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

Ryoga Ishida, Tatsuo Adachi, Aya Yokota, Hidehito Yoshihara, Kazuteru Aoki,
Yoshikazu Nakamura, and Michiaki Hamada*

A  Notations
Table S1 summarizes notations frequently utilized in the main manuscript.

| 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 \(s\) |
| \(\mathcal{C}\) | Set of all clusters |
| \(C\)    | Each cluster which is a set of subsequence, \(C \in \mathcal{C}\) |

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 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.

*To whom correspondence should be addressed. Department of Electrical Engineering and Bioscience Faculty of Science and Engineering, Waseda University 55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo 169-8555, Japan. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp
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

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

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

| Parameter       | Value                           |
|-----------------|---------------------------------|
| forward_primer  | TAATACGACTCCTATAGGGAGGAGGAGAGGTCAGATG |
| reverse_primer  | CCTATGCGTCCTAGGTGA             |
| minimum_length  | 25nt                           |
| maximum_length  | 35nt                           |
| add_forward_primer| GGGACCGAGGAGGAGGTCAGAG      |
| add_reverse_primer| CCTATGCGTCCTAGGTGA          |

The parameters used in the analysis of Data1. The function of each parameter is as follows. \( \text{forward\_primer}, \text{reverse\_primer} \): RaptRanker extracts only the sequences whose both primer binding regions are exactly match these parameters. \( \text{minimum\_length}, \text{maximum\_length} \): RaptRanker extracts only the random regions whose length \( l \) is \( \text{minimum\_length} \leq l \leq \text{maximum\_length} \). \( \text{add\_forward\_primer}, \text{add\_reverse\_primer} \): 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

| Parameter       | Value                           |
|-----------------|---------------------------------|
| forward_primer  | TAATACGACTCCTATAGGGAGAATTCGACCACAAG |
| reverse_primer  | TATGTGCGCATACATGGATCCCT        |
| minimum_length  | 35nt                           |
| maximum_length  | 45nt                           |
| add_forward_primer| GGGAAGACTTACGACCAGAAG        |
| add_reverse_primer| TATGTGCGCATACATGGATCCCT      |

The parameters used in the analysis of Data2. The function of each parameter is as follows. \( \text{forward\_primer}, \text{reverse\_primer} \): RaptRanker extracts only the sequences whose both primer binding regions are exactly match these parameters. \( \text{minimum\_length}, \text{maximum\_length} \): RaptRanker extracts only the random regions whose length \( l \) is \( \text{minimum\_length} \leq l \leq \text{maximum\_length} \). \( \text{add\_forward\_primer}, \text{add\_reverse\_primer} \): 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

