Fluorescence spectral shape analysis for nucleotide identification

Yun Huang[a,b,1], Zhiliang Li[a,b,1], April L. Risinger[c,2], Benjamin T. Enslow[d], Charles J. Zeman IV[a,b], Jiang Gong[e], Yajing Yang[a,b], and Kirk S. Schanze[a,2]

[a]Department of Chemistry, University of Texas at San Antonio, San Antonio, TX 78249; [b]Department of Chemistry, University of Florida, Gainesville, FL 32611-7200; [c]Department of Pharmacology, University of Texas Health Science Center, San Antonio, TX 78229; [d]Long School of Medicine, University of Texas Health Science Center, San Antonio, TX 78229; and [e]Key Laboratory for Material Chemistry of Energy Conversion and Storage, Ministry of Education, School of Chemistry and Chemical Engineering, Huazhong University of Science and Technology, 430074 Wuhan, People's Republic of China

Edited by Vivian Wing-Wah Yam, University of Hong Kong, Hong Kong, and approved June 19, 2019 (received for review December 10, 2018)

We report a conjugated polyelectrolyte fluorescence-based biosensor P-C-3 and a general methodology to evaluate spectral shape recognition to identify biomolecules using artificial intelligence. By using well-defined analytes, we demonstrate that the fluorescence spectral shape of P-C-3 is sensitive to minor structural changes and exhibits distinct signature patterns for different analytes. A method was also developed to select useful features to reduce computational complexity and prevent overfitting of the data. It was found that the normalized intensity of 3 to 5 selected wavelengths was sufficient for the fluorescence biosensor to classify 13 distinct nucleotides and distinguish as little as single base substitutions at distinct positions in the primary sequence of oligonucleotides rapidly with nearly 100% classification accuracy. Photophysical studies led to a model to explain the mechanism of these fluorescence spectral shape changes, which provides theoretical support for applying this method in complicated biological systems. Using the feature selection algorithm to measure the relative intensity of a few selected wavelengths significantly reduces measurement time, demonstrating the potential for fluorescence spectrum shape analysis in high-throughput and high-content screening.

machine learning | fluorescence | phosphate sensing | biosensor | nucleic acids

Over the past 2 decades, major technological innovations in screening as well as developments in artificial intelligence have rapidly advanced the field of biomedical research (1, 2). High-content screening has aided in the identification of novel drug candidates, therapeutic targets, prognostic markers, and has allowed for deeper insight into many different complex biological phenomena (3–5). A trade-off, however, exists between accuracy and universality for the biosensors used in such analyses. Most biosensors are specifically designed to recognize a single target and provide a one-dimensional signal, such as the intensity of a single wavelength from an emission spectrum (5). More complex multidimensional methodologies require the use of multiple probes and complex readouts to accurately assess complex biological systems. The selection of biosensors for such analyses often requires a priori knowledge about the target; therefore, the interpretation of results may be vulnerable to bias and limit the discovery of unanticipated observations, and the latent factors are difficult to capture. Higher-dimensional information of emission spectra, such as emission shape, remains insufficiently mined as it has proven difficult to rapidly quantify the shape change of spectra with high accuracy.

Conjugated polyelectrolytes (CPEs) are widely utilized in multiple applications (6, 7) such as organic solar cells, photodynamic therapy (8), and chemo- and biosensors (9, 10). Zhou and Swager (11) first demonstrated the effect of amplified quenching, which is able to amplify the fluorescence intensity change of conjugated polymers due to excitation energy delocalization and rapid transmission along the conjugated backbone. Both "turn-off" and "turn-on" CPE sensors have been developed based on amplified fluorescence intensity change (7, 12, 13). However, little attention has been paid to utilize the information contained in the shape of CPE fluorescence spectra. The development of single-molecule spectroscopy (14) and ultrafast time-resolved techniques (15, 16) has revealed that the fluorescence spectral shape is dependent upon polymer chain conformation and the emission of multiple segments with different energy gaps (17), which can provide useful structural information about the polymer's surroundings. Charged side chains and a hydrophobic backbone make the conformation of CPEs sensitive to electrostatic, hydrophobic, and steric interactions (7, 16), which make them good candidates for use as broad-spectrum biosensors (18–21). In 2014, Wu reported an array with 6 different CPEs for protein recognition and distinction via fluorescence correlation spectroscopy signal (22). Rana also reported using green fluorescent protein and a family of positively charged CPEs to identify 16 different cell types in vitro (23). However, both examples require arrays with multiple biosensors and a complicated design, which may limit the generality of the technique. The present report describes a sensitive and selective biosensor which relies on the response of a single fluorescent CPE.

Significance

Fluorescent biosensors are usually designed to recognize a single target analyte and provide a one-dimensional signal from an emission spectrum. Higher-dimensional information in emission spectra and latent factors remain insufficiently utilized. Here we report a broad-spectrum fluorescent biosensor and a general methodology to evaluate spectral shape recognition to classify biomolecules using machine learning. Using a feature selection algorithm to measure the relative intensity of a few selected wavelengths significantly reduces the measurement time, demonstrating the potential for fluorescence spectrum shape analysis in high-throughput technologies. By using well-defined analytes, we explain the mechanism of these fluorescence spectral shape changes, which is fundamental for applying this method for deeper insight into complex phenomena with correlated signals in biological systems.

Author contributions: Y.H., Z.L., and K.S.S. designed research; Y.H., Z.L., A.L.R., B.T.E., C.J.Z., J.G., and Y.Y. performed research; B.T.E. and Y.Y. contributed new reagents/analytic tools; Y.H., A.L.R., C.J.Z., J.G., and K.S.S. analyzed data; and Y.H., Z.L., A.L.R., and K.S.S. wrote the paper.

