






# Tool and Algorithm for the Determination of Aptamers in Nanopore Sequencing Data: AptaLong

*Maria A. Grigoryeva*<sup>1</sup> , *Maria G. Khrenova*<sup>1,2</sup> , *Maksim F. Subach*<sup>1</sup> ,  
*Vladimir V. Voevodin*<sup>1</sup> , *Maria I. Zvereva*<sup>1</sup> 

© The Authors 2024. This paper is published with open access at SuperFri.org

Nanopore sequencing is a third generation sequencing technology that allows direct, real-time sequencing of individual DNA or RNA molecules. It utilizes a nanopore – an extremely small pore – in a membrane to pass a single strand DNA or RNA. As the sequence passes through the nanopore, changes in electrical current are detected and used to determine the nucleotide sequence. Nanopore sequencing has several advantages. It offers long read lengths, allowing for the sequencing of difficult regions of the genome, such as repetitive regions. It also enables real-time sequencing, providing immediate data generation without the need for extensive library preparation. Many bioinformatics pipelines and tools have been developed specifically for nanopore sequencing data analysis, addressing the unique characteristics and challenges of this technology, while dealing with non-standard long reads, derived from the ligation process of shorter oligonucleotides, might be challenging. In this research we present a new algorithm that extracts an aptamer sequence from the results of nanopore sequencing of several SELEX experiments with single-stranded DNA. The algorithm is based on statistical methods, based on known primer sequences and length of searching aptamer. We used step-by-step displacement of the reference sequence with positional alignment and calculated the positional frequencies of each nucleotide. As a result, the nucleotide frequencies obtained at each step are averaged, and thus, we find the sequence that is more likely to represent the aptamer.

*Keywords: nanopore sequencing, aptamer, SELEX, primer, sequence alignment.*

## Introduction

Aptamers are a specific type of targeting ligands based on single-stranded nucleic acids that can bind to a target molecule with high affinity and specificity. Aptamers can bind to targets ranging from small molecules to complex structures such as protein complexes or cell surface. Due to these unique characteristics, as well as low immunogenicity, toxicity, ease of synthesis with minor variations, good stability, they are used for a variety of diagnostic and therapeutic applications. Aptamers are also used as molecular probes instead of antibodies [4]. After a quarter of a century of research aptamers undergo pharmacological revision as selective drug for a specific clinical need [5]. Aptamers are usually selected from synthetic nucleic acids libraries. There are different strategies to obtain aptamers, as well as approaches to design initial compound libraries based on pre-structured sequences and modified nucleotides for optimisation of properties after selection of the best sequence [7]. The process of aptamer identification known as systematic evolution of ligands by exponential enrichment (SELEX) involves numerous singular processes, each of which contributes to the success or failure of aptamer generation [3]. Usually, the library for SELEX is presented as a set of nucleic acids with different sequences, each consisting of possible aptamers sequence connected with flanking regions for amplification by PCR and primers hybridisation after every round of selection in SELEX. Every sequence in the library is a combination of A, C, G, T nucleotides. The choice of modification position to improve properties is carried out for an already selected aptamer sequence based on the structural data of the aptamer-target. Recently, the use of libraries with modified bases (an additional one is

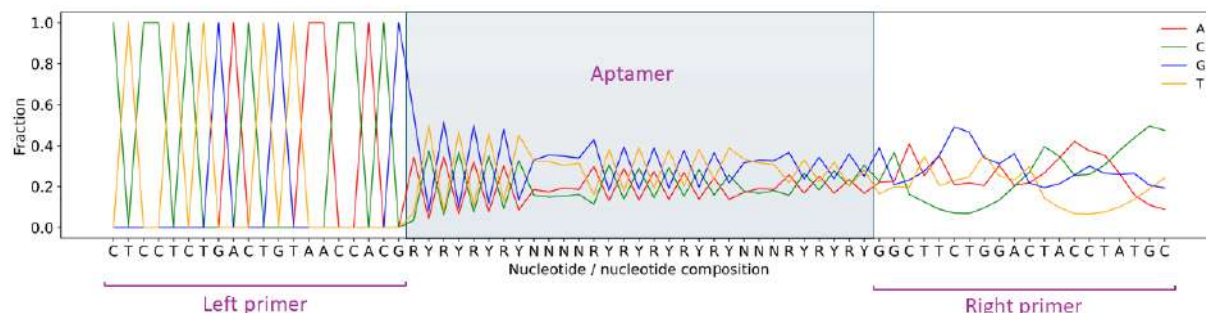
<sup>1</sup>Lomonosov Moscow State University, Moscow, Russian Federation

<sup>2</sup>Federal Research Centre “Fundamentals of Biotechnology” of the Russian Academy of Sciences

added to the four) has been proposed as methods for determining the sequence with modified bases directly have become available, in particular single-molecule nanopore sequencing [1]. The possibility of using nanopore sequencing for aptamer identification has recently been shown experimentally [2]. Since nanopore sequencing is focused on long reads, the selected sequences were combined into long reads for analysis. This required additional data analysis tools, which we offer.

By cutting out all the short sequences that include the left and right primers along with the aptamer between them, it theoretically becomes feasible to calculate the frequency of each nucleotide’s occurrence at every position. However, due to the sample preparation peculiarities, attempting a global alignment of sequences based on known primers, such as starting from the left primer shown in Fig. 1, reveals that the probabilities of the remaining nucleotides are too low to reliably identify the aptamer. Global alignment only provides insight into the prevailing positional probabilities of specific types of nucleotides. For instance, in Fig. 1, the notations R, Y, and N denote positions within the aptamer, indicating:

- R = 45:05:45:05 A/C/G/T – enriched with purine bases;
- Y = 05:45:05:45 A/C/G/T – enriched with pyrimidine bases;
- N = 25:25:25:25 A/C/G/T – equally probable.

