RaptRanker: *in silico* RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information

Ryoga Ishida[1], Tatsuo Adachi[2], Aya Yokota[1], Hidehito Yoshihara[2], Kazuteru Aoki[2], Yoshikazu Nakamura[2], and Michiaki Hamada *[1,3]

[1]Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan
[2]RIBOMIC, inc., Tokyo, Japan
[3]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

December 29, 2019

Abstract

Aptamers are short single-stranded RNA/DNA molecules that bind to specific target molecules. Aptamers have emerged as new target-specific drugs to replace antibodies due to their diversity, low immunogenicity, and ease of synthesis. Aptamers with high binding-affinity and target specificity are identified using an *in vitro* procedure called high throughput systematic evolution of ligands by exponential enrichment (HT-SELEX). However, the development of aptamer drugs takes a considerable amount of time and is costly because HT-SELEX produces a large dataset of candidate sequences, some of which have insufficient binding-affinity. Therefore, it is essential to further identify the aptamers with high binding-affinity from the HT-SELEX data. Here, we present RNA aptamer Ranker (RaptRanker), a novel *in silico* method for identifying truly high binding-affinity aptamers from HT-SELEX data by scoring and ranking. RaptRanker analyzes HT-SELEX data by evaluating the nucleotide sequence and secondary structure simultaneously, and by ranking according to scores reflecting local structure and sequence frequencies. For the HT-SELEX datasets tested, the performance of RaptRanker was superior to those of the existing methods.

Contents

1 Introduction 2

2 Materials and methods 4
2.1 Filtering 5
2.2 Removal of duplicates and computation of sequence-structure profiles (SSPs) 5
2.3 Determination of subsequences and computation of sub-sequence-structure profiles (SSSPs) . 7
2.4 Clustering subsequences 8
2.5 Calculation of score 8
2.6 Newly acquired experimental data in this study 9

3 Results and discussion 10
3.1 Comparison of performances for the identification of truly high binding-affinity aptamer . . 10
3.2 Validation of the effectiveness of analyzing both sequence and secondary structure 11

* To whom correspondence should be addressed. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp

3.3 Prediction of binding motifs in identified sequences . . . 12
3.4 Further discussion . . . 13

4 Conclusion 16

5 Availability 16

A Notations 19

B Calculation of Minimum Spanning Forest 19
B.1 Minimum Spanning Forest (MSF) . . . 19
B.2 Kruskal's algorithm . . . 19
B.3 UnionFind . . . 19

C Filtering parameters used in the analysis 20

D Calculation of Frequency and Enrichment 21
D.1 Frequency . . . 21
D.2 Enrichment . . . 21

E Truncation of sequences 22

F Calculation of Average K-mer Enrichment (AKE) 23

G ROC curves 24

H The number of sequences in each round of each dataset 28

I Effect of parameters 29

1 Introduction

Aptamers are chemically-synthesized short single-stranded RNA/DNA molecules that can bind various targets with high binding affinity and specificity. Aptamers have mainly two characteristics that distinguish them from other nucleic acid drugs. First, aptamers recognize the tertiary structure of target molecules, while other oligonucleotide drugs target mRNA molecules through their primary structures only. Second, aptamers can target extracellular molecules as well as intracellular molecules. Unlike other nucleic acid drugs, aptamers can be generated against various types of target molecules such as transcription factors [1, 2, 3, 4], proteins and their complexes [5], small organic molecules [6, 7], viruses [8], and cells [9, 10, 11]. Moreover, aptamer production by chemical synthesis offers easy manufacturing, batch-to-batch reproducibility, and long shelf life. Considering these features, aptamers show potential as replacements of antibodies in analytical, diagnostic, and therapeutic applications.

Aptamers bind their targets by forming tertiary structures that provide spatial complementarity [12]. Therefore, it is important to consider the structural features (e.g., RNA secondary structures) of aptamers when predicting their binding properties. The information regarding the secondary structure is widely utilized in aptamer development strategies including motif definition and strategic truncation. Although a variety of software is available for predicting secondary structures of aptamers (e.g., [13, 14]), only few analyze secondary structure information for the calculation of binding potentials of RNA aptamers.

Recently, aptamers with high binding-affinities were identified using HT-SELEX method [15, 1, 16]. HT-SELEX is an *in vitro* experimental method that improves the conventional SELEX method [17, 18] by including next generation sequencing (NGS) technology so that a larger amount of nucleotide sequence information can be processed. In (HT-)SELEX procedures, the aptamers with high binding-affinities are enriched by repeating the "Amplify only sequences that bind target" procedure, starting with large amounts of random RNA/DNA sequences. First, an RNA/DNA pool with 10^{12} to 10^{15} random nucleotide

sequences called round 0 pool (0R pool) or initial pool is generated. Then, high binding-affinity aptamers are identified by performing repeated rounds of a five-step procedure as shown in figure 1. In HT-SELEX method, aptamers obtained in not only the final round but also in some of the previous rounds including the 0R pool are sequenced. Since a vast number of sequences are generated by HT-SELEX [19, 20, 21], an efficient candidate selection method is required to reduce the number of non-binding clones, decrease experiment time, and minimize material cost.

There have been some previous *in silico* studies to identify the truly high binding-affinity sequences from HT-SELEX data by further analyses. Frequency, Enrichment [22], and MPBind [23] were developed to achieve this purpose by scoring and ranking the candidate sequences. Frequency calculates the frequency of appearance for each sequence. Enrichment calculates transition of frequency between consecutive rounds for each candidate sequence. MPBind performs statistical analysis based on changes in relative frequencies of all k-mer units. However, none of them includes aptamer structure data in their analyses. In contrast, AptaTRACE [24] traces the transition of secondary structure between HT-SELEX rounds for each k-mer and predicts the binding motif independent of frequency data. However, AptaTRACE does not provide a ranking for the candidates. Therefore, a computational approach that can manage large datasets of sequences, include secondary structure analysis, and rank the candidate sequences is lacking.

Another challenge in software development is that there are only a few public HT-SELEX datasets. In aptamer-related research, only the aptamers that show high binding affinity are reported; research reports rarely disclose the complete (HT-)SELEX data generated throughout the selection process. In many cases, even when the sequence data is made available, only the data from the final round is shared and the data generated in intermediate rounds of HT-SELEX procedure is not available. It should also be noted that there is little information about the candidate sequences that turn out to lack binding activity. This data of "false-positive aptamers" is highly valuable for evaluating the performances of *in silico* HT-SELEX analysis methods.

Here we present RNA aptamer Ranker (RaptRanker), a novel *in silico* method for identifying truly high binding-affinity aptamers from HT-SELEX data by scoring and ranking. RaptRanker determines unique sequences from all HT-SELEX rounds and clusters all subsequences of unique sequences by similarity based on both nucleotide sequence and secondary structure features. Then, RaptRanker identifies truly high binding-affinity aptamers by calculating the average motif enrichment (AME), which is a score assigned to each unique sequence based on frequency of subsequence clusters. We performed two new HT-SELEX experiments and evaluated sequence sets that include false-positive aptamers using surface plasmon resonance (SPR) assay. In both HT-SELEX dataset analyses, performance of RaptRanker was superior to those of existing methods.

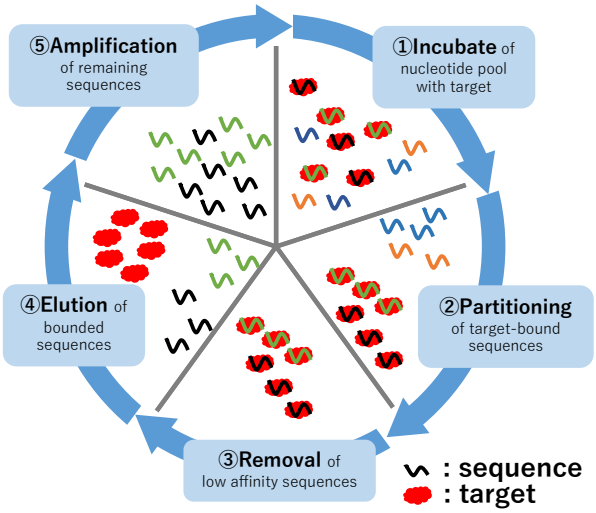


Fig. 1: Diagram of HT-SELEX method. HT-SELEX method identifies high binding-affinity aptamers by performing repeated rounds of a five-step procedure. 1. Addition of the target molecule to the RNA/DNA pool and incubation. 2. Partitioning of the target molecule-RNA/DNA complex and unbound RNA/DNA. 3. Removal of unbound RNA/DNA. 4. Elusion of target-bound RNA/DNA. 5. Amplification of selected RNA/DNA.

2 Materials and methods

RaptRanker clusters the subsequences of unique sequences based on similarity in both nucleotide sequence and secondary structure. Then, RaptRanker assigns a score for each subsequence and each unique sequence using the cluster data. Finally, RaptRanker ranks all unique sequences by score (Figure 2). Importantly, the score calculated by RaptRanker reflects particular characteristics of aptamers and SELEX experiment, such as nucleotide sequence similarity, secondary structure similarity, and frequency information.

Specifically, RaptRanker identifies truly high binding-affinity aptamers through the five steps shown in figure 2: (i) import of the sequence data from FASTA or FASTQ input file, (ii) determination of unique sequences by removing duplicates and computation of sequence-structure profiles (SSPs) representing nucleotide sequence and secondary structure features as predicted by CapR [13], (iii) derivation of subsequences sequentially from all unique sequences, and generation of sub-sequence-structure profiles (SSSPs) representing nucleotide sequence and secondary structure features for each subsequence, (iv) enumeration of all subsequence pairs matched based on similarity in sub-sequence-structure profiles using SketchSort [25] and clustering of subsequences by constructing minimum spanning forest (MSF) from all similar subsequence pairs, and (v) calculation of the scores and ranking unique sequences.

The details of each step are described in the following sections.

