Large Scale Analysis of the Mutational Landscape in HT-SELEX Improves Aptamer Discovery
Online Supplementary Materials

Jan Hoinka[1], Alexey Berezhnoy[2], Phuong Dao[1], Zuben E. Sauna[3], Eli Gilboa[2], and Teresa M. Przytycka[1]

[1] National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda MD 20894, USA, przytyck@ncbi.nlm.nih.gov
[2]Department of Microbiology & Immunology, University of Miami Miller School of Medicine, Miami, Florida 33101, USA
[3]Laboratory of Hemostasis, Division of Hematology, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring Maryland, 20993 USA

Supplementary Note 1: AptaCluster Details

The mathematical description below follows the preliminary conference report [1].

Data Representation

Each selection round is represented as an associative array with keys corresponding to the species in the pool and values representing their respective frequency counts. This operation scales linearly with N, the number of sequences in the pool. An aptamer $s = (s_i)_{i=1}^{n}$, with a randomized region of length n, is formally described by the sequence of nucleotides s_i over the alphabet $\Omega = \{A, C, G, T\}$ where the index i corresponds to the i-th position of the nucleotide sequence. Furthermore, let the set of unique aptamers for pool P be defined as $S = \{s^j \in P \parallel s^j \neq s^k \; \forall j, k \in [1, \ldots, |S|] \wedge \sum_{j=1}^{|S|} m(s^j) = N\}$, where $m(s^j)$ corresponds to the frequency of s^i.

Dimension reduction using Locality Sensitive Hashing

The principle of LSH relies on the fact that closely related, high-dimensional data points are likely to collapse to the same point after applying a probabilistic dimensionality reduction and will hence produce identical hash values [2].

AptaCluster capitalizes on this concept by representing each sequence $s^j \in S$ as an n-dimensional vector and reducing this vector into d dimensions ($d < n$). Our approach therefore produces a set I_d of d randomly sampled indices $i \in [1, \ldots, n]$ and restricts the number of nucleotides to be used for the hashing procedure to $s_{i \in I_d}$ for each sequence s^j. This establishes a strong correlation between the sequence similarity of a set of aptamers and the likelihood of producing the same mapping, where the choice of d guarantees members of the same set to differ in at most $n - d$ positions. In other words, our approach implicitly computes an upper bound to the edit distance.

AptaCluster then proceeds to iteratively refine this upper bound by applying a user defined number of distinct hash functions to the data set. Subsequently, two sequences are considered dissimilar with high probability if they are assigned to different sets in every iteration. The upper bound computation is modeled as $d_{lsh}^{k}(s^1, s^2)$, where k refers to the k^{th} iteration and $L^k(s)$ stands for the value of the k^{th} hash function for sequence s. By setting $d_{lsh}^{0}(s^1, s^2) = \infty$ for all pairs, the value of d_{lsh} is refined through the following recurrence.

$$d_{lsh}^{k}(s^1, s^2) = \begin{cases} n - d & L^k(s^1) = L^k(s^2) \\ d_{lsh}^{k-1}(s^1, s^2) & L^k(s^1) \neq L^k(s^2) \end{cases} \tag{1}$$

Here, only the assignment in the first line needs to be executed. Defining $L^k(s)$ involves selecting a mapping h from a family of functions F at random

$$F = \{h : \mathbb{N}^l \to \mathbb{N}^d \parallel h(I) = I_d\} \tag{2}$$

where $I = (1, \ldots, n)$ represents the nucleotide positions of an aptamer of size n, followed by applying the function

$$L = \{\Omega^l \to \Omega^d \parallel L(s) = (s_i) \ \forall\ i \in I_d\} \tag{3}$$

to each aptamer s. This produces a sub-string $\hat{s}$ constituting the concatenation of the nucleotides at the positions defined in I_d, which is subsequently used as input to the traditional hashing procedure. $I_d = (i_0, \ldots, i_d)$ can be efficiently computed as follows: Let $i_0 \in [1, n]$ be a randomly selected index of I and define $x \in [2, n-1]$ as a random number co-prime to n. Then, the remaining positions can be generated with

$$i_j = (i_{j-1} + x) \mod n, \quad j = 1, \ldots, d-1 \tag{4}$$

and

$$I_d = (i_j)_{j=0}^{d-1}, \quad i_j < i_{j+1} \ \forall\ j \tag{5}$$

corresponds to the sequence of indices after sorting these in ascending order. Designing I_d using the scheme described above assures each index in I to be selected exactly once and maximizes the probability that indices are chosen in a non-consecutive manner.

Cluster Extraction

AptaCluster establishes aptamer families in order of the species' frequency-counts in the pool, drawing on the assumption that this measure is related to the selective advantage of an aptamer due to its binding affinity. Until no unassigned aptamers are left, the most frequent sequence s not belonging to any group is designated to be the seed of a new cluster, which is consequently expanded by computing a k-mer based distance between the seed and those sequences for which d_{lsh} was finite and for which d_{kmer} is smaller that a user defined cutoff. In particular,

$$d_{kmer}(s^x, s^y) = \sum_{i=1}^{4k} \left| \frac{X_i}{|s^x| - k + 1} - \frac{Y_i}{|s^y| - k + 1} \right|^2 \tag{6}$$

where X_i and Y_i denotes the number of times the i-th k-mer occurs in sequence s^x and s^y respectively and $|s^i|$ corresponds to the length of the aptamer.

By circumventing the need to compute distances between species not related according to LSH, our method performs these steps in $\mathcal{O}(N * m * k)$ time, where m denotes the maximum number of seed sequences in a hash bucket which is bounded by the size of the largest bucket generated during LSH.

Parameters used in this study

For the experiments described in this paper, we performed a total of $r = 15$ iterations of LSH while sampling 60% of the randomized region (i.e. $n = 24$). The parameter $d_{cutoff} = 5$ is set in terms of the maximal number of point mutations any pair of sequences should have and is converted into the k-mer distance d_{kmer} cutoff by sampling a user defined number of aptamers from the pool (10000 by default), artificially mutating that sequence up to d_{cutoff} times, and averaging over all d_{kmer} between these mutants and the wild-type. Using the estimator we derived in Supplementary Note 2, the value of d_{cutoff} corresponds to an initial sequence similarity of $K = 87.5\%$, and thus to a probability of encountering sequences with at least the specified similarity of $1.3861 * 10^{-16}$ (see Supplementary Figure 2). Furthermore we set $k = 3$ for the computation of d_{kmer} which has shown to give reasonable results for aptamer-sized sequences.

Supplementary Note 2: Derivation and Convergence Analysis

The diversity of an initial pool in a SELEX experiment is a function of the sequence length. Here, we provide a mathematical estimation of the expected number of aptamers of size n with at least K % similarity.