Conflict of interest statement: A patent was submitted by Y.H., Z.L., and K.S.S.: Fluorescence Spectral Shape Analysis for Analyte Recognition PCT/US19/20109 28-Feb-2019.

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]Y.H. and Z.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: risingera@uthscsa.edu or Kirk.Schanze@utsa.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820713116/-/DCSupplemental.

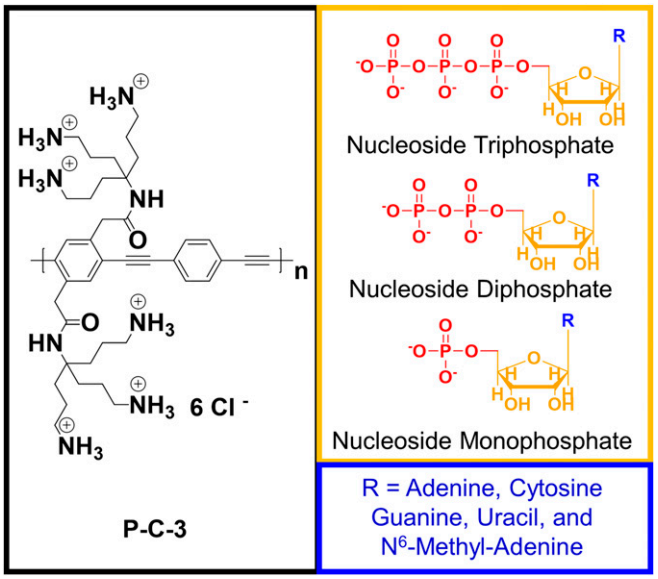

Fig. 1. Structures of P-C-3 and mono, di-, and trinucleotide phosphates.

Artificial intelligence is widely employed in diverse biomedical disciplines, such as drug discovery, genetics, and diagnostics (24) to extract patterns from data without explicit instructions (25). Data with very high dimensionality, like fluorescence, tend to be overfit by existing machine-learning methods, resulting in degeneration of classification performance. Therefore, dimensionality reduction is essential to recognize patterns from these high-dimensional data efficiently and correctly (26). Feature selection is a widely used method in machine learning to select a small subset of relevant features from a complex feature set. Compared with feature extraction, feature selection is not only able to decrease computational cost and increase classification performance, but also saves time during data harvesting by measuring only the most pertinent features. With the help of machine-learning algorithms and feature selection techniques, the disadvantages (26) of high-dimensional spectral data can be overcome, making it possible to develop a broad-spectrum biosensor using fluorescence spectral shape analysis. In previous studies, the selection of features is empirical with no apparent attempt to optimize the choice to improve the performance of classification algorithms and reduce data acquisition time. The use of machine-learning–based feature selection in this work sets it apart from previous studies that utilized linear discriminant analysis (LDA) to classify the chromic response of a conjugated polymer as it interacts with analytes (20, 21).

Herein we report a methodology for biosensing using the cationic CPE P-C-3 (Fig. 1). We have previously reported that the fluorescence of P-C-3 and related cationic CPEs is quenched by polyphosphate ions (27, 28). The current study goes beyond the previous work by demonstrating a general methodology to execute spectral shape recognition to classify nucleotide phosphates using machine-learning approaches. This is a report of selecting useful information contained in fluorescence spectral shapes of a CPE fluorescence-based biosensor associated with analyte structure for classification and distinguishes as little as single base substitutions at distinct positions in the primary sequence of oligonucleotides. With correlated signals, this design is more likely to capture the nonlinear collective effect of multiple factors and provide deeper insights than traditional species-specific probes.

Results and Discussion

The cationic conjugated polyelectrolyte P-C-3 (Fig. 1) was evaluated for its ability to serve as a fluorescence sensor for nucleotide recognition. P-C-3 has a fluorescence quantum yield of 27% in water and it provides 3 -NH_3^+ units on each side chain which can bind to anionic nucleotides via the phosphate moieties,