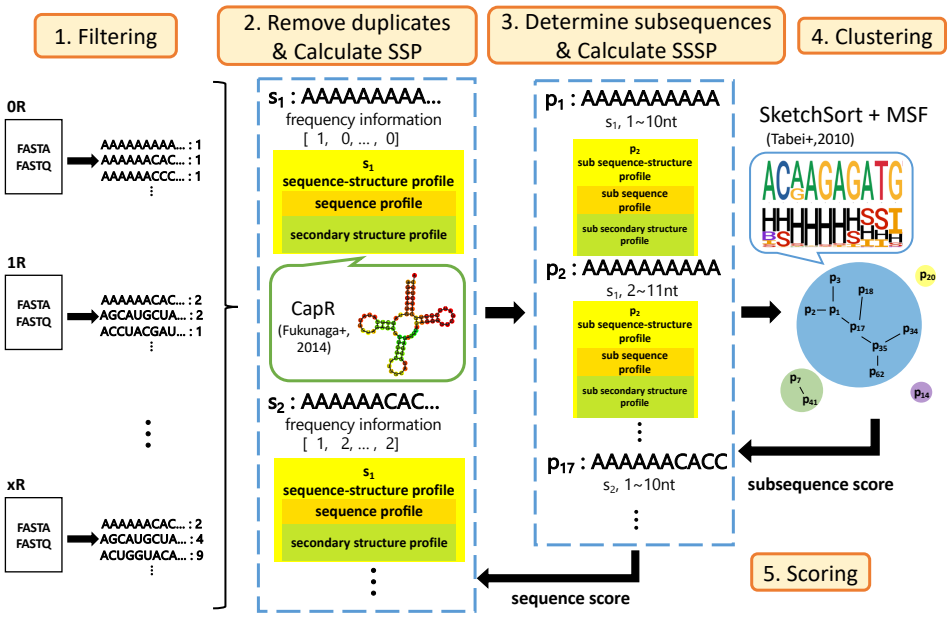


Fig. 2: An illustration of RaptRanker algorithm. First, RaptRanker determines the sequences to be analyzed from all the input FASTA/FASTQ files of HT-SELEX experiments by filtering. Second, RaptRanker extracts unique sequences by removing duplicates, and computes sequence-structure profiles (SSPs) which represents both sequence and structure information. Third, RaptRanker determines subsequences of unique sequences, and constructs subsequence-structure profile (SSSP) by decomposing SSPs. Forth, RaptRanker clusters subsequences (from all the rounds) by constructing Minimum Spanning Forest (MSF), based on similar pairs of SSSPs calculated by SketchSort. Finally, sequence scores are computed based on clustering. Each cluster obtained by RaptRanker can be regarded as a motif because it is a collection of similar subsequences of nucleotide sequence and secondary structure.

2.1 Filtering

RaptRanker determines sequences to be analyzed by two-step filtering. First, only the sequences whose primer-binding region is the same as the design are extracted. Then, the random regions whose length is within the user-defined limits are extracted. The sequences (i.e., the set of random regions) that pass the filtering criteria are subjected to further analyses.

In the first filtering, the sequences from FASTA or FASTQ files are imported and only the sequences whose primer binding regions exactly matches the design are extracted. Both ends of sequences obtained from the SELEX experiment include fixed nucleotide sequences necessary for amplification by PCR method. These sequences are called forward primer binding region (5' end) and reverse primer binding region (3' end). RaptRanker extracts only the sequences whose both forward and reverse primer binding regions exactly match the user-defined parameters *forward_primer* and *reverse_primer*.

In the second filtering, the random regions whose lengths are between the upper and lower limits. RaptRanker uses the user-defined parameters *minimum_length* and *maximum_length*, and extracts only the random regions whose lengths l are $minimum_length \leq l \leq maximum_length$.

2.2 Removal of duplicates and computation of sequence-structure profiles (SSPs)

2.2.1 Removal of duplicates

RaptRanker determines the unique sequences by removing duplicates present in the imported sequence dataset in advance to reduce the calculation cost. First, for each round, count the appearance of each

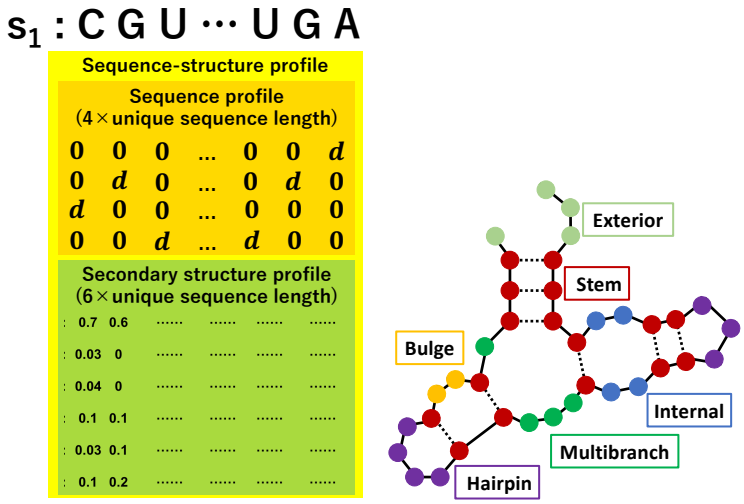
sequence and remove duplicates with radix sort [26]. After that, summarize the results of each round, and remove duplicates with radix sort again. This result is the unique sequences from all inputted HT-SELEX data, and each unique sequence has the frequency information on each round.

In the following, let S be the set of all unique sequences, and let s be a unique sequence in S.

2.2.2 Computation of sequence-structure profiles (SSPs)

For each unique sequence $s(\in S)$, RaptRanker computes sequence-structure profile (SSP) that represents nucleotide sequence and secondary structure features as predicted by CapR [13]. SSP consists of a sequence profile (4 $\times$ unique sequence length) representing the nucleotide sequence and a secondary structure profile (6 $\times$ unique sequence length) representing the secondary structure (Figure 3a).

The secondary structure is predicted for each unique sequence that includes a part of primer binding regions in addition to the random region. In this way, CapR can include the effect of the primer binding region on the secondary structure in the analysis. RaptRanker uses user-defined parameters *add_forward_primer* and *add_reverse_primer* in addition to and separate from parameters for filtering (Supplementary Tables S2 and S3).



(a) sequence-structure profile (SSP) (b) Secondary structure context

Fig. 3: (**a**) Diagram of sequence-structure profile (SSP). SSP is computed for each unique sequence, and consists of a sequence profile (4 $\times$ unique sequence length) representing the nucleotide sequence and a secondary structure profile (6 $\times$ unique sequence length) representing the secondary structure. The sequence profile is a matrix computed by converting the nucleotide sequence of each unique sequence into a one-hot vector using a real value $d \in \mathbb{R}$. The secondary structure profile is a matrix computed by converting the data obtained from CapR. (**b**) An example image of RNA secondary structure. The dashed lines represent base pairs. The secondary structure features predicted by CapR are classified into six types (Bulge, Exterior, Hairpin, Internal, Multibranch, Stem).

Sequence profile is a (4 $\times$ unique sequence length) matrix representing the nucleotide sequence information of each unique sequence $s(\in S)$. The sequence profile is computed by converting the nucleotide sequence data into a one-hot vector using a real value $d \in \mathbb{R}$ as $A = [d, 0, 0, 0]^T, G = [0, d, 0, 0]^T, C = [0, 0, d, 0]^T, U = [0, 0, 0, d]^T$ (Figure 3a). The real value d is a user-defined parameter *weight* representing the weight of the sequence information in subsequent clustering. If the value of d is large, the sequence information becomes more emphasized than the secondary structure information, and therefore more strict matching of the nucleotide sequences between entries is required to be clustered together.

Secondary structure profile is a (6 $\times$ unique sequence length) matrix representing the secondary structure information of each unique sequence $s(\in S)$. Secondary structure profile is computed by converting

the data obtained from CapR [13] into a matrix (Figure 3a). CapR is a software which calculates the probability of each base of RNA sequence being a part of a specific type of secondary structure. Secondary structure features predicted by CapR are classified into six types (Bulge, Exterior, Hairpin, Internal, Multibranch, Stem) (Figure 3b). Since the output of CapR is a probability value, the sum of six rows in each column of the secondary structure profile is 1.0.

2.3 Determination of subsequences and computation of sub-sequence-structure profiles (SSSPs)

2.3.1 Determination of subsequences

RaptRanker determines the list of subsequences. Subsequences are determined by cutting out a fragment from a unique sequence $s(\in S)$, and repeating this process, each time shifting the location of the fragment by one base. Each subsequence is indexed individually, and contains identity of the unique sequence it was obtained from and its original position on that unique sequence. Even if the nucleotide sequences are identical, they are distinguished by this information and treated separately. RaptRanker includes a user-defined parameter *wide* as window length.

In the following, let P be the set of all subsequences, and let p be a subsequence. In addition, let D_s be the set of subsequences generated from unique sequence s. In this case, $\cup_{s \in S} D_s = P$ holds. Also, since each subsequence $p(\in P)$ is treated separately, $D_s \cap D_{s'} = \emptyset$ holds for any set of two unique sequences $s, s'(\in S)$.

2.3.2 Computation of sub-sequence-structure profiles (SSSPs)

For each subsequence $p(\in P)$, RaptRanker computes sub-sequence-structure profile (SSSP) representing nucleotide sequence and secondary structure features of the subsequence by dividing SSPs into sub-profiles of a fixed length (Figure 4).

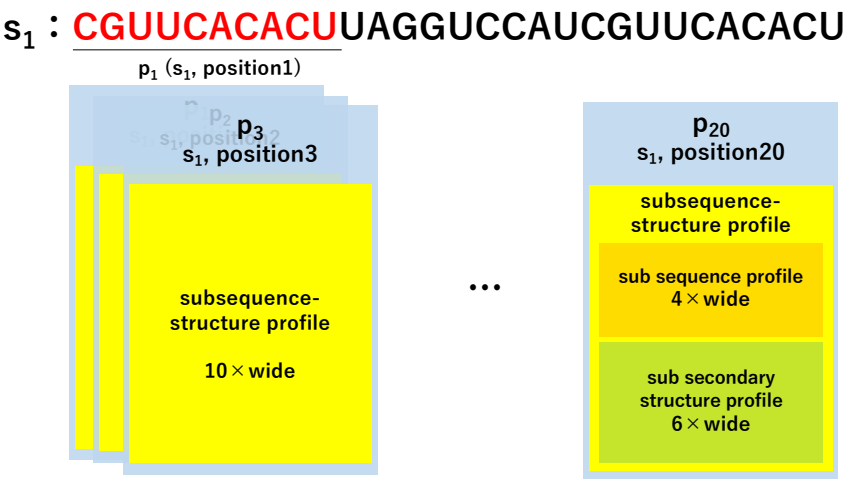


Fig. 4: Diagram of determination of sub-sequence-structure profile (SSSP) in case of $wide = 10$. Each subsequence $p(\in P)$ is indexed individually and contains identity of the unique sequence it was obtained from, its original position on that unique sequence, and SSSP. In this example, the nucleotide sequence of p_1 and p_{20} are identical ("`CGUUCACACU`"), but they are treated separately.

2.4 Clustering subsequences

2.4.1 Enumeration of all similar subsequence pairs

RaptRanker enumerates all similar subsequence pairs based on SSSP similarity using SketchSort [25]. SketchSort is a software that can enumerate similar pairs in a large number of high dimensional vectors rapidly and approximately. RaptRanker converts each SSSP into a row vector to be provided as an input to SketchSort. SketchSort requires two parameters; an upper limit of cosine distance between vectors *cosdist*, and an upper limit of the expectation value for false negatives *missing_ratio*.