Problem Statement

Let $\frac{1}{k},\ k \in \mathbb{N}$ be the threshold for the sequence dissimilarity according to the edit distance. Furthermore, we define n as the length of the aptamer and assume $\frac{n}{k} \in \mathbb{N}$. The expected fraction $F(n,k)$ of aptamers with at most k% dissimilarity can then be calculated the sum over all possible sequences with i variable nucleotides divided by the number of all permutations of sequences of size n:

$$F(n,k) = \sum_{i=1}^{\frac{n}{k}} f(i,n) \tag{7}$$

$$\text{where } f(i,n) = \frac{\binom{n}{i}3^i}{4^n} \tag{8}$$

Approximation

For $i \leq \frac{n}{2}$ we have

$$f(i,n) > 3 * f(i-1,n) \tag{9}$$

Proof:

$$f(i,n) = \frac{\binom{n}{i}3^i}{4^n} = \frac{\binom{n}{i-1}3^{i-1} * 3}{4^n} * \frac{n-i+1}{i} \tag{10}$$

$$= f(i-1,n) * \underbrace{\left(\frac{n+1}{i} - 1\right) * 3}_{\geq 1 \text{ for } i \leq \frac{1}{2}n} \geq 3 * f(i-1,n) \tag{11}$$

Thus, we can approximate an upper bound to (7) using the last term of the expansion:

$$F(n,k) \approx f(i, \frac{n}{k}) = \frac{\binom{n}{\frac{n}{k}}3^{\frac{n}{k}}}{4^n} = \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{n!}{\frac{n}{k}!(n-\frac{n}{k})!} \tag{12}$$

Substituting $x!$ with the Stirling approximation $x! \approx \sqrt{2\pi x} * \frac{x}{e}^x$ we get

$$F(n,k) \approx \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{\sqrt{2\pi n} * \left(\frac{n}{e}\right)^n}{\sqrt{2\pi\frac{n}{k}} * \left(\frac{\frac{n}{k}}{e}\right)^{\frac{n}{k}} * \sqrt{2\pi\left(n-\frac{n}{k}\right)} * \left(\frac{n-\frac{n}{k}}{e}\right)^{n-\frac{n}{k}}} \tag{13}$$

$$= \left(\frac{3^{\frac{1}{k}}}{4}\right)^n * \frac{\sqrt{2\pi n} * \left(\frac{n}{e}\right)^n}{\sqrt{2\pi \frac{n}{k}} * \left(\frac{n}{ke}\right)^{n\frac{1}{k}} * \sqrt{2\pi n\left(1-\frac{1}{k}\right)} * \left(\frac{n*(1-\frac{1}{k})}{e}\right)^{n*(1-\frac{1}{k})}} \tag{14}$$

$$= \underbrace{\left(\frac{3^{\frac{1}{k}}}{4}\right)^n}_{(A)} * \underbrace{\frac{\sqrt{2\pi n}}{\sqrt{2\pi n\frac{1}{k}} * \sqrt{2\pi n\left(1-\frac{1}{k}\right)}}}_{(B)} * \underbrace{\frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{ke}\right)^{n\frac{1}{k}} * \left(\frac{n(1-\frac{1}{k})}{e}\right)^{n(1-\frac{1}{k})}}}_{(C)} \tag{15}$$

We can rewrite (C) as follows:

$$\frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{ke}\right)^{n\frac{1}{k}} * \left(\frac{n(1-\frac{1}{k})}{e}\right)^{n(1-\frac{1}{k})}} = \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e} * \frac{1}{k}\right)^{n\frac{1}{k}} * \left(\frac{n}{e} * (1-\frac{1}{k})\right)^{n(1-\frac{1}{k})}} \tag{16}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n\frac{1}{k}} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(\frac{n}{e}\right)^{n(1-\frac{1}{k})} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{17}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n\frac{1}{k}+n(1-\frac{1}{k})} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{18}$$

$$= \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{n}{e}\right)^{n(\frac{1}{k}+1-\frac{1}{k})} * \left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{19}$$

$$= \frac{1}{\left(\frac{1}{k}\right)^{n\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{n(1-\frac{1}{k})}} \tag{20}$$

$$= \left(\frac{1}{\left(\frac{1}{k}\right)^{\frac{1}{k}} * \left(1-\frac{1}{k}\right)^{(1-\frac{1}{k})}}\right)^n \tag{21}$$

$$= \left(\frac{1}{\left(\frac{1}{k}\right)^{\frac{1}{k}} * \left(\frac{k-1}{k}\right)^{(1-\frac{1}{k})}}\right)^n \tag{22}$$

$$= \left(\frac{1}{\frac{1^{\frac{1}{k}} * (k-1)^{1-\frac{1}{k}}}{k^{\frac{1}{k}} * k^{1-\frac{1}{k}}}}\right)^n \tag{23}$$

$$= \left(\frac{1}{\frac{1^{\frac{1}{k}} * (k-1)^{1-\frac{1}{k}}}{k^{\frac{1}{k}=1-\frac{1}{k}}}}\right)^n \tag{24}$$

$$= \left(\frac{k}{(k-1)^{1-\frac{1}{k}}}\right)^n \tag{25}$$

$$= \left(\frac{k(k-1)^{\frac{1}{k}}}{k-1} \right)^n := (D) \tag{26}$$

Combining (A) and (D) yields:

$$\left(\frac{3^{\frac{1}{k}}}{4} \right)^n * \left(\frac{k(k-1)^{\frac{1}{k}}}{k-1} \right)^n = \underbrace{\left(\frac{(3(k-1))^{\frac{1}{k}} * k}{4k-4} \right)}_{(E)}^n \tag{27}$$

Hence, an estimator for $F(n,k)$ can be written in the form of (A) * (E):

$$F(n,k) = \frac{\sqrt{2\pi n}}{\sqrt{2\pi n \frac{1}{k}} * \sqrt{2\pi n \left(1 - \frac{1}{k}\right)}} * \left(\frac{(3(k-1))^{\frac{1}{k}} * k}{4(k-1)} \right)^n \tag{28}$$

Note, that for $k \geq 2$ it follows that (E) decreases with k and $(E) < 1$. Hence $F(n,k)$ decreases at least exponentially with n where the base of the exponent decreases with k.

Supplementary Note 3: AptaSim Details

AptaSim is a program, aimed at realistically recreating the selection process during SELEX using error-prone PCR. For our simulation, we represent a pool as a set of sequences in which each sequence is attributed with a count, representing its frequency, and a value between 0 and 100 simulating the binding affinity to a putative target. Given an initial pool, we then perform a user-defined number of iterations comprising of target affine selection followed by error prone amplification. The remaining sequences after the selection stage represent the sequenced portion of HT-SELEX and are stored for further analysis.

Initial Pool Generation

To allow for the inclusion of existing biases such as the base composition and nucleotide dependencies of a pool originating from an in-vitro SELEX experiment, the input set of sequences for the simulation is generated based on a first order Markov Chain that captures the conditional probabilities of randomly selecting one nucleotide given the choice of the previous. Each sequence is then assembled by randomly selecting the first nucleotide with respect to the base composition of the training data and iteratively sampling the remaining nucleotides according to the conditional distributions of the model. In addition, each sequence is assigned a random initial count (≤ 5) as well as a binding affinity (≤ 25). Finally, we simulate strong binders by choosing 100 arbitrary sequences for which the binding affinity is uniformly sampled between 80 and 100.

