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

Javier Perez Tobia[1], Po-Jung Jimmy Huang[2], Yuzhe Ding,[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, British Columbia, V1V 1V7, 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, Ontario, N6A 3K7, Canada

Email: apurva.narayan@uwo.ca

4. Systems Design Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, 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 the caffeine and theophylline selections. In the late rounds of the selections, when the library were converged, the predicted sequences matched well with the most abundant sequences. When the library was far from convergence and the sequences were deemed challenging for traditional analysis methods, this 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 biosensor and chemical biology 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 for biosensors and nanotechnology applications has been inspired by their competitive advantages to antibodies including higher stability, reversible denaturation, lower cost, programmable structures and easier modification.[4-6] The majority of DNA aptamers were obtained via combinatorial selections.[7, 8] Typical aptamer selections require over ten to twenty 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 highly desirable.

Computer-aided prediction of secondary structures of nucleic acids and sequence alignment have already become common tools in aptamer research.[9-11] The advent of high throughput sequencing allows the generation of overwhelming volume of information, which provides data for machine learning. One type of aptamer prediction is based on *de novo* design of aptamer secondary/tertiary structures and molecular docking. However, its success has been limited and few predicted aptamers were widely used.[12-14] Another approach uses previously reported aptamers as a training set to predict new aptamers.[15] Since many recent papers showed that not all published aptamer sequences are reliable,[16-20] the quality of the input data need to be carefully examined.

A third type of algorithm analyzed aptamer selection libraries.[21] 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 mainly in the later rounds of selections, and other factors such as PCR bias can also lead to preferentially amplified sequences that are not necessarily aptamers.[22] In addition, motif finding algorithms have been developed,[12, 21, 23] which rely on predicted secondary structures and dividing the structures into multiple subunits. A motif is defined based on the sequence of single-stranded regions,[24] 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 and coworkers 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).[25] This algorithm was applied to the selection libraries for human embryonic stem cells, epithelial cell adhesion molecules, and cell-surface vimentin. While secondary structure information was considered, calculation was mainly based on the

overall free energy of DNA folding. In addition, sequence 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 a DNA library and taking into account both conserved primary sequences and secondary structures. We call it conserved primary/secondary structure clustered pattern searching CPS^2, which can identify aptamers with sequencing data from only one round. Since we required one round of selection data, it is a one-cycle learning algorithm. Using Python as the coding language, CPS^2 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 CPS^2 and identified a few new aptamers that were previously neglected. We were able to identify aptamers when their sequence frequency was less than 0.1% of the library.

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.[26-30] Aptamer binding often does not need such a long sequence, and a common strategy for a library to evolve is to hide redundant nucleotides in a hairpin. The sequence of such a hairpin is not conserved, but the presence of such hairpins is conserved (Figure S1).[26, 31-33] For example, Figure 1A shows two aptamers that can bind uric acid.[31] 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 of the hairpins 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 the hairpins would not contribute to scoring, but we aim to count the hairpins as a positive score. To highlight the contribution of hairpins, hairpins are counted as extra scores in our analysis.

It needs to be noted that for any given length of random DNA sequence, forming a hairpin is statistically disfavored. For the example in Figure 1 with a 14-nt region to form a hairpin with a 3 bp stem, the probability is only $(1/4)^3$=1.6%. The chance of forming a 5 bp hairpin is even smaller. Therefore, the prevalence of such hairpins suggests an evolutionary advantage and can be taken as a feature of aptamers. Thermodynamically, pre-arranging a fraction of unimportant nucleotides in a hairpin

can reduce the entropic penalty of aptamer binding reactions. Kinetically, the nucleotides in a hairpin are unlikely to involve in misfolding of aptamers, allowing faster target binding. Both factors would favor the selection of better aptamers. Our CPS[2] algorithm aimed to capture such hairpin structures. We reason that focusing on such hairpins are more targeted than the analysis of global secondary structures.

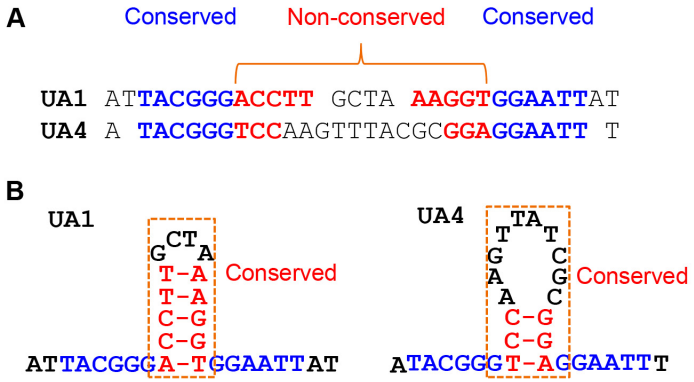


Figure 1. (A) The aligned primary sequences from the random region of two aptamers named UA1 and UA4, respectively.[31] The conserved sequences are in blue, and the middle 14 nucleotides are not conserved in the primary sequence. (B) Their folded secondary structures, where the conserved hairpins structures are highlighted.

An example to illustrate the CPS[2] algorithm

To illustrate our CPS[2] scoring system, we give the following example with sequences from the round 16 of the uric acid selection.[31] For Step 1, library processing, out of the entire 55985 sequences obtained, eight are shown in Figure 2A (on the left). The panel 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.[34] We use the bracket notation for expressing the sequences and highlight hairpin structures. 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[35] and added all the sequences with an alignment score higher than 0.8 to the cluster. 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 S2).

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.