2.4.2 Clustering of all subsequences

RaptRanker clusters all subsequences by constructing a MSF from all similar subsequence pairs. MSF can be calculated efficiently by Kruskal's algorithm [27] and UnionFind (Appendix B).

Since each MST from MSF becomes a cluster, clusters do not overlap each other in, i.e. a subsequence does not belong to more than one cluster.

In the following, let $\mathfrak{C}$ be the set of clusters, and let C be a cluster. In this case, since each cluster C is a set of subsequences and P is the set of all subsequences, $\cup_{C \in \mathfrak{C}} C = P$ holds. Also, since clusters do not overlap each other, $C \cap C' = \emptyset$ holds for any set of two clusters $C, C'(\in \mathfrak{C})$.

2.5 Calculation of score

RaptRanker calculates average motif enrichment (AME), a score for predicting binding affinity, for each unique sequence $s(\in S)$ based on the clustering results through three steps. (i) Calculation of subsequence frequency for each subsequence $p(\in P)$. (ii) Calculation of motif frequency and motif enrichment for each cluster $C(\in \mathfrak{C})$, and define motif enrichment of each cluster $C(\in \mathfrak{C})$ as the score of each subsequence in that cluster $p(\in C)$. (iii) Calculation of AME for each unique sequence $s(\in S)$ based on motif enrichment of subsequences $p(\in D_s)$.

2.5.1 Calculation of subsequence frequency

Subsequence frequency is calculated for each subsequence $p(\in P)$ and for each round. It is calculated by dividing the numebr of occurrences of each subsequence by total number of occurrences of subsequences. Subsequence frequency of a subsequence p at round x is

$$\text{subsequence_frequency}_x(p) = \frac{\text{count}_x(p)}{\sum_{p' \in P} \text{count}_x(p')} \tag{1}$$

where $\text{count}_x(p)$ is the occurrence of the subsequence p at round x.

2.5.2 Calculation of motif frequency and motif enrichment

Motif frequency is calculated for each cluster $C(\in \mathfrak{C})$ and for each round. Motif frequency is the sum of the subsequence frequencies of subsequences in a cluster. Motif frequency of a cluster C at round x is

$$\text{motif_frequency}_x(C) = \sum_{p \in C} \text{subsequence_frequency}_x(p). \tag{2}$$

Then, RaptRanker calculates motif enrichment for each cluster $C(\in \mathfrak{C})$ and for each round. Motif enrichment is calculated for a specific round by dividing the motif frequency in that round by the motif frequency in the previous round. Motif enrichment of a cluster C at round x is

$$\text{motif_enrichment}_x(C) = \frac{\text{motif_frequency}_x(C)}{\text{motif_frequency}_{x-1}(C)} \tag{3}$$

for $\text{motif_frequency}_{x-1}(C) \neq 0$ and round x is not the first round; the score is undefined otherwise.

RaptRanker defines motif enrichment of each cluster $C(\in \mathfrak{C})$ as the score of each subsequence in that cluster $p(\in C)$.

$$\text{motif_enrichment}_x(p) = \text{motif_enrichment}_x(C) \tag{4}$$

where $p \in C$.

2.5.3 Calculation of Average Motif Enrichment (AME)

AME is calculated for each unique sequence $s(\in S)$ and for each round. AME is the average of the motif enrichment values of subsequences from a unique sequence. The AME of a unique sequence s at round x is defined as

$$\text{AME}_x(s) = \frac{\sum_{p \in D_s} \text{motif_enrichment}_x(p)}{|D_s|}. \tag{5}$$

The score is undefined for the first round.

2.6 Newly acquired experimental data in this study

Because the number of existing datasets is not sufficient for evaluating our method, we performed two new HT-SELEX experiments, and selected evaluation sequence sets that include false-positive aptamers as determined by surface plasmon resonance (SPR) assay. The target molecules of HT-SELEX experiments were human recombinant transglutaminase 2 (TG2; Cat. No. 4376-TG) and human recombinant integrin alpha V beta 3 (αVβ3; Cat. No. 3050-AV). In the following, we refer the data from TG2 experiment as Data1, and the data from αVβ3 experiment as Data2. The sequence data are available on DRA009383 and DRA009384, and evaluation of sequence sets in each dataset are shown in supplementary information.

TG2 and αVβ3 were purchased from R&D systems. Z-Gln-Gly (Cat. No. C6154) and Hydroxylamine hydrochloride (Cat. No. 169417) were obtained from Sigma-Aldrich. All other buffers and chemicals were of special grade.

2.6.1 HT-SELEX

Selection of aptamers was followed by a previously-reported SELEX method [28] with some modifications. The sequence of the first template for Data1 and Data2 were `5'-- TCACACTAGC ACGCATAGG --30N-- CATCTGACCT CTCTCCTGCT CCC --3'` and `5'-- GAGGATCCAT GTATGCGCAC ATA --40N-- CTTCTGGTCG AAGTTCTCCC --3'`, respectively. The following primers were used. Data1-Forward, `5'-- TAATACGACTC ACTATAGGGAG CAGGAGAGAG GTCAGATG --3'`; Data1-Reverse, `5'-- TCACACTAGC ACGCATAGG --3'`; Data2-Forward, `5'-- CGGAATTCT AATACGACT CACTATAGGG AGAACTTCG ACCAGAA --3'`; Data2-Reverse, `5'-- GAGGATCCAT GTATGCGCA CATA --3'`. The oligonucleotide libraries containing 2'-fluoro-pyrimidine nucleotides were prepared by transcription using a mutant T7 RNA polymerase. The binding buffer used for selection contained 145 mM NaCl, 5.4 mM KCl, 5 mM MgCl2, 0.05% Tween20, and 20 mM Tris-HCl (pH 7.6). For Data2 SELEX, 1 mM MnCl2 was used instead of MgCl2. The recombinant proteins were immobilized on NHS-activated Sepharose (17-0906-01, GE Healthcare) and bound nucleotides were released with 6 M Urea. After selection, nucleotide pools were sequenced using an Ion PGM Hi-Q View Sequencing Kit (A30044, Thermo Fisher Scientific).

2.6.2 Surface plasmon resonance (SPR) assay

The SPR assays were performed essentially as described previously using the Biacore T200 instrument (GE Healthcare) [28]. Aptamers were prepared in 16-merpoly(A)-tailed form by in vitro transcription. A 5'-biotinylated dT16 oligomer was bound to the surface of the streptavidin sensor chip (BR100531, GE Healthcare) of active and reference flow cells. The poly(A)-tailed RNA was immobilized to the active flow cell by complementary hybridization to the dT16 oligomer. The recombinant protein in the binding buffer was injected to both flow cells of the sensor chip. Data was obtained by subtracting the reference flow cell data from the active flow cell data. Maximum response after injection was used for determination of true binding clone. Experimental parameters, including material concentrations, flow

rate, contact time, and criteria of true binding clone were optimized for each target (see Table 1). To regenerate the sensor chip, bound material was completely removed by injecting 6 M urea.

Table 1: Parameters of surface plasmon resonance (SPR) assay

Dataset	Aptamer (nM)	Recombinant protein (nM)	Flow rate (μL/min)	Contact time (s)	Criteria for True (RU)
Data1	300	300	20	60	> 25
Data2	100	100	30	60	> 25

2.6.3 Transglutaminase assay

Inhibitory activities of aptamers in Data1 was assessed by following procedures. Forty microliter of reaction mixture comprising 200 ng TG2, 50 mM Z-Gln-Gly, 100 mM hydroxylamine, 10 mM DTT, 10 mM CaCl2, and different concentrations of aptamer (0-1000 nM) in SELEX buffer was used in the assay. After incubation at 37 °C for 3 hours, 160 μL of assay reagent (370 mM FeCl3, 200 mM trichloroacetate, 670 mM HCl) was added and the solution was centrifuged for 2 min at 20,000 $\times g$. The supernatant was transferred to a microplate, and the resulting color was measured at 525 nm. Z-DON was used as a positive control. IC50 value was calculated by XLfit.

3 Results and discussion

We used data from rounds 0 to 4 on Data1, and 3 to 6 on Data2 as input for analysis. Data1 comprises data from rounds 0 to 8, but we did not include data from 5 to 8 in the analysis because we confirmed that the sequences evaluated as False were significantly amplified those rounds (see the Supplementary material, Section 3.4.3, for the details).

In the following, we used the parameters (*wide*=10, *weight*=0.5, *cosdist*=0.001, *missing_ratio*=0.00001) on RaptRanker for both data. Other data-specific parameters (related to filtering) are listed in supplementary tables S2 and S3.

3.1 Comparison of performances for the identification of truly high binding-affinity aptamer

To assess the performance of truly high binding-affinity aptamer identification, we compared "True Positive Rate (TPR) when False Positive Rate (FPR) is 0" metrics of two HT-SELEX datasets. This evaluation metrics corresponds to the rise of the ROC curve, and it considers the practical use that the binding-affinity is experimentally confirmed only for high scoring sequences (details are on section 3.4.2). Among the existing methods, Frequency and Enrichment were performed in addition to RaptRanker (Appendix D).

We compared each approach in the final round of both datasets. RaptRanker showed the highest TPR in both datasets. So, RaptRanker identified the truly binding-affinity aptamers with the highest performance among all three methods tested (Table 2). Here, the result of MPBind in Data1 was 0.0. MPBind is a method that can use control-round (SELEX round without adding target molecules) information. However, in this case, control-round information is not available in both data sets. This may be have caused the low performance of MPBind.

Table 2: Result of the performance comparison for the identification of truly high binding-affinity aptamers

Methods	Data1_4R TPR	Data2_6R TPR
RaptRanker	**0.625**	**0.378**
Enrichment	0.240	0.351
Frequency	0.080	0.216
MPBind	0.000	0.243

We compared the performance of truly high binding-affinity aptamers identifying on two HT-SELEX datasets. The evaluation metrics is "True Positive Rate (TPR) when False Positive Rate (FPR) is 0". We compared at 4R and 6R, the last round of inputs for each dataset. For each dataset, the highest values are shown in bold.

3.2 Validation of the effectiveness of analyzing both sequence and secondary structure

We validated the effectiveness of considering both nucleotide sequence and secondary structure. We compared binding-affinity prediction performances in a manner similar to that explained in the previous section. We included three types of analyses; first one is the analysis based on sequence similarity only, second one on secondary structure similarity only, and the last one on both parameters.

