Machine Learning Directed Aptamer Search from Conserved Primary Sequence and Secondary Structure

Javier Perez Tobia[1], Po-Jung Jimmy Huang[2], Runjhun Saran Narayan[2], Apurva Narayan[1,2,3,4]* and Juewen Liu[2]*

1. Department of Computer Science, University of British Columbia, Kelowna, BC, Canada

2. Department of Chemistry, Waterloo Institute for Nanotechnology, Water Institute, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
Email: liujw@uwaterloo.ca

3. Department of Computer Science, Western University, London, ON, Canada

Email: apurva.narayan@uwo.ca

4. Systems Design Engineering, University of Waterloo, Waterloo, ON Canada

JPT and PJJH contributed equally to this work

Abstract

Computer-aided prediction of aptamer sequences has been focused on primary sequence alignment and motif comparison. We observed that many aptamers have a conserved hairpin, yet the sequence of the hairpin can be highly variable. Taking such a secondary structure information into consideration, a new algorithm combining conserved primary sequences and secondary structures is developed, that combines three scores based on sequence abundance, stability, and structure, respectively. This algorithm was used in the prediction of aptamers from caffeine and theophylline selections. In the late rounds of the selection, when the library was converged, the predicted sequences matched well with the most abundant sequences. When the library was far from convergence and the sequences were deemed impossible for traditional analysis methods, the algorithm still predicted aptamer sequences that were experimentally verified by isothermal titration calorimetry. This algorithm paves a new way to look for patterns in aptamer selection libraries and mimics the sequence evolution process. It will help shorten the aptamer selection time and promote the biosensor application of aptamers.

Keywords: artificial intelligence, machine learning, SELEX, aptamers, isothermal titration calorimetry

Introduction

Aptamers are single-stranded nucleic acids that can selectively bind target molecules.[1-3] The interest in aptamers has been driven by their competitive advantages to antibodies including higher stability, reversible denaturation, lower cost, and easier modification. The majority of DNA aptamers were obtained via combinatorial selection.[4, 5] Typical aptamer selections involve a gradual decrease of target concentration over ten to twenty selection rounds. The more rounds of selection, the longer it takes and the more likely to make mistakes. Thus, a reliable method to extract aptamers from early rounds of selection is needed.

Computer-aided prediction of the secondary structure of nucleic acids and sequence alignment has already become common tools in aptamer research.[6-8] The advent of high throughput deep sequencing allows the generation of overwhelming amounts of information, which provides the data for machine learning. One type of aptamer prediction is based on de novo design of aptamer secondary/tertiary structures and molecular docking. However, the success has been limited and few predicted aptamers were used.[9-11] Another approach uses previously reported aptamers as a training set to predict new aptamers.[12] Since many recent papers showed that not all published aptamer sequences are reliable,[13-17] the quality of the input data need to be carefully ensured.

A third type of algorithm analyzed aptamer selection libraries.[18] The most common algorithms relied on clustering of primary sequences, and the most abundant sequence families are picked as aptamer candidates. However, aptamer enrichment happens only in the later rounds of selections, and other reasons such as PCR bias can also lead to preferentially amplified sequences that are not necessarily aptamers.[19] In addition, motif finding algorithms have been developed,[9, 18, 20] which rely on predicted secondary structures and divide the structures into multiple subunits. A motif is defined based on the sequence of single-stranded regions,[21] and it still looks for identical sequences. The same DNA can be folded into many possible secondary structures, and such methods focus too much on local structures that may or may not be important for target binding.

Finally, combined analysis of primary sequence and secondary structure has been attempted recently. Song et al took into consideration the secondary structure of DNA, which is a major conceptual advancement. The authors called their method Sequential Multidimensional Analysis algoRiThm for aptamer discovery (SMART-Aptamer).[22] This algorithm was applied to the selection libraries for human embryonic stem cells, epithelial cell adhesion molecules, and cell-surface vimentin. While the secondary structure information was considered, calculation was mainly based on the overall free energy released from DNA folding. In addition, sequencing data from multiple rounds were required. An empirical study

on our own SELEX data revealed that SMART-Aptamers' performance is dependent on the amount and diversity of aptamer libraries sequenced.

In this work, we aimed to advance the prediction by using a structural sequence pattern recognition algorithm mimicking the evolution of the library and taking into account both conserved primary sequences and secondary structures. We call it conserved primary and secondary pattern searching (CPSPS), which can identify aptamers with data from only one round. Using Python as the coding language, CPSPS readily incorporates a few well-developed software namely: RNAfold from ViennaRNA and the scikit-bio package. We analyze a few of our recently selected libraries using CPSPS and identified a few new aptamers that were previously neglected by manual analysis. We were able to identify aptamers when their sequence frequency was only 0.1% of the library.

Materials and Methods

Chemicals. The DNA samples used for the selection and sensing experiments were purchased from Integrated DNA Technologies (Coralville, IA, USA). The sequences are listed in Table S1. The chemicals including caffeine and theophylline were from Sigma-Aldrich. Milli-Q water was used to prepare all the buffers and solutions.

Computing methods. The detailed methods and codes are available at the following link https://github.com/Idsl-group/OptimalAptamerFinder. This website may be used for reproducing the results.

ITC. A MicroCal VP-ITC was used. DNA (9 μM or other concentrations, 2 mL) and target molecules (1 mM or other concentrations, 2 mL) were dissolved in buffer (500 mM NaCl, 10 mM $MgCl_2$, 50 mM HEPES pH7.5) and degassed for 10 min prior to measurement. Target (300 μL) was titrated into the cell chamber containing 1.4 mL aptamer. Except for an initial injection of 0.5 μL, 10 μL of target was titrated into the cell each time over 20 sec duration for a total of 20 to 28 injections at 25°C. The spacing was set for 360 sec between each injection. For some aptamers, the thermodynamic values were obtained by fitting the titration curves to a one-site binding model using the Origin software.

Results and Discussion

Conserved primary and secondary structures in aptamers.

Most aptamer selection libraries have a random region of 30 to 40 nucleotides or longer.[23-27] Aptamer binding often does not need such a long sequence, and a common strategy for a library to evolve is to hide the redundant nucleotides in a hairpin. The sequence of such a hairpin is not conserved, but the presence of such hairpins is conserved. An example is shown in Figure 1A for two aptamers that can bind uric acid.[28] Aside from the 6-mer conserved motifs in blue, the middle part can form a hairpin (the stem region marked in red). In this example, the nucleotides inside the hairpin differed a lot, but the secondary structure was conserved (Figure 1B). We reason that both the sequences in blue and the hairpins are evolutionarily conserved. For primary sequence clustering methods, the sequences in hairpins would not contribute to scoring, but our CPSPS algorithm counts it as a positive score. Such hairpins are more targeted than the analysis of global secondary structures as well.