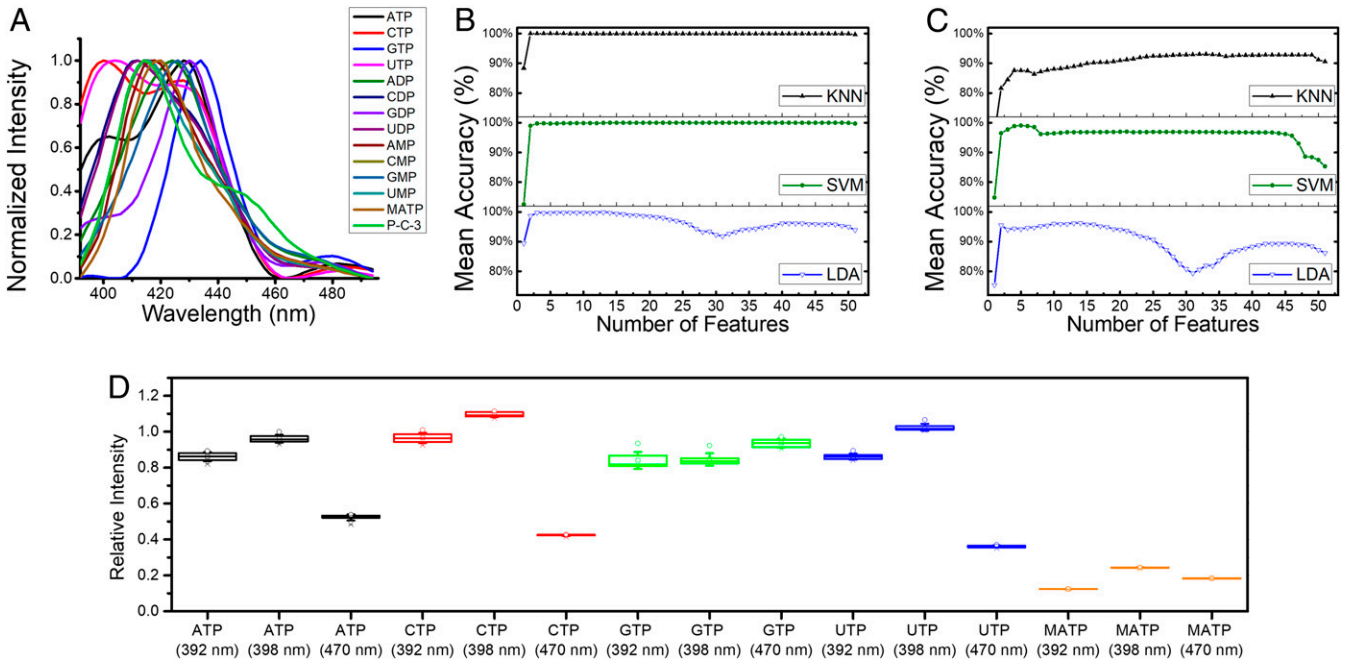

Fig. 2. (*A*) Normalized fluorescence emission spectra of all 13 P-C-3/nucleotide mixtures by scaling between 0 and 1. (*B*) The change of mean classification accuracy against the number of features during sequential floating backward search using LDA, SVM, and k-NN (k = 3) classification algorithms. The training set is randomly partitioned from the whole data, and it contains 5 incidences for each P-C-3/nucleotide mixture. The mean classification accuracy is the out-of-bag mean accuracy of 100 bootstrap iterations of the training set. (*C*) The change of mean classification accuracy against the number of features during sequential floating backward search with the training set plus 10% random error. (*D*) Boxplot of the 3 optimal features (relative intensities of 392, 398, and 470 nm) selected by SVM classification algorithm. The whisker labels represent the SDs.

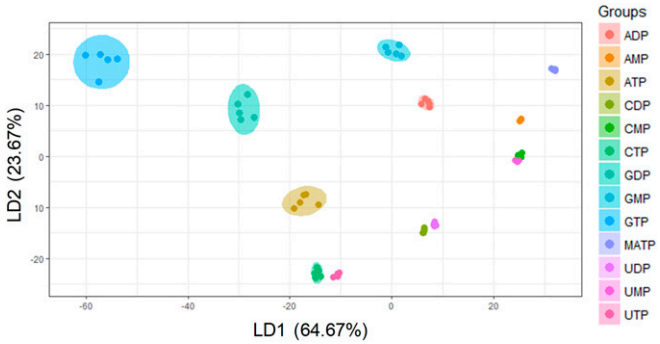

Fig. 3. LDA score plot of 3 selected features for 13 distinct nucleotides. The 3 features are the relative intensity of 398, 464, and 468 nm. LD1 and LD2 represent 64.67% and 23.69% of total variance; 95% confidence ellipses are shown for each nucleotide.

primarily through electrostatic interactions, as well as other nonspecific interactions with nitrogenous bases (27). Nucleotides serve as the biomolecular building blocks of DNA and RNA and are also critical for cellular energetics and signal transduction (29). P-C-3 was applied as a broad-spectrum biosensor to classify 13 distinct nucleotides: the mono-, di-, and triphosphate forms of adenosine (A), cytosine (C), guanine (G), and uracil (U) as well as methyl-ATP (m^6A), the most prevalent mammalian messenger RNA (mRNA) modification (30, 31).

P-C-3 (20 μM) was added to each nucleotide in 2-(*N*-morpholino)ethanesulfonic acid (MES) buffer (10 mM). The nucleotide concentrations were adjusted to give a consistent absorption at 260 nm. Data were acquired with samples at 25 °C. Normalization to absorption was utilized to allow for the future analysis of unknown or complex mixtures where absorbance, not concentration, could be used as a baseline. Each nucleotide was mixed with P-C-3 in 8 individual microplate wells to generate 104 samples (13 nucleotides × 8 replicates). The fluorescence emission spectrum of each P-C-3/nucleotide mixture was collected in 2-nm increments over a range of 392 to 494 nm at an excitation at 350 nm. The raw spectral data of each instance I_n have p dimensions:

$$I_n = (I_{n392\ \mathrm{nm}}, I_{n394\ \mathrm{nm}}, I_{n396\ \mathrm{nm}}, \ldots, I_{n494\ \mathrm{nm}}), \quad [1]$$

where $P = 52$. Unity-based normalization was performed for the fluorescence emission spectrum of each sample to remove the variable of intensity, leaving only spectral shape information. In Fig. 2*A*, it is notable that the normalized emission spectrum of each P-C-3/nucleotide combination has a distinct pattern of emission shape. However, collecting and analyzing the entire fluorescence emission spectrum is not optimal for high-throughput applications. Measuring the fluorescence intensity of every even wavelength from 392 to 494 nm generates 52 features, taking ~52× longer than measuring the intensity at a single wavelength. Besides longer data acquisition time, high-dimensional spectral data contain a high level of irrelevant and redundant features, which tend to decrease the performance of classification algorithms (26), especially with a small training set.

Feature selection was therefore applied to reduce the dimensionality of the original spectral data. For feature selection and model validation, the dataset was randomly partitioned into a training set (65/104) and test set (39/104) with the same instances under each category membership. For real-world applications, instead of using normalized fluorescence intensity, the relative fluorescence intensity at 414 nm of each instance is calculated as

$$I_{n\ relative} = I_n / I_{n414\ \mathrm{nm}}. \quad [2]$$

$I_{n414\mathrm{nm}\ relative}$ is always 1, so this element was removed, and the remaining vector was used for the analysis below. The whole feature set contains 51 features that express the relative fluorescence intensity of every even wavelength from 392 to 494 nm (except 414 nm) for each instance. These features are highly correlated, so using filter models for feature selection will result in a subset with features concentrated in one region of the spectra. Instead, wrapper models provide a simple way to select a feature subset considering the interaction of the algorithm and the training set (32, 33). Since the training set is small compared with the number of features, stratified bootstrap sampling was applied for adding randomness to improve the generalizability for selected features.