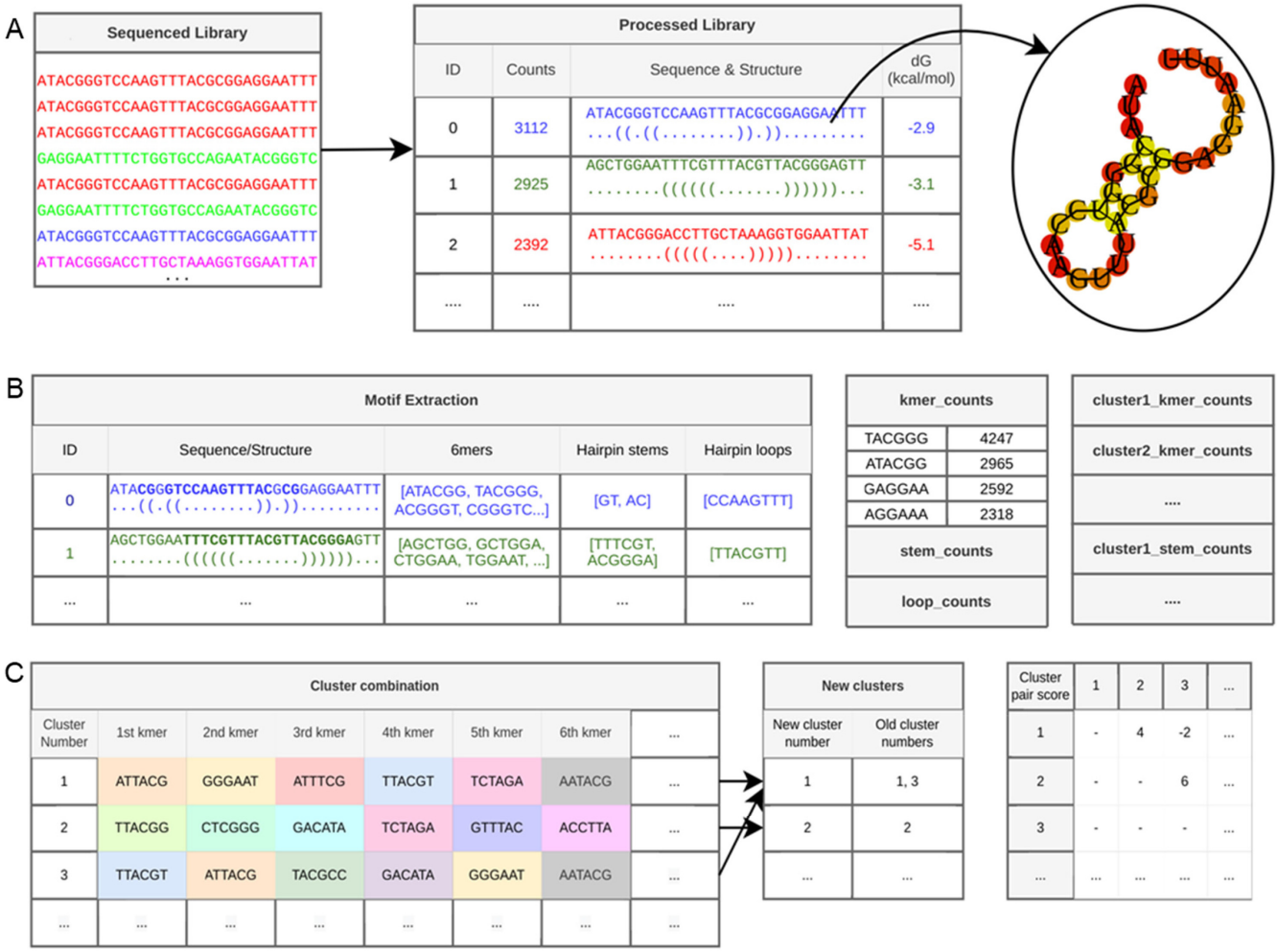


Figure 2. (A) An example showing top sequences forming top-3 clusters for illustration purposes. On the left, the same colored sequences have identical sequence. (B) Cluster merging based on our scoring mechanism as described in (C). (C) Clustering k-mer alignment score between two clusters.

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.

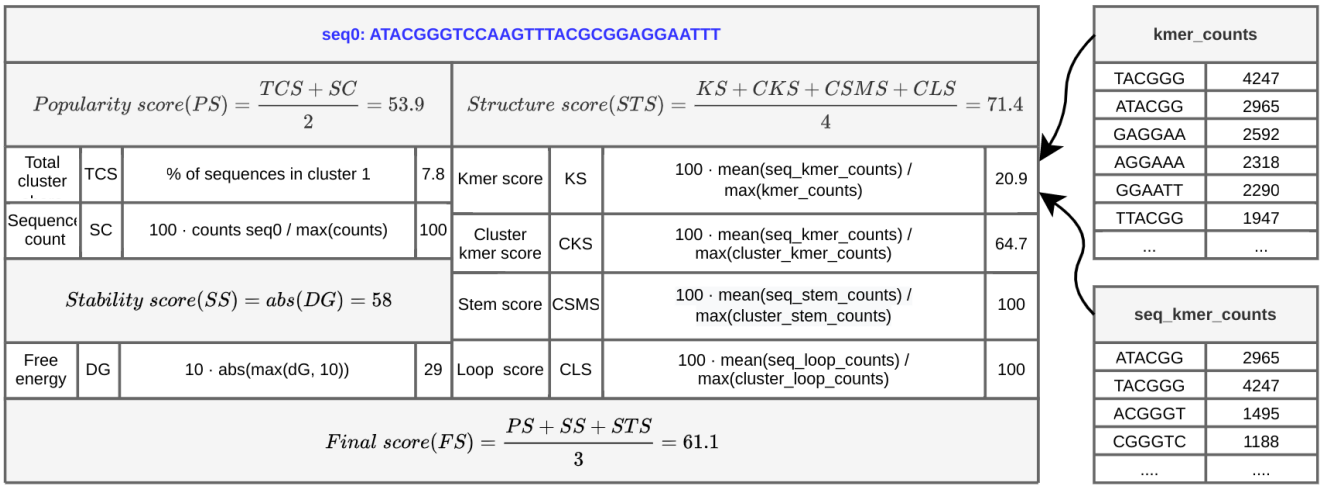


Figure 3. The evaluation of the final score for a single aptamer is evaluated using the popularity score, stability score, and the structure score.