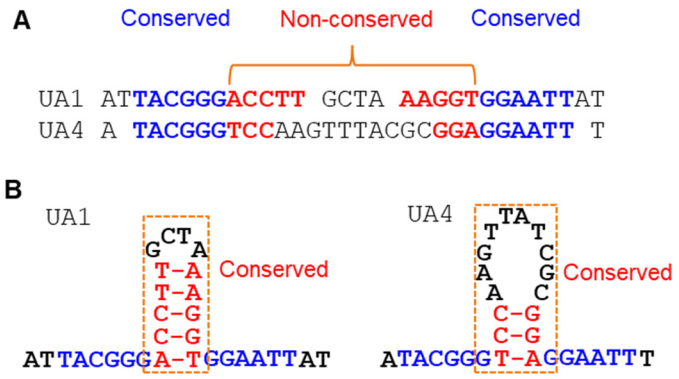

Figure 1. (A) The aligned primary sequences from the random region of two aptamers named UA1 and UA4, respectively. The conserved sequences are in blue. (B) The folded secondary structures, where the conserved hairpins structures are highlighted.

The CPSPS algorithm.

To reflect our idea of identifying conserved primary sequences and hairpins in a list of DNA sequences in a library, our approach works as follows.

1) **Library processing.** Calculate the number of occurrences of each sequence and remove all duplicate sequences from the library. Then, calculate the minimum free energy (MFE) structure and free energy (dG) of each sequence using the RNAfold tool from the ViennaRNA package.

2) **Clustering.** We generate 10 clusters by iterating the following procedure. We found empirically that 10 clusters were sufficient to incorporate a major portion of the sequences and it is also a size still amenable to experimental testing. i) A list of all the sequences in the library are generated. ii) The most common sequence in the list is added to a new cluster. iii) All remaining sequences in the list are aligned with respect to the leading sequence in the new cluster using the Striped Smith-Waterman alignment algorithm. iv) All sequences with an alignment score above 0.8 (1 is for identical sequences, 0 is for sequences that have nothing in common) is added to the new cluster and they are removed from the list.

3) **Collection of structural and motif information for every sequence.** For every sequence, we generate a list of all the 6-mers and hairpin structures present in it (6-mer is a widely used length for predicting aptamer sequences).[22] While doing this, we keep track of the total number of occurrences of each 6-mer and hairpin level on a round level (kmer_counts, stem_counts, loop_counts) and a cluster level (cluster_kmer_counts, cluster_stem_counts, cluster_loop_counts). To increase efficiency, we remove all motifs with frequencies below the 50th percentile frequency value since they do not contribute significantly to scores due to their low values.

4) **Cluster combination.** For every cluster and for every kind of motif (6-mer, hairpin stem, hairpin loop) we select their top 10 most common motifs. Then for every possible cluster pair, we compare their top motifs by initializing a similarity score to 0 and then adding -1 for every common motif and adding +1 for every non-common motif. Then the similarity scores for each motif kind for each cluster pair are added and if the score is 0 or less, the clusters are combined into a single cluster.

5) **Score calculation.** For every sequence, the following scores are calculated. i) Popularity score: average of the normalized cluster size (as a%) and the normalized frequency of the sequence relative to the most common sequence in the round (as a%). ii) Stability score: The absolute value of the free energy (dG) multiplied by 10 (with a maximum value of 100). iii) Structural score: average of cluster and round kmer score and hairpin score. These scores are calculated by normalizing the kmer and hairpin counts (with respect to the most common kmer and hairpins) and averaging over the kmers and hairpins that appear in each aptamer. These three scores are then combined to produce a final score for each aptamer.

An example.

To better illustrate our CPSPS scoring system, we give the following example with sequences from the round 16 of uric acid selection. For Step 1, library processing, out of the entire 55985 sequences obtained, eight are shown in Figure 2A (the table on the left). The table in the middle shows information about 3 sequences out of the 8633 unique sequences found in the processed library. The counts of each sequence are the number of times it appeared in the sequenced library and its structure (the stems represented by parenthesis as can be seen in the structure shown on the right) and dG values were calculated using ViennaRNA. We use the bracket notation for expressing the sequences. In this notation each character represents a nucleotide base. Brackets represent paired bases and dots represent unpaired ones. Each open bracket base pairs with a closed bracket base ahead of it (there are always the same number of open and closed brackets).

For Step 2, clustering, using the sequenced library from Step 1, the first cluster is initialized with seq0, which had 2392 copies. Thus, we aligned the rest of the sequences with respect to it using the Striped Smith-Waterman alignment algorithm.[29] Then, all these clustered sequences were removed from the library and the rest of the sequences were aligned using this method again. In this work, we repeated this process until we obtained 10 clusters (Figure S1).

Step 3 is the collection of structural and motif information for every sequence. Figure 2B (left) illustrates the first four 6-mers of the first two sequences in our library as well as the hairpins and loops found on them. The tables on the right side of Figure 2B display each motif counts for each motif type for the whole round or for each cluster.