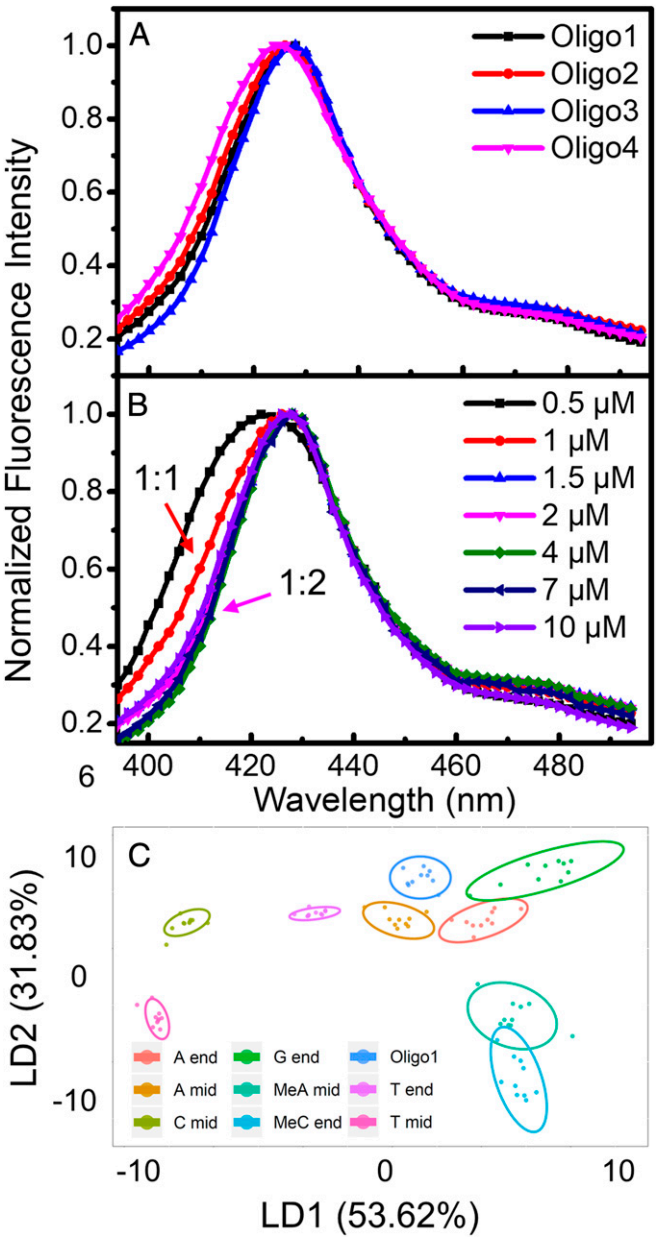

Fig. 4. (*A*) Normalized fluorescence emission spectra of all P-C-3/oligonucleotide mixtures at 10 μM (~1 μM in polymer chains). (*B*) Normalized fluorescence emission spectra of P-C-3/Oligo3 at 0.5 to 10 μM. The lines representing a ratio of P-C-3: oligonucleotide of 1:1 and 1:2 are marked by arrows. (*C*) LDA score plot of 5 selected features for 8 oligonucleotides that each vary by 1 nucleotide compared with difference of oligonucleotide 1. LD1 and LD2 represent 53.62% and 31.83% of total variance. Confidence ellipses (95%) are shown for each oligonucleotide.

Table 1. Size of P-C-3/nucleotide complexes and Stern–Volmer constants of nucleotides

Properties	P-C-3	ATP	CTP	GTP	UTP	MATP	ADP	AMP
Mean diameter, nm	20.2 ± 3.2	75.9 ± 0.1	394.5 ± 0.1	432 ± 0.01	165.9 ± 0.1	148.7 ± 0.4	57.3 ± 0.1	46.2 ± 0.2
K_{SV} (10^4 M^{-1})	/	8.3	17.0	66.	52.5	21.5	2.3	0.1

Size of P-C-3/nucleotide complexes are measured by DLS in MES solution. The size of P-C-3 is measured by AFM. Stern–Volmer constants for nucleotides are measured in 10 μM P-C-3 solution.

To evaluate the feature selection process, 3 widely used algorithms: LDA, support vector machines (SVM), and k-nearest neighbors (k-NN), were applied and the results were evaluated by classification accuracy. Fig. 2*B* illustrates the change of mean classification accuracy during the sequential floating backward search. With the full feature set, the mean classification accuracy of LDA is 94.42%. In the process of feature elimination, the mean classification accuracy decreased and reached the local minimum of 91.80% with 31 features, which indicates some irrelevant or redundant features degraded the classifier in accuracy. Then, the mean classification accuracy started to increase and reached 99.74% with only 3 features. For SVM and k-NN, the mean classification accuracies are 99.76% and 100% with 3 selected features, and they are both higher than the mean classification accuracies with the full feature set. By adding 10% random error to each feature, Fig. 2*C* exhibits a more pronounced trend of increasing accuracy with feature elimination and SVM reached 99.08% accuracy with 5 features. The results of sequential floating backward search demonstrate that a small optimal feature set can provide higher average classification accuracy than the original features with lower computational complexity and more generalizability. Since the optimal feature set is small, using forward feature selection from the empty feature set is more efficient. Fig. 2*D* illustrates the boxplot of the optimal feature set obtained by sequential floating forward search using LDA. Three optimal features are selected, and mean classification accuracies of out-of-bag bootstrap samples were close or equal to 100% for all 3 classification algorithms when the sequential floating forward search was completed.