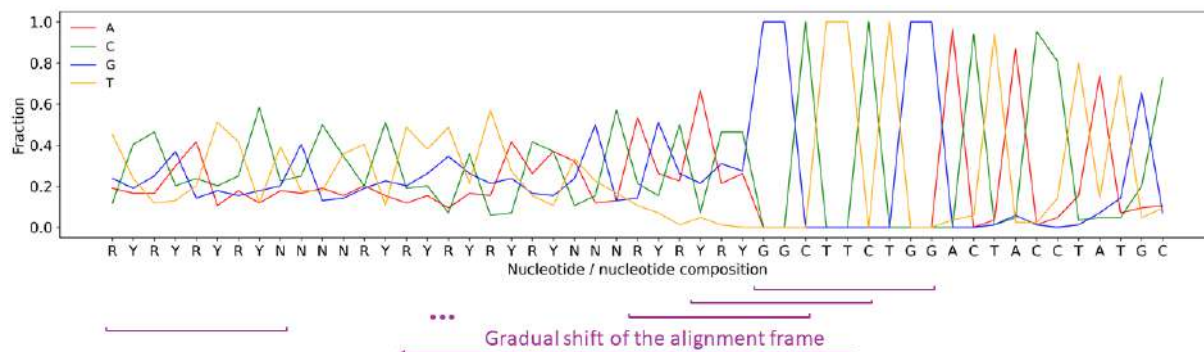


**Figure 1.** Globally aligned sequences: “Left primer – Aptamer – Right Primer”. Alignment from the left primer with the positional probabilities of nucleotide occurrences

Performing a global alignment of sequences starting from one of the primers does not enable the complete reconstruction of an aptamer in a single step. However, by selecting a shorter fragment of a primer for the alignment, a greater number of sequences can be included, providing more statistical data and increasing the likelihood of determining positional probabilities of nucleotides more effectively. Thus, we can take various fragments of a known primer, or fragments of an aptamer that were found in the immediate vicinity of the primer, in order to perform alignments step by step, clarifying previous statistical findings at each stage. By applying positional alignment to various known (or most likely known) fragments, averaging the probabilities obtained, it is possible to restore the aptamer.

The idea of the proposed algorithm is to incrementally shift some known reference fragment with sequence alignment and calculation of positional probabilities of nucleotide occurrence (see Fig. 2). At each stage, nucleotides with a high probability of occurrence will be added to the current reference fragment, thus the number of recognized nucleotides will increase. And as a result, we will only have to average the probabilities obtained in order to build an entire aptamer.

The paper is structured into five sections. Section 1 outlines current methods for aptamer detection. In Section 2, data representation is discussed. Section 3, which is further divided into five subsections, delves into the proposed algorithm, addressing data preparation, sequence extraction and alignment, description of the gradual movement of the reference sequence, pro-



**Figure 2.** Visualization of the general idea of the algorithm

cessing of collected statistics for all shifts, aggregation of results, and bifurcations. Section 4 is dedicated to statistical metrics for evaluating results. Lastly, Section 5 examines the validation of the AptaLong method.

## 1. Existing Approaches for Aptamer Search

### 1.1. Experimental Data Preparation

The preparation of SELEX samples for long-read nanopore and PacBio (sequencing of nucleic acids which involves real-time DNA replication with fluorescent nucleotide triphosphates producing long, accurate sequencing) requires the retrieval of long sequences, which is crucial for accurate characterization of aptamer candidates [6]. These methods can be widely used for detecting modified nucleic bases in aptamers. Currently there are two proposed methods for sample preparation after SELEX for the ability to analyze data using nanopore sequencing. The first method involves self-ligation, where SELEX library sequences after N rounds of selection are ligated with themselves, resulting in the formation of longer DNA molecules. This method was successfully used for obtaining aptamers for SARS-CoV-2 RBD protein with high affinity to the target [2]. Another method involves ligating the SELEX library after N rounds with a linearized plasmid vector (a circular DNA molecule that has been cut to form a linear piece). For this sequencing preparation TA cloning (a technique used to insert a piece of DNA into a plasmid vector) is used. The method takes advantage of a special feature where the DNA to be cloned has a single “A” (adenine) at each end, and the plasmid vector has a single “T” (thymine) at each end. These “A” and “T” ends naturally stick together, allowing the DNA to be easily inserted into the plasmid for replication and further use [9]. It is assumed that after N rounds of selection, the oligonucleotide pool becomes enriched with high-affinity aptamers. We can use this approach with further transformation of *E. coli* bacteria to yield a high amount of a plasmid containing most abundant aptamers for further sequencing. Moreover, ligation with a plasmid increases the length of aptamer sequences, making them suitable for nanopore sequencing [3].