Using the proposed algorithm (described in the materials section), we analyzed two of our previous aptamer selection results. We did not sequence the early rounds of the uric acid selection and its round 16 library was already highly converged. Therefore, we picked the caffeine and theophylline selections to better illustrate the algorithm especially in the early rounds of selections.[36, 37] 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.

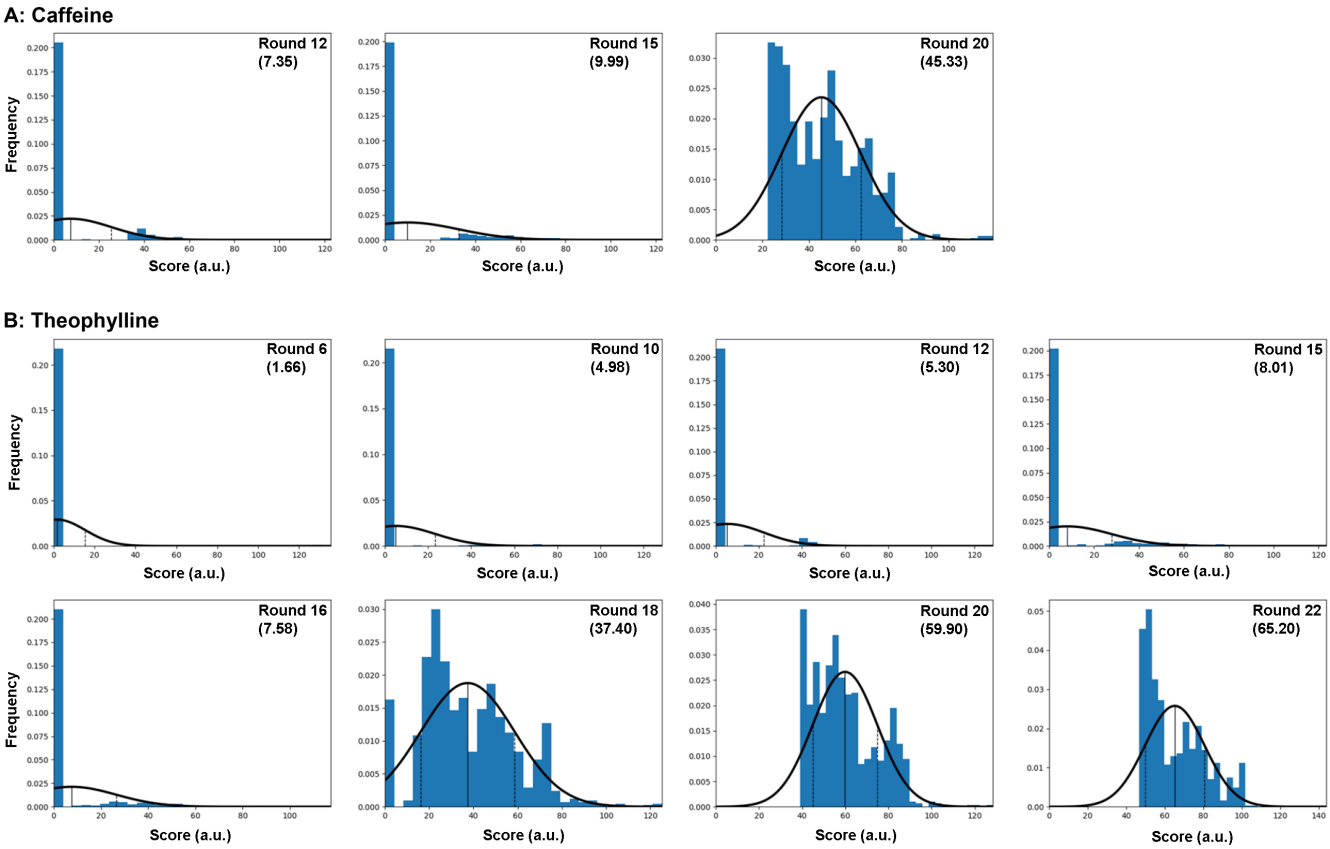


Figure 4. Histograms of sequence score distribution in the caffeine and theophylline selections. (A) Round 12, 15 and 20 of caffeine selection libraries. (B) Round 6, 10, 12, 15, 16, 18, 20 and 22 of theophylline selection libraries. The black curves are the fitted score distributions, and the mean scores are shown in the brackets on the top right corner in each panel and are also indicated by the solid vertical lines.

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.[37] 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 with dissociation constants (K_d) around 3 μM.[37] When we applied our CPS^2 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 S3), 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 second highest score). These ten sequences had scores ranged from 105.7 to 127.7. Interestingly, two of the 10 predicted sequences were in the top 10 sequences in round 20 (20-01 and 20-9). We then analyzed each of these top 10 predicted sequences in round 15. Seq0, Seq6, and Seq7 have the same conserved motifs and thus they belong to the same type. The fact that they switched sides in some sequences (see the sequence alignment in Figure 5A) indicated that these are likely to be real aptamers.[31, 38] Since they nearly fully disappeared in round 20, we did not test them in our previous study.[37] 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.