Fig. 3 demonstrates that the LDA score plot of the 3 selected features with the training set clustered into 13 distinct groups that represent each nucleotide. Uridine 5′-monophosphate (UMP) and cytidine 5′-monophosphate (CMP) clusters appear closest on the plot, but these 2 groups feature small variances and therefore can still be discriminated with the help of LD3 (11.64% of total variance). The test set split at first was utilized to evaluate the classifier with selected features. Fourfold cross-validation was conducted and repeated 30 rounds with selected features, resulting in a 100% overall classification accuracy for each classification algorithm. Relative intensities of 398, 464, and 468 nm were found to be optimal features for LDA; 392, 398, and 470 nm for SVM (linear kernel); and 392, 398, and 468 nm for k-NN with 100% accuracy for each classification algorithm (*SI Appendix*, Table S1). These results demonstrate that the intensities of 3 selected wavelengths plus the intensity of a reference wavelength (414 nm) of P-C-3 were sufficient to classify 13 distinct nucleotides within 1 min. Preliminary work shows that when multiple nucleotides are present in an analyte, the fluorescence response of P-C-3 is unique. As such, application of machine-learning algorithms allows identification of the components of the mixture with high accuracy.

The sensitivity and stability of fluorescence spectral shape analyses were further analyzed using DNA oligonucleotides. Four oligonucleotide sequences were evaluated: 2 sequences of 21 bases (oligo1 and oligo2) and 2 sequences of 34 bases (oligo3 and oligo 4), with each size-matched pair containing an equivalent purine to pyrimidine ratio, but with different individual nucleotides (see *SI Appendix*, Table S2 structure 1–4). Oligonucleotides (Sigma-Aldrich) were dissolved in DNase, RNase free, UltraPureTM Distilled Water (Invitrogen) to create a stock solution concentration of 100 μM (oligonucleotide chain concentration). Each oligonucleotide was mixed with 10 μM P-C-3 (~1 μM in polymer chains) in MES buffer and the fluorescence spectra were collected as described above for single nucleotides. Fig. 4*A* shows these 4 oligonucleotides have distinct fluorescence spectral shapes when mixed with P-C-3. As demonstrated for oligo3 in Fig. 4*B*, the saturation was detected at a 1:2 molar ratio between P-C-3 chains and oligonucleotide, which indicates stability of the fluorescence spectral shape is not affected by the variation of concentration as long as the oligonucleotide is 2× more concentrated than P-C-3. To further illustrate this point, 4 replicates of each oligonucleotide at 4, 7, and 10 μM were used in the dataset and assigned to the same membership, resulting in 12 instances for each oligonucleotide. Three features, the relative intensities of 402, 434, and 492 nm, were selected by LDA using the feature selection method described above. The 50× repeated 6-fold cross-validation gave a 100% accuracy (*SI Appendix*, Table S3). This perfect accuracy demonstrates that the fluorescence spectral shape analysis is sensitive to differences in complicated macromolecular analytes represented by distinct oligonucleotide sequences. The discovery of the saturation point enables spectral shape analysis using the signature pattern associated with the

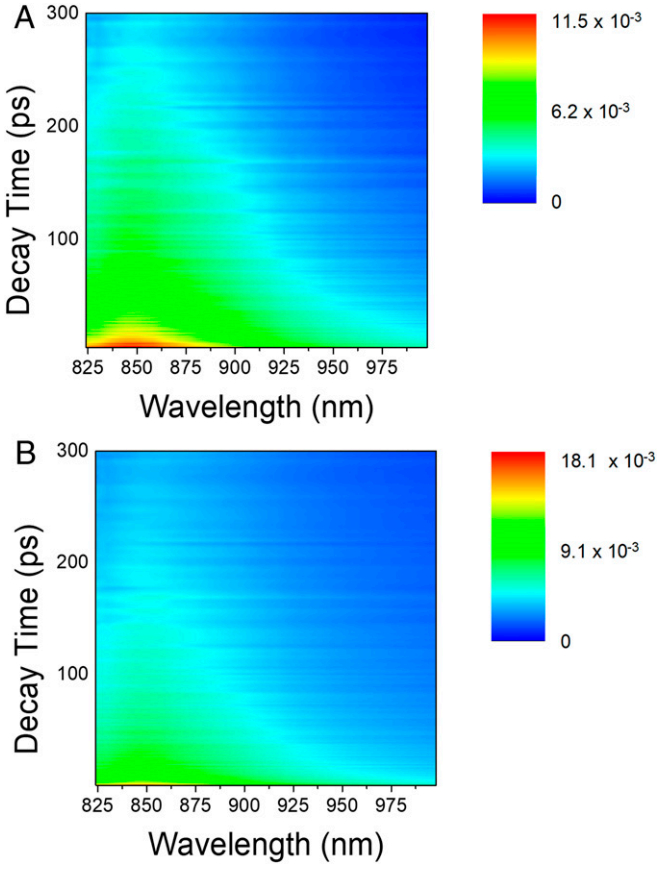

Fig. 5. Femtosecond transient absorption of singlet exciton on P-C-3. (*A*) P-C-3 alone, (*B*) P-C-3/ATP are excited at the pump wavelength of 390 nm with a pump energy of 115 μJ. The color reflects the value of ΔA.