### 1.2. Existing Algorithmical Methods for Aptamer Detection

There are many algorithms and tools for the analysis of nanopore sequencing, but in most cases these tools are aimed at searching for motifs. Motif commonly refers to a recurring pattern or sequence within a DNA/RNA molecule. An aptamer, on the other hand, is a specific type of sequence or molecule that can bind to a target molecule with high affinity and specificity. The

**Table 1.** Example representation of GAM for an aptamer consisting of 31 nucleotides

	0	1	2	3	4	5	6	7	8	9	10	...	26	27	28	29	30
<b>A</b>	0.45	0.05	0.45	0.05	0.45	0.05	0.45	0.05	0.25	0.25	0.25	...	0.05	0.45	0.05	0.45	0.05
<b>C</b>	0.05	0.45	0.05	0.45	0.05	0.45	0.05	0.45	0.25	0.25	0.25	...	0.45	0.05	0.45	0.05	0.45
<b>G</b>	0.45	0.05	0.45	0.05	0.45	0.05	0.45	0.05	0.25	0.25	0.25	...	0.05	0.45	0.05	0.45	0.05
<b>T</b>	0.05	0.45	0.05	0.45	0.05	0.45	0.05	0.45	0.25	0.25	0.25	...	0.45	0.05	0.45	0.05	0.45

specificity of aptamers lies in the method of their production – SELEX, which involves multiple binding and amplification operations, during which various clusters of compounds can be formed that are best bound to the target. Then, these compounds are ligated using primers, resulting in long chains of nucleotides. At the stages of amplification and ligation, various distortions of the sequences may occur, and therefore, the results can be significantly noisy, and the lengths of the sequences passing through the nanopores can also vary greatly. All this makes it much more difficult to use algorithms designed to find motifs.

One of the utilities that might be suitable is FASTAptamer<sup>3</sup>. It counts, normalizes and ranks read counts in a FASTQ file, compares populations for sequence distribution, generates clusters of sequence families, calculates fold-enrichment of sequences throughout the course of a selection and searches for degenerate sequence motifs. However, the sequences obtained from SELEX might not be in a form that is directly compatible with FASTAptamer. SELEX-derived aptamer sequences may contain modified bases, adapters, or linkers used in the SELEX process. These additional elements can make the sequences more complex and may not be recognized or handled properly by FASTAptamer, designed specifically for analyzing standard aptamer sequences. Therefore, while FASTAptamer may be useful for general analysis of aptamer sequences, it does not provide specialized functionalities for nanopore sequencing data or ligated sequences. For this reason, we decided to implement our own algorithm that takes into account the specifics of the data.

## 2. Data Representation

The initial data is represented in FASTQ files and several a priori known features.

- Left  $P_L$  and right  $P_R$  primers are known in advance, and are unique for a FASTQ file being analyzed.
- All sequences in a FASTQ file have variable lengths, but are supposed to contain left ( $P_L$ ) and right ( $P_R$ ) primers, and an aptamer  $A$  between them. In the sequencing process, primers and aptamers might be detected with errors reaching 10%. Moreover, during the ligation process, the sequences “Left Primer – Aptamer – Right Primer” might be disrupted.
- Global Alignment Matrix (GAM) with the prevailing positional probabilities of specific types of nucleotides.

## 3. Aptamer Search Method

The proposed algorithm, AptaLong, involves incrementally shifting the alignment frame while calculating the positional probabilities of each nucleotide. As the frame gradually moves,

<sup>3</sup><https://fastaptamer2.missouri.edu/>

nucleotide statistics are gathered at each shift. Upon reaching the shifting limit, these statistics are combined and summarized to reconstruct an aptamer.

The algorithm comprises multiple stages:

- data preparation;
- extraction and alignment of sequences;
- incremental movement of the reference sequence;
- consolidation of nucleotide probabilities over all reference shifts;
- traversal of bifurcations.

In the subsequent sections, we delve into a detailed explanation of each of these procedures.

### 3.1. Data Preparation

Complementary and direct forms of primers and aptamers often coexist within a FASTQ file. The analysis reveals a common occurrence of linked or “glued” complementary and direct primers. These connections, when abundant, can introduce statistical biases during sequence alignment due to shifts. In certain cases, as many as 50% of sequences in a FASTQ file may exhibit such fused primers. Hence, during the initial data preparation phase, it becomes critical to detect and segregate sequences with “glued” primers at their junctions. This step can notably increase the overall number of sequences initially.

#### Example:

Right primer ( $P_R$ ): GGCTTCTGGACTACCTATGC

Complementary right primer ( $P_{CR}$ ): GCATAGGTAGTCCAGAAGCC

Sequence with “glued”  $P_{CR}$  and  $P_R$ :

GACTGTAACACAGGATGTGTTCCCCTGTACGTTGTGCGTGTGCATAGGTAGTCCAGAAGCCGGCTTCTGGACTACCTATGCACACGAACACACTCTAACGACGCCACCGTGGTTACAGTCAGAGAGAATATACAGGCTAGAGAAGCAGTC

The resulting two sequences obtained by splitting the original sequence at the primer junction:

- GACTGTAACACAGGATGTGTTCCCCTGTACGTTGTGCGTGTGCATAGGTAGTCCAGAAGCC;
- GGCTTCTGGACTACCTATGCACACGAACACACTCTAACGACGCCACCGTGGTTACAGTCAGAGAGAATA  
TACAGGCTAGAGAAGCAGTC.

### 3.2. Sequences Extraction and Alignment

#### 3.2.1. Detection of the initial reference sequence

First of all, the initial reference sequence must be chosen. Since only primers are known in advance, the entire primer sequence could be utilized for alignment and aptamer discovery. However, given that primers might be distorted during ligation and nanopore sequencing, relying on the complete primer for alignment may not be advisable. Instead, certain primer fragments could be significantly represented in the data. Therefore, opting for a fragment from one of the primers as the starting point for the search is the most logical approach. This chosen fragment serves as the initial reference sequence and it should be linked to the sought aptamer. Experimental findings suggest that the optimal length of the reference sequence should not exceed half of the entire primer length.

