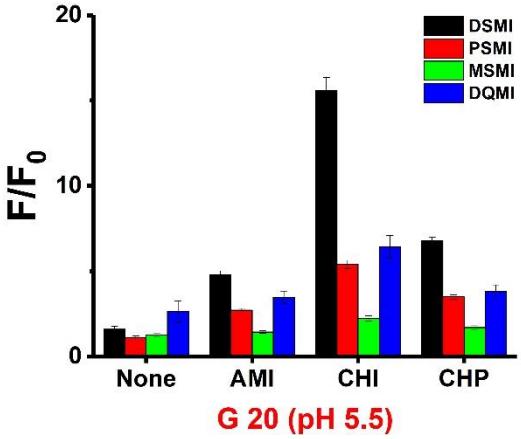
6. Folding prediction of unknown DNA structures

6.1 Fluorescence response plots of unknown DNA structures

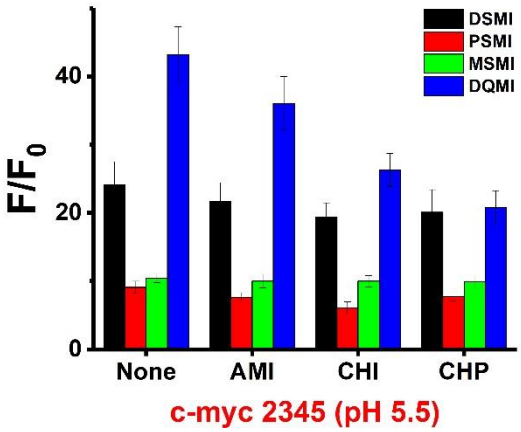
a) APE 1-4 track



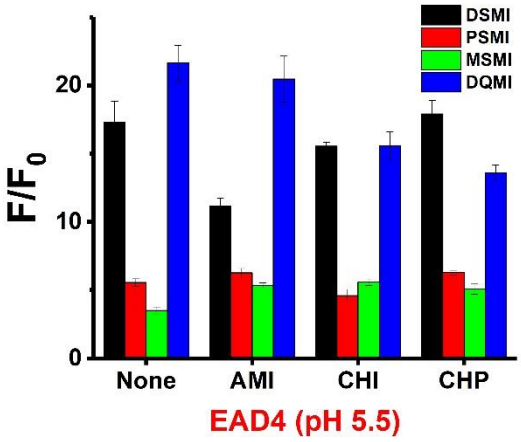
b) G 20



c) c-myc 2345



d) EAD4



e) Telo24

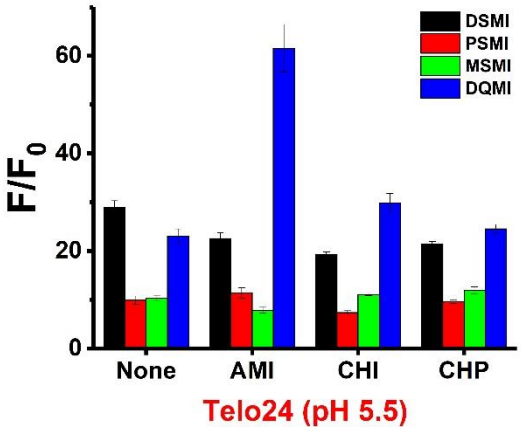


Figure *S*-17. Full fluorescence response plots of unknown DNA a) APE 1-4 track, b) G 20, c) c-myc 2345, d) EAD4, and e) Telo24 obtained with the full 16-element array. Sensor conditions identical to those described in Figure *S*-13.