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 CPS^2 algorithm can pick single-copy aptamer sequences in a library containing >50000 sequences. The extra scores to their hairpin structures are responsible for it.

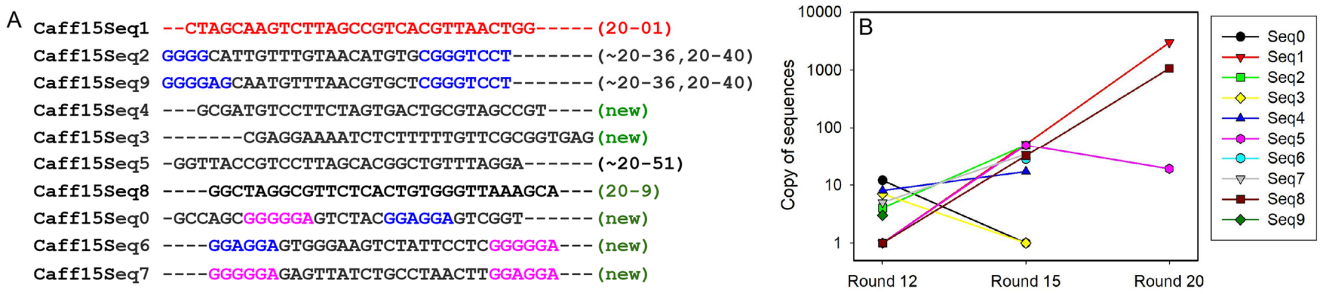


Figure 5. (A) Alignment of the top 10 sequences predicted by the CPS^2 algorithm from the round 15 caffeine selection. Seq0 means the highest prediction score by CPS^2 and 20-01 means the most abundant sequence in round 20. ~20-36, 20-40 means similar to the 36th and 40th most abundant sequences in round 20. (B) Counts of the top 10 predicted round 15 sequences in different rounds. In round 20, many sequences fully disappeared and only three survived.

We also tried our algorithm in the round 12 library, where the most abundant predicted sequence had less than 10 copies (Figure S4). Among the top 10 predicted sequences, seven contained GGGGGA

and GGAGGA. This is reminiscent of the classic adenosine and ATP aptamer.[39] For the rest three, they were the same as Seq3 and Seq4 in round 15. Yet, none of the most abundant sequences in round 20 were predicted. For the most abundant aptamer sequence in round 20 (appeared 2985 times), it appeared 52 times in round 15 and only once in round 12.

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).[40] These are either new sequences or sequences we did not measure before. For the sequences with the same conserved nucleotides, we only tested one of them. Their predicted secondary structures are also shown (Figure 6A-F). All of these aptamers could bind caffeine with K_d values ranged from ~10 μM to 140 μM (Figure 6G-L). As a control, we also tested a low score sequence from round 12 (Caf12Seq46638) and no binding was detected by ITC (Figure S5A).

Overall, the prediction was successful, since all the predicted sequences showed binding. Since our method used 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 are available.