For analyses relying on sequence similarity only, we calculated Average K-mer Enrichment (AKE) in a similar way we calculated Average Motif Enrichment (AME) based on RaptRanker's clustering (Appendix F); AKE is calculated based on k-mer, which is a set of subsequences with matching nucleotide sequences, while AME is calculated based on cluster RaptRanker clusters. Here, we calculated AKE for 10-mer because we set $wide = 10$ in RaptRanker.

As a method considering only secondary structure similarity, we calculated AME with $weight = 0.0$. By setting the parameter $weight = 0.0$, RaptRanker clusters the subsequences based on secondary structures similarity only. The AME calculated based on this clustering can be regarded as a score that predicts binding-affinity using secondary structure similarity only.

We evaluated the performances of each approach using the data from the final rounds of both datasets. Data generated by the method of analyzing similarities in both sequence and structural features showed the highest TPR in both datasets (Figure 3). In Data2, we obtained the same TPR values from the analysis by sequence similarity only showed and from the analysis by both parameters. In contrast, in Data1, analysis by sequence similarity only resulted in lower TPR values compares to those for analysis by secondary structure similarity only. These results indicate that analysis of secondary structure similarity is effective in HT-SELEX data processing, and that its importance varies among datasets. These results also suggest that RaptRanker is particularly useful in cases where the existing methods based on sequence similarity only cannot provide sufficient identification.

Table 3: Result of validation of the effectiveness

Methods	Data1_4R TPR	Data2_6R TPR
Both	**0.625**	**0.378**
sequence only	0.042	**0.378**
structure only	0.417	0.270

We compared binding-affinity prediction performances in three types of analyses; first one is the analysis based on sequence similarity only, second one on secondary structure similarity only, and the last one on both parameters. The evaluation metric was is "True Positive Rate (TPR) when False Positive Rate (FPR) is 0". We used data from rounds 4 and 6, the last round data from Data1 and Data2, respectively. For each dataset, the highest values are shown in bold.

3.3 Prediction of binding motifs in identified sequences

We acquired the truncated sequences from both datasets, and compared the truncated sequences with the binding motifs predicted by RaptRanker.

3.3.1 Truncation of sequences

For both Data1 and Data2, we generated several truncated forms of candidate aptamers according to Transglutaminase assay and SPR assay (Supplementary Figures S1 and S2). Minimal active sequences were defined as 25 nt and 36 nt (table 4). In addition, for Data2, we synthesized a series of single nucleotide deletion mutants of aptamers to predict binding motifs (Supplementary Figure S3). As a result, we estimated that "`UACGU---CUG`" is a binding motif .

3.3.2 Binding motif prediction with RaptRanker.

The highest scoring subsequence from the sequence identified by RaptRanker can be regarded as the binding motif. Since RaptRanker calculates $\text{motif_enrichment}_x(p)$ for each subsequence $p(\in D_s)$ from each unique sequence $s(\in S)$, the cluster C to which the highest scoring subsequence belongs p (the cluster C holds $C \ni p$ where the p holds $\max_{p \in D_s} \text{motif_enrichment}_x(p)$) can be regarded as the binding motif of the unique sequence s. Here, we selected the unique sequences with true binding-affinity (Table 5).

3.3.3 Comparison of motifs predicted by RaptRanker with experimentally-determined truncated sequences

We compared the truncated sequences and the binding motifs predicted by RaptRanker. In both datasets, the binding motifs predicted by RaptRanker were found to be parts of the truncated sequences (Figure 5ab). This result suggests that RaptRanker can predict the essential subsequence in each identified sequence. This information can be used for the optimization of identified sequences.

In addition, for Data2, we compared experimentally-determined motifs to the motifs predicted by RaptRanker, and found them to be similar (Figure 5b). However, short and discontinuous motifs such as those shown in figure 5b may affect the binding-affinity prediction performance of RaptRanker negatively because such discontinuous motifs are recognized as two different motifs in RaptRanker. Moreover, short motifs are harder to detect when the actual motif length is shorter than the value of *wide* in RaptRanker. Considering these shortcomings, it is conceivable that the binding affinity prediction performance of RaptRanker was sluggish in Data2 due to presence of discontinuous and short motifs.

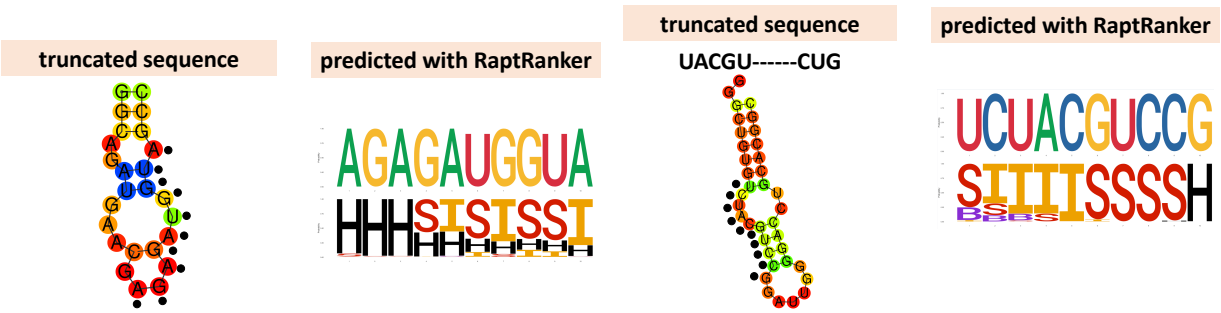
Table 4: Truncated sequences

Dataset	truncated sequence
Data1	`GGCAGAUGAACGAGAGAUGGUAGCC`
Data2	`GGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGG`

Table 5: The unique sequences for binding motif prediction

Dataset	sequence (random region only)
Data1	`AACGAGAGAUGGUAGGCCUUUGAUGAUGCU`
Data2	`GCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUG`

We selected the unique sequences exhibiting true binding-affinity and containing the binding motifs predicted by *in vitro* experiments.



(a) Data1 (b) Data2

Fig. 5: We compared the truncated sequences and the binding motifs predicted by RaptRanker. Black dots in truncated sequences show the nucleotide sequences that are identical to binding motifs predicted with RaptRanker. H, S, I, and B indicate hairpin, stem, internal loop, and bulge loop, respectively (cf. Figure 3b).

3.4 Further discussion

3.4.1 Significance of clustering all subsequences simultaneously

RaptRanker enables the computation of motif enrichment by clustering all subsequences obtained in all rounds simultaneously. That is, RaptRanker takes advantage of data generated in multiple rounds of HT-SELEX by clustering all subsequences at once.

In the case of clustering all subsequences at once, clustering is not affected differences among rounds even if "Subsequences from sequences not observed in some rounds" are present. RaptRanker determines all unique sequences S from all rounds, and clusters all their subsequences P. In this case, those subsequences only affect the calculation of motif frequency and motif enrichment (Figure 6a).

In contrast, in the case of clustering subsequences on each round, it is necessary to follow individual clusters through rounds. Cluster fragmentation or integration may occur between rounds in some cases. Such events make cluster mapping and computation of motif enrichment difficult (Figure 6b).

Since RaptRanker clusters all subsequences from all rounds simultaneously, it can track the transition of motif frequency and compute motif enrichment. As an example, we visualized the transition of motif frequency for the motifs shown in figure 5ab (Figure 7). The motif on Data1 is not observed in rounds 0 to 2, but it was highly enriched between rounds 3 and 4 (Figure 7a). In contrast, the motif in Data2 was enriched smoothly between rounds 3 and 6 (Figure 7b).

3.4.2 Evaluation metrics of binding affinity prediction performance

We used TPR when FPR is 0 as the metric for evaluation of binding-affinity prediction performance. This metric is based on the approach of ROC enrichment [29]. ROC enrichment is used for evaluation of protein-ligand virtual screening and calculated by dividing TPR by FPR. According to Jain and Nicholls, ROC enrichment should be reported in four points of FPR={ 0.5%, 1.0%, 2.0%, 5.0% } [30]. However, since there are few false positives in both datasets, the FPR is 7.1% (Data1) and 8.3% (Data2) when False = 1. So, we used the TPR when FPR is 0 as an alternative metric.

3.4.3 Analysis with ROC curves

The analysis using ROC curve is another way for evaluating binding affinity prediction performance. For the data analysis in section 3.1, the ROC curves for data from rounds 0 to 4 in Data1 are shown in figure S4 and the ROC curves for Data2 are shown in figure S5, respectively. ROC curves corresponding to analysis in section 3.2 are shown in figure S6. In addition, the ROC curves for data for rounds 0 to 8 in Data1 are shown in Figure S7.

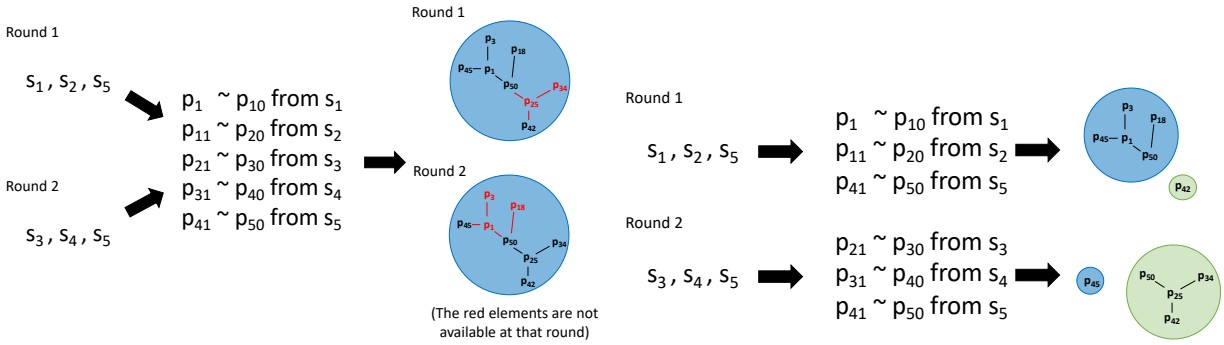
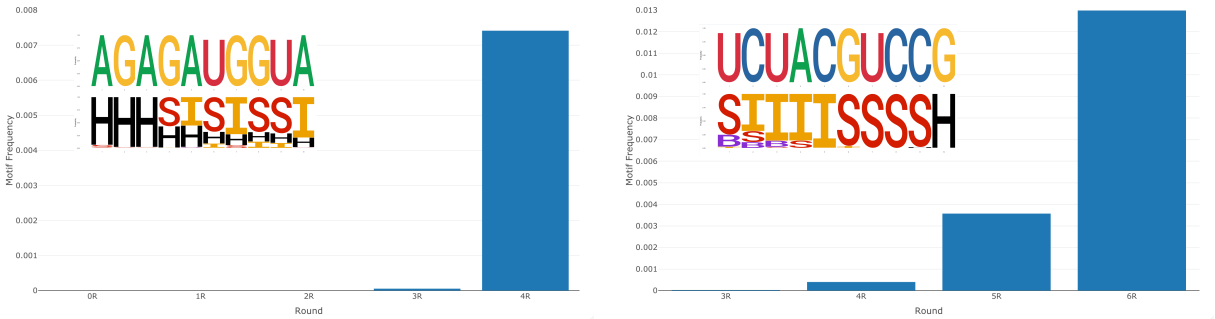


Fig. 6: **(a)** In the case of clustering subsequences from all rounds simultaneously, the clustering is not affected by differences among rounds even if "Subsequences from sequences not observed in some rounds" are present. **(b)** In the case of clustering subsequences on each round, it is necessary to follow individual clusters across rounds. In this case, cluster fragmentation or integration may occur between rounds.