6.2 SVM decision boundary plot for prediction results

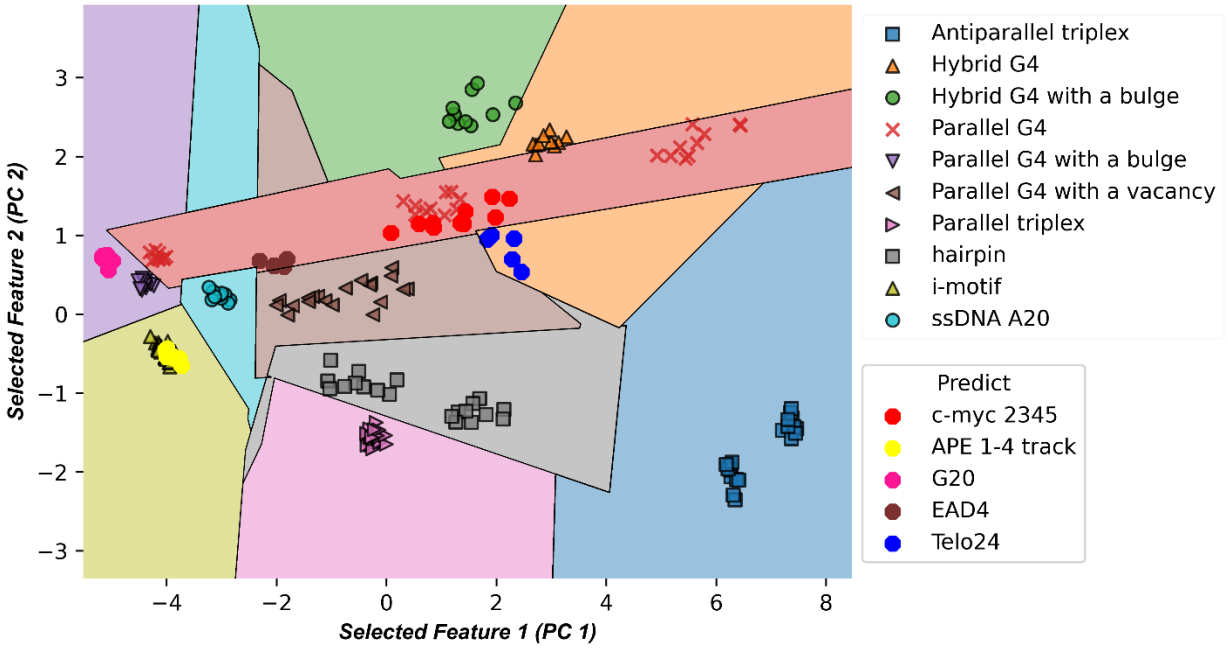


Figure *S*-18. Prediction of unknown DNAs by using the model of 18 DNA training set with 16-element Host:Guest Array. Sensor conditions identical to those described in Figure *S*-13, and the image is identical to that shown in Figure 5, with the DNAs labeled by folding type, not individual strand.