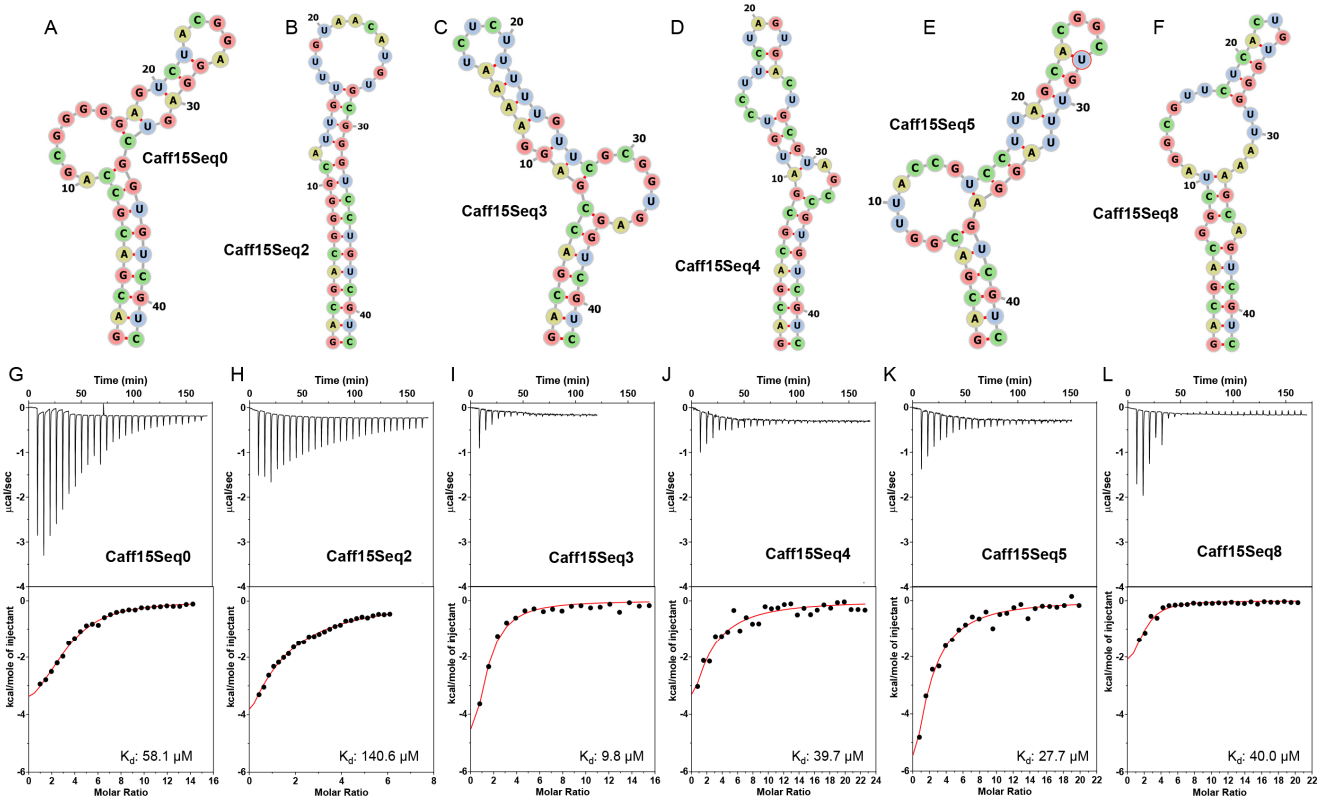


Figure 6. The secondary structures of some predicted aptamers for caffeine from the round 15 library drawn by ViennaRNA: (A) Caff15Seq0, (B) Caff15Seq2, (C) Caff15Seq3, (D) Caff15Seq4, (E) Caff15Seq5, and (F) Caff15Seq8. In these structures, U needs to be replaced by T. ITC results of these caffeine 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.[36] 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 (circular permutation sequences).

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), the scores ranged from 92.4 to 129.6, while the majority of the sequences had scores close to zero (Figure 4B). 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. Therefore, 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.

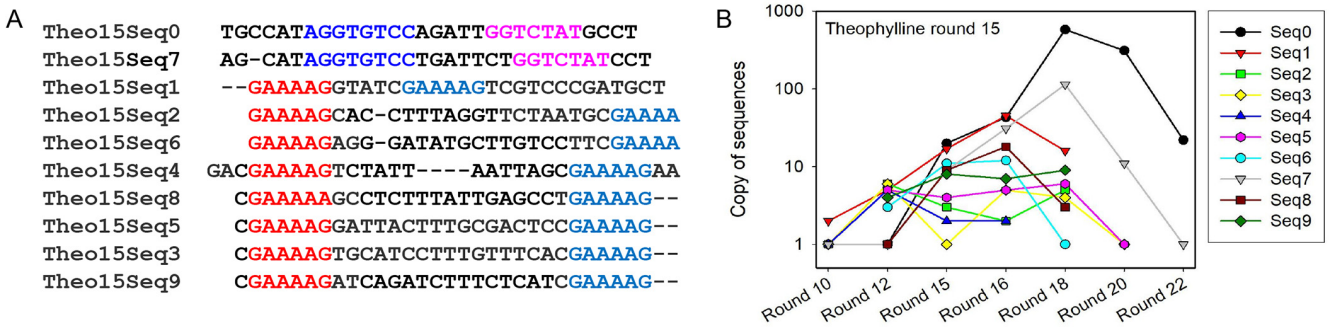


Figure 7. (A) Sequence alignment of the top 10 algorithm predicted sequences from the round 15 theophylline selection library. (B) Evolution of the top 10 predicted round 15 sequences in different rounds.

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, 8F, the third most abundant sequence in round 22). As a reference

point, the tested aptamers from the round 22 selection had K_d values in the 0.5 to 1 μM range.[36] 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 influenced the binding affinity. Since the conserved regions are the same, we reasoned that Theo15Seq0 and Theo2203 are evolutionarily related.

The remaining eight of the predicted round 15 sequences are also very interesting, since they all belong to one family with conserved GAAAA and GAAAA. The ViennaRNA[34] predicted secondary structures show that these two conserved motifs connected two stems, and one of the stems belong to a hairpin with are highly variable sequences. Therefore, this algorithm indeed can reflect conserved hairpin structures. We measured the ITC of round 15 Seq6 (Figure 8C, Figure 8G), and binding was measured with a K_d of 22.7 μM. This is an aptamer motif for theophylline that was not researched previously, which fully disappeared in round 22. As a negative control, we also tested a low score sequence (Theo10Seq48097) and no binding was detected by ITC (Figure S5B).

In round 10, most of the top 10 predicted sequences had the GGGGA and GGAGGA (Figure S6), which was similar to that found in the round 12 of the caffeine selection. We tested Theo10Seq3 (Figure 8D, 8H), and a K_d of 96 μM was obtained. In round 6, many purine-rich sequences were also predicted, when the library was extremely diverse (Figure S7). 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 purine-rich sequences are low-affinity aptamers for theophylline.

