Base Motif Recognition and Design of DNA Templates for Fluorescent Silver Clusters by Machine Learning

*Stacy M. Copp, Petko Bogdanov, Mark Debord, Ambuj Singh, and Elisabeth Gwinn**

DNA-encapsulated silver clusters[1] (Ag_N-DNA) possess unique fluorescence properties that depend on the specific DNA template that stabilizes the cluster.[2–4] Particular template choices select for fluorescent silver clusters with peak emission wavelengths throughout the visible and near-IR spectrum,[5,6] corresponding to 10–20 silver atoms bound to the DNA.[7] This wide color palette, combined with proposed low toxicity, high quantum yields of some clusters,[7] low synthesis costs, small cluster sizes and compatibility with DNA have enabled many applications employing Ag_N-DNA. Colorimetric sensing for single base mutations,[4,8] metal ions,[9,10] and microRNAs[11,12] are just a few of these exciting applications. Ag_N-DNA have also been incorporated into logic devices[13] and used for cell labeling.[14,15] Nevertheless, despite the rapid growth of applications for Ag_N-DNA, it is not understood why certain sequences produce brightly fluorescent solutions while other apparently similar sequences do not. This question of the influence of biopolymer host sequence on the structure of templated inorganic particles[16–18] is scientifically rich, with potential for high payoff if approaches can be developed to address the complexities introduced by sequence diversity and flexibility of polymeric templates. In the specific context of Ag_N-DNA, robust design methods for DNA template strands are sorely needed to aid future applications development.

In the emerging physical picture of silver cluster formation on DNA host molecules, silver cations (Ag^+) assist attachment of bases in a DNA template to the silver cluster, subject to backbone-imposed geometrical constraints. This view is based on the presence of both neutral (Ag^0) and cationic silver atoms in Ag_N-DNA[7] and on the known affinity of Ag^+ for the nucleobases.[19] Prior studies showed that homopolymers of cytosine (C) or guanine (G) stabilize fluorescent Ag_N-DNA, albeit often with poor temporal stability, while adenine (A) or thymine (T) homopolymers produce negligible yields of fluorescent products.[20] Beyond the recognition that C- and G-rich DNA strands favor fluorescent cluster solutions,[2,21,22] very little is known about how the composition of mixed-base templates relates to fluorescent cluster formation. For this reason, template design[11] relies on experimentally testing oligomers with sequences selected by informed guessing.

More-systematic template prediction is challenged by the huge space of possibilities: for templates of base length L, there exist 4^L distinct sequences of the four canonical bases, with typical values of $L \approx 10–20$ corresponding to $10^6–10^{12}$ unique sequences. Even for an individual template, DNA flexibility and the ability of Ag^+ to reconfigure native Watson–Crick pairing typically result in many distinct cluster products in a single solution environment, most of them non-fluorescent.[23] As understanding improves of the interactions among DNA, Ag^+, and Ag^0, and of the interactions of all constituents with solvent molecules in modes that favor radiative over non-radiative decay, first-principle theoretical methods like density functional theory may eventually become predictive. However, given these undetermined fundamental issues, a systematic data-driven approach to template selection is currently necessary.

To develop predictive power for selecting DNA templates that stabilize fluorescent Ag_N-DNA, we adopt an approach that combines large experimental data sets with computational machine learning tools for pattern recognition (**Figure 1**). We identify short, consecutive sets of bases, called "base motifs," that preferentially select for fluorescent Ag_N-DNA. We then demonstrate the power of a motif-based method to design DNA templates producing brightly fluorescent Ag_N-DNA solutions. This new understanding of the connection between base motifs and fluorescence brightness builds a basis for statistically predictive design methods for templates that stabilize Ag_N-DNA with desired characteristics for specific applications.

To limit the scope of the problem, we consider only L = 10 templates. Specifically, we examine clusters stabilized by 684 distinct 10-base oligomers with random sequences of the canonical bases, corresponding to <0.07% of all possible 10-base oligomers. Because C and G are important for stabilizing fluorescent silver clusters, random sequences containing fewer than three C plus G bases were excluded. Standard Ag_N-DNA synthesis by $NaBH_4$ reduction was performed on templates in parallel using robotic pipetting in well plate format. Emission spectra were measured by a well plate fluorimeter, using 280 nm light to simultaneously excite all fluorescent Ag_N-DNA products formed by a given template, *via* UV absorbance of the bases.[6] Products were measured 1 day, 1 week, and 4 weeks after synthesis to test fluorescent product stability. Stable products typically peaked in brightness at 1 week and decayed only