(a) Data1 (b) Data2

Fig. 7: Transition of motif frequency across rounds. In each subfigure, horizontal axis and vertical axis show round number and motif frequencies, respectively. The motifs are predicted by RaptRanker. RaptRanker can track the transition of motif frequency by clustering subsequences from all rounds simultaneously.

Here, ROC curves for data from rounds 0 to 3 in Data1 and round 3 in Data2 are not shown because they were not informative. In these rounds, AME and Enrichment scores were undefined because these scores calculate relative increase of frequency compared to the previous round (Equation 5 and Equation S2) and in these instances either no previous round was present or the sequence was not observed in the previous rounds (Table S4, Table S5). In other words, unless a sequence is observed in successive rounds, AME and Enrichment scores are undefined and the ROC curve is almost a diagonal line. For the same reason, AME and Enrichment values for round 4 data in both Data1 and Data 2 were undefined for some sequences (Figure S4a, Figure S5a).

Besides, since the ROC curves obtained for analysis with RaptRanker and all the existing methods were below the diagonal line for data from rounds 5 to 8 in Data1, we only used data from rounds 0 to 4 for analysis. We also confirmed significant amplification of the unique sequences whose binding-affinity is False data from rounds 5 to 8 in Data1. Frequency, Enrichment, MPBind, and RaptRanker take frequency information into account, so if the False sequences were amplified significantly in the experiment, they are ranked higher. The amplification of the False sequences in Data1, rounds 5–8 might be caused by some experimental errors.

3.4.4 Compare sequence score definitions

As AME is defined as the average motif enrichment of the subsequences from each unique sequence, we can consider best motif enrichment (BME) as the maximum value of motif enrichment. BME of a unique sequence s at round x is

$$\mathrm{BME}_x(s) = \max_{p \in D_s} \mathrm{motif_enrichment}_x(p). \tag{6}$$

AME assesses the conservation of features over a wider range, whereas BME assesses the conservation of features locally. Since AME is the average, "unique sequences whose nucleotide sequence and secondary structure are conserved as a whole" receive higher scores. In contrast, since BME takes the maximum value, "unique sequences whose nucleotide sequence and secondary are conserved in a specific part" receive higher scores.

We compared AME and BME for the data from the final rounds in both datasets. AME showed higher TPR than BME in both datasets (Table 6). The ROC curve is shown in figure S8. These results confirm that the binding-affinity prediction performance of AME is higher than that of BME. AME is also convenient in identifying high-binding affinity sequences because multiple unique sequences that share the same motif score the same in BME.

3.4.5 Effect of parameters

We examined the effects changing the values of *wide*, *weight* and *cosdist* on the performance. We used the parameters (*wide*=10, *weight*=0.5, *cosdist*=0.001, *missing_ratio*=0.00001) on RaptRanker for both data. and observed that *wide*, *weight*, and *cosdist* greatly affected the clustering result. Since the clustering of RaptRanker uses SketchSort, which is an approximation algorithm, we measured the TPR when FPR is 0 for each combination of parameters three times and calculated the average.

The parameters used in this study showed high TPR on average for both datasets, and no significant change in the results was observed when parameters were modified in a range (Supplementary Tables S6 and S7).

3.4.6 Future works

RaptRanker calculates scores for each round and each unique sequence. As a result, RaptRanker, just like the existing methods, cannot detect *bad* rounds (such as rounds 5–8 in Data1). Since RaptRanker uses data from all rounds as the input, we are planning to realize identification of truly high binding-affinity aptamers in the presence of *bad* rounds.

In the clustering method of RaptRanker, features of subsequences are transformed into sub-sequence-structure profiles (SSSPs), which are clustered using SketchSort and MSF. Since SSSPs need only a matrix of real values, various extensions are possible. For example, RaptRanker can identify DNA aptamers by using DNA secondary structure prediction software instead of CapR. Alternatively, RaptRanker can add new features like G-quadruplex [31] by adding new rows to SSSPs. Generally, this clustering technique can be applied to identification of nucleic acid motifs not only in aptamers but also in non-coding RNAs and/or messenger RNAs. In particular, compared to general k-mer analysis, it is more effective in searching long motifs.

Table 6: Result of sequence score definitions comparisons

Methods	Data1_4R TPR	Data2_6R TPR
Average Motif Enrichment (AME)	**0.625**	**0.378**
Best Motif Enrichment (BME)	0.583	0.243

We compared AME and BME in for the data from the final rounds in datasets. AME showed higher TPR than BME in both datasets.

4 Conclusion

We developed RaptRanker, a novel *in silico* method for identifying truly high binding-affinity aptamers from HT-SELEX data based on local sequence and structural information. RaptRanker determines unique sequences by analyzing data obtained in *all* HT-SELEX rounds, and clusters all subsequences of unique sequences based on similarity in both nucleotide sequence and secondary structure features. Then, RaptRanker identifies truly high binding-affinity aptamers by calculating AME, which is a score for each unique sequence based on inclusion of clusters.

To evaluate the performance of RaptRanker in identifying truly high binding-affinity aptamers, we performed two new HT-SELEX experiments, and evaluated sequence sets that include false-positive aptamers as detected by surface plasmon resonance (SPR) assay. The sequence data which includes the intermediate rounds and evaluation sequence sets are freely available, which will be useful in future aptamer researches.

In both HT-SELEX datasets, RaptRanker showed the best performance in identifying truly high binding-affinity sequences among all methods tested: Frequency, Enrichment, and MPBind. Moreover, we confirmed the effectiveness of including both nucleotide sequence and secondary structure data in HT-SELEX analysis by examining the identification accuracy of different approaches based on similarity in (i) nucleotide sequence only, (i) secondary structure only, or (iii) both parameters. In addition, we confirmed that RaptRanker correctly predicts the essential subsequence in each identified sequence. RaptRanker is particularly useful when the existing methods analyzing sequence similarity only cannot provide sufficient identification.

5 Availability

An implementation of the proposed method, RaptRanker, is available on GitHub `https://github.com/hmdlab/RaptRanker`, and newly acquired HT-SELEX data is available on DRA009383 and DRA009384.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR1881, Japan. AY is currently working at a private company in Japan. Computation for this study was partially performed on the NIG supercomputer at ROIS National Institute of Genetics. RI, AY and MH thank Tsukasa Fukunaga and members of Hamada Laboratory for valuable comments. We are also grateful to Satoko Yamazaki, Hisako Ikeda and Hisayo Yasumoto for their early contributions in obtaining part of experimental data. We would like to thank Editage (www.editage.com) for English language editing.

Conflict of interest statement

None declared.

References

[1] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. Sillanpää, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R. Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, June 2010.

[2] Giomar Rivera-Cancel, Laura B. Motta-Mena, and Kevin H. Gardner. Identification of natural and artificial DNA substrates for the light-activated LOV-HTH transcription factor EL222. *Biochemistry*, 51(50):10024–10034, December 2012.

[3] Tatsuaki Kurata, Akira Katayama, Masakazu Hiramatsu, Yuya Kiguchi, Masamitsu Takeuchi, Tomoyuki Watanabe, Hiroshi Ogasawara, Akira Ishihama, and Kaneyoshi Yamamoto. Identification of the Set of Genes, Including Nonannotated morA, under the Direct Control of ModE in Escherichia coli. *Journal of Bacteriology*, 195(19):4496–4505, October 2013.

[4] Akira Ishihama, Ayako Kori, Etsuko Koshio, Kayoko Yamada, Hiroto Maeda, Tomohiro Shimada, Hideki Makinoshima, Akira Iwata, and Nobuyuki Fujita. Intracellular Concentrations of 65 Species of Transcription Factors with Known Regulatory Functions in Escherichia coli. *Journal of Bacteriology*, 196(15):2718–2727, August 2014.

[5] Alexey Berezhnoy, C. Andrew Stewart, James O. Mcnamara II, William Thiel, Paloma Giangrande, Giorgio Trinchieri, and Eli Gilboa. Isolation and Optimization of Murine IL-10 Receptor Blocking Oligonucleotide Aptamers Using High-throughput Sequencing. *Molecular Therapy*, 20(6):1242–1250, June 2012.

[6] Brian R. Baker, Rebecca Y. Lai, McCall S. Wood, Elaine H. Doctor, Alan J. Heeger, and Kevin W. Plaxco. An Electronic, Aptamer-Based Small-Molecule Sensor for the Rapid, Label-Free Detection of Cocaine in Adulterated Samples and Biological Fluids. *Journal of the American Chemical Society*, 128(10):3138–3139, March 2006.

[7] Xiaolei Zuo, Yi Xiao, and Kevin W. Plaxco. High Specificity, Electrochemical Sandwich Assays Based on Single Aptamer Sequences and Suitable for the Direct Detection of Small-Molecule Targets in Blood and Other Complex Matrices. *Journal of the American Chemical Society*, 131(20):6944–6945, May 2009.

[8] Jennifer M. Binning, Tianjiao Wang, Priya Luthra, Reed S. Shabman, Dominika M. Borek, Gai Liu, Wei Xu, Daisy W. Leung, Christopher F. Basler, and Gaya K. Amarasinghe. Development of RNA Aptamers Targeting Ebola Virus VP35. *Biochemistry*, 52(47):8406–8419, November 2013.

[9] Kevin N. Morris, Kirk B. Jensen, Carol M. Julin, Michael Weil, and Larry Gold. High affinity ligands from in vitro selection: Complex targets. *Proceedings of the National Academy of Sciences*, 95(6):2902–2907, March 1998.

[10] Dion A. Daniels, Hang Chen, Brian J. Hicke, Kristine M. Swiderek, and Larry Gold. A tenascin-C aptamer identified by tumor cell SELEX: Systematic evolution of ligands by exponential enrichment. *Proceedings of the National Academy of Sciences*, 100(26):15416–15421, December 2003.