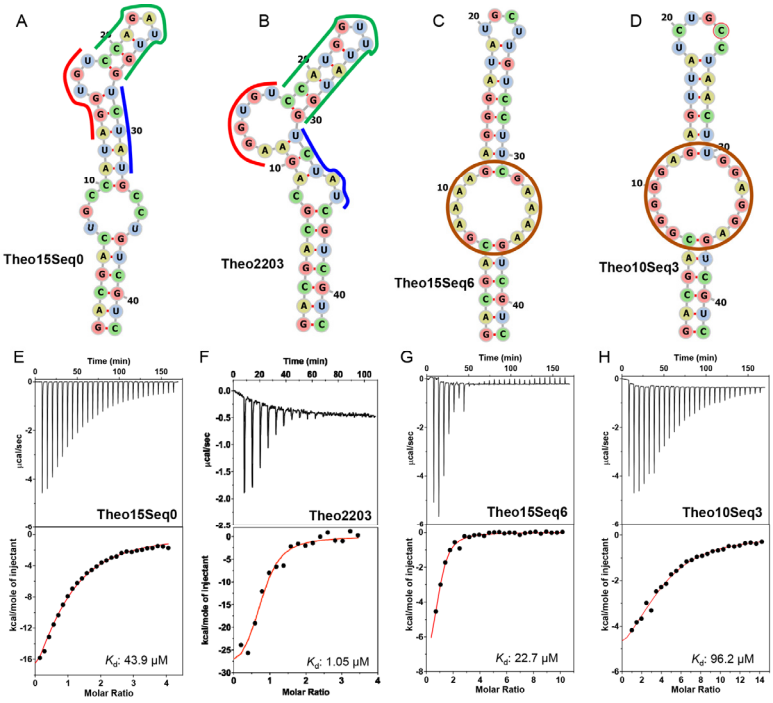


Figure 8. ITC traces and the secondary structures of (A) Seq0 of round 15, (B) Theo2203, the third most abundant sequence in round 22 theophylline selection, (C) Seq6 of round 15, and (D) Seq3 of round 10. In these structures, U needs to be replaced by T. In (A) and (B), the red and blue lines highlight 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. (F) Adapted from ref.[36] with permission. Copyright 2022 American Chemical Society.

Additional discussion

With the above sequence analysis and ITC experiments, we have made a number of interesting observations. Our CPS^2 algorithm predicted aptamers that were previously ignored. In addition, by identifying the conserved primary sequences and secondary hairpin structures, our algorithm was able to predict aptamers from early rounds of selections. It works independently on each round of sequencing results based on the intrinsic combined primary and secondary structure patterns expected for aptamers in the libraries. Based on our observations in this work, the following points are discussed.

1. Contributions of hairpin structures to final scores. In the example in Figure 2, the hairpin contributed 70% of the structure score in this particular sequence, and the structure score carried the most weight in all the three score components. This extra hairpin score sets our algorithm apart from the previous work focused on primary sequence alignment.[12, 21, 23, 25] Different hairpin sequences might be predicted in different clusters (e.g. the purine-rich sequences in Figure 5A and 7A), although they may belong to the same family as illustrated in Figure 1. The presence of multiple sequences with different hairpins is an indication of successful selection and prediction. We noticed that some low score sequences also contained the same conserved regions as some high score ones. A future development could be to pool all these sequences as one family, just like what one would do with manual analysis.

2. For a typical selection experiment, early rounds of selections are often done with a high concentration of target analytes.[26, 29] Thus, enriched sequences may contain both high affinity and low-affinity aptamers. CPS^2 would not know if an aptamer is high or low affinity based on a single round of data. For example, the highest score sequence in round 15 of the theophylline selection (Theo15Seq0) is a low affinity sequence, although it has the same conserved region with a high affinity sequence. We searched the top five most abundant round 22 sequences in the round 15 library, and found that they presented in very low copy numbers (1, 6, 0, 1, and 1 copies for the top five sequences). These top five round 22

sequences were confirmed to be high affinity aptamers. Theo15Seq0 has 28 copies in the round 15 library, and thus Theo15Seq0 had the highest score due to its relatively high popularity. In addition, Theo15Seq0 and the most abundant sequences in round 22 have similar conserved regions. Therefore, at a low selection pressure (e.g. round 15), it is difficult to tell high and low affinity sequences based on the scores alone.

When a few rounds of results are compared, those dropped in 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. In the case where the high affinity and low affinity aptamers do not have evolutionary relationship, direct use of a low concentration of target is recommended. On the other hand, if there is evolutionary relationship, using gradually decreased concentrations might be a safer way (since sometimes it is hard to know what an appropriate low concentration to try is). That being said, it is difficult to know beforehand regarding the outcome of sequences. The overall number of rounds may not need to be high since the algorithm can help identify aptamer sequences. For example, the most abundant sequence in round 20 of the caffeine selection (a high affinity aptamer) was predicted in round 15 as the sequence with the second highest score. At round 15, the majority of the sequences still had a low score close to zero (Figure 4A). Therefore, applying this algorithm to round 15 of the caffeine selection would be a success since it can save five rounds of the selection yet give a high affinity aptamer.

3. Regarding the score provided by the algorithm, the maximum score can be greater than 100, and the lower end approaches zero. Instead of waiting for most sequences in a library to reach a high score, one might stop a selection when a small fraction of sequences reach a high score. Still using the caffeine example, applying CPS^2 to round 12 did not result in high affinity aptamers (although the predicted sequences can still bind caffeine). At round 20, when the entire library had high scores, this algorithm becomes less valuable since good aptamers can be obtained simply based on abundance. So, round 15 might be a good place to stop the selection. We recommend to combine the CPS^2 score with target concentration. A high score obtained in a low target concentration (e.g. <100 μM) is likely to result in better aptamers than a high score obtained a high target concentration (e.g. > 1 mM). Of note, the high and low target concentrations are different for different targets, depending their possible interactions with DNA. For example, 25 mM glucose might be a low concentration (e.g. aptamer K_d 10 mM), while 1 mM dopamine might be high (aptamer K_d 150 nM).[26] So, it is important to rationally use target concentration based on the structure of target molecules.

4. For any algorithm to work, the selection needs to enrich actual aptamers. If selection experiments cannot achieve this goal, no algorithm would work. As shown in a few recent publications, some

published aptamers failed to bind to their intended targets,[16, 18-20, 41] and applying this algorithm to such selections may not produce correct sequences. 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. The chance for a random sequence to form a stable hairpin is statistically quite low. We applied this algorithm to two separate SELEX experiments containing both highly converged and highly diverse libraries. In each case, the algorithm was able to predict aptamer sequences that were verified by experiments. In particular, high affinity aptamers were predicted in the early rounds when their abundance was still low. In addition, new sequences that were previously ignored were also predicted and validated. This algorithm can be applicable to other aptamer selection experiments and can be valuable when a library is still quite diverse. There is still room for further improvement. For example, currently, the algorithm cannot pool sequences in the same families. Many of the predicted sequences actually belong to the same family and thus have similar binding. Future efforts will be made to assign sequences to families based on their conserved primary sequences and secondary structures.