S. M. Copp, M. Debord
Physics Department, UCSB
Santa Barbara, CA 93106, USA
Dr. P. Bogdanov, Prof. A. Singh
Department of Computer Science
and Biomolecular Science
and Engineering Program, UCSB
Santa Barbara, CA 93106, USA
Prof. E. G. Gwinn
Physics Department and California
Nanosystems Institute, UCSB
Santa Barbara, CA 93106, USA
E-mail: bgwinn@physics.ucsb.edu

DOI: 10.1002/adma.201401402

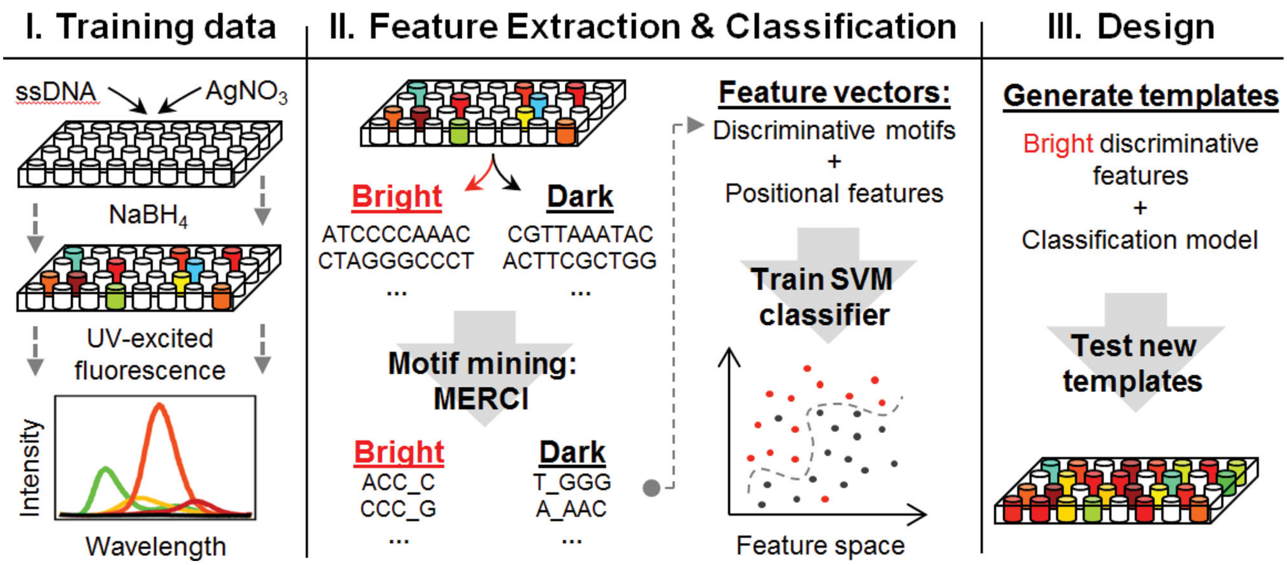

Figure 1. Schematic of methods used to recognize discriminative base motifs within single-stranded templates (ssDNA) for fluorescent Ag$_N$-DNA and to construct new templates for solutions with increased brightness.

slowly thereafter. Here we focus on the data sets collected one week after synthesis.

The fluorescence intensity integrated across 450–850 nm, I_{int}, is used as a brightness metric for products formed on a given template (data provided in Table S3). I_{int} represents the product of chemical yields, fluorescence quantum yields and extinction coefficients summed over all the different fluorescent Ag$_N$-DNA produced in a single solution. About 25% of spectra corresponding to fluorescent solutions exhibited two distinct peaks, typically with one dominant, which arise when the template stabilizes more than one main fluorescent Ag$_N$-DNA species. To include fluorescence from templates producing high yields of multiple silver clusters in the brightness metric, we chose I_{int} over peak intensity I_{peak}.

To elucidate trends connecting template sequence to Ag$_N$-DNA brightness, we considered several pattern recognition schemes that are widely used in machine learning and data mining: artificial neural networks, support vector machines (SVM), random forest, and logistic regression, all available in the Weka library.[24] While all schemes had comparable performance, we selected SVMs due to slight gains. SMVs are classifiers that learn to separate two classes of training data, which are represented in a high-dimensional "feature space" discussed below, by fitting an optimal hyperplane between the two classes.[25] SVMs are widely employed in bioinformatics tasks such as protein-protein binding site prediction[26] and gene classification.[27] Here we use SVMs to categorize base sequences favorable for forming fluorescent Ag$_N$-DNA. We use the random template data to train the SVM to make predictions of the probability of brightness for new, untested DNA templates. The two training data classes correspond to "bright" DNA templates that stabilize fluorescent Ag$_N$-DNA and "dark" templates that do not. Each template is represented by a feature vector that includes information on sequence, as well as Ag$_N$-DNA solution brightness.