In the context of optimal primer-aptamer ligation, in the sequence structure denoted as Left Primer - Aptamer - Right Primer ( $P_L - A - P_R$ ), the positioning of the reference

sequence at both ends of the aptamer is predetermined. In the provided illustration, with identified primers and the aptamer length specified, the initial reference sequence started at position  $idx$ , with the left portion of the right primer highlighted in red:

CTCCTCTGACTGTAACCACG\*\*\*\*\*GGCTTCTGGACTACCTATGC  
 0----- $idx$ ----- $L_A + 2 * L_P$

Another option, is to select the initial reference as the right part of the left primer:

CTCCTCTGACTGTAACCACG\*\*\*\*\*GGCTTCTGGACTACCTATGC  
 0----- $idx$ ----- $L_A + 2 * L_P$

### 3.2.2. Input data

- FastQ sequences – a list of DNA sequences of variable lengths;
- GAM, represented as a set of vectors with the prevailing probabilities of all nucleotides for a DNA library:

$$GAM = \begin{pmatrix} p(A_i)_G \\ p(C_i)_G \\ p(G_i)_G \\ p(T_i)_G \end{pmatrix};$$

- Left  $P_L$  and right  $P_R$  primers;
- $Ref_{idx} = [X]\{L_{Ref}\}$  – reference sequence – a fragment from the  $P_L$  or  $P_R$ , where  $L_{Ref}$  is the length of the reference sequence,  $X = [ACGT]$  – one of nucleotides A, C, G and T, and  $idx$  – the index of the occurrence of the reference;
- $L_P + L_A$  – the length of the fragments, where  $L_A$  – the length of an aptamer,  $L_P$  – the length of a primer;
- $2 * L_P + L_A$  – the total length of an ideally ligated sequence.

### 3.2.3. Fragments extraction and alignment by reference

At this stage, a set of fragments having equal length of  $L_P + L_A$ , aligned to the reference fragment  $Ref$  at its start position  $idx$ , are extracted from the initial sequences. If  $Ref$  belongs to the  $P_R$ , then the extracted sequence  $S_i$  can be represented as a potential aptamer, followed by the reference and the remaining part of the right primer:

$$S_i = [X]\{L_A\}[X]\{L_{Ref}\}[X]\{L_P - L_{Ref}\}.$$

If  $Ref$  belongs to the  $P_L$ , then the sequence consists of the initial part of the left primer, the reference and an aptamer:

$$S_i = [X]\{L_P - L_{Ref}\}[X]\{L_{Ref}\}[X]\{L_A\}.$$

For instance, shown below is a sequence comprised of two segments with the reference  $Ref = GGCTTCTGG$  beginning from the 31st position and  $P_R = GGCTTCTGGACTACCTATGC$ . The potential 31-base aptamers are marked in red, the reference segment in blue, and the remaining part of the potential right primer in gray:

ACCCTCCTCGGCTGCTTGGTCGCGGGCGGGTGTGGA**TACACTGGAGGTGCGCATAGGTGGTCAAGCCGGCTTCTGG**ACT  
 ACCTATGCAGACATCAAACGTACAGGTACCCATGCATCGTGGTTACAGTCAGAGAAGCTTCTGGACTTTACTATGCATA  
 TACAAATGTAAAGAGAGAAATTGCTT**TACACAACGTGGATTTACCAGTCAGAGGAGGCTTCTGG**ACTGCCTATTAGCAC

Two fragments can be found and aligned by the position of the reference GGCTTCTGG:

- TACTACTGGAGGTGCGCATAGGTGGTCAAGCCGGCTTCTGGACTACCTATGC;
- TTTACACAACGTGGATTTACCAGTCAGAGGAGGCTTCTGGACTGCCTATTA.

Consequently, a series of aligned fragments are obtained from the FASTQ file. These fragments can be presented in a tabular form (Tab. 2), where the columns represent position numbers and the rows signify the sequential numbers of the fragments.

The extracted sequences indicate considerable diversity among the aptamers, even when they originate from the same extended sequence, and the residual portions of the right primer also display variations.

A set of the extracted and aligned sequences can be represented as:  $Sequences = \langle S_1, S_2, S_3, \dots, S_N \rangle$ , where  $N$  is the number of extracted sequences.

**Table 2.** Table with the results of the initial alignment: columns are the numbers of positions, rows – numbers of extracted fragments. Position of the right primer is highlighted as gray. At the beginning of each extracted sequence there is an aptamer of  $L_A$  length, then there is a reference sequence and the remaining part of primer  $P_R$ . And the total length of the extracted sequences is  $L_A + L_P$

	0	1	2	3	4	5	...	LA	idx	idx	idx	idx	idx	...	LA
										+	+	+	+		+
										1	2	3	4		LP
0	T	G	G	T	T	A	...	...	G	G	C	T	T	...	G
1	A	G	A	C	A	T	...	...	G	G	C	T	T	...	G
2	A	G	T	G	C	T	...	...	G	G	C	T	T	...	G
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
N	C	T	G	G	G	T	...	...	G	G	C	T	T	...	G

**Filtration of the extracted sequences.** Another vital stage in the alignment process is the exclusion of sequences with incorrect nucleotides at primer positions. This becomes crucial, especially as the alignment extends beyond the primer region. If the reference sequence strays too far from the primer, there might be fragments with entirely different sequences at the primer positions. Such sequences can arise from amplification process nuances during SELEX and distortions in sequences during ligation. Misalignment of primer positions with the reference sequence can result in missing or distorted primers, complicating the confirmation process. It is recommended to remove these sequences from the analysis to maintain the accuracy of statistical evaluations.