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.

The CPS^2 algorithm design.

1) Library processing. For every round sequenced results for the given binding target we do the following: a) Clip the non-interesting regions of every aptamer by selecting only the relevant indices, which are indicated as a parameter (from here onwards we refer to the remaining interesting regions as the aptamers themselves, which are typically the random regions). b) Count

the number of occurrences of each unique aptamer sequences before removing all duplicates to avoid computing them twice. c) Using the RNAFold tool from the ViennaRNA package, the secondary structure and free energy (dG) of each sequence were computed. d) For every different aptamer the k-mers (6-mer is a widely used length for predicting aptamer sequences)[25] present in the sequence were searched. In addition, the stems and hairpin loops presence in the sequence were searched. e) Store all this information in an object that we refer to as SequenceLibrary.

2) Clustering: For each round, our algorithm clusters the sequences in the library as follows: a) A maximum number of clusters is be specified, in our case, 10 empirically proved to be a sufficient number of clusters. b) We generate a list of all unique sequences present in the library ordered by their number of occurrences and follow these steps iteratively until we have reached the maximum number of clusters:
    i. Create a new cluster and add the most common sequence in the remaining list as the leading sequence for the cluster.
    ii. All remaining sequences in the list are aligned with respect to the leading sequence of the cluster using the Striped Smith-Waterman alignment algorithm.
    iii. All sequences with an alignment score above the specified threshold, which in our case was 0.8 (1 is for identical sequences, 0 is for sequences that have nothing in common), are added to the cluster.
    iv. All sequences in the current cluster are removed from the list and we start over from step (i) with the remaining list.
3) Motif scoring: Once we have the structural information, free energy, cluster and motifs for every sequence we can start calculating some scores. For each round of each binding target we calculate the following: a) kmer_counts: Keeps track of the number of occurrences of each k-mer in the round, that is, for each k-mer we calculate the sum of the number of times each aptamer sequence that contains the k-mer appears in the round. b) stem_counts: Keeps track of the number of occurrences of each hairpin stem sequence in the round, using a similar logic to kmer_counts. c) loop_counts: Keeps track of the number of occurrences of each hairpin loop sequence in the round, using a similar logic to kmer_counts. d) cluster_kmer_counts: Keeps track of the number of occurrences of each k-mer in each cluster in the round, that is, for each k-mer we calculate the sum of the number of times each aptamer sequence that contains the k-mer appears in each cluster of the round. e) cluster_stem_counts: Similar to cluster_kmer_counts but for hairpin stem sequences. f) cluster_loop_counts: Similar to cluster_kmer_counts but for hairpin loop sequences.

4) Cluster combination: Before computing scores, we perform another clustering step in which we try to combine clusters which are too similar. To do this we use the _cluster_counts: a) For each of cluster_kmer_counts, cluster_stem_counts and cluster_loop_counts we perform the following steps: a_i) We select the n most common motifs (in our experiments n was set to 10) from each of the clusters found in the library. a_ii) For each cluster pair, we compare the selected motifs and compute a similarity score by adding -1 for every common motif and +1 for every different motif that is not in both clusters. b) For each cluster pair, we average the similarity scores compute from cluster_kmer_counts, cluster_stem_counts and cluster_loop_counts. If the average similarity score between any 2 clusters is 0 or less the clusters are combined. c) The cluster_kmer_counts, cluster_stem_counts and cluster_loop_counts are recalculated (if needed) for the new clusters.
5) Score calculation: Once all the information for each aptamer sequence is collected, we calculate their score. The final score is the average of three scores: Popularity Score (PS), Stability Score (SS) and Structure Score (STS). a) Popularity Score (PS): It is the average ot two scores: the total cluster score (TCS) and the sequence count score (SC). a_i) Total cluster score (TCS): The percentage of the total number of sequences in the round that the corresponding cluster of the sequence accounts for. a_ii) Sequence count score (SC): The quotient of the number of occurrences of the sequence in the round over the maximum number of times a sequence occurred in the given round multiplied by 100. b) Stability Score (SS): Given by taking the maximum (in absolute value) of the dG energy and 10 and multiplying it by 10. c) Structure Score (SS): The structure score is the average of four scores: the kmer score (KS), cluster kmer score (CKS), stem score (CSMS) and loop score (CLS). c_i) Kmer score (KS): This score is calculated by first calculating the scaled frequency of each kmer. To do this, we divide all the counts in kmer_counts over the maximum number of occurrences of a single k-mer. Then we take the average of the scaled frequencies of each k-mer that appears in the sequence and multiply it by 100. c_ii) Cluster kmer score (CKS): This score is calculated by first calculating the scaled frequency of each kmer with respect to its cluster. To do this, we divide all the counts in cluster_kmer_counts over the maximum number of occurrences of a single k-mer in that cluster. Then we take the average of the scaled frequencies of each k-mer that appears in the sequence and multiply it by 100. c_iii) Stem score (CSMS): similar to kmer_score but for hairpin stems using stem_counts. c_iv) Loop score (CLS): similar to kmer_score but for hairpin loops using loop_counts.

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. Briefly, this program is designed to run on a Linux environment. The following packages were installed including pyfastaq, numpy, pandas, termcolor, matplotlib, scipy, scikit-bio and tqdm. In addition, RNAFold was installed. The subsequent detailed steps are also supplied in Supporting Information.