Any classification scheme requires categorization criteria. We chose to classify DNA templates stabilizing Ag$_N$-DNA with I_{int} values in the top 30% as "bright" and the bottom 30% as "dark." The middle 40% of templates are excluded from analysis to avoid an arbitrary single threshold distinguishing "bright" from "dark."

SVMs rely on selecting feature vectors that successfully capture the class-determining characteristics. To compare different choices of feature vector composition and thus elucidate template features that are most important for selecting for Ag$_N$-DNA brightness, we use the SVM accuracy, $A = (t_B + t_D)/(t_B + f_B + t_D + f_D)$, where t_B is the number of true predictions the SVM makes for bright strands, f_B is the number of false bright predictions, and t_D and f_D are the number of true and false dark predictions, respectively. A is the fraction of test template sequences that the SVM correctly selects as "bright" or "dark". To evaluate A for a given choice of feature vector space, the data is divided into a training set (85% of the templates) and a test set. The SVM chooses the optimal hyperplane dividing bright and dark sequences in the training set, and A is evaluated using the remaining 15% of the data. Multiple subdivisions of the data are used to obtain stable values for A.

Initially we chose feature vectors containing the entire template sequence, coding each base as an integer ({A,C,T,G}~{1,2,3,4}). Trained SVMs using these feature vectors gave poor accuracy, $A \approx 60\%$, for bright-dark predictions, indicating that use of sequence in the feature vectors gives rather poor separation between bright and dark. This poor separation results from an insufficient representation of the features of a sequence that actually determine its fitness as a template for fluorescent Ag$_N$-DNA. Feature vectors containing only integer-encoded sequence appear not to capture salient features that are invariant with position in the sequence, such as multibase motif patterns. For example, consider two distinct sequences containing the same multibase motif. If the motif occurs in different positions in the two sequences, these sequences can be arbitrarily distant from one another in a sequence-only feature space. Now, if that motif plays an important role in determining if sequences stabilize fluorescent Ag$_N$-DNA, regardless

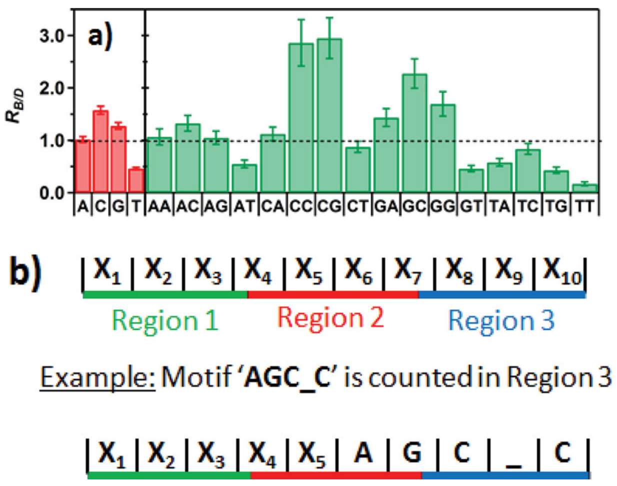

Figure 2. a) Ratios of average motif counts, $R_{B/D}$, per strand in bright to dark templates, with standard error bars. Single base counts (red) show that C and G occur more frequently in templates stabilizing fluorescent clusters while T occurs twice as frequently in dark templates than in bright templates. The spread of $R_{B/D}$ values increases for 2-base motifs (green) and still more for 3-base motifs (Figure S2, Supporting Information). b) Schematic of the three regions considered within a template. Positional features are summed in each region, and motifs are counted in the region containing their middle base(s) (see example).

of position within a sequence, then this crucial information will not be captured by the feature vectors and will not be conveyed to the SVM. In such a poorly chosen feature vector space, the fitted hyperplate will not separate bright and dark classes very efficiently, resulting in low predictive power, i.e., low accuracy, *A*. Therefore, to gain insight into the defining aspects of bright *versus* dark templates, we use *A* as a metric to assess different choices of feature vector.

Under the hypothesis that there exist certain multibase patterns that determine a template's fitness for hosting fluorescent Ag clusters, regardless of these patterns' positions within the template, we next considered vectors composed of information about base motifs instead of only information about base sequence. **Figure 2**a further motivates this choice, displaying average ratios, $R_{B/D}$, of the total counts of individual bases and of 2-base motifs in bright templates to the counts in dark templates. For individual bases (Figure 2, red) $R_{B/D}$ confirms the expectation that bright templates contain higher C and G counts and lower T counts than dark templates. A counts do not, on average, distinguish bright *versus* dark. For longer motifs, the range of $R_{B/D}$ values also increases. Two-base motifs (Figure 2a, green) show a 10:1 spread, and 3-base motifs (Supporting Information, Figure S2) show a 25:1 $R_{B/D}$ spread, suggesting that predictive power of motifs will increase with length until motif length is sufficient to fully define a fluorescent cluster. SVM-based prediction methods may thus benefit from considering motifs at least 3–4 bases in length. Another recent study has also related 4–5 base subsets of DNA sequences to characteristics of Ag_N-DNA fluorescence.[28]