Target Affine Sampling

The sampling step simulates incubation, binding, partitioning, and washing of a selection cycle during a SELEX experiment. Assuming enriched and target affine species to have a higher probability of selection, we sample, without replacement, 20% of the current pool according to the distribution of the sequence counts and accept a sequence with the probability corresponding to its binding affinity. Hence, the probability of selecting a sequence from the pool is proportional to its frequency and affinity.

Amplification

In order to restore the pool to its original size, we simulate a number of PCR cycles in which the amplification efficiency $e \in [0,1]$ as well as the mutation probability $p \in [0,1]$ can be specified. The number of required PCR cycles c is computed as follows:

$$c = \left\lceil \frac{\log\left(\frac{i}{x}\right)}{\log(1+e)} \right\rceil$$

where i and x correspond to the sizes of the initial and current pool respectively. In each PCR cycle, every aptamer is then subject to amplification as many times as its current count and in dependency of the specified probability of amplification e. If accepted, and based on the mutation probability p, the sequence is either duplicated or a mutant, differing by one base from the original at a random position, is introduced into the pool.

Default Parameters

We trained our Markov Model with all sequences from selection round 2 of our IL10 experiment. An initial pool of 100 million unique sequences of size 40nt was then generated containing approximately 100 high affinity binders. We then performed a total of 10 cycles. During each sampling step, 20% of the pool was retained whereas for amplification, we found a mutation probability of $p = 0.05$ and an amplification efficiency of $e = 0.95$ as suitable parameters for realistically recreating the pool characteristics of our in-vitro experiment.

Supplementary Note 4: AptaMut Details

AptaMut aims at extracting favorable mutants by recognizing that at each cycle, the sequenced aptamers represent a fraction of the true pool size (Figure 1b in paper). This is achieved by first identifying potentially favorable mutants in each cluster (sequence extraction) and consequently scoring each mutant, in which the likelihood of observing a divergent enrichment compared to the clusters' seed sequence is computed. A log-score near zero indicates a neutral mutant while significantly positive (respectively negative) log scores indicate a possibility of beneficial (respectively detrimental) mutants. In the following, we describe the details of each of these steps.

Sequence Extraction

We define a mutant as a sequence present in a particular cycle X and the next cycle $X + 1$, but that has never been encountered in the earliest sequenced pool. Consequently, a potentially favorable mutant is expected to have higher enrichment as compared to its parent sequence. Here, enrichment refers to the ratio between the mutant's frequency in cycles $X+1$ and X normalized by their respective pool sizes. We used the clustering results of the four largest aptamer families of selection cycle 5 and extracted potential favorable mutants. These mutants are subsequently scores as described below.

Scoring

In order to compute a score for each mutant reflecting the significance of the fold-change in enrichment between cycles X and $X+1$, we developed a generative model mirroring the experimental design of the HT-SELEX protocol. The model is based on the notion that the sequenced aptamers at each cycle only represent a fraction of the true pool size and that the process of selecting these sequences from the pool can be described in terms of a Bernoulli experiment. In addition we assume that the enrichment and the amplification processes are subject to noise modeled by a normal distribution. The model is parametrized by the expected sequence enrichment so that different sequence enrichments correspond to different models. Given the model built using the enrichment equal to the enrichment of the seed, we compute the probability that the mutant's counts in X and $X + 1$ could have been generated by this model. This probability is then normalized by the probability of the optimal counts, as described below.

We divide each selection round into three distinct sets denoted as *pool*, representing the remaining sequences after selection and amplification, *sample*, describing the established, sequenced portion of this pool, and *experiment*, standing for the unknown species forming the input for the next cycle (see Fig. 1b in paper). Furthermore, let m_s^x, m_e^x, and $m_p^x = m_s^x + m_e^x$, be the frequency of a sequence in the sets *sample*, *experiment*, and *pool* respectively. We define the enrichment of a sequence between selection cycles X and $X+1$

as $f_s^{x \to x+1}$ for the sample sets, and as $f_e^{x \to x+1}$ for the experiment sets. Similarly, we define the enrichment of the parent of the sequence between selection cycles as $\hat{f}_s^{x \to x+1}$ for the sample sets and as $\hat{f}_e^{x \to x+1}$ for the experiment sets. Finally, for a mutant that is neutral w.r.t. its parent sequence, its expected frequency in the pool $X+1$ can be described as

$$m_p^{x+1} \approx c * \hat{f}_s^{x \to x+1} * \underbrace{(m_p^x - m_s^x)}_{=m_e^x} \tag{29}$$

for any unknown count m_p^x in pool X. Here, we use the constant c to model both, the amplification stage (PCR) after each selection round and for normalization purposes.

Our model then aims at comparing the probability of observing frequencies m_s^x, m_s^{x+1} in sample sets X, $X+1$ of a mutant with the probability of observing the expected frequencies m_s^x, $m_s^x * \hat{f}_s^{x \to x+1}$. Let $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ refer to the probability of simultaneously observing m_s^x in sample set X and m_s^{x+1} in sample set $X+1$ by chance given the expected abundance of the mutant in the experiment sets can be described as a function of the enrichment of the parent sequence between the sample sets. Similarly, $P(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1})$ refers to the probability of the mutant being neutral, i.e. observing m_s^x and $\hat{f}_s^{x \to x+1} * m_s^x$ in the sample sets of X and $X+1$ respectively and under the assumption that their actual enrichment is identical to the seed's $\hat{f}_s^{x \to x+1}$. Then, we aim to compare $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ and $P(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1})$. We therefore define a significance score $S(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ for a mutant as the probability of the mutant's observed enrichment being higher than its parent, normalized by the probability of the mutant being neutral i.e. exhibiting an enrichment rate equal to its parent sequence:

$$S(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1}) = \frac{P\left(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1}\right)}{P\left(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1}\right)} \tag{30}$$

In what follows, we show how to compute $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x \to x+1})$ and $P(m_s^x, \hat{f}_s^{x \to x+1} * m_s^x, \hat{f}_s^{x \to x+1})$. The observations m_s^x and m_s^{x+1} in the sample sets can be interpreted as the result of partitioning pools X and $X+1$ into *sample* and *experiment* sets and hence as random variables following binomial distributions, in which m_s^x and m_s^{x+1} correspond to a known number of successes out of m_p^x and m_p^{x+1} unknown trails respectively. For any frequency of a mutant m_p in each pool, the probability of observing exactly m_s mutants in the sample set is then given by the probability mass function (*pmf*)

$$f_B(m_s, m_p, p) = Pr(X = m_s) = \binom{m_p}{m_s} p^{m_s} (1-p)^{m_p - m_s} \tag{31}$$

of the Binomial distribution $B(m_p, p)$ and the probability of simultaneously observing both frequencies in the sampled pools corresponds to the product of their respective *pmfs*. Since the original number of mutants in pool X is an unknown quantity, we have to consider all possible pool sizes in order to estimate $P(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1})$:

$$P\left(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}\right) = \sum_{m_p^x = m_s^x}^{\infty} f_B\left(m_p^x, m_s^x, p\right) * f_B\left(m_p^{x+1}, m_s^{x+1}, p\right) \tag{32}$$