7. UV-Vis Absorbance Spectra

7.1 UV-Vis Spectra of Dye-DNA

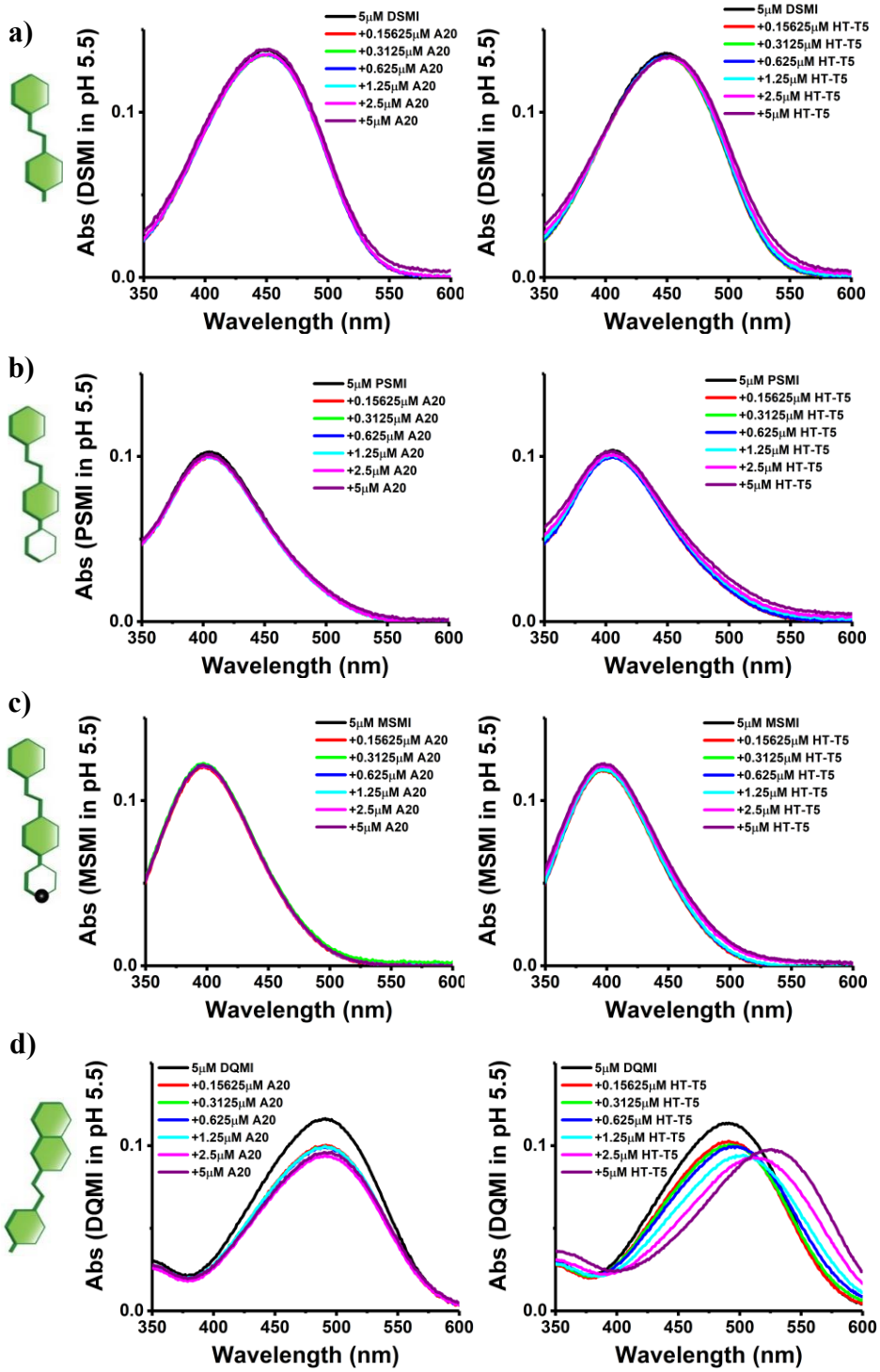


Figure *S*-19. UV spectra of Dyes 4 dyes a) **DSMI**, b) **PSMI**, c) **MSMI**, and d) **DQMI** with increasing concentration of DNA. Left: Dye+A20; Right: Dye+HT-T5. [**Dye**] = 5 μM, [DNA] = 0-5 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