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

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

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

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 Evaluated sequence sets

#### Table S4: Evaluated sequence sets for Data1

| ID    | Sequence                                | Acquisition date | RU   | Positive/Negative |
|-------|-----------------------------------------|------------------|------|-------------------|
| Data1-1 | GTCGAGATTTCTGAGGGTTCTCCTGGTGCTG      | 161007           | 340.2| 1                 |
| Data1-2 | TGTAGGGAATGACCCGGGTCCTTTAAGCTT       | 161007           | 315.7| 1                 |
| Data1-3 | TGTAGGGAATGACCCGGGTCCTTTAAGCTT       | 161007           | 277.0| 1                 |
| Data1-4 | AACGAGAGATGCTGACCTTCATAAACAGTC       | 161007           | 244.6| 1                 |
| Data1-5 | AACGAGAGATGCTGACCTTCATAAACAGTC       | 161007           | 211.3| 1                 |
| Data1-6 | TTAGGAGATTTACGAGATGGCTGTATG         | 161007           | 210.5| 1                 |
| Data1-7 | GTACGAGATTTCTGAGGGCTTCTTGTTGAC       | 161007           | 134.7| 1                 |
| Data1-8 | GTACGAGATTTCTGAGGGCTTCTTGTTGAC       | 161007           | 134.1| 1                 |
| Data1-9 | AACGAGAGATGCTGACCTTCATAAACAGTC       | 161007           | 127.8| 1                 |
| Data1-10 | AACGAGAGATGCTGACCTTCATAAACAGTC      | 161007           | 127.4| 1                 |
| Data1-11 | AACGAGAGATGCTGACCTTCATAAACAGTC      | 161007           | 115.9| 1                 |
| Data1-12 | ACACAGAGAGATGCTGACCTTCATAAACAGTC   | 161007           | 105.5| 1                 |
| Data1-13 | ACACAGAGAGATGCTGACCTTCATAAACAGTC   | 161007           | 85.1 | 1                 |
| Data1-14 | ACACAGAGAGATGCTGACCTTCATAAACAGTC   | 161007           | 82.2 | 1                 |
| Data1-15 | ACACAGAGAGATGCTGACCTTCATAAACAGTC   | 161007           | 59.8 | 1                 |
| Data1-16 | ACACAGAGAGATGCTGACCTTCATAAACAGTC   | 161007           | 51.8 | 1                 |
| Data1-17 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 46.8 | 1                 |
| Data1-18 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 46.2 | 1                 |
| Data1-19 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 44.6 | 1                 |
| Data1-20 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 40.4 | 1                 |
| Data1-21 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 38.1 | 1                 |
| Data1-22 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 34.4 | 1                 |
| Data1-23 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | 0.7  | 0                 |
| Data1-24 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -4.5 | 0                 |
| Data1-25 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -7.7 | 0                 |
| Data1-26 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -8.2 | 0                 |
| Data1-27 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -10.3| 0                 |
| Data1-28 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -13.4| 0                 |
| Data1-29 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -14.1| 0                 |
| Data1-30 | ATACGAGATGCTGACCTTCATAAACAGTC      | 161007           | -24.1| 0                 |

The list of evaluated sequence sets of Data1. Sequences are presented in descending order of the RU value. The RU values at the end of protein injection were used for the binding level of aptamers (Figure S1a).
### Table S5: Evaluated sequence sets for Data2