So far, our model is only concerned with possible variations in the number of mutant sequences in the pool and does not take into account any biases that might affect the enrichment value of the seed sequence. These noises, such as artifacts during PCR and sequencing errors, might lead to an overestimation or underestimation of the true enrichment value. We therefore extend our approach with a continuous random variable f to model the observed seed enrichment $\hat{f}_s^{x\to x+1}$ in the sequenced portion of the pool. More specifically, we assume that f follows a normal distribution $\mathcal{N}(\hat{f}_s^{x\to x+1}, \hat{f}_s^{x\to x+1}/3)$ with mean $\hat{f}_s^{x\to x+1}$ and standard deviation $\hat{f}_s^{x\to x+1}/3$. We then express the probability of observing frequencies of a mutant in sample sets X, $X+1$ and the probability of observing its expected frequencies as functions of f. It follows that the new significance score of a mutant, denoted as $\hat{S}$, corresponds to ratio of the expected values of these functions:

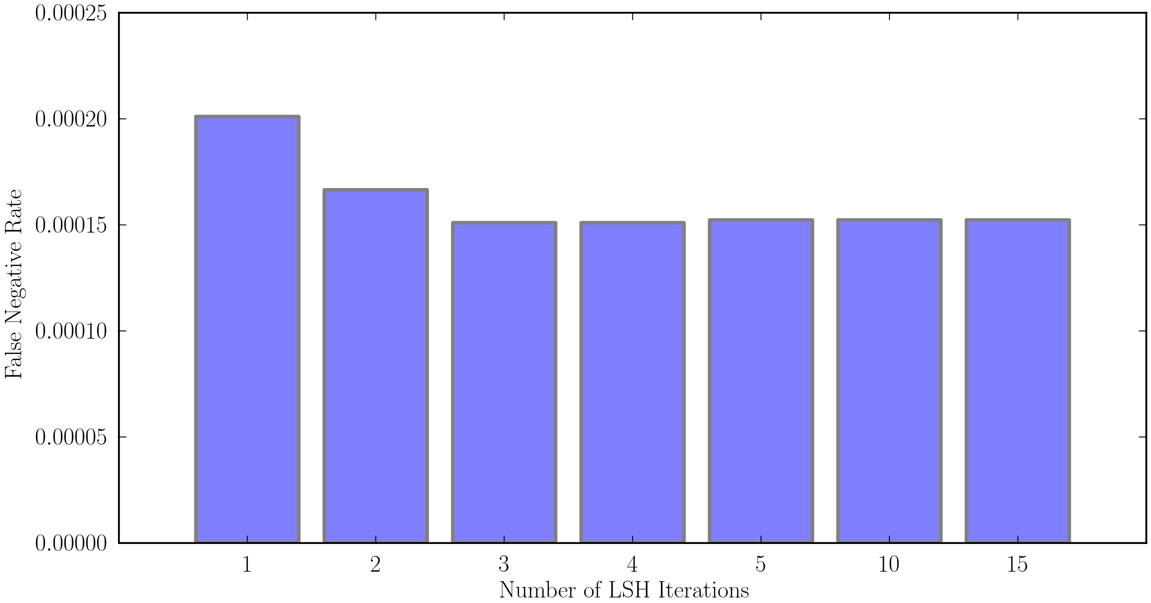
$$\hat{S}(m_s^x, m_s^{x+1}, \hat{f}_s^{x\to x+1}) = \frac{\int_0^{\infty} P\left(m_s^x, m_s^{x+1}, f\right) p(f) df}{\int_0^{\infty} P\left(m_s^x, \hat{f}_s^{x\to x+1} * m_s^x, f\right) p(f) df} \tag{33}$$

Here, $p(f)$ is the probability density function of the normal distribution $\mathcal{N}(\hat{f}_s^{x\to x+1}), \hat{f}_s^{x\to x+1}/3)$. Finally, we approximate each integral within three standard deviations from the mean by discretizing $p(f)$ into equidistant intervals of length d denoted as $k = \lfloor 2\hat{f}_s^{x\to x+1}/d \rfloor$-1. Below, we show how to approximate the integral on the example of the numerator and note that the denominator is approximated in a similar manner.

$$\begin{aligned}\int_0^{\infty} P\left(m_s^x, m_s^{x+1}, f\right) p(f) df &\approx \int_0^{2\hat{f}_s^{x\to x+1}} P\left(m_s^x, m_s^{x+1}, f\right) p(f) df \\ &\approx \sum_{i=1}^{k} P\left(m_s^x, m_s^{x+1}, f\right) * P\left(i * d \leq f < (i+1) * d\right)\end{aligned} \tag{34}$$

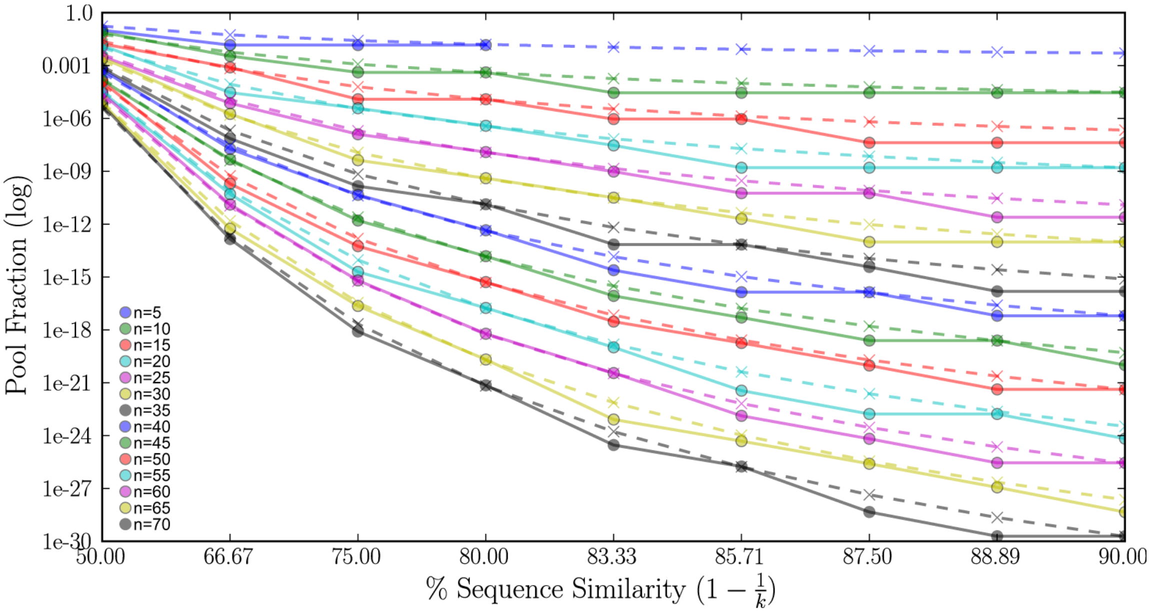
Setting $p = 0.5$, therefore assuming that each mutant has equal chance of being selected for sequencing, allows for the computation of the significance score $\hat{S}$ for all favorable mutants identified during the sequence extraction step. Analogous to p, we set $c = 2$, hence assuming that after selection, each pool is amplified back to its original size. For the discretization step, we found that a value of $d = 0.5$ as the width of intervals yielded the desired accuracy for our purposes. The main text we show $log(\hat{S})$ values of the so computed scores.