[11] Hui Shi, Wensi Cui, Xiaoxiao He, Qiuping Guo, Kemin Wang, Xiaosheng Ye, and Jinlu Tang. Whole Cell-SELEX Aptamers for Highly Specific Fluorescence Molecular Imaging of Carcinomas In Vivo. *PLOS ONE*, 8(8):e70476, August 2013.

[12] Yusuke Nomura, Shigeru Sugiyama, Taiichi Sakamoto, Shin Miyakawa, Hiroaki Adachi, Kazufumi Takano, Satoshi Murakami, Tsuyoshi Inoue, Yusuke Mori, Yoshikazu Nakamura, and Hiroyoshi Matsumura. Conformational plasticity of RNA for target recognition as revealed by the 2.15 Å crystal structure of a human IgG–aptamer complex. *Nucleic Acids Research*, 38(21):7822–7829, November 2010.

[13] Tsukasa Fukunaga, Haruka Ozaki, Goro Terai, Kiyoshi Asai, Wataru Iwasaki, and Hisanori Kiryu. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biology*, 15(1):R16, January 2014.

[14] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, Feb 2009.

[15] Yue Zhao, David Granas, and Gary D. Stormo. Inferring Binding Energies from Selected Binding Sites. *PLOS Computational Biology*, 5(12):e1000590, December 2009.

[16] Gillian V. Kupakuwana, James E. Crill Ii, Mark P. McPike, and Philip N. Borer. Acyclic Identification of Aptamers for Human alpha-Thrombin Using Over-Represented Libraries and Deep Sequencing. *PLOS ONE*, 6(5):e19395, May 2011.

[17] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510, August 1990.

[18] Andrew D. Ellington and Jack W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–822, August 1990.

[19] William H. Thiel, Thomas Bair, Kristina Wyatt Thiel, Justin P. Dassie, William M. Rockey, Craig A. Howell, Xiuying Y. Liu, Adam J. Dupuy, Lingyan Huang, Richard Owczarzy, Mark A. Behlke,

James O. McNamara, and Paloma H. Giangrande. Nucleotide Bias Observed with a Short SELEX RNA Aptamer Library. *Nucleic Acid Therapeutics*, 21(4):253–263, August 2011.

[20] Mayumi Takahashi, Xiwei Wu, Michelle Ho, Pritsana Chomchan, John J. Rossi, John C. Burnett, and Jiehua Zhou. High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency. *Scientific Reports*, 6, September 2016.

[21] Tao Wang, Changying Chen, Leon M. Larcher, Roberto A. Barrero, and Rakesh N. Veedu. Three decades of nucleic acid aptamer technologies: Lessons learned, progress and opportunities on aptamer development. *Biotechnology Advances*, 37(1):28–50, January 2019.

[22] Jan Hoinka, Alexey Berezhnoy, Phuong Dao, Zuben E. Sauna, Eli Gilboa, and Teresa M. Przytycka. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Research*, 43(12):5699–5707, July 2015.

[23] Peng Jiang, Susanne Meyer, Zhonggang Hou, Nicholas E. Propson, H. Tom Soh, James A. Thomson, and Ron Stewart. MPBind: a Meta-motif-based statistical framework and pipeline to Predict Binding potential of SELEX-derived aptamers. *Bioinformatics*, 30(18):2665–2667, September 2014.

[24] Phuong Dao, Jan Hoinka, Mayumi Takahashi, Jiehua Zhou, Michelle Ho, Yijie Wang, Fabrizio Costa, John J. Rossi, Rolf Backofen, John Burnett, and Teresa M. Przytycka. AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell Systems*, 3(1):62–70, July 2016.

[25] Yasuo Tabei, Takeaki Uno, Masashi Sugiyama, and Koji Tsuda. Single versus Multiple Sorting in All Pairs Similarity Search. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 145–160, October 2010.

[26] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.

[27] Joseph B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

[28] Shin Miyakawa, Yusuke Nomura, Taiichi Sakamoto, Yoshiki Yamaguchi, Koichi Kato, Satoko Yamazaki, and Yoshikazu Nakamura. Structural and molecular basis for hyperspecificity of RNA aptamer to human immunoglobulin G. *RNA*, 14(6):1154–1163, June 2008.

[29] Anthony Nicholls. What do we know and when do we know it? *Journal of Computer-Aided Molecular Design*, 22(3-4):239–255, March 2008.

[30] Ajay N. Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design*, 22(3-4):133–139, March 2008.

[31] Carolina Roxo, Weronika Kotkowiak, and Anna Pasternak. G-Quadruplex-Forming Aptamers—Characteristics, Applications, and Perspectives. *Molecules*, 24(20):3781, January 2019.

Supplementary Information of "RaptRanker: *in silico* RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information"

A Notations

Table S1 summarize notations frequently utilized in the main manuscript.

Table S1: Notations utilized in the main text

notation	description
S	Set of all unique sequences (in all rounds)
s	Each unique sequence, $s \in S$
P	Set of all subsequences
p	Each subsequence, $p \in P$
D_s	Set of subsequences from the unique sequence s
$\mathfrak{C}$	Set of all clusters
C	Each cluster which is a set of subsequence, $C \in \mathfrak{C}$

B Calculation of Minimum Spanning Forest

B.1 Minimum Spanning Forest (MSF)

Minimum Spanning Forest (MSF) is a weighted unconnected graph in which each connected component becomes a Minimum Spanning Tree (MST). The weighted unconnected graph is one type of graph in which the edges have weights, and not all vertices are necessarily connected. In contrast, a graph in which all vertices must be connected with weights is called weighted connected graph. MST is one type of weighted connected graph in which the sum of the weights of the edges is minimum, and which does not have any closed path.

B.2 Kruskal's algorithm

Kruskal's algorithm [27] is an algorithm for calculating MSF. Kruskal's algorithm calculates MSF by using only the edges which do not form a closed path in ascending order of weight. That is, use only the edges whose two vertices belong to different trees (Algorithm S1).

Generally, the time complexity of Kruskal's algorithm is $O(|E| \log |E|)$ for the set of edges E. However, the time complexity can be reduced to $O(|E|a(|V|))$ by sorting E in ascending order of its weight beforehand, and using a data structure called UnionFind. $a(n)$ is the inverse function of Ackermann function, which is a very slow increasing function.

B.3 UnionFind

UnionFind is one type of disjoint-set data structure. Disjoint-set data structure holds a set of data divided into tree structures in which elements do not overlap each other. UnionFind can perform the following two operations rapidly.

- Union: Unite two trees into one tree.
- Find: Find the tree to which the element belongs.

Algorithm S1 Kruskal's algorithm

Require: A set of edges $E = \{e\}$
Ensure: Minimum Spanning Forest F of graph $G(E, V)$ ($V = \{v\}$ is a set of vertices)
  $F = \emptyset$
  while $E \neq \emptyset$ **do**
    Let edge $e = \{u, v\}$ whose weight is minimum in E
    Remove e from E
    if Vertices u and v belong to different trees in F **then**
      Add e to F
    end if
  end while

The amortized execution time of each operation can be reduced to $a(n)$ by the following contrivances. $a(n)$ is the inverse function of Ackermann function, which is a very slow increasing function.

- Union: Unite the trees where the larger tree as root.
- Find: Replace all references to the parent node which is traced recursively to the references to the root node.

C Filtering parameters used in the analysis

Table S2: Filtering parameters for Data1

forward_primer	TAATACGACTCACTATAGGGAGCAGGAGAGAGGTCAGATG
reverse_primer	CCTATGCGTGCTAGTGTGA
minimum_length	25nt
maximum_length	35nt
add_forward_primer	GGGAGCAGGAGAGAGGTCAGATG
add_reverse_primer	CCTATGCGTGCTAGTGTGA

The parameters used in the analysis of Data1. The function of each parameter is as follows. *forward_primer*, *reverse_primer* : RaptRanker extracts only the sequences whose both primer binding regions are exactly match these parameters. *minimum_length*, *maximum_length* : RaptRanker extracts only the random regions whose length l is *minimum_length* $\leq l \leq$ *maximum_length*. *add_forward_primer*, *add_reverse_primer* : The secondary structure is predicted to each unique sequence which include these sequences to consider the effect of the primer binding region on the secondary structure.

Table S3: Filtering parameters for Data2

forward_primer	TAATACGACTCACTATAGGGAGAACTTCGACCAGAAG
reverse_primer	TATGTGCGCATACATGGATCCTC
minimum_length	35nt
maximum_length	45nt
add_forward_primer	GGGAGAACTTCGACCAGAAG
add_reverse_primer	TATGTGCGCATACATGGATCCTC

The parameters used in the analysis of Data2. The function of each parameter is as follows. *forward_primer*, *reverse_primer* : RaptRanker extracts only the sequences whose both primer binding regions are exactly match these parameters. *minimum_length*, *maximum_length* : RaptRanker extracts only the random regions whose length l is *minimum_length* $\leq l \leq$ *maximum_length*. *add_forward_primer*, *add_reverse_primer* : The secondary structure is predicted to each unique sequence which include these sequences to consider the effect of the primer binding region on the secondary structure.

D Calculation of Frequency and Enrichment

D.1 Frequency

Frequency value is calculated for each unique sequence $s(\in S)$ and for each round by dividing the number of occurrences of a unique sequence s in a particular round by the number of occurrences of all unique sequences in that round. Frequency of a unique sequence s at round x is

$$\text{Frequency}_x(s) = \frac{\text{count}_x(s)}{\sum_{s' \in S} \text{count}_x(s')} \tag{S1}$$

where $\text{count}_x(s)$ is the occurrence of the sequence s at round x.

D.2 Enrichment

RaptRanker calculates Enrichment value for each unique sequence $s(\in S)$ and for each round by dividing the Frequency value in a round by the Frequency value in the previous round. Enrichment of a unique sequence s at round x is

$$\text{Enrichment}_x(s) = \frac{\text{Frequency}_x(s)}{\text{Frequency}_{x-1}(s)} \tag{S2}$$

for $\text{Frequency}_{x-1}(s) \neq 0$ and round x is not the first round of inputs; The score is undefined for the other cases.