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.

Author contributions

AN and JL designed the research; JPT, PJJH, YD and RSN performed the research; AN, JL, JPT analyzed the data; JL and JPT wrote the article with contributions from other authors.

Conflicts of Interest: None

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, ITC for control DNA sequences, 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. *Front. Chem.* **2020**, 8, 434.
(4) Xing, H.; Wong, N. Y.; Xiang, Y.; Lu, Y., DNA Aptamer Functionalized Nanomaterials for Intracellular Analysis, Cancer Cell Imaging and Drug Delivery. *Curr. Opin. Chem. Biol.* **2012**, 16, 429-435.
(5) Huang, Z.; Qiu, L.; Zhang, T.; Tan, W., Integrating DNA Nanotechnology with Aptamers for Biological and Biomedical Applications. *Matter* **2021**, 4, 461-489.
(6) Panigaj, M.; Johnson, M. B.; Ke, W.; McMillan, J.; Goncharova, E. A.; Chandler, M.; Afonin, K. A., Aptamers as Modular Components of Therapeutic Nucleic Acid Nanotechnology. *ACS Nano* **2019**, 13, 12301-12321.
(7) Yüce, M.; Ullah, N.; Budak, H., Trends in Aptamer Selection Methods and Applications. *Analyst* **2015**, 140, 5379-5399.
(8) Lyu, C.; Khan, I. M.; Wang, Z., Capture-Selex for Aptamer Selection: A Short Review. *Talanta* **2021**, 229, 122274.
(9) Zuker, M., Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, 31, 3406-3415.
(10) Chowdhury, B.; Garai, G., A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, 109, 419-431.
(11) 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. *Nat. Commun.* **2021**, 12, 2366.
(12) Chen, Z.; Hu, L.; Zhang, B.-T.; Lu, A.; Wang, Y.; Yu, Y.; Zhang, G., Artificial Intelligence in Aptamer − Target Binding Prediction. *Int. J. Mol. Sci.* **2021**, 22, 3605.
(13) 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.
(14) Buglak, A. A.; Samokhvalov, A. V.; Zherdev, A. V.; Dzantiev, B. B., Methods and Applications of in Silico Aptamer Design and Modeling. *Int. J. Mol. Sci.* **2020**, 21, 8420.
(15) 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.
(16) 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.
(17) 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.
(18) Zong, C.; Liu, J., The Arsenic-Binding Aptamer Cannot Bind Arsenic: Critical Evaluation of Aptamer Selection and Binding. *Anal. Chem.* **2019**, 91, 10887-10893.
(19) Zhao, Y.; Yavari, K.; Liu, J., Critical Evaluation of Aptamer Binding for Biosensor Designs. *TrAC, Trends Anal. Chem.* **2022**, 146, 116480.
(20) 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.

(21) Heredia, F. L.; Roche-Lima, A.; Parés-Matos, E. I., A Novel Artificial Intelligence-Based Approach for Identification of Deoxynucleotide Aptamers. *PLOS Comput. Biol.* **2021**, 17, e1009247.
(22) 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. *Sci. Rep.* **2016**, 6, 33697.
(23) Iwano, N.; Adachi, T.; Aoki, K.; Nakamura, Y.; Hamada, M., Generative Aptamer Discovery Using Raptgen. *Nat. Comput. Sci.* **2022**, 2, 378-386.
(24) 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.
(25) 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.
(26) 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.
(27) 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 Chem. Biol.* **2017**, 12, 3103-3112.
(28) 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.
(29) 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.
(30) 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.
(31) Liu, Y.; Liu, J., Selection of DNA Aptamers for Sensing Uric Acid in Simulated Tears. *Analy. Sens.* **2022**, 2, e202200010.
(32) Liu, J.; Brown, A. K.; Meng, X.; Cropek, D. M.; Istok, J. D.; Watson, D. B.; Lu, Y., A Catalytic Beacon Sensor for Uranium with Parts-Per-Trillion Sensitivity and Millionfold Selectivity. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, 104, 2056-2061.
(33) He, J.; Wang, J.; Zhang, M.; Shi, G., Selection of a Structure-Switching Aptamer for the Specific Methotrexate Detection. *ACS Sensors* **2021**, 6, 2436-2441.
(34) Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L., Viennarna Package 2.0. *Algorithms Mol. Biol.* **2011**, 6, 26.
(35) Smith, T. F.; Waterman, M. S., Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, 147, 195-197.
(36) Huang, P.-J. J.; Liu, J., A DNA Aptamer for Theophylline with Ultrahigh Selectivity Reminiscent of the Classic RNA Aptamer. *ACS Chem. Biol.* **2022**, 17, 2121 − 2129.
(37) 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.
(38) Jenison, R. D.; Gill, S. C.; Pardi, A.; Polisky, B., High-Resolution Molecular Discrimination by RNA. *Science* **1994**, 263, 1425-1429.
(39) Huizenga, D. E.; Szostak, J. W., A DNA Aptamer That Binds Adenosine and Atp. *Biochemistry* **1995**, 34, 656-665.
(40) Slavkovic, S.; Johnson, P. E., Isothermal Titration Calorimetry Studies of Aptamer-Small Molecule Interactions: Practicalities and Pitfalls. *Aptamers* **2018**, 2, 45-51.

(41) Ding, Y.; Liu, X.; Huang, P.-J. J.; Liu, J., Homogeneous Assays for Aptamer-Based Ethanolamine Sensing: No Indication of Target Binding. *Analyst* **2022**, 147, 1348-1356.

For TOC graphics only