Supplementary Figure 1: False Negative Rates



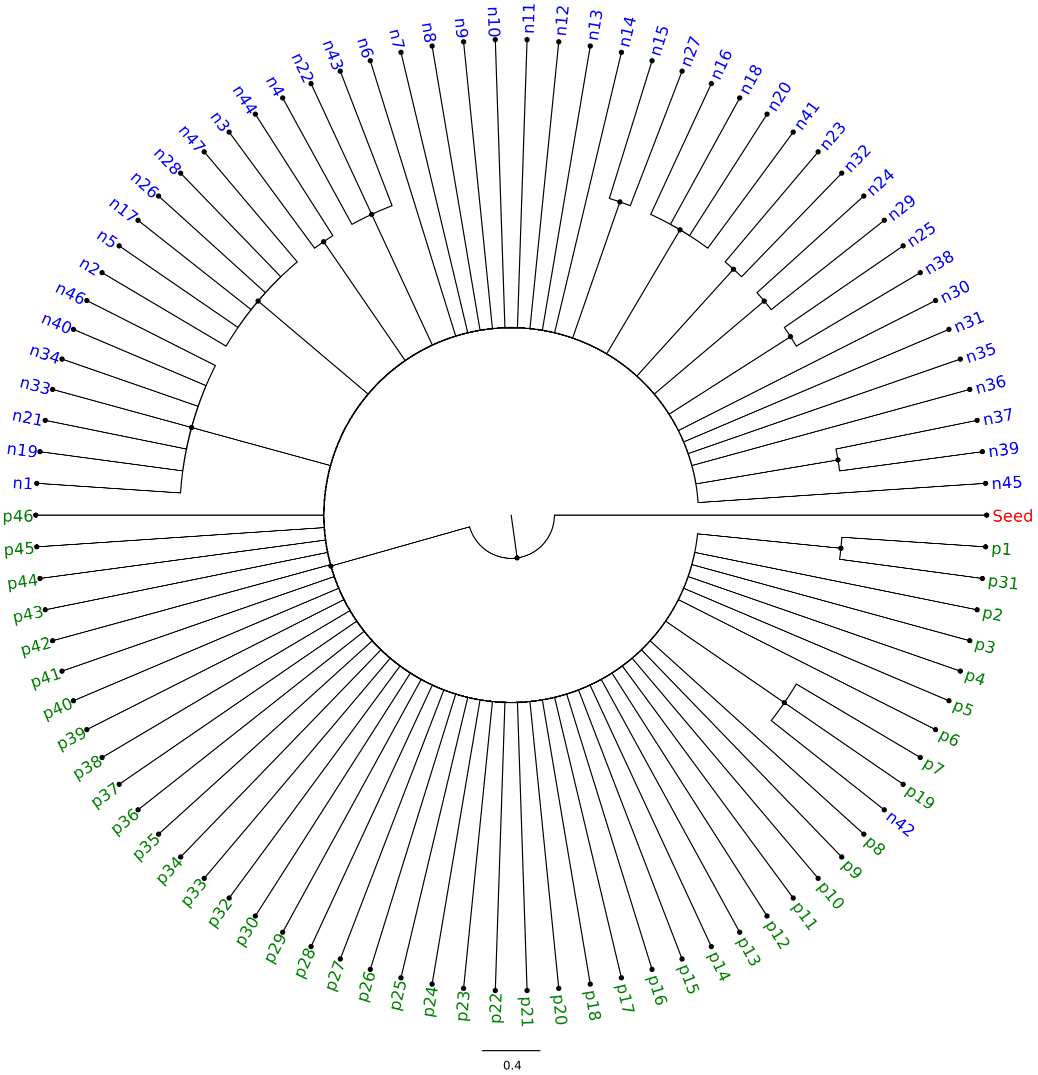
The false negative rates for the 20 largest clusters for different numbers of locality sensitive hashing iterations in selection cycle 5. The graph shows that LSH is stable for even a small number of iterations.

Supplementary Figure 2: Initial Pool Composition



Comparison of the predicted pool fraction of sequences with an expected sequence similarity K between our estimator (dashed lines) and the exact formula (continuous lines). Our estimator provides a reasonable upper bound for the expected fraction of sequences in an initial pool with at most K% sequence similarity.

Supplementary Figure 3: Phylogenetic Tree of Cluster ID 2



Phylogenetic tree of the mutants from cluster ID 2. Leafs labeled with **p** (green) correspond to the set of beneficial mutants in the order as depicted in Supplementary Table 3, where leafs starting with **n** (blue) stand for the degenerative species. The tree was constructed with PAUP* version 4.0 beta using the heuristic search option (1 Mio iterations) and an initial tree construction by adding species in the order of increasing log-score.

Supplementary Table 1: Sequences Introduced due to Mutagenesis

Number of species with counts 1 to 5 present in the top 20 clusters of selection round 5 compared to the frequency of their occurrence in selection round 2. The overwhelming majority of the sequences are not present in the latter.

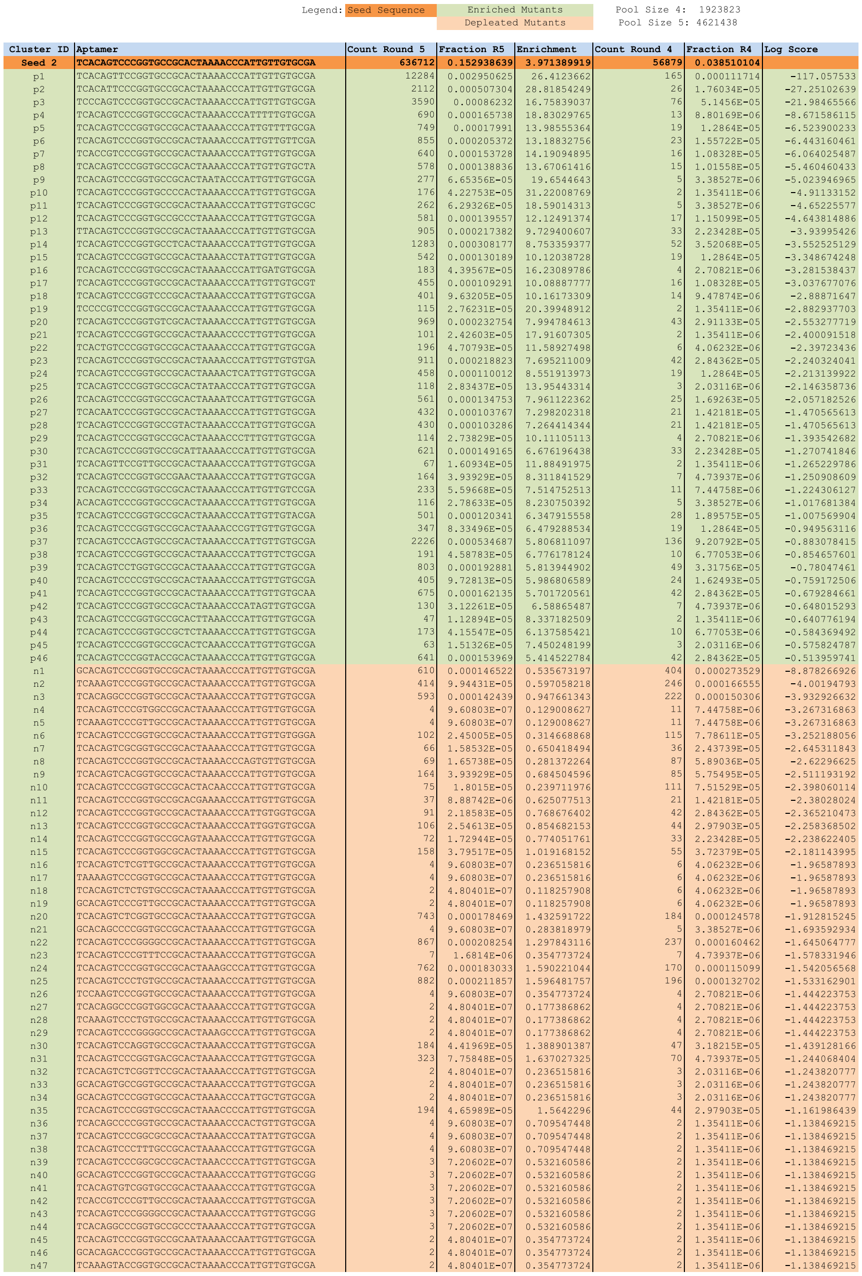
	Nr. of aptamers with frequency				
	1	2	3	4	5
Top 20, cycle 5	8529	2202	1074	614	465
Found in cycle 2	61	36	27	18	16