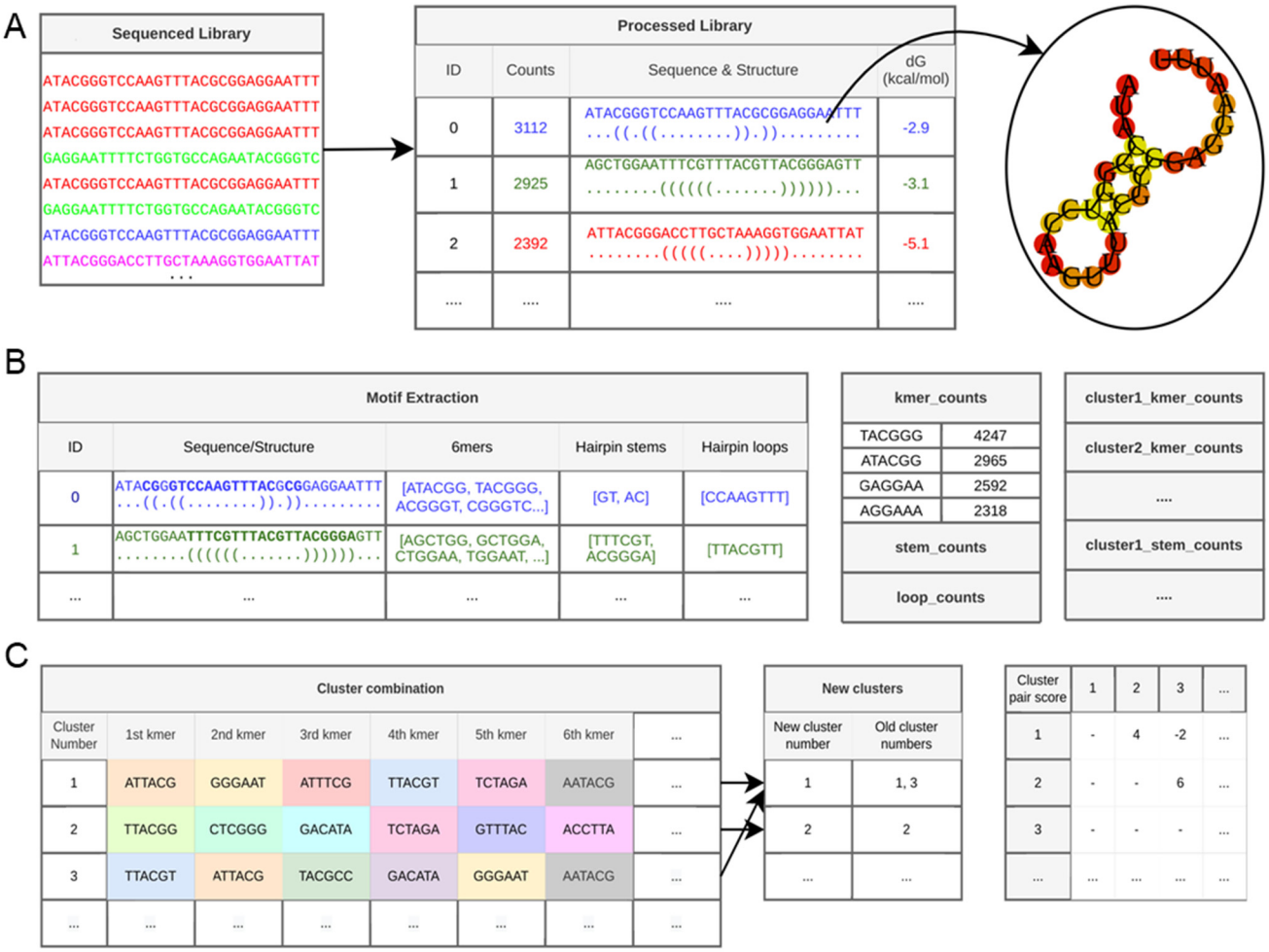

Figure 2. (A) A table showing top 6-mers for top-3 cluster for illustration purposes. (B) Cluster merging based on our scoring mechanism as described in (C). (C) Clustering k-mer alignment score between two clusters.

For Step 4, cluster combination, clusters 1 and 3 have several kmers in common so they can be combined into a single cluster, while cluster 2 remains as an independent cluster (Figure 2C). The table on the far right of Figure 2C shows the cluster pairs scores based on the top six kmers of the first three clusters shown on the left. Finally, for Step 5, score calculation, Figure 3 shows how the final score was calculated for the first sequence of this round. We can now select the highest-scoring aptamers and test them experimentally.

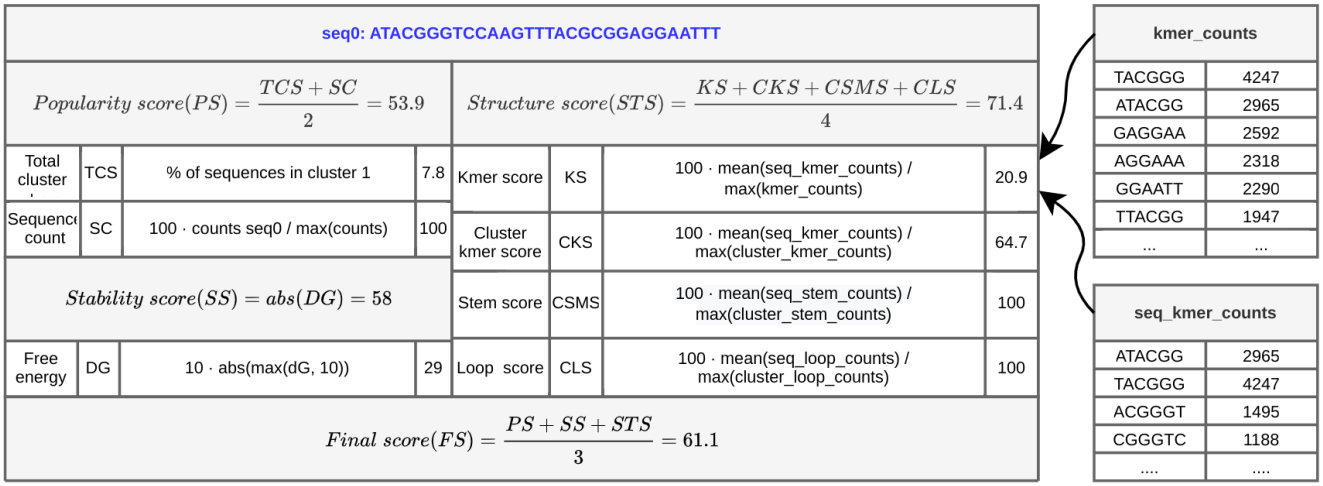

Figure 3. The evaluation of the final score for a single aptamer is evaluated using the popularity score, stability score, and the structure score.

Using this algorithm, we analyzed two of our previous aptamer selection results.[30, 31] Figure 4 lists the score distributions of the top 1000 highest scoring aptamers for each sequenced round for each target molecule. For the caffeine selection, the majority of the sequences have a score close to zero for the round 12 and 15 libraries, yet the score shifts to above 20 for the round 20 library. For the theophylline selection, a similar trend was observed, and high score sequences dominate only since around 18. For these two selections, the score distribution shifts towards highest scores at later rounds, suggesting successful selections. Nevertheless, a small fraction of high score sequences are still present for caffeine round 12 and theophylline round 10. When the library is still so diverse, it is difficult to perform rational manual analysis, but our algorithm has given high score to a few sequences. In the following sections, we analyzed individual selections and experimentally tested some predicted sequences.