7.2 UV-Vis Spectra of DNA Addition to Dye•Host Complexes

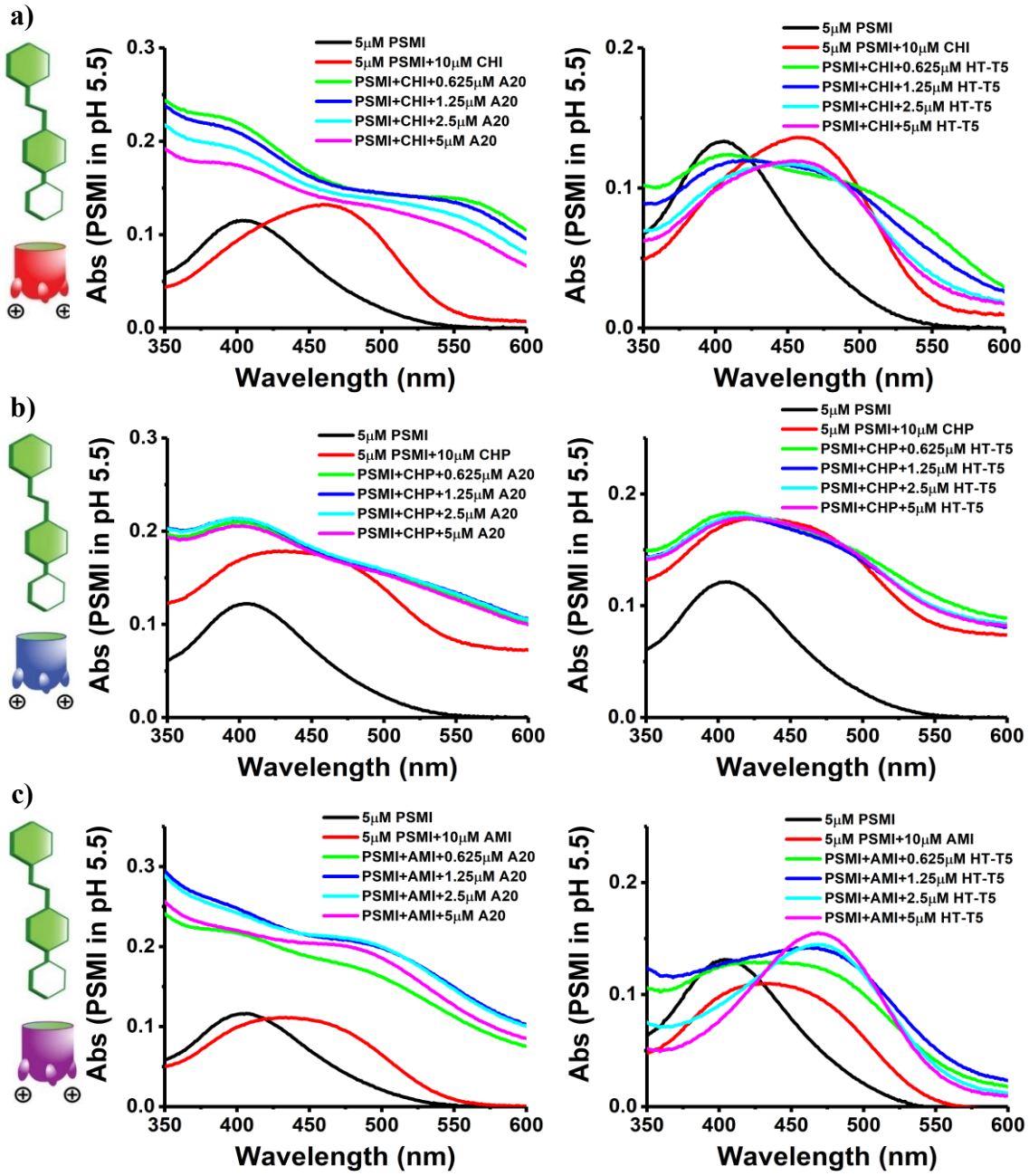


Figure *S*-20. UV spectra of **PSMI**•Host complexes a) **CHI**, b) **CHP**, and c) **AMI** upon titration of DNA. Left: **PSMI**+Host+A20; Right: **PSMI**+Host+HT-T5. [**PSMI**] = 5 μM, [Host] = 10 μM, [DNA] = 0-5 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