| ID      | Sequence                                                                 | Acquisition date | RU      | Positive/Negative |
|---------|--------------------------------------------------------------------------|------------------|---------|-------------------|
| Data2-1 | GCTGTGTCCTACGTCCGGATTGGGACCTGCAGCCGCCATG                                 | 170314           | 260.1   | 1                 |
| Data2-2 | ACACAACACACGCTTTTGGAGCATACCTCAATCAACGAT                                  | 170314           | 239.0   | 1                 |
| Data2-3 | TCTCTTAGTCCTTTTGGGGCGACCGAATTACGGGAGACAC                                 | 170314           | 155.8   | 1                 |
| Data2-4 | TTCTCTACGGCGAGCTTACGGGAGCTTACCAATG                                      | 170314           | 153.0   | 1                 |
| Data2-5 | TCTCTTAGTCCTTTTGGGGCGACCGAATTACGGGAGACAC                                 | 170314           | 120.2   | 1                 |
| Data2-6 | TTTCTCTACGGCGAGCTTACGGGAGCTTACCAATG                                      | 170314           | 113.6   | 1                 |
| Data2-7 | TTTCTCTACGGCGAGCTTACGGGAGCTTACCAATG                                      | 170314           | 109.2   | 1                 |
| Data2-8 | TCTCTTAGTCCTTTTGGGGCGACCGAATTACGGGAGACAC                                 | 170314           | 102.9   | 1                 |
| Data2-9 | TTTCTCTACGGCGAGCTTACGGGAGCTTACCAATG                                      | 170314           | 99.7    | 1                 |
| Data2-10| TTTCTCTACGGCGAGCTTACGGGAGCTTACCAATG                                      | 170314           | 92.2    | 1                 |
| Data2-11| GACACTGGCAGGATTACGTACCCAGCAAATTCCTCCGCTGTTCGCTTACCTCTCGCC                | 170314           | 90.8    | 1                 |
| Data2-12| Data2-11                                                                | 170314           | 85.5    | 1                 |
| Data2-13| Data2-11                                                                | 170314           | 83.2    | 1                 |
| Data2-14| Data2-11                                                                | 170314           | 73.9    | 1                 |
| Data2-15| Data2-11                                                                | 170314           | 73.5    | 1                 |
| Data2-16| Data2-11                                                                | 170314           | 73.0    | 1                 |
| Data2-17| Data2-11                                                                | 170314           | 65.1    | 1                 |
| Data2-18| Data2-11                                                                | 170314           | 62.6    | 1                 |
| Data2-19| Data2-11                                                                | 170314           | 57.2    | 1                 |
| Data2-20| Data2-11                                                                | 170314           | 55.5    | 1                 |
| Data2-21| Data2-11                                                                | 170314           | 52.3    | 1                 |
| Data2-22| Data2-11                                                                | 170314           | 51.1    | 1                 |
| Data2-23| Data2-11                                                                | 170314           | 47.0    | 1                 |
| Data2-24| Data2-11                                                                | 170314           | 46.9    | 1                 |
| Data2-25| Data2-11                                                                | 170314           | 45.8    | 1                 |
| Data2-26| Data2-11                                                                | 170314           | 45.2    | 1                 |
| Data2-27| Data2-11                                                                | 170314           | 40.6    | 1                 |
| Data2-28| Data2-11                                                                | 170314           | 39.1    | 1                 |
| Data2-29| Data2-11                                                                | 170314           | 38.5    | 1                 |
| Data2-30| Data2-11                                                                | 170314           | 29.5    | 1                 |
| Data2-31| Data2-11                                                                | 170314           | 27.6    | 1                 |
| Data2-32| Data2-11                                                                | 170314           | 24.8    | 0                 |
| Data2-33| Data2-11                                                                | 170314           | 24.1    | 0                 |
| Data2-34| Data2-11                                                                | 170314           | 21.9    | 0                 |
| Data2-35| Data2-11                                                                | 170314           | 21.6    | 0                 |
| Data2-36| Data2-11                                                                | 170314           | 18.6    | 0                 |
| Data2-37| Data2-11                                                                | 170314           | 18.2    | 0                 |
| Data2-38| Data2-11                                                                | 170314           | 15.4    | 0                 |
| Data2-39| Data2-11                                                                | 170314           | 15.0    | 0                 |
| Data2-40| Data2-11                                                                | 170314           | 8.8     | 0                 |
| Data2-41| Data2-11                                                                | 170314           | 5.0     | 0                 |
| Data2-42| Data2-11                                                                | 170314           | 3.5     | 0                 |
| Data2-43| Data2-11                                                                | 170314           | -0.1    | 0                 |
| Data2-44| Data2-11                                                                | 170314           | -4.2    | 0                 |

The list of evaluated sequence sets of Data2. Sequences are presented in descending order of the RU value. The RU values at the end of protein injection were used for the binding level of aptamers (Figure S2a).
Figure S1: The figures show SPR sensorgrams of the association and dissociation phases of aptamers. The RU values at the end of protein injection were used for the binding level of aptamers. (a) SPR sensorgrams for all aptamers from the evaluated sequence sets for Data1. (b) List of the top 3 aptamers from each method (Frequency, Enrichment, MPBind, and RaptRanker) used to the evaluated sequence sets for Data1. (c) SPR sensorgrams for the top 3 aptamers from each method used to the evaluated sequence sets for Data1. These sensorgrams are re-displayed from Figure S1a.
Figure S2: The figures show SPR sensorgrams of the association and dissociation phases of aptamers. The RU values at the end of protein injection were used for the binding level of aptamers. (a) SPR sensorgrams for all aptamers from the evaluated sequence sets for Data2. (b) List of the top 3 aptamers from each method (Frequency, Enrichment, MPBind, and RaptRanker) used to the evaluated sequence sets for Data2. (c) SPR sensorgrams for the top 3 aptamers from each method used to the evaluated sequence sets for Data2. These sensorgrams are re-displayed from Figure S2a.
Truncation of sequences