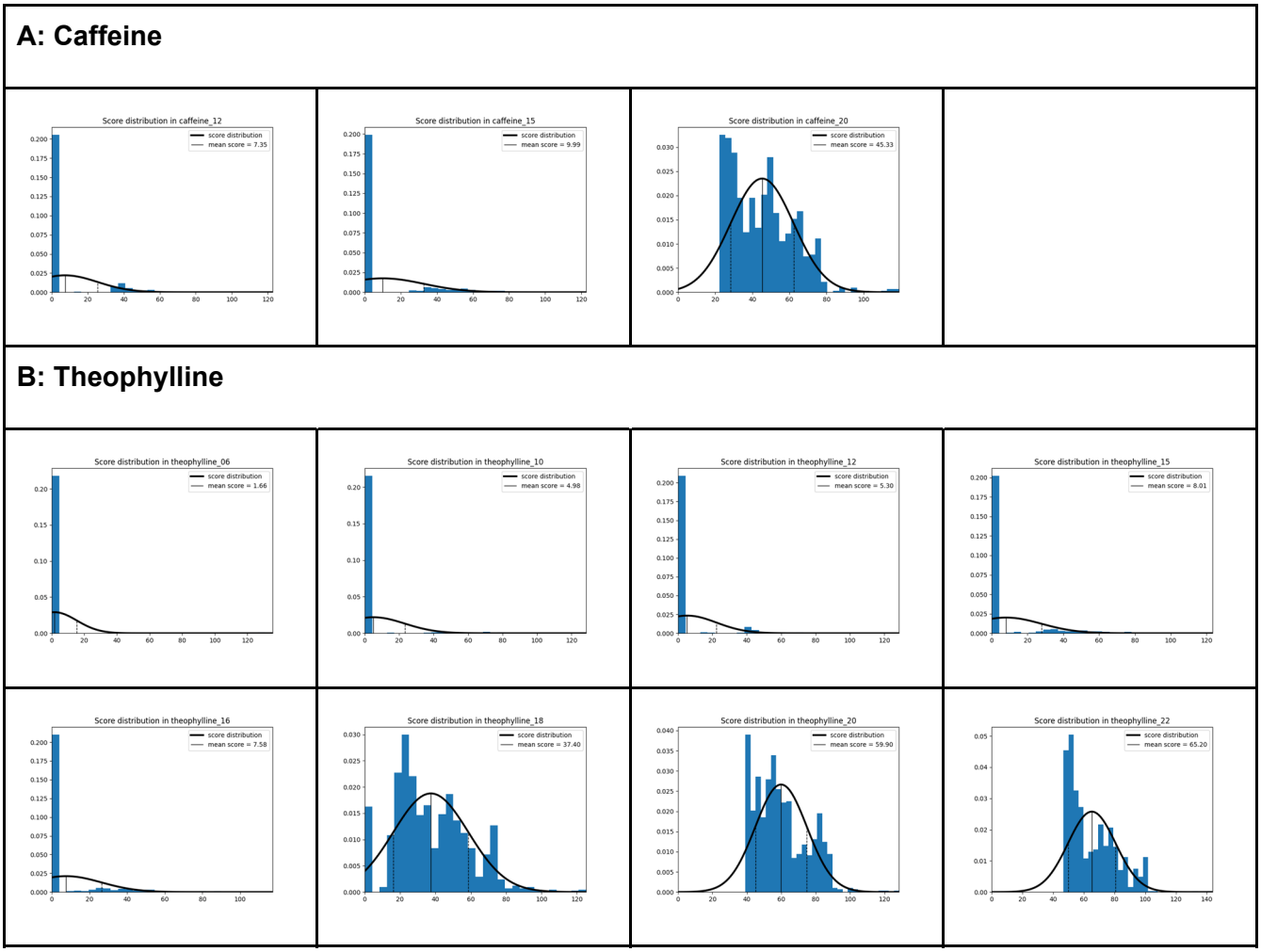

Figure 4. Histograms of sequence score distribution in caffeine and theophylline selections. (A) Round 12, 15 and 20 of caffeine selection libraries. (B) Round 6, 10, 12, 15, 16, 18, 20 and 22 of theophylline selection libraries.

Prediction of caffeine aptamers

The round 20 library of the caffeine selection contained at least five major families along with numerous orphan sequences in the top 70 most abundant sequences.[31] The most abundant family of round 20 contained over 16.59% of the sequenced library, and the other major families (with 5.48%, 3.83%, 2.76% …) also contained a few percent, and thus identification of the binding sequences was quite easy by testing the most abundant sequence in each family. Based on our previous results, all the tested sequences showed binding to caffeine.[31] When we applied our CPSPS algorithm to the round 20 library, eight out of the ten predicted sequences were in the top 10 most abundant sequences in the library (Figure S2), suggesting that the popularity scores might have dominated.

We then challenged our algorithm to the round 15 library, which was still far from converging, since even the most abundant sequence (52 copies) was below 0.1% of the entire library. The top ten predicted sequences are shown in Figure 5A (Seq0 means the highest score and Seq1 has the next score). Interestingly, two of the 10 predicted sequences were in the top 10 sequences in round 20. We then analyzed each of these top 10 predicted sequences in round 15. Seq0, Seq6, and Seq7 have the GGGGGA and GGAGGA conserved sequences and thus they belong to the same type. The fact that they switched sides in some sequences (see the sequence alignment in Figure 5B) indicated that these are likely to be real aptamers.[28, 32] Since they were not present in the top 70 sequences in round 20, we did not test them in our previous studies. Seq8 is the same as the 9th most abundant sequence in round 20, and its binding to caffeine was previously confirmed. Seq2 and Seq9 are similar to the top 36th and 40th sequences in round 20. Seq3 and Seq4 are not in the top 70 sequences of round 20 and thus they were not studied before.

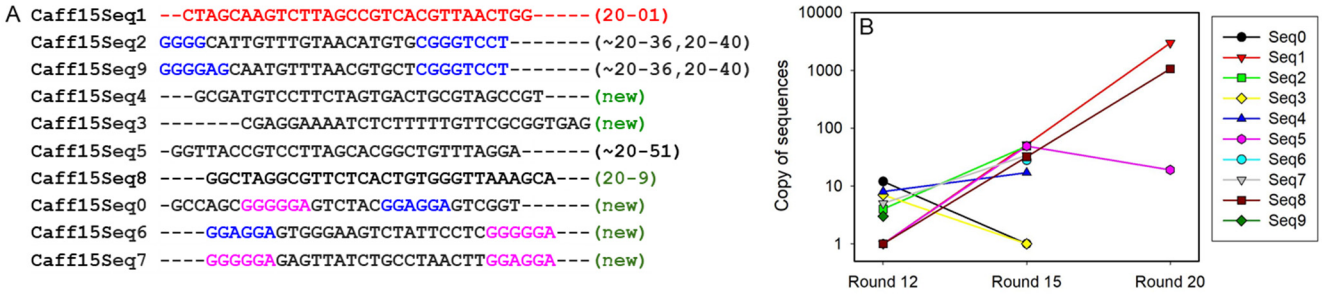

Figure 5. (A) Sequence alignment of the top 10 predicted sequences from the round 15 caffeine selection library. Seq0 means the highest prediction score by CPSPS, and 20-01 means round 20 the most abundant sequence. ~20-36, 20-40 means similar to the 36th and 40th most abundant sequences in round 20. (B) Evolution of the top 10 predicted round 15 sequences in different rounds.