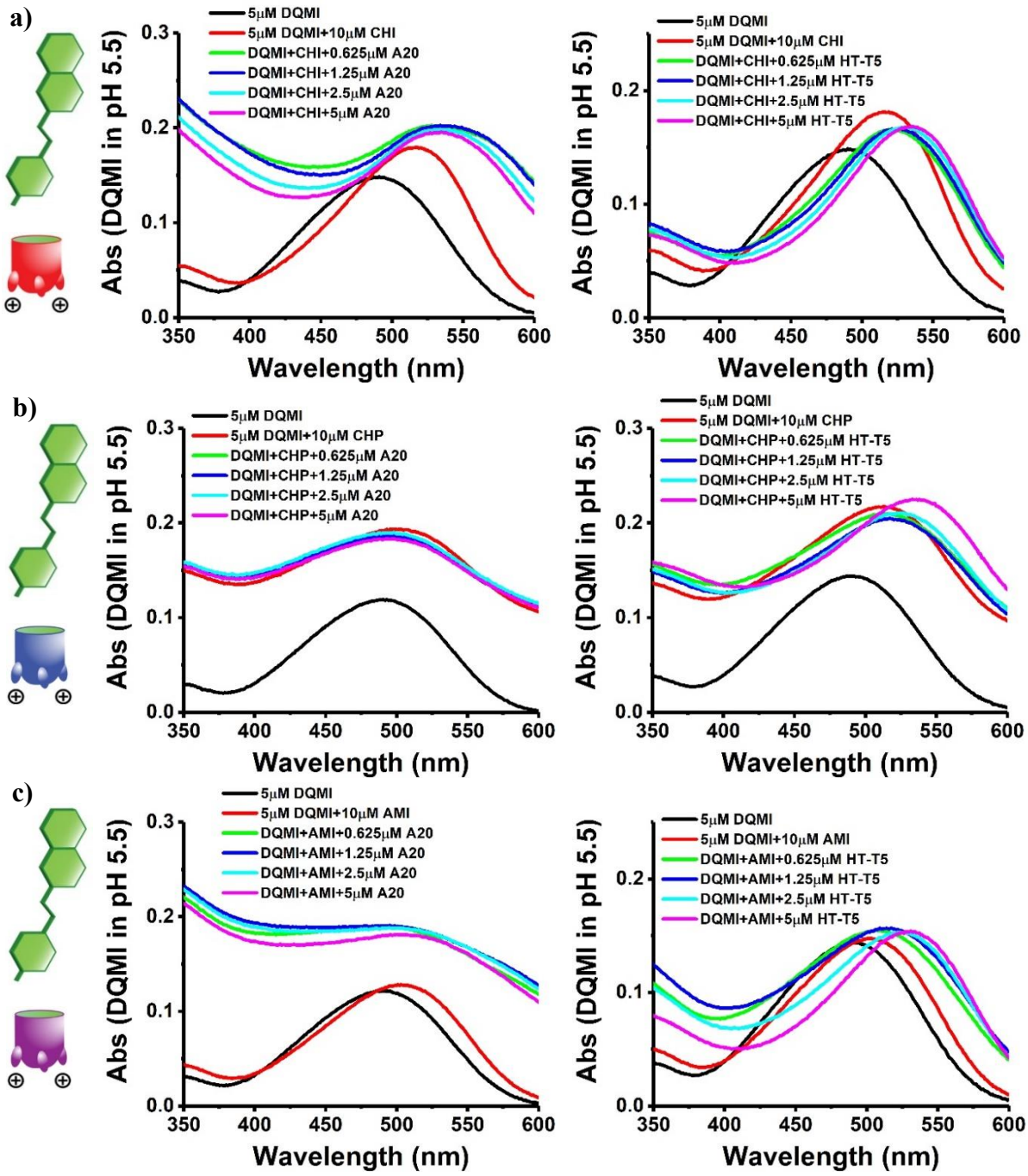


Figure *S*-21. UV spectra of **DQMI**•Host complexes a) **CHI**, b) **CHP**, and c) **AMI** upon titration of DNA. Left: **DQMI**+Host+A20; Right: **DQMI**+Host+HT-T5. [**DQMI**] = 5 μM, [Host] = 10 μM, [DNA] = 0-5 μM, 20mM CH_3COOK, 5mM $MgCl_2$, pH 5.5 buffer.

8. References

1. He, Y.-D.; Zheng, K.-W.; Wen, C.-J.; Li, X.-M.; Gong, J.-Y.; Hao, Y.-H.; Zhao, Y.; Tan, Z. Selective Targeting of Guanine-Vacancy-Bearing G-Quadruplexes by G-Quartet Complementation and Stabilization with a Guanine–Peptide Conjugate. *J. Am. Chem. Soc.* **2020**, *142*, 11394-11403.
2. Li, X.-M.; Zheng, K.-W.; Zhang, J.-Y.; Liu, H.-H.; He, Y.-D.; Yuan, B.-F.; Hao, Y.-H.; Tan, Z. Guanine-vacancy–bearing G-quadruplexes Responsive to Guanine Derivatives. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 14581-14586.
3. Meier, M.; Moya-Torres, A.; Krahn, N. J.; McDougall, M. D.; Orriss, G. L.; McRae, E. K. S.; Booy, E. P.; McEleney, K.; Patel, T. R.; McKenna, S. A.; Stetefeld, J. Structure and Hydrodynamics of a DNA G-quadruplex with a Cytosine Bulge. *Nucleic Acids Res.* **2018**, *46*, 5319-5331.
4. Mukundan, V. T.; Phan, A. T. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *J. Am. Chem. Soc.* **2013**, *135*, 5017-5028.
5. Phan, A. T.; Guéron, M.; Leroy, J.-L. The Solution Structure and Internal Motions of a Fragment of the Cytidine-rich Strand of the Human Telomere. *J. Mol. Biol.* **2000**, *299*, 123-144.
6. Day, H. A.; Pavlou, P.; Waller, Z. A. E., i-Motif DNA: Structure, Stability and Targeting with Ligands. *Bioorg. Med. Chem.* **2014**, *22*, 4407-4418.
7. Choi, J.; Kim, S.; Tachikawa, T.; Fujitsuka, M.; Majima, T. pH-Induced Intramolecular Folding Dynamics of i-Motif DNA. *J. Am. Chem. Soc.* **2011**, *133*, 16146-16153.
8. Sklenář, V.; Felgon, J. Formation of a Stable Triplex from a Single DNA Strand. *Nature* **1990**, *345*, 836-838.
9. Macaya, R.; Wang, E.; Schultze, P.; Sklenář, V.; Feigon, J., Proton Nuclear Magnetic Resonance Assignments and Structural Characterization of an Intramolecular DNA Triplex. *J. Mol. Biol.* **1992**, *225*, 755-773.
10. Radhakrishnan, I.; de los Santos, C.; Patel, D. J. Nuclear Magnetic Resonance Structural Studies of Intramolecular Purine · Purine · Pyrimidine DNA Triplexes in Solution: Base Triple Pairing Alignments and Strand Direction. *J. Mol. Biol.* **1991**, *221*, 1403-1418.
11. Ma, H.; Wan, C.; Wu, A.; Zewail, A. H. DNA Folding and Melting Observed in Real Time Redefine the Energy Landscape. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 712-716.
12. Shiflett, P. R.; Taylor-McCabe, K. J.; Michalczyk, R.; Silks, L. A. P.; Gupta, G. Structural Studies on the Hairpins at the 3‘ Untranslated Region of an Anthrax Toxin Gene. *Biochemistry* **2003**, *42*, 6078-6089.
13. Greve, J.; Maestre, M. F.; Levin, A. Circular Dichroism of Adenine and Thymine Containing Synthetic Polynucleotides. *Biopolymers* **1977**, *16*, 1489-1504.
14. Chakraborty, S.; Sharma, S.; Maiti, P. K.; Krishnan, Y. The Poly dA Helix: a New Structural Motif for High Performance DNA-based Molecular Switches. *Nucleic Acids Res.* **2009**, *37*, 2810-2817.
15. Rogers, R. A.; Fleming, A. M.; Burrows, C. J. Unusual Isothermal Hysteresis in DNA i-Motif pH Transitions: A Study of the RAD17 Promoter Sequence. *Biophys. J.* **2018**, *114*, 1804-1815.
16. Rogers, R. A.; Fleming, A. M.; Burrows, C. J. Rapid Screen of Potential i-Motif Forming Sequences in DNA Repair Gene Promoters. *ACS Omega* **2018**, *3*, 9630-9635.
17. Jin, M.; Liu, X.; Zhang, X.; Wang, L.; Bing, T.; Zhang, N.; Zhang, Y.; Shangguan, D. Thiazole Orange-Modified Carbon Dots for Ratiometric Fluorescence Detection of G-