Also apparent in Figure 2 and Figure S2 in the Supporting Information are motifs with $R_{B/D}$ near unity, which do not, on average, distinguish between bright and dark templates. To test whether including these possibly non-discriminative motifs would degrade *A*, we constructed 64-element motif feature vectors composed of the occurrence numbers of *all* 3-base motifs in the template. For example, the feature vector for the template AAAAATTTTT contains '3' for the two entries corresponding to motifs AAA and TTT, '1' for entries corresponding to AAT and ATT, and '0' for the remaining 60 elements. These 64-element motif feature vectors gave even poorer predictive accuracy, $A \approx 50\%$, than the feature vectors containing the entire sequences. Because inclusion of irrelevant information in feature vectors reduces SVM accuracy, we then selected only the ten, 3-base motifs with $R_{B/D}$ farthest from unity. The 10-element feature vectors increased *A* to ~ 75%. Thus, many motifs are apparently non-discriminative for brightness: removing them dramatically improves SVM performance for DNA template classification.

To efficiently and systematically select discriminative motif features, we employ a motif-miner called MERCI[29] that has been applied to various bioinformatics applications.[30,31] For a set of sequences separated into positive (P) and negative (N) classes, MERCI identifies motifs occurring with frequency $\geq F_P$ in class P and $\leq F_N$ in class N. Maximum motif lengths and the maximum length and number of gaps that act as wildcards are specified as parameters. For example, 'AA_GC,' where the gap '_' represents any base or no base, is either a 5-base motif with a 1-base gap or the 4-base motif AAGC. (Gaps add additional flexibility to the features representing template sequences by accounting for possible similarities of multibase patterns with similar function.) Our feature vectors describe the set of bright:dark discriminative motifs selected by MERCI by using a binary variable for every identified motif. The feature vector entry for a given motif is '1' if that motif occurs in the template and '0' otherwise. We found that using motifs of at most 10 bases in length with gaps of at most 1 base and setting $F_P = 10$, $F_N = 10$ results in optimal training and testing accuracies for the 684, ten-base training templates. Since the frequency constraints (F_P, F_N) are not symmetric for bright and dark classes, we run MERCI twice, using bright and then dark as the positive class, to separately extract motifs that discriminate for bright and for dark templates.

With the above optimal parameters, all discriminative motifs identified by MERCI contained 3 to 5 bases, with 4 and 5 base motifs making up 98% of those identified (Supporting Information, Table S4 and S5). Thus, motifs 3–5 bases in length appear to define Ag clusters with the requisite structures for emission within the detection bandwidth of our well plate reader (roughly 400–850 nm). It is notable that MERCI does not identify any discriminative 2-base motifs under the conditions necessary for optimal SVM training and testing accuracies, indicating that 2-base motifs are too small to independently template fluorescent Ag clusters.

The 10 most common motifs associated to bright and dark templates are listed in **Table 1**. Several of the bright motifs contain consecutive C bases, consistent with previous findings of fluorescent Ag_N-DNA formed on C-rich templates.[5,11,28] Multiple G bases are also common in bright motifs.[2,8] The particular combinations of G with C bases and bright motifs containing A bases were unanticipated (Table 1).

Prior work that focused on small sets of "patterned" templates, obtained by single to few-base mutations of certain C-rich parent sequences, could not test for global significance

Table 1. Top 10 most frequently occurring discriminative motifs, as identified by MERCI, for both bright and dark random templates, with motif sequence, the number of occurrences in bright and dark templates, and the average $I_{\rm int}$ with standard error for templates containing each motif. All motifs identified for random templates are tabulated in the Supporting Information (Table S4 and S5).