For these 10 sequences predicted in round 15, we plotted their copy number in round 12 and round 20 (Figure 5B). Among them, Seq1 and Seq8 increased exponentially across these rounds, which fits the concept of exponential aptamer enrichment. Seq0 and Seq3 dropped to a single copy in round 15 and completely disappeared in round 20, yet they both had high scores in round 15. Thus, our CPSPS algorithm can pick up single-copy aptamer sequences in a library containing >50000 sequences.

We also tried our algorithm in the round 12 library, where the most abundant predicted sequence had less than 10 copies (Figure S3). Among the top 10 predicted, the sequences containing GGGGGA

and GGAGGA represent seven of them. For the rest three, they were the same as Seq3 and Seq4 in round 15. Yet, none of the top sequences in round 20 were predicted. For the most abundant aptamer sequence in round 20, which appeared 2985 times, it appeared 52 times in round 15 and only once in round 12. In the earlier rounds, the dominating predicted sequences completely disappeared in the round 20 predictions.

Based on the above discussion, we tested seq0, 2, 3, 4, 5, and 8 from round 15 for binding to caffeine using isothermal titration calorimetry (ITC, Figure 6). Their predicted secondary structures are also shown. All of these aptamers could bind caffeine with K_d values ranged from ~10 μM to 140 μM.

Overall, the prediction was successful and reliable, since all the predicted sequences worked. Since our method uses data from a single round and does not rely on exploring the evolution of kmers across rounds, we can still maintain good performance when data from very few rounds is available.

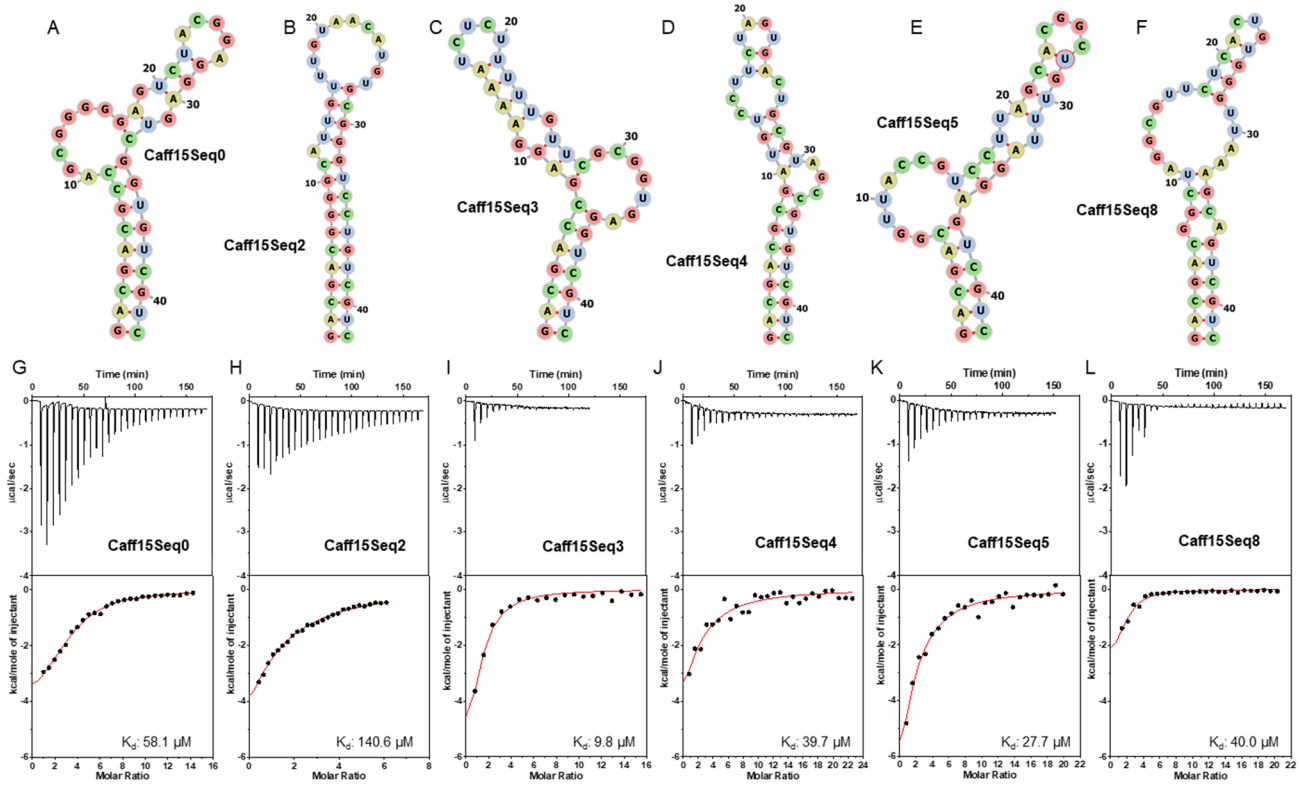

Figure 6. The secondary structures of some predicted aptamers for caffeine from the round 15 library: (A) Caff15Seq0, (B) Caff15Seq2, (C) Caff15Seq3, (D) Caff15Seq4, (E) Caff15Seq5, and (F) Caff15Seq8. In these structures, U needs to be replaced by T. ITC results of these caffeine aptamers: (G) titrating 3.5 mM caffeine into 50 μM Caff15Seq0, (H) titrating 1.5 mM caffeine into 50 μM Caff15Seq2, (I) titrating 1 mM caffeine into 9 μM Caff15Seq3, (J) titrating 1 mM caffeine into 9 μM Caff15Seq4, (K) titrating 1 mM caffeine into 9 μM Caff15Seq5, and (L) titrating 5 mM caffeine into 50 μM

Caff15Seq8.

Prediction of theophylline aptamers

To further test the algorithm, we also analyzed the theophylline selection.[30] When we did the theophylline selection, we initially stopped at round 15. However, the round 15 library was still very diverse, and we had to perform additional seven rounds of selections with gradually decreased theophylline concentrations. The round 22 library was highly converged, and its top 20 sequences represented 58% of the library. These top 20 sequences can be assigned into two families, which actually bind theophylline in the same way (just flipping the orientation).