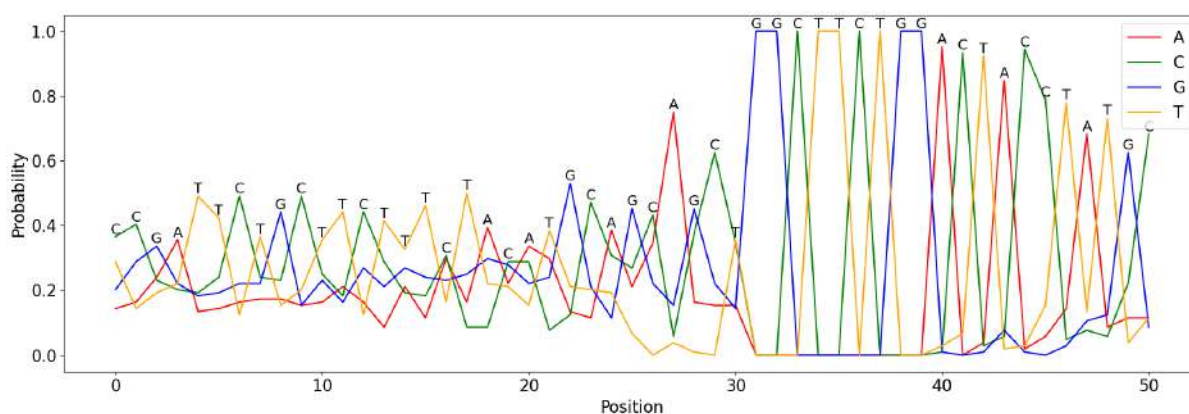
Various methods can be employed to measure the similarity between a primer and its corresponding sequence fragment. In our study, we utilized the Levenshtein distance algorithm for this purpose. The similarity is computed for each sequence within a shift and subsequently averaged.

### 3.2.4. Position-specific probability of nucleotide occurrences

Positional occurrence probabilities are calculated for each nucleotide from the collected fragments, as shown in Tab. 3. Subsequently, Fig. 3 illustrates the graphical representation of the distribution of occurrence probabilities for all nucleotides at each position.

**Table 3.** Positional probabilities of nucleotide occurrences. At the initial alignment the highest probability is at the region of the primer, and it decreases as moving further from the reference

	0	1	2	3	4	5	...	LA	idx	idx	idx	idx	idx	...	LA
										+	+	+	+		+
										1	2	3	4		LP
<b>A</b>	0.14	0.16	0.24	0.35	0.13	0.14	...	...	0	0	0	0	0	...	0
<b>C</b>	0.36	0.40	0.23	0.20	0.19	0.24	...	...	0	0	1	0	0	...	0
<b>G</b>	0.20	0.28	0.33	0.22	0.18	0.19	...	...	1	1	0	0	0	...	1
<b>T</b>	0.22	0.14	0.19	0.22	0.49	0.42	...	...	0	0	0	1	1	...	0
<b>C</b>	<b>C</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>T</b>	<b>...</b>	<b>...</b>	<b>G</b>	<b>G</b>	<b>C</b>	<b>T</b>	<b>T</b>	<b>...</b>	<b>G</b>	



**Figure 3.** Graphical representation of the distribution of positional probabilities of nucleotide occurrences for the alignment by the reference sequence GGCTTCTGG

At each position, the probabilities for A, C, G, and T are determined by the ratio of the specific nucleotide count at that position among all sequences to the total number of the aligned sequence fragments, calculated as:

$$p(X_i) = \frac{N_{X_i}}{N_i} .$$

Here,  $N_{X_i}$  represents the number of nucleotides X (A, C, G, or T) at position  $i$  across all  $N$  sequences, and  $N_i$  – number of all nucleotides at position  $i$ .

Consequently, for each position  $i$ , a vector consisting of four probabilities is generated:

$$p_i = \begin{pmatrix} p(A_i) \\ p(C_i) \\ p(G_i) \\ p(T_i) \end{pmatrix} .$$

The whole representation of probabilities is a set of such vectors for each position:  $Probabilities = \langle p_i \rangle$ , where  $i = 0$  to  $L_A + L_P$ .

The result sequence for the current reference alignment would be a set of nucleotides with the maximum positional probabilities.



### 3.3. Gradual Movement of Reference Sequence

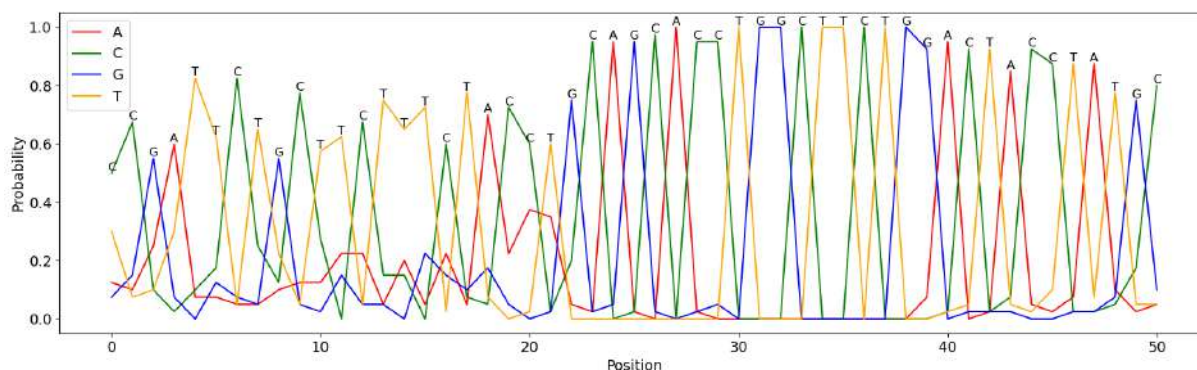
The proposed method implies a gradual movement of a reference sequence in both directions with the evaluation of the positional probabilities of occurrence for each nucleotide. The length of the reference sequence remains the same at each shift, but the combination of nucleotides might be different and is determined along the way depending on the nucleotide statistics. And the optimal choice of the reference for the next shift is the main challenge in this algorithm.