Bright Motifs				Dark Motifs			
Motif	#Bright	#Dark	Avg. $I_{\rm int}$	Motif	#Bright	#Dark	Avg. $I_{\rm int}$
CC_C	54	9	$(4.7 \pm 0.9) \times 10^5$	T_TT	8	64	$(7 \pm 3) \times 10^4$
C_CC	52	5	$(4.8 \pm 0.9) \times 10^5$	TT_T	6	62	$(7 \pm 3) \times 10^4$
GCG	42	9	$(1.8 \pm 0.3) \times 10^5$	AT_T	10	52	$(3.3 \pm 0.6) \times 10^4$
CCG	42	8	$(2.9 \pm 0.5) \times 10^5$	A_TT	8	52	$(4 \pm 2) \times 10^4$
GCC	42	7	$(3.0 \pm 0.6) \times 10^5$	TTG	9	47	$(5 \pm 2) \times 10^4$
CGC	40	6	$(2.2 \pm 0.2) \times 10^5$	TTT	3	46	$(7 \pm 4) \times 10^4$
CCC	36	2	$(6 \pm 1) \times 10^5$	TTC	8	38	$(1.1 \pm 0.5) \times 10^4$
GG_AC	22	8	$(2.0 \pm 0.4) \times 10^5$	CTT	9	36	$(1.2 \pm 0.5) \times 10^4$
G_GAA	19	10	$(3.0 \pm 0.9) \times 10^5$	TTA	7	36	$(1.1 \pm 0.5) \times 10^4$
AGC_G	18	10	$(1.2 \pm 0.3) \times 10^5$	ATT	4	38	$(2.5 \pm 0.5) \times 10^4$

of specific motifs.[5,11,28] In contrast, our application of pattern-recognition algorithms to large random template sets establishes that particular 3–5 base motifs participate in stabilizing Ag$_N$-DNA. Apparently these motifs have special Ag-binding characteristics that favor formation of clusters with visible fluorescence wavelengths.

In addition to motif composition, the number of bright motifs required to stabilize an emissive Ag cluster is important. Prior studies identified the numbers of Ag atoms and DNA strands contained in fluorescent Ag$_N$-DNA with 15–34 base templates.[7,23] For templates with 19 or more bases, the Ag$_N$–DNA contained just one DNA strand, but for shorter, 15–16 base templates, two copies of the same strand simultaneously stabilized the clusters. This implies that at least two bright motifs are required to stabilize fluorescent clusters. In longer templates, this cluster "sandwiching" between bright motifs can be achieved by folding the strand around the cluster. With shorter templates, it appears that clusters engage multiple bright motifs by simultaneously attaching to two strands. For 10-base templates, the fluorescent clusters with known composition were indeed found to be stabilized by two copies of the template strand.[32]

We therefore expect the Ag$_N$-DNAs in this study to be stabilized via attachment to two 10-base template copies. For these "strand dimer" Ag$_N$-DNAs, a cluster would engage a 3–5 base motif in each strand, with Ag-base bonds holding strands together around the encapsulated cluster. Because the lengths of bright motifs are well below half the single-stranded DNA persistence length,[33] they can act as stiff, linear cluster scaffolds, rationalizing the elongated, rod-like cluster shapes indicated by Ag$_N$-DNA optical properties.[7] The cluster size range of 4–6 neutral silver atoms in most visible-emitting Ag$_N$-DNA[32] may also arise from the requirement that atoms arranged in an elongated cluster make contact the 3–5 bases of linear motif scaffolds within each template. For Ag$_N$-DNA stabilized by single, longer template strands, we expect that these templates must contain at least two bright motifs as well as sufficiently long, flexible runs of intervening bases to allow the strand to present both bright motifs to the cluster.

In addition to motif inclusion features, we consider positional features describing motif location along the 10-base random template sequences. Templates are partitioned into 3 equal regions, and dark and bright discriminative motifs are counted in each region (Figure 2b); a motif is counted in the region containing its middle symbol(s). We also tested use of positional features describing average nucleotide size and "stickiness", a metric of each base's interaction strength with Ag clusters (parameters in Table S1). The single base dependence in Figure 2a ($R_{\rm B/D} > 1$ for G and C bases, $R_{\rm B/D} < 1$ for T, $R_{\rm B/D} \approx 1$ for A) is parameterized by this rough "stickiness" categorization, for comparison to results assuming equal stickiness. The LIBSVM library[34] was used for classification. While including the positional features improved SVM accuracy from 82% to 86%, with stable accuracies across multiple SVM runs, base size and stickiness had little effect. Improved SVM accuracy upon inclusion of positional motif information suggests that locations of 3–5 base motifs within a template are also somewhat important. In summary, the propensity of a DNA template strand to stabilize fluorescent Ag clusters is determined by certain select 3-base to 5-base motifs within the template as well as by the relative positions of these motifs with respect to one another.

To generate new DNA templates for bright Ag$_N$-DNA solutions, we adopt a simple model that draws on the extracted motif features and the $I_{\rm int}$ values measured for Ag$_N$-DNA solutions. Let $p(M,P)$ be a probability density function describing the probability of motif M being incorporated at position P ($P = \{1,2,3\}$ our case) in a bright template. For the set of training templates, T, and the set of MERCI-identified motifs, M, we define the probability of bright template inclusion for every position p and motif m as:

$$p(M=m,P=p)=\frac{\sum_{t\in T} I_t x(t,m,p)}{\sum_{n\in M}\sum_{t\in T} I_t x(t,n,p)} \quad (1)$$

where $I_{\rm t}$ is the intensity of template t and $x(t,m,p)$ is '1' if motif m occurs in training template t at position p and '0' otherwise.