Thus, the round 15 library was a good starting point for the analysis, in which the most abundant sequence was only 0.03% of the entire library. Among the top 10 predicted sequences (Figure 7A), seq0 and seq7 belong to the dominating families in round 22, although they contained only 20 and 9 copies (out of over 60,000 sequences), respectively. Thus, this is another example of predicting the right sequences when the library was far from convergence. Interestingly, these two particular sequences increased exponentially until round 18, after which they dropped significantly. Nevertheless, their family became dominating in round 22.

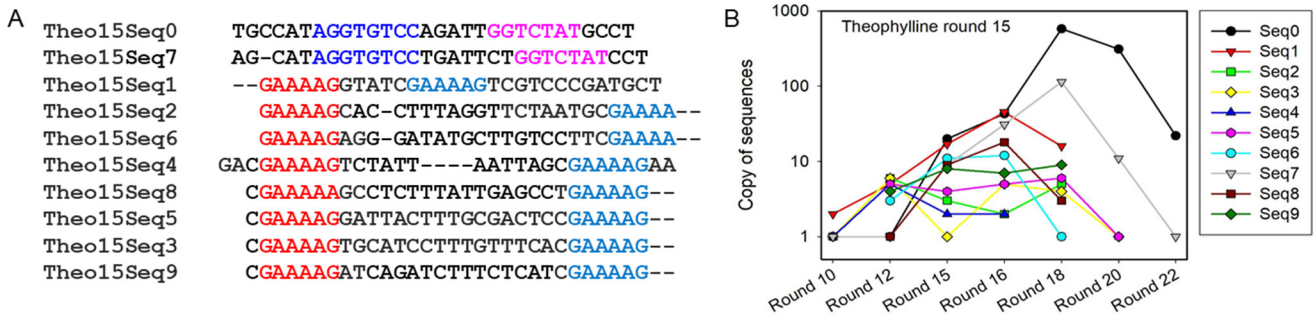

Figure 7. (A) Sequence alignment of the top 10 predicted sequences from the round 15 theophylline selection library. (B) Evolution of the top 10 predicted round 15 sequences in different rounds.

We tested the Theo15Seq0 (the highest score aptamer in round 15 Seq0, see Figure 8A for secondary structure) using ITC (Figure 8E). While it showed binding, the K_d value was 43.9 μM, 42-fold higher than Theo2203 (Figure 8B, 8E, the third most abundant sequence in round 22). It is understandable that the Theo15Seq0 dropped to 20 copies in round 22 since it has a low binding affinity. Although these two sequences have the same conserved regions and a hairpin connecting these regions, their differences outside these regions can also significantly influence binding affinity. Since the conserved regions are

the same, we can consider Theo15Seq0 and Theo2203 are evolutionarily related.

The remaining eight of the predicted round 15 sequences are also very interesting, and they all belong to one family with conserved GAAAAG and GAAAAG. Using ViennaRNA[33] to predict their secondary structures shows that these two conserved motifs are distributed on either side of a hairpin, and the hairpin sequences are highly variable. We measured the ITC of round 15 Seq6 (Figure 8C, Figure 8F). Indeed, binding was measured with a K_d of 22.7 μM. This is a new aptamer motif for theophylline that was not researched previously, which was fully disappeared in round 22.

In round 10, the top 10 predicted sequences had one of the GAAAAG regions replaced by GGAGGA. We tested Theo10Seq3 (Figure 8D, 8H), and a K_d of 96 μM was obtained. Given their purine-rich sequences (see the regions in the circle in Figure 8C and 8D), they might bind theophylline in the same way. One of such purine-rich sequences was even predicted in round 6, when the library was extremely diverse (Figure S4). Such sequences however were nearly fully eliminated in the later rounds when the theophylline concentration was dropped from 1 mM (round 15) to 0.05 mM (round 22) (Figure 7B). Therefore, the GAAAAG containing sequences are low-affinity aptamers for theophylline.

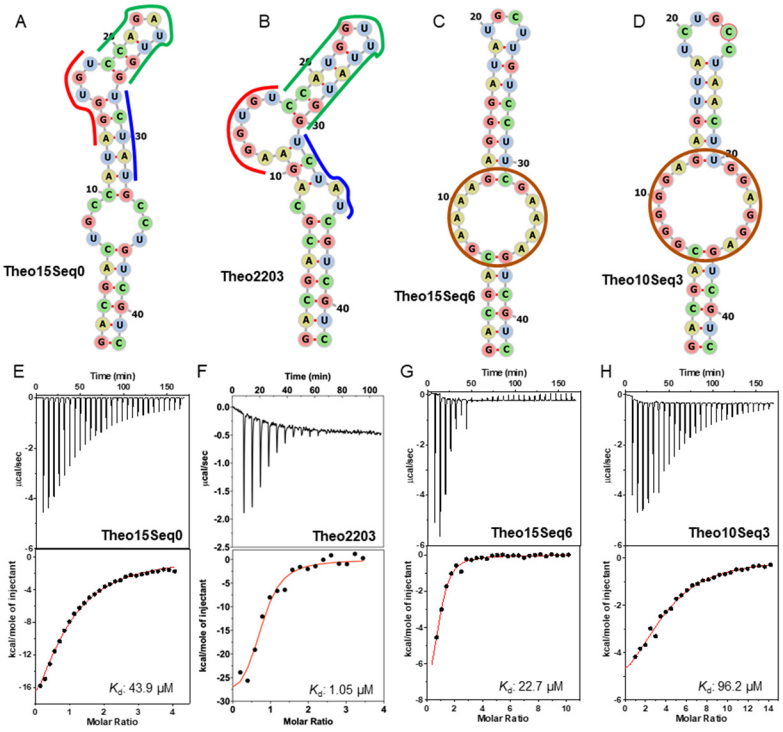

Figure 8. ITC traces and the secondary structures of (A) Seq6 of round 15, (B) Theo2203, the third most abundant sequence in round 22 theophylline selection, (C) Seq3 of round 10, and (D) Seq0 of round 15. In these structures, U needs to be replaced by T. In (A) and (B), the red and blue lines highlight the

identical sequences and the green lines show the hairpin structures. ITC traces of (E) titrating 1 mM theophylline into 50 μM Theo15Seq0, (F) titrating 1 mM theophylline into 9 μM Theo2203, (G) titrating 5 mM theophylline into 100 μM Theo15Seq5, and (H) titrating 3.5 mM theophylline into 50 μM Theo10Seq3.