The shift can be uniquely identified by the combination of nucleotides,  $Ref$ , and its position within an ideally ligated primer and aptamer:  $idx: RefID = \{Ref, idx\}$ . At each shift  $N$  sequence fragments are extracted in accordance with the position of the reference and a length equal to primer plus aptamer (a set of sequences ( $Sequences_{RefID}$ )) and a matrix with nucleotides probabilities ( $Probabilities_{RefID}$ ) are calculated and saved for further processing.

#### 3.3.1. Reference sequence offset

After the distributions of positional probabilities for the initial reference alignment have been obtained at step 3.2.4, the reference sequence must be shifted to the left or to the right.

To determine how many positions  $k$  to move, it is necessary to estimate the maximum probabilities of nucleotides in the immediate vicinity of the current reference sequence. If there are nucleotides with a high probability (for example, more than 85%) of occurrence near the current reference, then this reference sequence shifts to the last highly probable nucleotide. It is illustrated in Fig. 4. If there is not a single nucleotide with a high positional probability in



**Figure 4.** Graphical representation of the distribution of positional probabilities of nucleotide occurrences for the alignment by the reference TGGCTTCTG with 7 highly probable nucleotides to the left

the immediate vicinity, then the displacement occurs by one step (as it is shown in Fig. 3 in Section 3.2.4).

With each subsequent shift, the next nucleotide in the reference sequence is not known in advance. In order to determine the most optimal nucleotide for the next displacement, it is necessary to perform some additional steps to evaluate the following characteristics for each variant of the reference sequence.

Thus, for each shift there might be four possible options, in accordance with the number of nucleotides (A, C, G and T). If the initial reference sequence, represented as a regular expression, is  $Ref_{idx} = (X)\{L_{Ref}\}$ , then the options shifted by one nucleotide to the left will be  $Ref_{idx-x} =$

$(X)(X)\{L_{Ref} - 1\}$ , where  $X = A|C|G|T$ . And if the movement in another direction, to the right, then the options will be:  $Ref_{idx+x} = (X)\{L_{Ref} - 1\}(X)$ .

For each of these options, the actions outlined in sections 3.2.3 and 3.2.4 are executed. Subsequently, based on the outcomes, the following metrics are calculated for each variation of the subsequent reference:

- $N_{RefID}$  – the number of sequences, extracted and aligned with a reference  $RefID$ ;
- $Sim_P$  – average similarity of  $P_L$  or  $P_R$  with the corresponding fragments of the extracted sequences:  $S[L_A : ]$  for comparison with the right primer, and  $S[: L_P]$  – for the left. The similarity can be measured as a number between 0 and 1;
- $N_{hits} = N_{RefID} * Sim_P$  – the number of hits as to the product of the number of sequences and similarity with primer.

Finally, the option with the highest value of  $N_{hits}$  is chosen as a candidate for the next shift.

The reference sequence displacement to the left or to the right is repeated until the index of the reference sequence reaches its limit. At each shift the following data is obtained:

- $Sequences_{RefID}$  – a set of sequences aligned by the current reference  $Ref$  at  $idx$  index of the start position in a sequence;
- $Probabilities_{RefID}$  – position-based probability of occurrences of each nucleotide in  $RefID$ ;
- $Bifurcations$  – an array of equally (or almost equally) probable reference sequences detected as a candidates for the next shifts.
  - If several sequences with close values of  $N_{hits}$  are found, and candidates for the next shift are equally likely, one of them is selected (with the maximum value of  $N_{hits}$ ), and the others are written to the array of bifurcations. In the next pass, an aptamer search will be started from the bifurcation.

### 3.4. Processing of the Collected Statistics for All Shifts

The data obtained at each shift can be represented as a list of sequences and positional probabilities of nucleotides:  $Shift_{RefID} = \langle Sequences_{RefID}, Probabilities_{RefID} \rangle$ .

And the final stage of the algorithm is the aggregation of data obtained at all shifts. Therefore, for each nucleotide its total positional representation is produced based on the  $Probabilities$  from all shifts.

The probability representation for nucleotide X can be expressed as a vector of probabilities of this nucleotide at each position for all shifts:

$$p(X) = \left\langle \begin{pmatrix} p(X_{1_1}) \\ p(X_{1_2}) \\ \dots \\ p(X_{1_{N_{shifts}}}) \end{pmatrix}, \begin{pmatrix} p(X_{2_1}) \\ p(X_{2_2}) \\ \dots \\ p(X_{2_{N_{shifts}}}) \end{pmatrix}, \dots, \begin{pmatrix} p(X_{(L_A+LP)_1}) \\ p(X_{(L_A+LP)_2}) \\ \dots \\ p(X_{(L_A+LP)_{N_{shifts}}}) \end{pmatrix} \right\rangle = \langle p(X_{i_j}) \rangle,$$

where  $i$  is the index of the nucleotide in a sequence,  $j$  is the index of the offset or shift.

The overall representation of probabilities for all nucleotides is:

$$p = \begin{pmatrix} p(A) \\ p(C) \\ p(G) \\ p(T) \end{pmatrix}.$$

After calculating the positional probabilities for all nucleotides, these values can be aggregated to derive the sequence encrypted by these probabilities. Thus, for a nucleotide X the average probability at each position is calculated as:  $\overline{p(X_i)}$ , where  $i$  is an index of the nucleotide in a sequence.