E Truncation of sequences

sequence	IC_{50} (μM)
GGGAGCAGGAGAGAGGUCAGAUGAACGAGAGAUGGUAGGCCUUUGAUGAUGCUCCUAUGCGUGCUAGUGUGA	0.140±0.016
GGGAGCAGGAGAGAGGUCAGAUGAACGAGAGAUGGUAGGCCUUUGAUGAUGCUCCC	0.267±0.029
GGGAGGAGAGAGGUCAGAUGAACGAGAGAUGGUAGGCCUUUGAUGAUCCC	0.243±0.055
GGGAGAGAGGUCAGAUGAACGAGAGAUGGUAGGCCUUUGAUCCC	0.350±0.098
GGAGAGGUCAGAUGAACGAGAGAUGGUAGGCCUUUCC	0.167±0.024
GGAGGUCAGAUGAACGAGAGAUGGUAGGCCUCC	0.185±0.041
GGGGUCAGAUGAACGAGAGAUGGUAGGCCCC	0.284±0.026
GGGUCAGAUGAACGAGAGAUGGUAGGCCC	0.231±0.029
GGUCAGAUGAACGAGAGAUGGUAGGCC	0.192±0.029
GGCAGAUGAACGAGAGAUGGUAGCC	0.427±0.173

Fig. S1: We truncated a truly high binding aptamer sequence of Data1 based on Transglutaminase assay. The sequence in top row is the original sequence, and the sequence in bottom row is the minimal active sequence.

sequence	Relative binding (%)
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	100
GGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	20
GGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	10
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUAC	15
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAU	13
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGC	31
GGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGC	11

Fig. S2: We truncated a truly high binding aptamer sequence of Data2 based on SPR assay. The sequence in top row is the original sequence, and the sequence in bottom row is the minimal active sequence.

sequence	Relative binding (%)
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	100
GGGAGAACUUCGACCAGAAGGCUGUGU UACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	127
GGGAGAACUUCGACCAGAAGGCUGUGUC ACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	1
GGGAGAACUUCGACCAGAAGGCUGUGUCU CGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUA GUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUAC UCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUACG CCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGU CGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	15
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCC GAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	169
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGG UUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	125
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGA UGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	113
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUU GGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	162
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGG CCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	38
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGAC UGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACC GCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	0
GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCU CACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC	3
(putative motif) **UACGU** **CUG**	N/A

Fig. S3: We predict a binding motif of Data2 based on SPR assay for a series of mutant aptamers with single nucleotide deletion. The sequence in top row is the original sequence, and the putative motif is shown in bottom row.

F Calculation of Average K-mer Enrichment (AKE)

As a method analyzing only sequence similarity, we calculated AKE in a similar way we calculated AME based on RaptRanker's clustering. Let $\mathfrak{K}$ be the set of all k-mers, and let K be a k-mer. In this case, since P is a set of all subsequences, $\cup_{K\in\mathfrak{K}}K = P$ holds. Also, since k-mers do not overlap each other, $K \cap K' = \emptyset$ holds for any pair of k-mers $K, K'(\in \mathfrak{K})$. AKE is calculated in the same way as AME by replacing the cluster C in equation 2, equation 3, equation 4, and equation 5 with k-mer K.

RaptRanker calculates k-mer frequency for each k-mer $K(\in \mathfrak{K})$and for each round by summing up the subsequence frequencies of subsequences that belong to that k-mer. K-mer frequency of a k-mer K at round x is

$$\text{kmer_frequency}_x(K) = \sum_{p\in K} \text{subsequence_frequency}_x(p). \tag{S3}$$

Then, RaptRanker calculates k-mer enrichment for each k-mer $K(\in \mathfrak{K})$ and for each round by dividing the k-mer frequency in a round by the k-mer frequency in the previous round. k-mer enrichment of a k-mer K at round x is

$$\text{kmer_enrichment}_x(K) = \frac{\text{kmer_frequency}_x(K)}{\text{kmer_frequency}_{x-1}(K)} \tag{S4}$$

for $\text{kmer_frequency}_{x-1}(K) \neq 0$ and round x is not the first round; otherwise the score is undefined.

RaptRanker defines k-mer enrichment of each k-mer $K(\in \mathfrak{K})$ as the score of each subsequence belonging to that k-mer $p(\in K)$, such that,

$$\text{kmer_enrichment}_x(p) = \text{kmer_enrichment}_x(K) \tag{S5}$$

where $p \in K$.

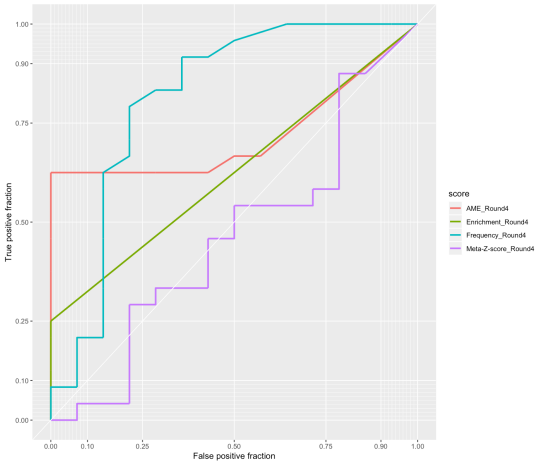
RaptRanker calculates AKE for each unique sequence $s(\in S)$ and for each round by taking the average of the k-mer enrichment values of subsequences from a unique sequence. The AKE of a unique sequence

s at round x is defined as

$$\mathrm{AKE}_x(s) = \frac{\sum_{p \in D_s} \mathrm{kmer_enrichment}_x(p)}{|D_s|}. \tag{S6}$$

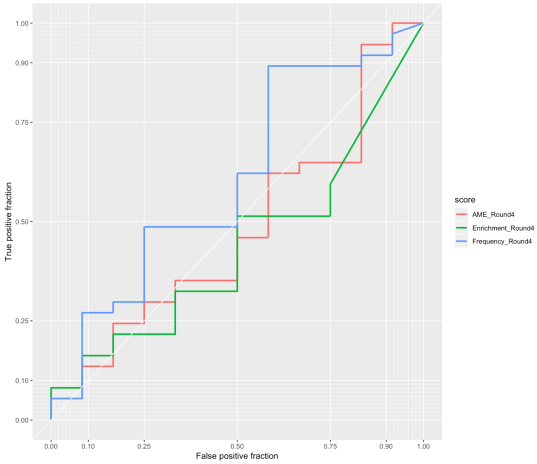
If round x is the first round, the score is undefined.

G ROC curves

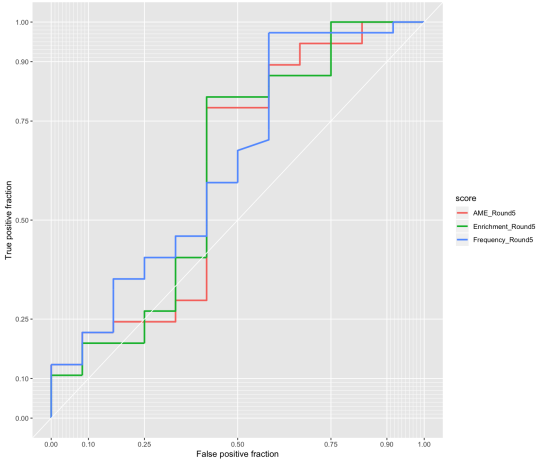


(a) 4R

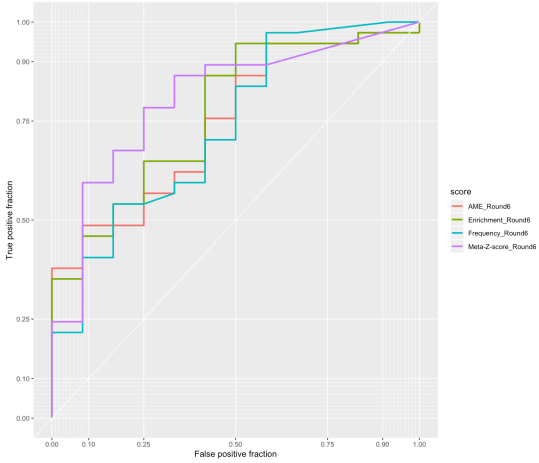
Fig. S4: The ROC curves for Data1 rounds 0–4R. The ROC curves for rounds 0–3R are not listed because they are not informative because for those rounds, AME and Enrichment values of almost all sequences were undefined due to insufficient enrichments of pools .



(a) 4R

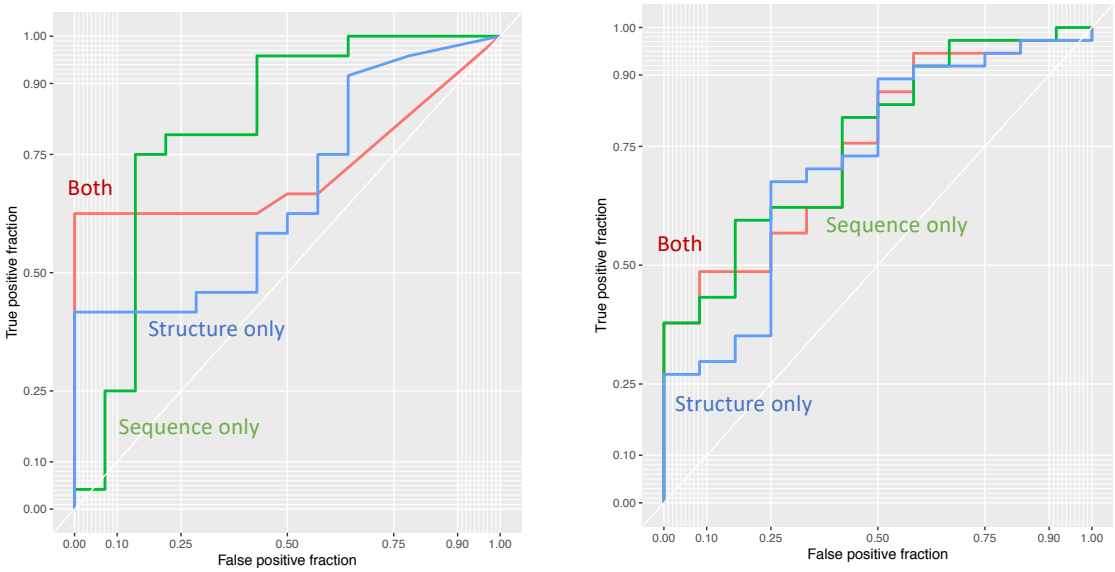


(b) 5R



(c) 6R

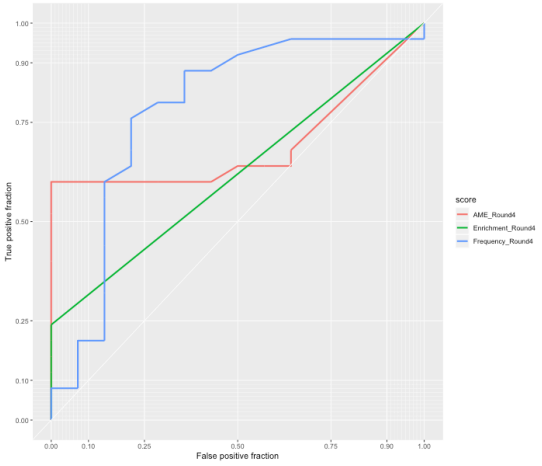
Fig. S5: The ROC curves for Data2. The ROC curves for round 3 are not shown because they were not informative; AME and Enrichment values of all sequences were undefined for round 3 since it was the first round included in the analysis.