Table 2. Photoinduced absorption decay lifetimes of P-C-3/nucleotide complexes at 850 nm when excited by 390-nm pulse

Species	A_1	τ_1, ps	A_2	τ_2, ps
P-C-3	0.62	320 ± 15	0.38	8.7 ± 0.8
ATP	0.49	360 ± 12	0.51	6.7 ± 0.3
CTP	0.60	368 ± 6	0.40	7.2 ± 0.3
GTP	0.40	252 ± 9	0.60	7.2 ± 0.3
UTP	0.40	245 ± 18	0.60	4.4 ± 0.4
MATP	0.28	139 ± 20	0.72	2.4 ± 0.4
ADP	0.59	363 ± 6	0.41	6.6 ± 0.2
AMP	0.58	329 ± 6	0.42	7.9 ± 0.3

structure independent of analyte concentration, which constitutes a major advantage over intensity-based sensors. Furthermore, P-C-3 was mixed with 10 μM oligonucleotides that differ by a single nucleotide at the end or middle of oligonucleotide 1, including methyladenine and methylcytidine (see *SI Appendix*, Table S2, structure 1 and 5–12) with 10 replicates. Fig. 4*C* shows the LDA score plot with 5 selected relative intensities of 390, 398, 420, 430, and 482 nm, and the 50× repeated 5-fold cross-validation gave a 99.1% accuracy (*SI Appendix*, Table S3c), which demonstrates that this method is able to distinguish a single base substitution either at the end or the middle of the sequence in a 21-base oligonucleotide.

Previous studies have shown that the interaction of P-C-3 with polyphosphate ions in aqueous solution leads to the formation of aggregates (27). The interchain interactions that occur within these aggregates give rise to the fluorescence quenching and band-shape changes that are the basis of the analysis presented above. The quenching studies are shown in *SI Appendix*, Fig. S1. To gain insight into the interactions of P-C-3 with nucleotides and the factors determining the fluorescence emission shape of the conjugated polyelectrolyte, dynamic light scattering (DLS) and atomic force microscopy (AFM) were used to study the size of the P-C-3/nucleotide complexes. The sizes of all P-C-3/nucleotide complexes were measured by DLS in MES solution with the same concentration used for emission shape analysis. The size of pure P-C-3 was measured by AFM since it is too small to measure by DLS. The mean diameters of all of the P-C-3/nucleotide complexes are larger than that of pure P-C-3, which confirms the formation of aggregates (Table 1). Among these complexes, P-C-3/nucleotide triphosphate complexes are larger than P-C-3/nucleotide di- and monophosphate complexes. This can be explained by the fact that negatively charged phosphate groups of nucleoside triphosphates have stronger electrostatic attractions with the branched amine side groups of P-C-3 than nucleoside diphosphates and monophosphates, leading to larger aggregates. This trend is also supported by Stern–Volmer quenching constants (K_{sv}) of nucleotides with P-C-3 (Table 1). The nucleoside triphosphates caused more pronounced aggregation-induced quenching and showed higher K_{sv} values. The nitrogenous base structures also impact aggregate size and fluorescence quenching, indicating that, in addition to electrostatic interactions, hydrophobic interactions and steric interactions between P-C-3 and the nucleotides also influence the self-assembled structures of P-C-3/nucleotide complexes. Interestingly, the N^6 methyl derivative of adenosine 5′-triphosphate (MATP) has a Stern–Volmer constant (K_{sv}) more than 2× that of ATP, illustrating that P-C-3 can produce an amplified signal change in response to a small structural difference in the analyte. A reasonable explanation for the significant increase in Ksv is that the N^6 methylation of ATP weakens hydrogen bonding between the amine and water, leading to an increase in its hydrophobicity and thus a stronger aggregation with the hydrophobic backbone of P-C-3.

In addition to the aggregation, the overall fluorescence spectral shape is strongly influenced by the conformation of the polymer chains (34). Femtosecond transient absorption was applied to study the rapid photophysical processes of the system. According to the literature (35), photoinduced absorption can be utilized to study the exciton dynamics on the polymer chain. Fig. 5 shows the photoinduced absorption (PA) global time-resolved spectrum from 825 to 1,000 nm of P-C-3 and P-C-3/ATP. The PA in this region is assigned to the PA of the singlet exciton of P-C-3, which is comparable with previous reports (35) and the absorption maximum is around 850 nm, such that the decay of the PA intensity at 850 nm is fitted using the biexponential decay function to obtain 2 decay lifetimes:

$$I(t) - I_\infty = A_1 e^{t/\tau_1} + A_2 e^{t/\tau_2}, \quad [3]$$

where A_i represents the weight of each rate constant τ_i and I_∞ is the PA amplitude at time infinity. A kinetic model is proposed to assign these 2 lifetimes to different kinetic pathways. In Tables 2 and 3, the decay component (τ_2) of P-C-3 is assigned to the rapid energy transfer from isolated polymer chains which have no interchain π-electron delocalization, to aggregated chains where π-electron density is delocalized by interchain interactions (16). The assignment well explained the decrease in τ_2 and the increase of the amplitude A_2 in Table 2, after the addition of nucleotides which promote the formation of aggregates. The energy transfer process competes with the emission process of "isolated" chains, which influences both the quantum yield and the fluorescence emission shape of P-C-3/nucleotide complexes. Not only does the energy transfer to the aggregates, but the decay processes of isolated chains also influence the fluorescence emission shape. This kinetic model links the long lifetime component (τ_1) to the fluorescence emission and other decay channels in isolated chains. After excitation, energy is transferred from the high-energy sites to energy traps before emitting or decaying nonradiatively (16, 34). The variation of the long lifetime component (τ_1) between different P-C-3/nucleotide complexes implies that the conformations of isolated chains are also affected by interactions with nucleotides, which influence the fluorescence shape together with the formation of aggregates.

The kinetic model also gives an insight to the signature pattern of fluorescence shape independent of analyte concentration. In Table 3, after the concentration of oligonucleotide exceeded 2× the molar ratio of P-C-3 chains, the amplitudes of both decay pathways remain stable. This result demonstrates that the interactions between P-C-3 and the analyte are saturated, and addition of more analyte does not induce additional aggregation or affect the conformation of the isolated polymer chains. This explanation is also supported by the maintenance of the particle sizes of P-C-3/nucleotide complexes when adding more oligonucleotide after the saturation point (*SI Appendix*).