Supplementary Table 2: Cluster ID 1 - Selected Aptamers for Structural Analysis

Legend:	Seed Sequence	Enriched Mutants	Pool Size 4: 1923823
		Depleated Mutants	Pool Size 5: 4621438

Cluster ID	Aptamer	Count Round 5	Fraction R5	Enrichment	Count Round 4	Fraction R4	Log Score
SEED 1	**TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC**	**351921**	**7.61E-02**	**30.31212404**	**4833**	**2.51E-03**	
p1	TAACACTCGATTCTCCTAGCCCTCTAGAAATTCCCCTCCC	3097	0.000670138	429.7421301	3	1.56E-06	-19.16434217
p2	TAACACTTGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	1093	0.000236506	227.4982959	2	1.04E-06	-6.862747998
p3	TAACACTCGATTCTCCTAGCCCGCAAGAAATTCCCCTCCC	336	7.27E-05	139.8708644	1	5.20E-07	-2.642604337
p4	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCCCCC	1131	0.000244729	117.7038301	4	2.08E-06	-4.484570278
p5	TAACACTCTATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	472	0.000102133	98.24263097	2	1.04E-06	-2.042432047
p6	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTTCC	432	9.35E-05	89.91698428	2	1.04E-06	-1.754571899
p7	TAACACTCGATTCTCCTAGCCCGCTAGAGATTCCCCTCCC	622	0.00013459	86.30920405	3	1.56E-06	-2.21397266
p8	TAACACTCGATTCTCCTATCTCGCTAGAAATTCCCCTCCC	204	4.41E-05	84.92159627	1	5.20E-07	-1.387821997
p9	TAACACTCGATTCTCCTAGCCCGCTAGATATTCCCCTCCC	404	8.74E-05	84.0890316	2	1.04E-06	-1.53965282
p10	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCTTCCC	388	8.40E-05	80.75877292	2	1.04E-06	-1.433547752
p11	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTACC	188	4.07E-05	78.26107891	1	5.20E-07	-1.228048933
p12	TAACACTCGATTCTCCTAGCCCGCTATAAATTCCCCTCCC	1445	0.000312673	75.19099669	8	4.16E-06	-2.98686202
p13	TAACACTCGATTCTCCTAGCCCGCTACAAATTCCCCTCCC	174	3.77E-05	72.43312623	1	5.20E-07	-1.130789174
p14	TGACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	167	3.61E-05	69.51914988	1	5.20E-07	-1.108986469
p15	TAACACTCAATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	645	0.000139567	67.12552646	4	2.08E-06	-1.582980065
p16	TACCACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	315	6.82E-05	65.5644677	2	1.04E-06	-0.901286456
p17	TAACACTCGATTCTTCTAGCCCGCTAGAAATTCCCCTCCC	305	6.60E-05	63.48305603	2	1.04E-06	-0.837647438
p18	TAACACTCGCTTCTCCTAGCCCGCTAGAAATTCCCCTCCC	152	3.29E-05	63.27491486	1	5.20E-07	-1.067580847
p19	TAACACTCGATTCTCCTAGTCCGCTAGAAATTCCCCTCCC	443	9.59E-05	61.47102475	3	1.56E-06	-1.091191611
p20	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCTCCTCCC	424	9.17E-05	58.83456996	3	1.56E-06	-0.99261414
p21	TAACACTCGATTCTCCTAGCCCACTAGAAATTCCCCTCCC	540	0.000116847	56.19811518	4	2.08E-06	-1.022545259
p22	TAACATTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	266	5.76E-05	55.36555051	2	1.04E-06	-0.595583672
p23	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCT	660	0.000142813	54.94926817	5	2.60E-06	-1.07750217
p24	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCACTCCC	372	8.05E-05	51.61900949	3	1.56E-06	-0.69557297
p25	TAACACTCGATTCTCCTAGCCCGCTAGAAATTACCCTCCC	123	2.66E-05	51.20272716	1	5.20E-07	-0.891868083
p26	TAACACTCGATTCTCCTATCCCGCTAGAAATTCCCCTCCC	1428	0.000308995	45.72701337	13	6.76E-06	-0.898079417
p27	TAACACTCGATTCTCCTAGCCCGCTAGAAATCCCCCTCCC	867	0.000187604	45.11459802	8	4.16E-06	-0.705078901
p28	TAACACTCGATTCTCCTAGCCCGCTGGAAATTCCCCTCCC	431	9.33E-05	44.85442156	4	2.08E-06	-0.502529456
p29	TAACACTCGATTCTCCTAGCCGGCTAGAAATTCCCCTCCC	106	2.29E-05	44.12592747	1	5.20E-07	-0.524085542
p30	TAACACTCGATTCTACTAGCCCGCTAGAAATTCCCCTCCC	105	2.27E-05	43.70964514	1	5.20E-07	-0.524085542
n1	TAACCCTCTATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	30	6.49E-06	12.48847004	1	5.20E-07	-2.136210865
n2	TAACACTCGATACTCCTAGCCCGCTAGAAATTCCCCTCCC	24	5.19E-06	9.990776031	1	5.20E-07	-2.628904498
n3	TAACACTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCG	125	2.70E-05	8.672548638	6	3.12E-06	-0.545491351
n4	TAACACTCGATTGTCCTAGCCCGCTAGAAATTCCCCTCCC	40	8.66E-06	8.325646693	2	1.04E-06	-0.535680901
n5	TAACACTCGATTCTCCTAGCCCGCTAGAAAATCCCCTCCC	36	7.79E-06	7.493082023	2	1.04E-06	-0.891422702
n6	TAACACTCGATTCTCCTAGCCCGCTAGAAAGTCCCCTCCC	36	7.79E-06	7.493082023	2	1.04E-06	-0.891422702
n7	TAACTCTCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	330	7.14E-05	6.868658521	20	1.04E-05	-2.56278133
n8	TAACCCTCAATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	10	2.16E-06	4.162823346	1	5.20E-07	-4.397795847
n9	TAACACGCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	30	6.49E-06	3.12211751	4	2.08E-06	-2.075661839
n10	TAACGCCCGATTCTCCTAGCCCGCTAGAAATTCCCCTCCC	5	1.08E-06	2.081411673	1	5.20E-07	-5.049623067