| sequence | IC<sub>50</sub> (μM) |
|----------|-------------------|
| GGGACAGGAGAGGCUAGAGAAAGAUGAUGCUGCUUUGCUAGUAGCUGUAGUGA | 0.140 ± 0.016 |
| GGGACAGGAGAGGCUAGAGAAAGAUGAUGCUGCUUUGCUAGUCGUCUC | 0.267 ± 0.029 |
| GGGAGGAGGAGAUGAUGCUGUAGGCUUUGCUAGUAGGCU | 0.243 ± 0.055 |
| GGGAGAGGAGAUGAUGCUGUAGGCUUUGCUAGUAGGCU | 0.350 ± 0.098 |
| GGGAGGAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.167 ± 0.024 |
| GGGUCAGAUGCAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.185 ± 0.041 |
| GGGUCAGAUGCAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.266 ± 0.026 |
| GGGUCAGAUGCAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.231 ± 0.029 |
| GGGUCAGAUGCAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.189 ± 0.029 |
| GGGUCAGAUGCAGAUGAUGCUGUAGGCUUUGCUAGUAGGCGUCU | 0.427 ± 0.173 |

![Graphs](image1.png)

![Graphs](image2.png)

![Graphs](image3.png)

![Graphs](image4.png)
Figure S3: The truncation of a high binding-affinity aptamer sequence of Data1. We truncated a high binding-affinity aptamer sequence of Data1 based on the Transglutaminase assay. The sequence in the top row is the original sequence, and the sequence in the bottom row is the minimal active sequence. The figures show TG2 activity in the presence of indicated aptamers. Experimental procedures are described in the “Transglutaminase assay” section. Broken lines indicate half maximal inhibitory concentration (IC₅₀).
Figure S4: **The truncation of a high binding-affinity aptamer sequence of Data2.** We truncated a high binding-affinity aptamer sequence of Data2 based on SPR assay. In the table, the sequence of S2-1 is the original sequence, and the sequence of S2-7 is the minimal active sequence. The SPR sensorgrams in the bottom figure show the association and dissociation phases for each aptamer.
### Table

| ID         | Sequence                                                                 | Relative binding (%) |
|------------|--------------------------------------------------------------------------|----------------------|
| Mother     | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 100                  |
| S3-1       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCU CACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 139                  |
| S3-2       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 1                    |
| S3-3       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-4       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-5       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-6       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-7       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 16                   |
| S3-8       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 182                  |
| S3-9       | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 130                  |
| S3-10      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 120                  |
| S3-11      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 174                  |
| S3-12      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 43                   |
| S3-13      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-14      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 0                    |
| S3-15      | GGGAGAACUUCGACCAGAAGGCUGUGUCUACGUCCGGAUUGGGGACCUGCACGGCCCAUGUAUGUGCGCAUACAUGGAUCCUC | 3                    |

### Figure S5: Binding motif evaluation.

We predict a binding motif of Data2 based on SPR assay for a series of mutant aptamers with single nucleotide deletion (shown in the top Table). The sequence of “Mother” is the original sequence, and the putative motif is shown in the bottom row of the table. The SPR sensorgrams in the bottom figure show the association and dissociation phases for each aptamer.
**H 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 $\mathcal{R}$ 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, $\bigcup_{K \in \mathcal{R}} K = P$ holds. Also, since k-mers do not overlap each other, $K \cap K' = \emptyset$ holds for any pair of k-mers $K, K'\in \mathcal{R}$. 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 $K \in \mathcal{R}$ 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

\[
\text{kmer}_{\text{frequency}}_x(K) = \sum_{p \in K} \text{subsequence}_{\text{frequency}}_x(p). \tag{S3}
\]

Then, RaptRanker calculates k-mer enrichment for each k-mer $K \in \mathcal{R}$ 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

\[
\text{kmer}_{\text{enrichment}}_x(K) = \frac{\text{kmer}_{\text{frequency}}_x(K)}{\text{kmer}_{\text{frequency}}_{x-1}(K)} \tag{S4}
\]

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

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

\[
\text{kmer}_{\text{enrichment}}_x(p) = \text{kmer}_{\text{enrichment}}_x(K) \tag{S5}
\]

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

\[
\text{AKE}_x(s) = \frac{\sum_{p \in D_s} \text{kmer}_{\text{enrichment}}_x(p)}{|D_s|}. \tag{S6}
\]

If round $x$ is the first round, the score is undefined.
Figure S6: 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.
Figure S7: 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.
Figure S8: 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.
Figure S9: 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.
Figure S10: 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.