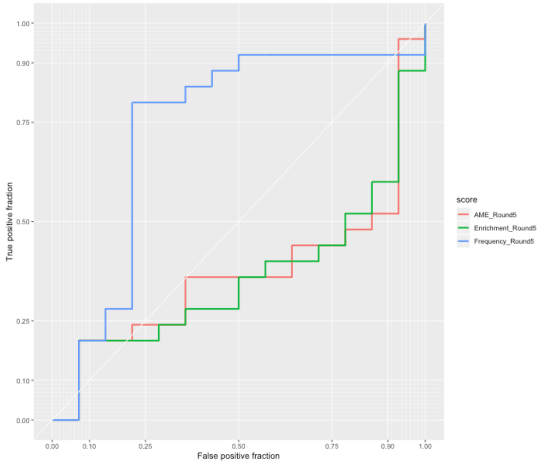
(a) RAPT1_4R

(b) RAPT3_6R

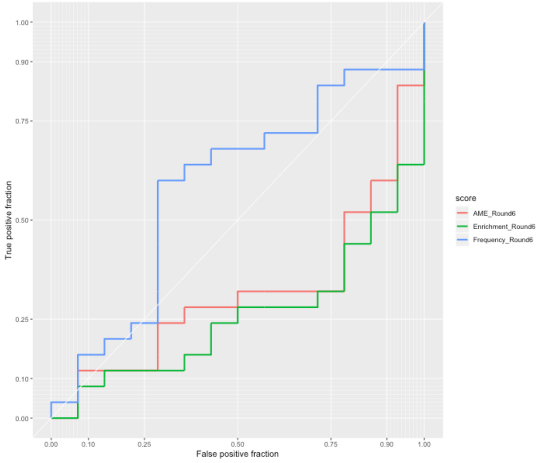
Fig. S6: ROC curves corresponding to analysis in section "Validation of the effectiveness of analyzing both sequence and secondary structure". Only the results from rounds 4 and 6 are shown because they are the last rounds of each dataset.



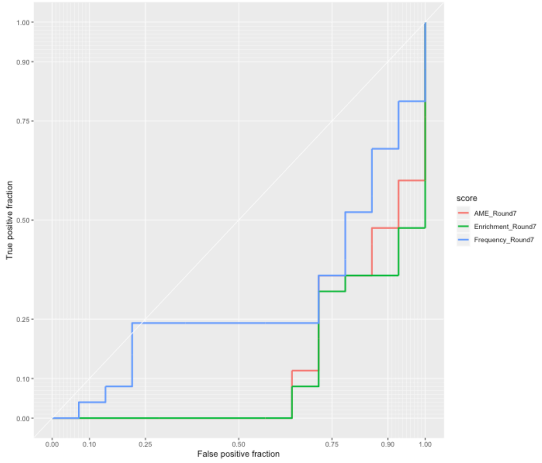
(a) 4R



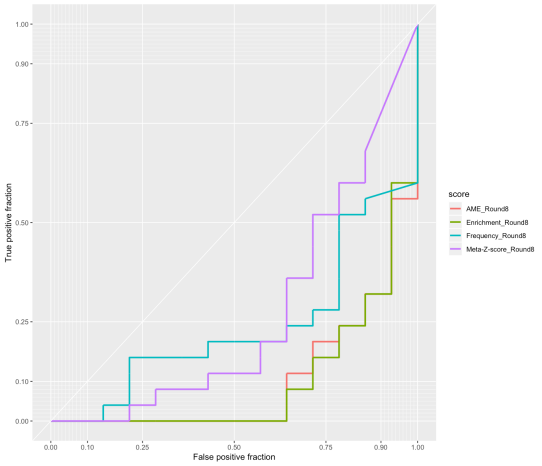
(b) 5R



(c) 6R

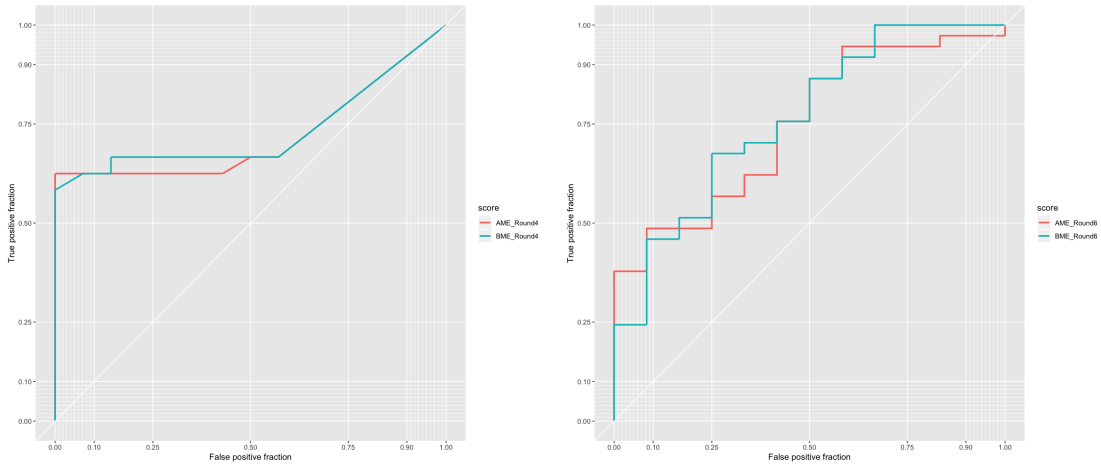


(d) 7R



(e) 8R

Fig. S7: ROC curves for Data1, rounds 0–8R. The ROC curves for rounds 0–3R were not informative because, for those rounds, AME and Enrichment values of almost all sequences were undefined due to insufficient enrichments of pools.



(a) Data1_4R (b) Data2_6R

Fig. S8: ROC curves corresponding to analysis in section "Compare sequence score definitions". Only the results from rounds 4 and 6 are shown because they are the last rounds of each dataset.

H The number of sequences in each round of each dataset

Table S4: The number of sequences in each round of Data1

round	raw	filtered	unique	filtered/raw	unique/raw	unique/filtered
0R	162003	114663	113106	70.78%	69.82%	98.64%
1R	91610	72555	71453	79.20%	78.00%	98.48%
2R	45431	36583	36071	80.52%	79.40%	98.60%
3R	91441	72758	71719	79.57%	78.43%	98.57%
4R	80864	64327	49260	79.55%	60.92%	76.58%
5R	108428	86805	28582	80.06%	26.36%	32.93%
6R	98237	80920	11825	82.37%	12.04%	14.61%
7R	49469	40341	5092	81.55%	10.29%	12.62%
8R	113137	81161	6029	71.74%	5.33%	7.43%
to 4R total	471349	360886	341415	76.56%	72.43%	94.60%
to 8R total	840620	650113	379534	77.34%	45.15%	58.34%

Titles "raw", "filtered", "unique" indicate the number of sequences inputted, the number of sequences to be analyzed, and the number of unique sequences, respectively.

Table S5: The number of sequences in each round of Data2

round	raw	filtered	unique	filtered/raw	unique/raw	unique/filtered
3R	146505	98861	96491	67.48%	65.86%	97.60%
4R	121185	80218	5889	66.19%	4.86%	7.34%
5R	116917	80114	6954	68.52%	5.95%	8.68%
6R	82488	47360	4813	57.41%	5.83%	10.16%
total	467095	306553	169924	65.63%	36.38%	55.43%

Titles "raw", "filtered", "unique" indicate the number of sequences inputted, the number of sequences to be analyzed, and the number of unique sequences, respectively.

I Effect of parameters

Table S6: Effect of parameters on Data1 (0R to 4R)

cosdist	*wide*	*weight*	repeat_1	repeat_2	repeat_3	Average TPR
		0.1	0.542	0.542	0.542	0.542
	8	0.5	0.542	0.542	0.542	0.542
		1.0	0.625	0.625	0.625	0.625
		0.1	0.542	0.542	0.542	0.542
0.0001	10	0.5	0.542	0.542	0.542	0.542
		1.0	0.625	0.625	0.625	0.625
		0.1	0.500	0.500	0.500	0.500
	12	0.5	0.542	0.542	0.542	0.542
		1.0	0.542	0.542	0.542	0.542
		0.1	0.667	0.667	0.667	0.667
	8	0.5	0.667	0.667	0.667	0.667
		1.0	0.625	0.625	0.625	0.625
		0.1	0.625	0.625	0.625	0.625
0.001	**10**	**0.5**	**0.625**	**0.625**	**0.625**	**0.625**
		1.0	0.625	0.625	0.625	0.625
		0.1	0.625	0.625	0.625	0.625
	12	0.5	0.625	0.625	0.625	0.625
		1.0	0.625	0.625	0.625	0.625
		0.1	0.375	0.375	0.375	0.375
	8	0.5	0.708	0.708	0.708	0.708
		1.0	0.583	0.583	0.583	0.583
		0.1	0.500	0.500	0.500	0.500
0.01	10	0.5	0.750	0.750	0.750	0.750
		1.0	0.625	0.625	0.625	0.625
		0.1	0.542	0.542	0.542	0.542
	12	0.5	0.708	0.708	0.708	0.708
		1.0	0.625	0.625	0.625	0.625

We averaged the values of three runs for each parameter combination (i.e., repeat_1, repeat_2 and repeat_3). The results for the default parameters are shown in bold.

Table S7: Effect of parameters on Data2

cosdist	*wide*	*weight*	repeat_1	repeat_2	repeat_3	Average TPR
0.0001	8	0.1	0.378	0.378	0.378	0.378
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
	10	0.1	0.378	0.378	0.378	0.378
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
	12	0.1	0.405	0.405	0.405	0.405
		0.5	0.405	0.405	0.405	0.405
		1.0	0.405	0.405	0.405	0.405
0.001	8	0.1	0.378	0.378	0.378	0.378
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
	10	0.1	0.378	0.378	0.378	0.378
		0.5	**0.378**	**0.378**	**0.378**	**0.378**
		1.0	0.378	0.378	0.378	0.378
	12	0.1	0.405	0.405	0.405	0.405
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
0.01	8	0.1	0.189	0.189	0.189	0.189
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
	10	0.1	0.297	0.297	0.297	0.297
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378
	12	0.1	0.270	0.270	0.270	0.270
		0.5	0.378	0.378	0.378	0.378
		1.0	0.378	0.378	0.378	0.378

We averaged the values of three runs for each parameter combination (i.e., repeat_1, repeat_2 and repeat_3). The results for the default parameters are shown in bold.