(Here $p(m,p)$ can be interpreted as the intensity-weighted probability that motif m occurs at position p, across all training sequences.) Starting with an empty sequence, we iteratively sample motifs to include in consecutive regions of the new template according to the motif's value of $p(m,p)$, rejecting motifs that are incompatible with the previously included motifs. This process continues until all base positions are assigned. Only sequences differing by at least two and at most three mutations from any of the 684 templates in the training data set are retained, corresponding to ca. 10^6 distinct sequences. We then classify each newly generated template as bright or dark using our previously trained SVM.

We tested the effectiveness of this motif-based design method by experimentally testing the 374 template sequences to which the SVM assigned the highest brightness probabilities (Table S6). The average I_{int} value for Ag_N-DNA solutions synthesized with this designed template set is much brighter, by a factor of > 3 at one week after synthesis, than for the random template set used to train the SVM. **Figure 3**a demonstrates this shift to higher solution fluorescence by comparing probability distribution functions (PDFs) for the top 30% of I_{int} values from random and designed templates. For the random template set, low-fluorescence solutions were by far the most probable outcome, while in striking contrast, none of the top 30% of designed templates produced solutions in the lowest brightness bin (Figure 3a; note the scale break in the PDF axis). By the I_{int} threshold definitions used to classify random template sequences as “bright” or “dark”, 295 of the 374 generated templates are bright, while only 3 are dark (the remaining 76 lie between bright and dark thresholds). This dramatic increase in the ability to select bright DNA templates corresponds to an accuracy A = 78% for the motif-based design method, somewhat below the SVM accuracy, A = 86%, obtained from the random sequence training data. The decrease in realized *versus* predicted accuracy may reflect the small training set size, <0.07% of the possible space of templates. Larger training sets may yield even higher accuracies. Accuracy may also be reduced by formation of larger, infrared clusters (not detectable due to the plate reader's low sensitivity above ca. 750 nm), as might be expected from the inclusion of higher numbers of “bright” motifs in the designed *versus* random templates.

The hypothesis that greater numbers of “bright” base motifs in template sequences lead to formation of larger Ag clusters is consistent with the red-shift in the Ag_N-DNA color distribution for the designed versus random template set (Figure 3b). Ag_N-DNA color reddens with increasing numbers of Ag atoms in the cluster,[23] which has been attributed to elongation of the free-electron path in rod-shaped[7,32,35] clusters. We expect the higher average number of bright motifs in the designed templates to increase the numbers of Ag^+ associated to the template's bases before reduction, causing an increase in the average neutral silver cluster size formed by reduction and thus longer fluorescence wavelength.

An examination of the 10 brightest template sequences (Supporting Information, Table S2) hints at why motifs in Table 1 are predictive of fluorescence intensity and which nucleotides interact with Ag clusters. Nine of the 10 brightest templates contain two or more adjacent C bases, matching cytosine's accepted role in forming fluorescent Ag_N-DNA. The “brightest” sequence, however, contains no adjacent C's, and of the 498 random and generated “bright” templates (Supporting Information, Table S3 and S6), 27% contain no consecutive C's, and 60% contain no more than two consecutive C's. The discriminative 3–5 base motifs leading to optimal SVM performance suggest that Ag clusters interact with at least 3 consecutive bases on each template. Thus, while C's have often appeared to play a dominant role in stabilizing fluorescent Ag_N-DNAs, our motif mining and SVM results suggest that Ag clusters must also be engaged to at least one A, G, or T in the majority of these bright templates. Previous studies suggest that thymines do not associate with Ag clusters at neutral pH,[20] and the role of adenine is still unclear. It may be that thymines act as termination sites to limit Ag cluster growth during synthesis and prevent cluster migration after synthesis, important aspects for visible and near-IR emission wavelengths and time stability. This would explain the presence of T's in 15% of identified bright motifs.