Discussion

Using our CPSPS algorithm, we predicted aptamers that were previously ignored. In addition, by identifying conserved the primary sequence and secondary structures, our algorithm was able to predict aptamers from early rounds of selections. Based on our observations in this work, the following points are discussed.

1. Our CPSPS algorithm works independently on each round of sequencing results. It is based on the intrinsic combined primary and secondary structure patterns expected for aptamers in the libraries. This feature sets our algorithm apart from most previous methods that either rely on sequence evolution in different rounds or based solely on primary sequence alignment.
2. For a typical selection experiment, early rounds of selections are often done with a high concentration of target analytes. Thus, the enriched aptamers may contain both high affinity and low-affinity ones. CPSPS would not know if an aptamer is high or low affinity based on a single round of data. When a few rounds of results are compared, then high and low-affinity aptamers are more obvious. Those dropped in population in the later rounds are likely to be low-affinity aptamers. Therefore, it is important to experimentally push for high-affinity aptamers by using a few rounds of low analyte concentration selection. The overall number of rounds may not need to be high since the algorithm can help identify sequences.
3. From the caffeine and theophylline example, the sequences predicted in the early rounds (low affinity sequences) may not have evolutionary relationship to the ones in the later rounds (high affinity sequences). Thus, an implication is that it is possible to use a low target concentration to begin with.
4. For any algorithm to work, the selection needs to enrich actual aptamers. If the design of the selection experiment cannot achieve this goal, no algorithm would work. Thus, it is important to have well-executed aptamer selection experiments.

Conclusions

In summary, we developed a new algorithm to recognize the pattern of evolved aptamers based on both 6-mer primary sequence motifs and hairpin secondary structures. While such hairpins are often considered as random sequences by primary sequence clustering, they have extra scores in our algorithm since they also reflect the evolution of the libraries. We applied this algorithm to two separate SELEX experiments with a total of eight sequenced libraries. They contain both highly converged and highly diverse libraries. In each case, the algorithm was able to predict aptamer sequences that were verified by experiments. This algorithm can be applicable to other aptamer selection experiments and can be highly valuable when the library is still quite diverse. Such predicted aptamers can be of great interest in designing analytical biosensors.

Acknowledgements

Funding for this work was from NSERC and a WIN-NRC Nanotechnology Joint Seed Funding Program.

Supporting Information

DNA sequences tested in this work and additional aptamer alignment data.

References

(1) Yu, H.; Alkhamis, O.; Canoura, J.; Liu, Y.; Xiao, Y., Advances and Challenges in Small-Molecule DNA Aptamer Isolation, Characterization, and Sensor Development. *Angew. Chem. Int. Ed.* **2021**, 60, 16800-16823.

(2) Wu, L.; Wang, Y.; Xu, X.; Liu, Y.; Lin, B.; Zhang, M.; Zhang, J.; Wan, S.; Yang, C.; Tan, W., Aptamer-Based Detection of Circulating Targets for Precision Medicine. *Chem. Rev.* **2021**, 121, 12035–12105.

(3) McConnell, E. M.; Nguyen, J.; Li, Y., Aptamer-Based Biosensors for Environmental Monitoring. *Frontiers in Chemistry* **2020**, 8.

(4) Roth, A.; Breaker, R. R., The Structural and Functional Diversity of Metabolite-Binding Riboswitches. *Annu. Rev. Biochem* **2009**, 78, 305-334.

(5) Zhang, J.; Jensen, M. K.; Keasling, J. D., Development of Biosensors and Their Application in Metabolic Engineering. *Curr. Opin. Chem. Biol.* **2015**, 28, 1-8.

(6) Yüce, M.; Ullah, N.; Budak, H., Trends in Aptamer Selection Methods and Applications. *Analyst* **2015**, 140, 5379-5399.

(7) Lyu, C.; Khan, I. M.; Wang, Z., Capture-Selex for Aptamer Selection: A Short Review. *Talanta* **2021**, 229, 122274.

(8) Zuker, M., Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, 31, 3406-3415.

(9) Chowdhury, B.; Garai, G., A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, 109, 419-431.

(10) Song, J.; Zheng, Y.; Huang, M.; Wu, L.; Wang, W.; Zhu, Z.; Song, Y.; Yang, C., A Sequential Multidimensional Analysis Algorithm for Aptamer Identification Based on Structure Analysis and Machine Learning. *Anal. Chem.* **2020**, 92, 3307-3314.

(1) Yu, H.; Alkhamis, O.; Canoura, J.; Liu, Y.; Xiao, Y., Advances and Challenges in Small-Molecule DNA Aptamer Isolation, Characterization, and Sensor Development. *Angew. Chem. Int. Ed.* **2021**, 60, 16800-16823.

(2) Wu, L.; Wang, Y.; Xu, X.; Liu, Y.; Lin, B.; Zhang, M.; Zhang, J.; Wan, S.; Yang, C.; Tan, W., Aptamer-Based Detection of Circulating Targets for Precision Medicine. *Chem. Rev.* **2021**, 121, 12035−12105.

(3) McConnell, E. M.; Nguyen, J.; Li, Y., Aptamer-Based Biosensors for Environmental Monitoring. *Frontiers in Chemistry* **2020**, 8, 434.

(4) Yüce, M.; Ullah, N.; Budak, H., Trends in Aptamer Selection Methods and Applications. *Analyst* **2015**, 140, 5379-5399.

(5) Lyu, C.; Khan, I. M.; Wang, Z., Capture-Selex for Aptamer Selection: A Short Review. *Talanta* **2021**, 229, 122274.

(6) Zuker, M., Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, 31, 3406-3415.

(7) Chowdhury, B.; Garai, G., A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, 109, 419-431.

(8) Bashir, A.; Yang, Q.; Wang, J.; Hoyer, S.; Chou, W.; McLean, C.; Davis, G.; Gong, Q.; Armstrong, Z.; Jang, J.; Kang, H.; Pawlosky, A.; Scott, A.; Dahl, G. E.; Berndl, M.; Dimon, M.; Ferguson,

B. S., Machine Learning Guided Aptamer Refinement and Discovery. *Nature Communications* **2021**, 12, 2366.

(9) Chen, Z.; Hu, L.; Zhang, B.-T.; Lu, A.; Wang, Y.; Yu, Y.; Zhang, G., Artificial Intelligence in Aptamer–Target Binding Prediction. *International Journal of Molecular Sciences* **2021**, 22, 3605.