Quadruplex and Double-Stranded DNA. *ACS Appl. Mater. Interfaces* **2018**, *10*, 25166-25173.

18. Yang, Q.; Xiang, J.; Yang, S.; Li, Q.; Zhou, Q.; Guan, A.; Zhang, X.; Zhang, H.; Tang, Y.; Xu, G. Verification of Specific G-quadruplex Structure by Using a Novel Cyanine Dye Supramolecular Assembly: II. The Binding Characterization with Specific Intramolecular G-quadruplex and the Recognizing Mechanism. *Nucleic Acids Res.* **2010**, *38*, 1022-1033.
19. Lin, D.; Fei, X.; Gu, Y.; Wang, C.; Tang, Y.; Li, R.; Zhou, J. A Benzindole Substituted Carbazole Cyanine Dye: a Novel Targeting Fluorescent Probe for Parallel c-myc G-quadruplexes. *Analyst* **2015**, *140*, 5772-5780.
20. Sengar, A.; Heddi, B.; Phan, A. T. Formation of G-Quadruplexes in Poly-G Sequences: Structure of a Propeller-Type Parallel-Stranded G-Quadruplex Formed by a G15 Stretch. *Biochemistry* **2014**, *53*, 7718-7723.
21. Panyutin, I. G.; Kovalsky, O. I.; Budowsky, E. I.; Dickerson, R. E.; Rikhirev, M. E.; Lipanov, A. A. G-DNA: a Twice-folded DNA Structure Adopted by Single-stranded Oligo(dG) and Its Implications for Telomeres. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 867-870.
22. Joly, L.; Rosu, F.; Gabelica, V. d(TG_nT) DNA Sequences Do Not Necessarily Form Tetramolecular G-quadruplexes. *Chem. Commun.* **2012**, *48*, 8386-8388.
23. Cheng, X.; Liu, X.; Bing, T.; Cao, Z.; Shangguan, D. General Peroxidase Activity of G-Quadruplex−Hemin Complexes and Its Application in Ligand Screening. *Biochemistry* **2010**, *48*, 7817-7823.
24. Ma, Y.; Tsushima, Y.; Sakuma, M.; Sasaki, S.; Iida, K.; Okabe, S.; Seimiya, H.; Hirokawa, T.; Nagasawa, K. Development of G-quadruplex ligands for selective induction of a parallel-type topology. *Org. Biomol. Chem.* **2018**, *16*, 7375-7382.