Thus, a plausible model of Ag cluster association with DNA template(s) consists of at least two 3–5 base motifs, rich in C's, G's, and/or A's. These motifs occur either in two separate template strands or in a single longer strand separated by a dark linker long enough to allow two bright motif regions to encapsulate the cluster. T's likely serve to limit cluster size. Variations

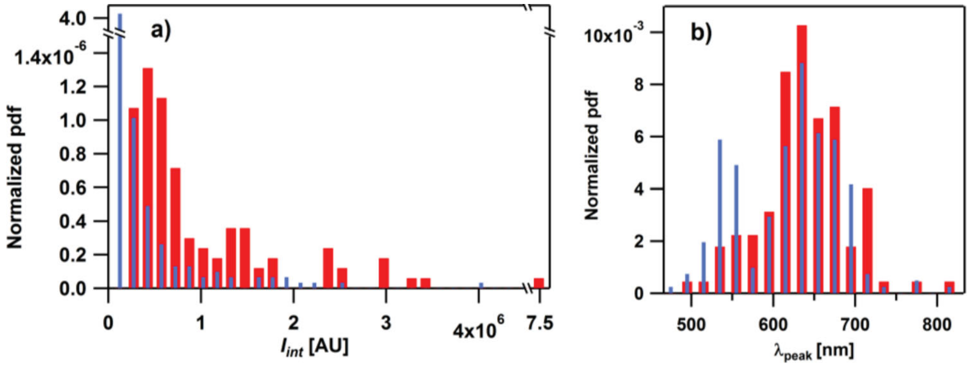

Figure 3. a) Normalized probability density (pdf) functions of the top 30% of I_{int} values for strands with randomly generated sequences (blue) and with sequences designed by motif selection (red). The distribution of I_{int} shifts towards higher values for designed templates. b) PDFs of peak wavelengths for bright clusters stabilized by random (blue) and designed templates (red).

in C, G, and perhaps also A content likely further control cluster size and thus wavelength.

In conclusion, we have used large array data sets and machine learning tools to show that multibase motifs govern the fluorescence brightness of Ag_N-DNA solutions formed on DNA templates. We separately identify sets of motifs that select for brightness and sets of motifs that discriminate against bright products. Both motif types will be important for realizing designed multicluster constructs. By selecting motifs that discriminate between bright and dark templates and representing templates with feature vectors composed of information about these discriminative motifs, SVM-based classifiers can be trained to predict template brightness with high accuracy. Combining this predictive power with an intensity-weighted, motif-based generative model, we experimentally demonstrate ca. 80% accuracy for generating templates that produce bright Ag_N-DNA solutions. Even higher accuracies may be achieved in the future by using larger training data sets. The lengths of identified discriminative motifs suggest that Ag clusters engage with regions of DNA templates containing 3–5 consecutive bases. While many of the identified bright motifs are rich in C and G bases, more complex roles of multibase motifs still need to be investigated. Motifs identified as discriminative for low-fluorescence Ag_N-DNA solutions may also be utitilized in sensing schemes to create regions of DNA that do not stabilize fluorescent Ag clusters. The techniques used here, motif selection by the MERCI algorithm and classification by SVM techniques, were developed for use in contexts quite different from stabilization of fluorescent silver clusters. Thus, in terms of the broader context of biopolymer-templated inorganic nanomaterials, we expect that the approaches developed here may be generally useful for studies of polypeptides, proteins, and RNA as templates for other inorganic materials such as metal and semiconductor clusters.

Experimental Section

Robotic Well Plate Synthesis: 10-base DNA template strands were ordered from Integrated DNA Technologies (Coralville, IA) with standard desalting, suspended in RNase-free H_2O in well plate format. A Beckman Coulter Biomek 2000 pipetting robot was used to synthesize Ag clusters on each template strand in parallel. DNA was mixed with a solution of $AgNO_3$ in NH_4OAc buffer and incubated at room temperature for 20 min before $NaBH_4$ reduction, for final concentrations of 20 μM DNA, 100 μM $AgNO_3$, and 50 μM $NaBH_4$ in 10 mM NH_4OAc. Of the synthesis conditions tested, these conditions produced the largest number of brightly fluorescent Ag_N-DNA solutions. Products were stored at 4 °C until measurement.

Spectral Characterization: Ag_N-DNA fluorescence spectra were measured using a Tecan Infinite 200 PRO well plate reader. 280 nm light was used to excite all clusters simultaneously,[6] and emission was measured from 400–850 nm. Because scattered excitation light was detected from 400–430 nm, integrated intensity values were integrated from 450–850 nm.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors thank Monte Radeke and Lincoln Johnson for training and use of a pipetting robot acquired with support from NIH-NEI 5-R24-EY14799. S.M.C. acknowledges support from NSF-DGE-1144085 (NSF Fellowship). This work was supported by NSF-CHE-1213895 and NSF-DMR-1309410.