Supplementary Table 3: Cluster ID 2 - Selected Aptamers for Structural Analysis



Legend: Seed Sequence | Enriched Mutants | Pool Size 4: 1923823

Depleated Mutants | Pool Size 5: 4621438

Cluster ID	Aptamer	Count Round 5	Fraction R5	Enrichment	Count Round 4	Fraction R4	Log Score
Seed 2	**TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA**	**636712**	**0.152938639**	**3.971389919**	**56879**	**0.038510104**	
p1	TCACAGTTCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	12284	0.002950625	26.4123662	165	0.000111714	-117.057533
p2	TCACATTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2112	0.000507304	28.81854249	26	1.76034E-05	-27.25102639
p3	TCCCAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	3590	0.00086232	16.75839037	76	5.1456E-05	-21.98465566
p4	TCACAGTCCCGGTGCCGCACTAAAACCCATTTTTGTGCGA	690	0.000165738	18.83029765	13	8.80169E-06	-8.671586115
p5	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTTTGCGA	749	0.00017991	13.98555364	19	1.2864E-05	-6.523900233
p6	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTTCGA	855	0.000205372	13.18832756	23	1.55722E-05	-6.443160461
p7	TCACCGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	640	0.000153728	14.19094895	16	1.08328E-05	-6.064025487
p8	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCTA	578	0.000138836	13.67061416	15	1.01558E-05	-5.460460433
p9	TCACAGTCCCGGTGCCGCACTAATACCCATTGTTGTGCGA	277	6.65356E-05	19.6544643	5	3.38527E-06	-5.023946965
p10	TCACAGTCCCGGTGCCCCACTAAAACCCATTGTTGTGCGA	176	4.22753E-05	31.22008769	2	1.35411E-06	-4.91133152
p11	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGC	262	6.29326E-05	18.59014313	5	3.38527E-06	-4.65225577
p12	TCACAGTCCCGGTGCCGCCCTAAAACCCATTGTTGTGCGA	581	0.000139557	12.12491374	17	1.15099E-05	-4.643814886
p13	TTACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	905	0.000217382	9.729400607	33	2.23428E-05	-3.93995426
p14	TCACAGTCCCGGTGCCTCACTAAAACCCATTGTTGTGCGA	1283	0.000308177	8.753359377	52	3.52068E-05	-3.552525129
p15	TCACAGTCCCGGTGCCGCACTAAAACCTATTGTTGTGCGA	542	0.000130189	10.12038728	19	1.2864E-05	-3.348674248
p16	TCACAGTCCCGGTGCCGCACTAAAACCCATTGATGTGCGA	183	4.39567E-05	16.23089786	4	2.70821E-06	-3.281538437
p17	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGT	455	0.000109291	10.08887777	16	1.08328E-05	-3.037677076
p18	TCACAGTCCCGGTCCCGCACTAAAACCCATTGTTGTGCGA	401	9.63205E-05	10.16173309	14	9.47874E-06	-2.88871647
p19	TCCCCGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	115	2.76231E-05	20.39948912	2	1.35411E-06	-2.882937703
p20	TCACAGTCCCGGTGTCGCACTAAAACCCATTGTTGTGCGA	969	0.000232754	7.994784613	43	2.91133E-05	-2.553277719
p21	TCACAGTCCCGGTGCCGCACTAAAACCCCTTGTTGTGCGA	101	2.42603E-05	17.91607305	2	1.35411E-06	-2.400091518
p22	TCACTGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	196	4.70793E-05	11.58927498	6	4.06232E-06	-2.39723436
p23	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGTGA	911	0.000218823	7.695211009	42	2.84362E-05	-2.240324041
p24	TCACAGTCCCGGTGCCGCACTAAAACTCATTGTTGTGCGA	458	0.000110012	8.551913973	19	1.2864E-05	-2.213139922
p25	TCACAGTCCCGGTGCCGCACTATAACCCATTGTTGTGCGA	118	2.83437E-05	13.95443314	3	2.03116E-06	-2.146358736
p26	TCACAGTCCCGGTGCCGCACTAAAATCCATTGTTGTGCGA	561	0.000134753	7.961122362	25	1.69263E-05	-2.057182526
p27	TCACAATCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	432	0.000103767	7.298202318	21	1.42181E-05	-1.470565613
p28	TCACAGTCCCGGTGCCGTACTAAAACCCATTGTTGTGCGA	430	0.000103286	7.264414344	21	1.42181E-05	-1.470565613
p29	TCACAGTCCCGGTGCCGCACTAAAACCCTTTGTTGTGCGA	114	2.73829E-05	10.11105113	4	2.70821E-06	-1.393542682
p30	TCACAGTCCCGGTGCCGCATTAAAACCCATTGTTGTGCGA	621	0.000149165	6.676196438	33	2.23428E-05	-1.270741846
p31	TCACAGTTCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	67	1.60934E-05	11.88491975	2	1.35411E-06	-1.265229786
p32	TCACAGTCCCGGTGCCGAACTAAAACCCATTGTTGTGCGA	164	3.93929E-05	8.311841529	7	4.73937E-06	-1.250908609
p33	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTCCGA	233	5.59668E-05	7.514752513	11	7.44758E-06	-1.224306127
p34	ACACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	116	2.78633E-05	8.230750392	5	3.38527E-06	-1.017681384
p35	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTACGA	501	0.000120341	6.347915558	28	1.89575E-05	-1.007569904
p36	TCACAGTCCCGGTGCCGCACTAAAACCCGTTGTTGTGCGA	347	8.33496E-05	6.479288534	19	1.2864E-05	-0.949563116
p37	TCACAGTCCCAGTGCCGCACTAAAACCCATTGTTGTGCGA	2226	0.000534687	5.806811097	136	9.20792E-05	-0.883078415
p38	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTCTGCGA	191	4.58783E-05	6.776178124	10	6.77053E-06	-0.854657601
p39	TCACAGTCCTGGTGCCGCACTAAAACCCATTGTTGTGCGA	803	0.000192881	5.813944902	49	3.31756E-05	-0.78047461
p40	TCACAGTCCCCGTGCCGCACTAAAACCCATTGTTGTGCGA	405	9.72813E-05	5.986806589	24	1.62493E-05	-0.759172506
p41	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCAA	675	0.000162135	5.701720561	42	2.84362E-05	-0.679284661
p42	TCACAGTCCCGGTGCCGCACTAAAACCCATAGTTGTGCGA	130	3.12261E-05	6.58865487	7	4.73937E-06	-0.648015293
p43	TCACAGTCCCGGTGCCGCACTTAAACCCATTGTTGTGCGA	47	1.12894E-05	8.337182509	2	1.35411E-06	-0.640776194
p44	TCACAGTCCCGGTGCCGCTCTAAAACCCATTGTTGTGCGA	173	4.15547E-05	6.137585421	10	6.77053E-06	-0.584369492