(10) Knight, C. G.; Platt, M.; Rowe, W.; Wedge, D. C.; Khan, F.; Day, P. J. R.; McShea, A.; Knowles, J.; Kell, D. B., Array-Based Evolution of DNA Aptamers Allows Modelling of an Explicit Sequence-Fitness Landscape. *Nucleic Acids Res.* **2008**, 37, e6-e6.

(11) Buglak, A. A.; Samokhvalov, A. V.; Zherdev, A. V.; Dzantiev, B. B., Methods and Applications of in Silico Aptamer Design and Modeling. *International Journal of Molecular Sciences* **2020**, 21, 8420.

(12) Lee, G.; Jang, G. H.; Kang, H. Y.; Song, G., Predicting Aptamer Sequences That Interact with Target Proteins Using an Aptamer-Protein Interaction Classifier and a Monte Carlo Tree Search Approach. *PLOS ONE* **2021**, 16, e0253760.

(13) Bottari, F.; Daems, E.; de Vries, A.-M.; Van Wielendaele, P.; Trashin, S.; Blust, R.; Sobott, F.; Madder, A.; Martins, J. C.; De Wael, K., Do Aptamers Always Bind? The Need for a Multifaceted Analytical Approach When Demonstrating Binding Affinity between Aptamer and Low Molecular Weight Compounds. *J. Am. Chem. Soc.* **2020**, 142, 19622–19630.

(14) Daems, E.; Moro, G.; Campos, R.; De Wael, K., Mapping the Gaps in Chemical Analysis for the Characterisation of Aptamer-Target Interactions. *TrAC, Trends Anal. Chem.* **2021**, 142, 116311.

(15) Zong, C.; Liu, J., The Arsenic-Binding Aptamer Cannot Bind Arsenic: Critical Evaluation of Aptamer Selection and Binding. *Anal. Chem.* **2019**, 91, 10887-10893.

(16) Zhao, Y.; Yavari, K.; Liu, J., Critical Evaluation of Aptamer Binding for Biosensor Designs. *TrAC, Trends Anal. Chem.* **2022**, 146, 116480.

(17) Zara, L.; Achilli, S.; Chovelon, B.; Fiore, E.; Toulmé, J.-J.; Peyrin, E.; Ravelet, C., Anti-Pesticide DNA Aptamers Fail to Recognize Their Targets with Asserted Micromolar Dissociation Constants. *Anal. Chim. Acta* **2021**, 1159, 338382.

(18) Heredia, F. L.; Roche-Lima, A.; Parés-Matos, E. I., A Novel Artificial Intelligence-Based Approach for Identification of Deoxynucleotide Aptamers. *PLOS Computational Biology* **2021**, 17, e1009247.

(19) Takahashi, M.; Wu, X.; Ho, M.; Chomchan, P.; Rossi, J. J.; Burnett, J. C.; Zhou, J., High Throughput Sequencing Analysis of RNA Libraries Reveals the Influences of Initial Library and Pcr Methods on Selex Efficiency. *Scientific Reports* **2016**, 6, 33697.

(20) Iwano, N.; Adachi, T.; Aoki, K.; Nakamura, Y.; Hamada, M., Generative Aptamer Discovery Using Raptgen. *Nature Computational Science* **2022**, 2, 378-386.

(21) Hoinka, J.; Zotenko, E.; Friedman, A.; Sauna, Z. E.; Przytycka, T. M., Identification of Sequence–Structure RNA Binding Motifs for Selex-Derived Aptamers. *Bioinformatics* **2012**, 28, i215-i223.

(22) Song, J.; Zheng, Y.; Huang, M.; Wu, L.; Wang, W.; Zhu, Z.; Song, Y.; Yang, C., A Sequential Multidimensional Analysis Algorithm for Aptamer Identification Based on Structure Analysis and Machine Learning. *Anal. Chem.* **2020**, 92, 3307-3314.

(23) Nakatsuka, N.; Yang, K.-A.; Abendroth, J. M.; Cheung, K. M.; Xu, X.; Yang, H.; Zhao, C.; Zhu, B.; Rim, Y. S.; Yang, Y.; Weiss, P. S.; Stojanović, M. N.; Andrews, A. M., Aptamer–Field-Effect Transistors Overcome Debye Length Limitations for Small-Molecule Sensing. *Science* **2018**, 362, 319-324.

(24) Yang, K.-A.; Chun, H.; Zhang, Y.; Pecic, S.; Nakatsuka, N.; Andrews, A. M.; Worgall, T. S.; Stojanovic, M. N., High-Affinity Nucleic-Acid-Based Receptors for Steroids. *ACS Chemical Biology* **2017**, 12, 3103-3112.

(25) Yu, H.; Luo, Y.; Alkhamis, O.; Canoura, J.; Yu, B.; Xiao, Y., Isolation of Natural DNA Aptamers for Challenging Small-Molecule Targets, Cannabinoids. *Anal. Chem.* **2021**, 93, 3172-3180.

(26) Zhao, Y.; Ong, S.; Chen, Y.; Jimmy Huang, P.-J.; Liu, J., Label-Free and Dye-Free Fluorescent Sensing of Tetracyclines Using a Capture-Selected DNA Aptamer. *Anal. Chem.* **2022**, 94, 10175–10182.

(27) Luo, Y.; Jin, Z.; Wang, J.; Ding, P.; Pei, R., The Isolation of a DNA Aptamer to Develop a Fluorescent Aptasensor for the Thiamethoxam Pesticide. *Analyst* **2021**, 146, 1986-1995.

(28) Liu, Y.; Liu, J., Selection of DNA Aptamers for Sensing Uric Acid in Simulated Tears. *Analysis & Sensing* **2022**, 2, e202200010.

(29) Smith, T. F.; Waterman, M. S., Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, 147, 195-197.

(30) Huang, P.-J. J.; Liu, J., A DNA Aptamer for Theophylline with Ultrahigh Selectivity Reminiscent of the Classic RNA Aptamer. *ACS Chemical Biology* **2022**, https://doi.org/10.1021/acschembio.2c00179.

(31) Huang, P.-J. J.; Liu, J., Selection of Aptamers for Sensing Caffeine and Discrimination of Its Three Single Demethylated Analog. *Anal. Chem.* **2022**, 94, 3142–3149.

(32) Jenison, R. D.; Gill, S. C.; Pardi, A.; Polisky, B., High-Resolution Molecular Discrimination by RNA. *Science (Washington, DC, United States)* **1994**, 263, 1425-1429.

(33) Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L., Viennarna Package 2.0. *Algorithms for Molecular Biology* **2011**, 6, 26.