Received: March 28, 2014
Revised: May 20, 2014
Published online: July 14, 2014

[1] J. T. Petty, J. Zheng, N. V Hud, R. M. Dickson, *J. Am. Chem. Soc.* **2004**, *126*, 5207.
[2] E. G. Gwinn, P. O'Neill, A. J. Guerrero, D. Bouwmeester, D. K. Fygenson, *Adv. Mater.* **2008**, *20*, 279.
[3] B. Sengupta, C. M. Ritchie, J. G. Buckman, K. R. Johnsen, P. M. Goodwin, J. T. Petty, *J. Phys. Chem. C* **2008**, 18776.
[4] W. Guo, J. Yuan, Q. Dong, E. Wang, *J. Am. Chem. Soc.* **2010**, *132*, 932.
[5] J. T. Petty, C. Fan, S. P. Story, B. Sengupta, M. Sartin, J.-C. Hsiang, J. W. Perry, R. M. Dickson, *J. Phys. Chem. B* **2011**, *115*, 7996.
[6] P. R. O'Neill, E. G. Gwinn, D. K. Fygenson, *J. Phys. Chem. C* **2011**, *115*, 24061.
[7] D. Schultz, K. Gardner, S. S. R. Oemrawsingh, N. Markešević, K. Olsson, M. Debord, D. Bouwmeester, E. Gwinn, *Adv. Mater.* **2013**, *25*, 2797.
[8] H.-C. Yeh, J. Sharma, I.-M. Shih, D. M. Vu, J. S. Martinez, J. H. Werner, *J. Am. Chem. Soc.* **2012**, *134*, 11550.
[9] W. Guo, J. Yuan, E. Wang, *Chem. Commun. (Camb).* **2009**, 3395.
[10] G.-Y. Lan, C.-C. Huang, H.-T. Chang, *Chem. Commun. (Camb).* **2010**, *46*, 1257.
[11] P. Shah, A. Rørvig-Lund, S. Ben Chaabane, P. W. Thulstrup, H. G. Kjaergaard, E. Fron, J. Hofkens, S. W. Yang, T. Vosch, *ACS Nano* **2012**, *6*, 8803.
[12] Y. Liu, M. Zhang, B. Yin, B. Ye, *Anal. Chem.* **2012**, *84*, 5165.
[13] Z. Huang, Y. Tao, F. Pu, J. Ren, X. Qu, *Chemistry* **2012**, *18*, 6663.
[14] J. Yu, S. Choi, C. I. Richards, Y. Antoku, R. M. Dickson, *Photochem. Photobiol.* **2008**, *84*, 1435.
[15] Y. W. Zhou, C. M. Li, Y. Liu, C. Z. Huang, *Analyst* **2013**, *138*, 873.
[16] Y. Ofir, B. Samanta, V. M. Rotello, *Chem. Soc. Rev.* **2008**, *37*, 1814.
[17] A. Houlton, A. R. Pike, M. Angel Galindo, B. R. Horrocks, *Chem. Commun. (Camb).* **2009**, 1797.
[18] J. Liu, B. Uprety, S. Gyawali, A. T. Woolley, N. V. Myung, J. N. Harb, *Langmuir* **2013**, *29*, 11176.
[19] S. Menzer, M. Sabat, B. Lippert, *J. Am. Chem. Soc.* **1992**, *114*, 4644.
[20] D. Schultz, E. Gwinn, *Chem. Commun. (Camb).* **2011**, *47*, 4715.
[21] C. I. Richards, S. Choi, J.-C. Hsiang, Y. Antoku, T. Vosch, A. Bongiorno, Y.-L. Tzeng, R. M. Dickson, *J. Am. Chem. Soc.* **2008**, *130*, 5038.
[22] H.-C. Yeh, J. Sharma, J. J. Han, J. S. Martinez, J. H. Werner, *Nano Lett.* **2010**, *10*, 3106.
[23] D. Schultz, E. G. Gwinn, *Chem. Commun. (Camb).* **2012**, *48*, 5748.
[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10.
[25] C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.
[26] J. R. Bradford, D. R. Westhead, *Bioinformatics* **2005**, *21*, 1487.
[27] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, D. Haussler, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 262.
[28] Y. Teng, X. Yang, L. Han, E. Wang, *Chemistry* **2014**, *20*, 1111.
[29] C. Vens, M.-N. Rosso, E. G. J. Danchin, *Bioinformatics* **2011**, *27*, 1231.

[30] S. K. Dhanda, P. Vir, G. P. S. Raghava, *Biol. Direct* **2013**, *8*, 30.

[31] G. Cilingir, A. O. T. Lau, S. L. Broschat, *J. Microbiol. Methods* **2013**, *95*, 313.

[32] S. M. Copp, D. Schultz, S. Swasey, J. Pavlovich, M. Debord, A. Chiu, K. Olsson, E. Gwinn, *J. Phys. Chem. Lett.* **2014**, *5*, 959.

[33] B. Tinland, A. Pluen, J. Sturm, G. Weill, *Macromolecules* **1997**, *30*, 5763.

[34] C.-C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1.

[35] R. R. Ramazanov, A. I. Kononov, *J. Phys. Chem. C* **2013**, *117*, 18681.