p45	TCACAGTCCCGGTGCCGCACTCAAACCCATTGTTGTGCGA	63	1.51326E-05	7.450248199	3	2.03116E-06	-0.575824787
p46	TCACAGTCCCGGTACCGCACTAAAACCCATTGTTGTGCGA	641	0.000153969	5.414522784	42	2.84362E-05	-0.513959741
n1	GCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	610	0.000146522	0.535673197	404	0.000273529	-8.878266926
n2	TCAAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	414	9.94431E-05	0.597058218	246	0.000166555	-4.00194793
n3	TCACAGGCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	593	0.000142439	0.947661343	222	0.000150306	-3.932926632
n4	TCACAGTCCCGTGGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.129008627	11	7.44758E-06	-3.267316863
n5	TCAAAGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.129008627	11	7.44758E-06	-3.267316863
n6	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGGGA	102	2.45005E-05	0.314668868	115	7.78611E-05	-3.252188056
n7	TCACAGTCGCGGTGCCGCACTAAAACCCATTGTTGTGCGA	66	1.58532E-05	0.650418494	36	2.43739E-05	-2.645311843
n8	TCACAGTCCCGGTGCCGCACTAAAACCCAGTGTTGTGCGA	69	1.65738E-05	0.281372264	87	5.89036E-05	-2.62296625
n9	TCACAGTCACGGTGCCGCACTAAAACCCATTGTTGTGCGA	164	3.93929E-05	0.684504596	85	5.75495E-05	-2.511193192
n10	TCACAGTCCCGGTGCCGCACTACAACCCATTGTTGTGCGA	75	1.8015E-05	0.239711976	111	7.51529E-05	-2.398060114
n11	TCACAGTCCCGGTGCCGCACGAAAACCCATTGTTGTGCGA	37	8.88742E-06	0.625077513	21	1.42181E-05	-2.38028024
n12	TCACAGTCCCGGTGCCGCACTAAAACCCATTGTGGTGCGA	91	2.18583E-05	0.768676402	42	2.84362E-05	-2.365210473
n13	TCACAGTCCCGGTGCCGCACTAAAACCCATTGGTGTGCGA	106	2.54613E-05	0.854682153	44	2.97903E-05	-2.258368502
n14	TCACAGTCCCGGTGCCGCAGTAAAACCCATTGTTGTGCGA	72	1.72944E-05	0.774051761	33	2.23428E-05	-2.238622405
n15	TCACAGTCCCGGTGGCGCACTAAAACCCATTGTTGTGCGA	158	3.79517E-05	1.019168152	55	3.72379E-05	-2.181143995
n16	TCACAGTCTCGTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.236515816	6	4.06232E-06	-1.96587893
n17	TAAAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.236515816	6	4.06232E-06	-1.96587893
n18	TCACAGTCTCTGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.118257908	6	4.06232E-06	-1.96587893
n19	GCACAGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.118257908	6	4.06232E-06	-1.96587893
n20	TCACAGTCTCGGTGCCGCACTAAAACCCATTGTTGTGCGA	743	0.000178469	1.432591722	184	0.000124578	-1.912815245
n21	GCACAGCCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.283818979	5	3.38527E-06	-1.693592934
n22	TCACAGTCCCGGGGCCGCACTAAAACCCATTGTTGTGCGA	867	0.000208254	1.297843116	237	0.000160462	-1.645064777
n23	TCACAGTCCCGTTTCCGCACTAAAACCCATTGTTGTGCGA	7	1.6814E-06	0.354773724	7	4.73937E-06	-1.578331946
n24	TCACAGTCCCGGTGCCGCACTAAAGCCCATTGTTGTGCGA	762	0.000183033	1.590221044	170	0.000115099	-1.542056568
n25	TCACAGTCCCTGTGCCGCACTAAAACCCATTGTTGTGCGA	882	0.000211857	1.596481757	196	0.000132702	-1.533162901
n26	TCCAAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.354773724	4	2.70821E-06	-1.444223753
n27	TCACAGGCCCGGTGGCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n28	TCAAAGTCCCTGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n29	TCACAGTCCCGGGGCCGCACTAAAGCCCATTGTTGTGCGA	2	4.80401E-07	0.177386862	4	2.70821E-06	-1.444223753
n30	TCACAGTCCAGGTGCCGCACTAAAACCCATTGTTGTGCGA	184	4.41969E-05	1.388901387	47	3.18215E-05	-1.439128166
n31	TCACAGTCCCGGTGACGCACTAAAACCCATTGTTGTGCGA	323	7.75848E-05	1.637027325	70	4.73937E-05	-1.244068404
n32	TCACAGTCTCGGTTCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n33	GCACAGTGCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n34	GCACAGTCCCGGTGCCGCACTAAAACCCATTGCTGTGCGA	2	4.80401E-07	0.236515816	3	2.03116E-06	-1.243820777
n35	TCACAGTCCCGGTGCCGCACTAAACCCCATTGTTGTGCGA	194	4.65989E-05	1.5642296	44	2.97903E-05	-1.161986439
n36	TCACAGCCCCGGTGCCGCACTAAAACCCACTGTTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n37	TCACAGTCCCGGCGCCGCACTAAAACCCATTATTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n38	TCACAGTCCCTTTGCCGCACTAAAACCCATTGTTGTGCGA	4	9.60803E-07	0.709547448	2	1.35411E-06	-1.138469215
n39	TCACAGTCCCGGCGCCGCACTAAACCCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n40	GCACAGTCCCGGTGCCGCACTAAAACCCATTGTTGTGCGG	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n41	TCACAGTGTCGGTGCCGCACTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n42	TCACCGTCCCGTTGCCGCACTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n43	TCACAGTCCCGGGGCCGCACTAAAACCCATTGTTGTGCGG	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n44	TCACAGGCCCGGTGCCGCCCTAAAACCCATTGTTGTGCGA	3	7.20602E-07	0.532160586	2	1.35411E-06	-1.138469215
n45	TCACAGTCCCGGTGCCGCAATAAAACCAATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215
n46	GCACAGACCCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215
n47	TCAAAGTACCGGTGCCGCACTAAAACCCATTGTTGTGCGA	2	4.80401E-07	0.354773724	2	1.35411E-06	-1.138469215

References

[1] Jan Hoinka, Alexey Berezhnoy, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Aptacluster–a method to cluster ht-selex aptamer pools and lessons from its application. In *Research in Computational Molecular Biology*, pages 115–128. Springer.

[2] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. *Similarity Search in High Dimensions via Hashing*, page 518–529. VLDB '99. Morgan Kaufmann Publishers Inc., 1999.