For each nucleotide, a sequence of positional probabilities can be calculated by averaging the values of all positional probabilities across all shifts, resulting in a probability representation of an aptamer:

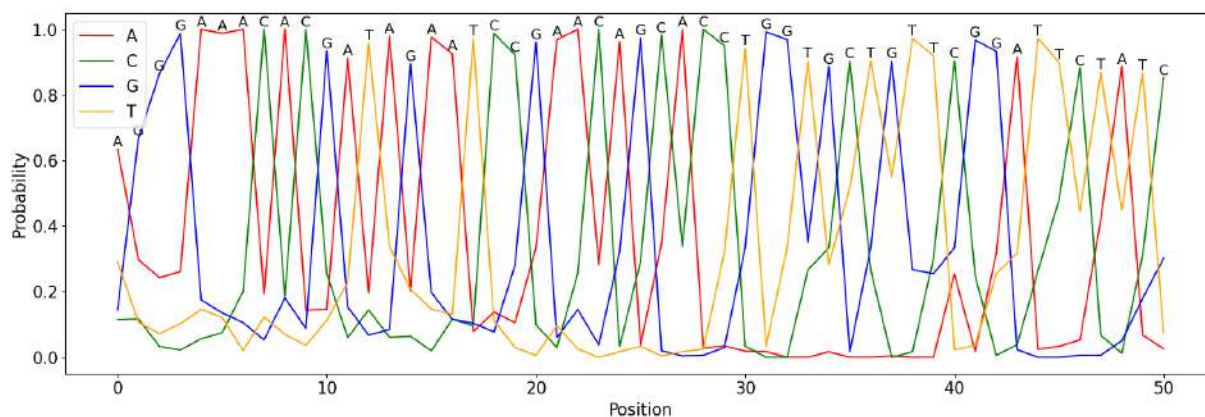
$$\overline{p(X)} = \left\langle \overline{\sum_{j=1}^{N_{shifts}} p(X_1)}, \overline{\sum_{j=1}^{N_{shifts}} p(X_2)}, \dots, \overline{\sum_{j=1}^{N_{shifts}} p(X_{L_A+L_P})} \right\rangle .$$

Finally, to obtain an aptamer sequence, the nucleotides corresponding to the highest probability are chosen at each position:

$$Aptamer = \left\langle \max_i \begin{pmatrix} \overline{p(A)} \\ \overline{p(C)} \\ \overline{p(G)} \\ \overline{p(T)} \end{pmatrix} \right\rangle ,$$

where  $i = 0$  to  $L_A + L_P$ .

Figure 5 demonstrates an example of the final distribution of positional probabilities of all nucleotides.



**Figure 5.** Aggregated table with positional probabilities of occurrence of each nucleotide: the first 31 nucleotides represent an aptamer, and remaining – right primer

### 3.5. Bifurcations

In the process of stepwise displacement of the reference sequence, bifurcations may appear at various stages, namely, nucleotides with similar values of positional probabilities of occurrence. In this case, the algorithm selects the most likely nucleotide, and writes the one closest to it to the list of bifurcations.

The table below (Tab. 4) shows a fragment of sequences aligned with the reference belonging to the right primer starting at position 31. It is necessary to select the nucleotide for the next shift one step to the left, that is, to position 30. C and T nucleotides in this position are almost equally likely (34 and 35%, respectively). Consequently, the algorithm will opt for T as

the nucleotide for the next shift, while recording C in the bifurcation list. Therefore, starting from the original reference GGCTTCTGG, applying an offset of one results in TGGCTTCTG. Given the new starting position with C, CGGCTTCTG, this change is noted at position 30 as a bifurcation.

**Table 4.** Nucleotides probabilities with a bifurcation: nucleotides C and T are equally probable at position 30

		25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
<b>A</b>	...	0.21	0.34	0.75	0.16	0.15	0.15	0	0	0	0	0	0	0	0	0
<b>C</b>	...	0.26	0.43	0.05	0.37	0.62	0.34	0	0	1	0	0	1	0	0	0
<b>G</b>	...	0.45	0.22	0.15	0.45	0.22	0.14	1	1	0	0	0	0	0	1	1
<b>T</b>	...	0.06	0	0.03	0.00	0	0.35	0	0	0	1	1	0	1	0	0
	...	<b>G</b>	<b>C</b>	<b>A</b>	<b>G</b>	<b>C</b>	<b>C</b> or <b>T</b>	<b>G</b>	<b>G</b>	<b>C</b>	<b>T</b>	<b>T</b>	<b>C</b>	<b>T</b>	<b>G</b>	<b>G</b>

Thus, the array of bifurcations consists of tuples, representing the index of the occurrence of the reference and the reference itself:

$$Bifurcations = \langle (Idx, Ref) \rangle .$$

The concept of preserving these bifurcations involves repeating all stages of aptamer search multiple times, starting from each identified bifurcation as an initial reference sequence. To prevent looping, the maximum number of bifurcations can be limited by the user.

#### 4. Statistical Metrics for Evaluation of the Results

To ensure that the obtained aptamer is statistically significant and to evaluate how well it aligns with the GAM and its corresponding primer position, several metrics are calculated for the output of each bifurcation.