Although this model can explain the general trends of our observations, it is still difficult to predict the lifetime of P-C-3/nucleotide complexes and the fluorescence emission spectral shapes by a simple theoretical model. With the help of machine-learning methods, these multimodal changes in P-C-3 florescence can be utilized for single nucleotide and oligonucleotide classification, which reveals the advantage of using machine learning in the analysis of this complicated system as a complementary

Table 3. Photoinduced absorption decay lifetimes of P-C-3 with oligonucleotide sequences at 850 nm when excited by 390-nm pulse

Molar ratio	A_1	τ_1, ps	A_2	τ_2, ps
1:1	0.38	142 ± 8	0.62	3.9 ± 0.1
2:1	0.20	57 ± 5	0.80	3.8 ± 0.2
4:1	0.24	68 ± 9	0.76	3.8 ± 0.2

Traces were fit from the intensity maximum by exponential decay with 2 lifetimes, and the sum of A_1 and A_2 are normalized to 1 to reflect their amplitudes.

methodology in evaluating nonspecific interactions in broad-spectrum biosensor design and analysis.

Conclusions

In summary, this work demonstrated a methodology of designing a fluorescence-based biosensor for spectral shape analysis by artificial intelligence with a model explaining the change of the fluorescence shape. The CPE-based biosensor, P-C-3, was found to distinguish structurally similar nucleotides and oligonucleotides with high precision and specificity using a methodology of selecting optimal features and analyzing the fluorescence spectral shape. This fluorescence biosensor was able to classify 13 distinct single nucleotides and as well as 12 oligonucleotides differing by as little as 1 base within 1 min with nearly 100% classification accuracy. Saturation points were also revealed for oligonucleotide sequences which enable spectral shape analysis using the signature pattern independent of analyte concentration. A method was also developed to select useful features to reduce computational complexity, improve classification performance, and reduce data acquisition time.

The size and fluorescence lifetime change of the distinct P-C-3/nucleotide complexes arise due to the formation of aggregates between nucleotides and P-C-3 solution. Ultrafast photoinduced absorption studies of P-C-3/nucleotide complexes reveal the singlet excitation decay can be fit to 2 lifetime components. The short lifetime component is assigned to the rapid energy transfer from isolated polymer chains to aggregated chains and the longer lifetime component is linked to the exciton relaxation from isolated chains, which is related to the conformation of the polymer chains. Both the conformation change of the isolated polymer chain and the formation of aggregates influence the fluorescence emission shape which is the basis for classification. This demonstrates P-C-3 can detect changes in electrostatic, hydrophobic, and steric interactions in a self-assembling system in response to small change in analyte structure. These nonspecific interactions result in fluorescence emission shape changes that can detect differences as subtle as methylation of ATP without more sophisticated chemical processes or a complicated sensor array.

In the current study, the sensor is limited to detection of single target nucleotides or polynucleotides in vitro. However, given the high sensitivity and selectivity of P-C-3 spectral response, it is likely that future work will enable the sensor response to be extended to detection and/or identification of more than a single target analyte in biologically relevant environments. Furthermore, the high classification accuracy and fast detection speed reveal the potential for spectral shape analysis in high-throughput and high-content screening, as well as the application of this technique in mammalian cells.

Materials and Methods

A brief summary is provided here; see *SI Appendix* for details. The synthesis and characterization of P-C-3 were described previously (27). The molecular weight (M_n) was 12 kDa with polydispersity index (PDI) = 2.2. Fluorescence spectra for Stern–Volmer quenching studies were obtained on a PTI Quanta Master spectrometer and spectra for classification were obtained on a Bio-Tek Synergy H1 Hybrid Multi-Mode Plate Reader. UV-visible absorption spectra were obtained on a Varian Cary 100 dual-beam spectrophotometer. All analytical codes were written in R programming language (open source) with MLR package and other basic packages.

ACKNOWLEDGMENTS. This work was supported by the National Science Foundation (Grant CHE-1737714). K.S.S. also acknowledges the Welch Foundation for support through the Welch Chair at the University of Texas at San Antonio (Award AX-0045-20110629). A.L.R. acknowledges support from the Voelcker Foundation.