### J The Number of Sequences in Each Round of Each Dataset

#### Table S6: 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% |

#### Table S7: 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.
K Effect of parameters

Table S8: Effect of parameters on Data1 (0R to 4R)

| cosdist | wide  | weight | repeat_1 | repeat_2 | repeat_3 | Average TPR |
|---------|-------|--------|----------|----------|----------|-------------|
| 0.1     | 0.524 | 0.524  | 0.524    | 0.524    | 0.524    | 0.524       |
| 8       | 0.5   | 0.524  | 0.524    | 0.524    | 0.524    | 0.524       |
| 1       | 0.619 | 0.619  | 0.619    | 0.619    | 0.619    | 0.619       |
| 1.00E-04| 10    | 0.5    | 0.286    | 0.286    | 0.286    | 0.286       |
|         | 1     | 0.429  | 0.429    | 0.429    | 0.429    | 0.429       |
| 0.1     | 0.524 | 0.524  | 0.524    | 0.524    | 0.524    | 0.524       |
| 12      | 0.5   | 0.571  | 0.571    | 0.571    | 0.571    | 0.571       |
|         | 1     | 0.619  | 0.619    | 0.619    | 0.619    | 0.619       |
| 0.001   | 10    | 0.5    | 0.524    | 0.524    | 0.524    | 0.524       |
|         | 1     | 0.524  | 0.524    | 0.524    | 0.524    | 0.524       |
| 0.1     | 0.667 | 0.667  | 0.667    | 0.667    | 0.667    | 0.667       |
| 12      | 0.5   | 0.667  | 0.667    | 0.667    | 0.667    | 0.667       |
|         | 1     | 0.667  | 0.667    | 0.667    | 0.667    | 0.667       |
| 0.01    | 10    | 0.5    | 0.238    | 0.238    | 0.238    | 0.238       |
|         | 1     | 0.333  | 0.333    | 0.333    | 0.333    | 0.333       |
| 0.1     | 0.619 | 0.619  | 0.619    | 0.619    | 0.619    | 0.619       |
| 12      | 0.5   | 0.762  | 0.762    | 0.762    | 0.762    | 0.762       |
|         | 1     | 0.333  | 0.333    | 0.333    | 0.333    | 0.333       |

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 S9: Effect of parameters on Data2

| cosdist | wide | weight | repeat_1 | repeat_2 | repeat_3 | Average TPR |
|---------|------|--------|----------|----------|----------|-------------|
| 0.1     | 0.361| 0.361  | 0.361    |          |          | 0.361       |
| 8       | 0.361| 0.361  | 0.361    |          |          | 0.361       |
| 1       | 0.361| 0.361  | 0.361    |          |          | 0.361       |
| 0.0001  | 10   | 0.5    | 0.361    | 0.361    | 0.361    | 0.361       |
|         | 1    | 0.361  | 0.361    | 0.361    |          | 0.361       |
| 0.1     | 0.389| 0.389  | 0.389    |          |          | 0.389       |
| 8       | 0.5  | 0.389  | 0.389    |          |          | 0.389       |
| 1       | 0.389| 0.389  | 0.389    |          |          | 0.389       |
| 0.001   | 10   | 0.5    | 0.361    | 0.361    | 0.361    | 0.361       |
|         | 1    | 0.361  | 0.361    | 0.361    |          | 0.361       |
| 0.1     | 0.389| 0.389  | 0.389    |          |          | 0.389       |
| 8       | 0.5  | 0.389  | 0.389    |          |          | 0.389       |
| 1       | 0.389| 0.389  | 0.389    |          |          | 0.389       |
| 0.01    | 10   | 0.5    | 0.361    | 0.361    | 0.361    | 0.361       |
|         | 1    | 0.361  | 0.361    | 0.361    |          | 0.361       |
| 0.1     | 0.278| 0.278  | 0.278    |          |          | 0.278       |
| 8       | 0.5  | 0.278  | 0.278    |          |          | 0.278       |
| 1       | 0.278| 0.278  | 0.278    |          |          | 0.278       |
| 0.1     | 0.250| 0.250  | 0.250    |          |          | 0.250       |
| 12      | 0.5  | 0.361  | 0.361    | 0.361    |          | 0.361       |
| 1       | 0.361| 0.361  | 0.361    |          |          | 0.361       |

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.