1. Average number of sequences at all shifts:  $\overline{N_s} = \overline{\sum_{j=1}^{N_{shifts}} N_{RefID}}$  .
2. Overall probability of the determined aptamer:

$$\overline{p(Z)} = \overline{\sum_{i=1}^{L_P+L_A} \max(p(X)_i)} .$$

3. Similarity with the GAM:

First of all, the maximum probability of the GAM is calculated:

$$\max(GM) = \overline{\sum_{i=1}^{L_P+L_A} \max(p(X_i)_G)} .$$

Then, the average position probability of each statistically found nucleotide in accordance with the GAM is determined:

$$\overline{Z_{GM}} = \overline{\sum_{i=1}^{L_P+L_A} p(X_i|Z_i)_G} ,$$

where  $Z$  represents the sequence referred to the detected aptamer.

Finally, the similarity of this aptamer with the global alignment is calculated as:  $Z_{GM}/\max(GM)$ .

4. Similarity with the primer – calculated as the percentage of the similarity between the sequence that was obtained in the searching process and the primer.

## 5. Validation of the Algorithm

Validation of AptaLong was carried out based on the research aimed to identify and characterize a novel DNA aptamer, named MEZ, that binds to the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein. Key steps in the research included the generation of aptamers through the SELEX method, specifically targeting the RBD of the SARS-CoV-2 spike protein from the Wuhan-Hu-1 strain. Aptamer sequences were identified with the novel methodology based on nanopore sequencing, described in this paper [2].

MEZ, the best candidate aptamer detected by the developed algorithms, was chemically synthesized and tested for its binding affinity to the SARS-CoV-2 Spike RBD domain from different strains. The research found that MEZ had a comparable binding affinity to known aptamers, along with a shorter length of only 31 nucleotides. Experimental data and computational simulations showed that the 3'-end of the aptamer plays a crucial role in binding to the SARS-CoV-2 spike protein and strain identification.

## Conclusion

The developed algorithm, AptaLong, facilitates the exploration of aptamers within custom sequences derived from the ligated outcomes of SELEX experiments. The algorithm functions by initially selecting a known segment of the sequence (referred to as the reference fragment) and conducting multiple alignments based on the predetermined position of this reference. Alignments proceed through the stepwise shifting of the reference in both left and right directions. Subsequently, the positional probabilities of all nucleotides are computed across all shifts, ultimately leading to aptamer identification.

This tool enables the analysis of a variety of FASTQ files containing diverse types of aptamers, provided they share identical primers. Through the utilization of bifurcation in the search process, diverse aptamers can be effectively identified. The results are presented visually through graphical representations showcasing nucleotide probabilities, along with detailed information in Excel files for each shift and bifurcation stage, ensuring straightforward interpretation of the results.

The AptaLong tool can also be utilized for aptamer sequence determination from the data obtained on highthroughput variant of the nanopore sequencer, PrometION. In this case, a set of 96 samples, or even more, can be processed simultaneously and rapidly. Such extensive data analysis demands access to supercomputing facilities. Future enhancements to the AptaLong will incorporate a parallel implementation strategy to further optimize and scale the tool for efficient processing of large volumes of data.

## Acknowledgements

This work was supported by the project of the Interdisciplinary Scientific and Educational School of Moscow State University “Molecular technologies of living systems and synthetic biology” (№ 23-Sh04-45). The study was performed using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University [8].

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Berkovich, A., Pyshkina, O., Zorina, A., *et al.*: Direct determination of the structure of single biopolymer molecules using nanopore sequencing. *Biochemistry (Moscow)* 89, S234–S248 (2024). <https://doi.org/10.1134/S000629792414013X>
2. Khrenova, M., Nikiforova, L.A., Grabovenko, F., *et al.*: Highly specific aptamer for SARS-CoV-2 Spike protein from the authentic strain. *Org. Biomol. Chem.* 22, 5936–5947 (2024). <https://doi.org/10.1039/D4OB00645C>
3. Kohlberger, M., Gadermaier, G.: SELEX: Critical factors and optimization strategies for successful aptamer selection. *Biotechnol. Appl. Biochem.* 69(5), 1771–1792 (2022). <https://doi.org/10.1002/bab.2244>
4. Kumar Kulabhusan, P., Hussain, B., Yüce, M.: Current perspectives on aptamers as diagnostic tools and therapeutic agents. *Pharmaceutics* 12(7), 646 (2020). <https://doi.org/10.3390/pharmaceutics12070646>
5. Nimjee, S.M., White, R.R., Becker, R.C., *et al.*: Aptamers as therapeutics. *Annu. Rev. Pharmacol. Toxicol.* 57, 61–79 (2017). <https://doi.org/https://doi.org/10.1146/annurev-pharmtox-010716-104558>
6. Rhoads, A., Au, K.F.: PacBio Sequencing and its Applications. *Genomics Proteomics Bioinformatics* 13(5), 278–289 (2015). <https://doi.org/10.1016/j.gpb.2015.08.002>
7. Subach, M.F., Khrenova, M.G., Zvereva, M.I.: Modern methods of aptamer chemical modification and principles of aptamer library selection. *Moscow Univ. Chem. Bull.* 79, 79–85 (2024). <https://doi.org/10.3103/S002713142470010X>
8. Voevodin, V.V., Antonov, A.S., Nikitenko, D.A., Shvets, P.A., Sobolev, S.I., Sidorov, I.Y., Stefanov, K.S., Voevodin, V.V., Zhumatiy, S.A.: Supercomputer Lomonosov-2: Large Scale, Deep Monitoring and Fine Analytics for the User Community. *Supercomput. Front. Innov.* 6(2), 4–11 (2019). <https://doi.org/10.14529/jsfi190201>
9. Zhou, M.Y., Gomez-Sanchez, C.E.: Universal ta cloning. *Curr. Issues Mol. Biol.* 2(1), 1–7 (2000). <https://doi.org/10.21775/cimb.002.001>