1. A. M. Darcy, A. K. Louie, L. W. Roberts, Machine learning and the profession of medicine. *JAMA* **315**, 551–552 (2016).
2. G. Schneider, Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
3. M. Boutros, F. Heigwer, C. Laufer, Microscopy-based high-content screening. *Cell* **163**, 1314–1325 (2015).
4. M. Mattiazzi Usaj *et al*., High-content screening for quantitative cell biology. *Trends Cell Biol.* **26**, 598–611 (2016).
5. S. Singh, A. E. Carpenter, A. Genovesio, Increasing the content of high-content screening: An overview. *J. Biomol. Screen.* **19**, 640–650 (2014).
6. C. Zhu, L. Liu, Q. Yang, F. Lv, S. Wang, Water-soluble conjugated polymers for imaging, diagnosis, and therapy. *Chem. Rev.* **112**, 4687–4735 (2012).
7. H. Jiang, P. Taranekar, J. R. Reynolds, K. S. Schanze, Conjugated polyelectrolytes: Synthesis, photophysics, and applications. *Angew. Chem. Int. Ed. Engl.* **48**, 4300–4316 (2009).
8. Y. Huang *et al*., Selective imaging and inactivation of bacteria over mammalian cells by imidazolium-substituted polythiophene. *Chem. Mater.* **29**, 6389–6395 (2017).
9. S. Wang *et al*., Two-photon absorption of cationic conjugated polyelectrolytes: Effects of aggregation and application to 2-photon-sensitized fluorescence from green fluorescent protein. *Chem. Mater.* **29**, 3295–3303 (2017).
10. C. Zhu *et al*., Multifunctional cationic poly(p-phenylene vinylene) polyelectrolytes for selective recognition, imaging, and killing of bacteria over mammalian cells. *Adv. Mater.* **23**, 4805–4810 (2011).
11. Q. Zhou, T. M. Swager, Method for enhancing the sensitivity of fluorescent chemosensors: Energy migration in conjugated polymers. *J. Am. Chem. Soc.* **117**, 7017–7018 (1995).
12. A. Satrijo, T. M. Swager, Anthryl-doped conjugated polyelectrolytes as aggregation-based sensors for nonquenching multicationic analytes. *J. Am. Chem. Soc.* **129**, 16020–16028 (2007).
13. A. Duarte, K.-Y. Pu, B. Liu, G. C. Bazan, Recent advances in conjugated polyelectrolytes for emerging optoelectronic applications. *Chem. Mater.* **23**, 501–515 (2011).
14. P. F. Barbara, A. J. Gesquiere, S.-J. Park, Y. J. Lee, Single-molecule spectroscopy of conjugated polymers. *Acc. Chem. Res.* **38**, 602–610 (2005).
15. A. J. Wise, J. K. Grey, Understanding the structural evolution of single conjugated polymer chain conformers. *Polymers (Basel)* **8**, 388 (2016).
16. L. M. Hardison, X. Zhao, H. Jiang, K. S. Schanze, V. D. Kleiman, Energy transfer dynamics in a series of conjugated polyelectrolytes with varying chain length. *J. Phys. Chem. C* **112**, 16140–16147 (2008).
17. T. Huser, M. Yan, L. J. Rothberg, Single chain spectroscopy of conformational dependence of conjugated polymer photophysics. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11187–11191 (2000).
18. C. Prével, M. Pellerano, T. N. Van, M. C. Morris, Fluorescent biosensors for high throughput screening of protein kinase inhibitors. *Biotechnol. J.* **9**, 253–265 (2014).
19. E. Ghanem *et al*., Differentiation and identification of cachaça wood extracts using peptide-based receptors and multivariate data analysis. *ACS Sens.* **2**, 641–647 (2017).
20. M. S. Maynor, T. L. Nelson, C. O'Sullivan, J. J. Lavigne, A food freshness sensor using the multistate response from analyte-induced aggregation of a cross-reactive poly(thiophene). *Org. Lett.* **9**, 3217–3220 (2007).
21. T. L. Nelson, C. O'Sullivan, N. T. Greene, M. S. Maynor, J. J. Lavigne, Cross-reactive conjugated polymers: Analyte-specific aggregative response for structurally similar diamines. *J. Am. Chem. Soc.* **128**, 5640–5641 (2006).
22. D. Wu, K. S. Schanze, Protein induced aggregation of conjugated polyelectrolytes probed with fluorescence correlation spectroscopy: Application to protein identification. *ACS Appl. Mater. Interfaces* **6**, 7643–7651 (2014).
23. S. Rana *et al*., Ratiometric array of conjugated polymers-fluorescent protein provides a robust mammalian cell sensor. *J. Am. Chem. Soc.* **138**, 4522–4529 (2016).
24. M. W. Freyer, E. A. Lewis, "Isothermal titration calorimetry: Experimental design, data analysis, and probing Macromolecule/Ligand binding and kinetic interactions" in *Biophysical Tools for Biologists: In Vitro Techniques*, J. J. Correia, H. W. Detrich, Eds. (Methods in Cell Biology, Academic Press, Cambridge, MA, 2008), vol 84, pp. 79–113.
25. D. Bzdok, M. Krzywinski, N. Altman, Points of significance: Machine learning: A primer. *Nat. Methods* **14**, 1119–1120 (2017).
26. J. Tang, S. Alelyani, H. Liu, "Feature selection for classification: A review" in *Data Classification: Algorithms and Applications*, C. C. Aggarwal, Ed. (CRC Press, Boca Raton, FL, 2014).
27. Z. Li, R. Acharya, S. Wang, K. S. Schanze, Photophysics and phosphate fluorescence sensing by poly(phenylene ethynylene) conjugated polyelectrolytes with branched ammonium side groups. *J. Mater. Chem. C Mater. Opt. Electron. Devices* **6**, 3722–3730 (2018).
28. X. Zhao, K. S. Schanze, Fluorescent ratiometric sensing of pyrophosphate via induced aggregation of a conjugated polyelectrolyte. *Chem. Commun. (Camb.)* **46**, 6075–6077 (2010).
29. F. Du *et al*., Tightly coupled brain activity and cerebral ATP metabolic rate. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6409–6414 (2008).
30. Y. Xiang *et al*., RNA m^6A methylation regulates the ultraviolet-induced DNA damage response. *Nature* **543**, 573–576 (2017).
31. Y. Fu, D. Dominissini, G. Rechavi, C. He, Gene expression regulation mediated through reversible m^6A RNA methylation. *Nat. Rev. Genet.* **15**, 293–306 (2014).
32. R. Kohavi, G. H. John, Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).
33. I. Inza, P. Larrañaga, R. Blanco, A. J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* **31**, 91–103 (2004).
34. J. Yu, D. Hu, P. F. Barbara, Unmasking electronic energy transfer of conjugated polymers by suppression of $O(_2)$ quenching. *Science* **289**, 1327–1330 (2000).
35. C. X. Sheng, M. Tong, S. Singh, Z. V. Vardeny, Experimental determination of the charge/neutral branching ratio η in the photoexcitation of π-conjugated polymers by broadband ultrafast spectroscopy. *Phys. Rev. B Condens. Matter Mater. Phys.* **75**, 085206 (2007).