ORIGINAL ARTICLE

Galaxy-compatible Tool for Rapid Aptamer Clustering and HT-SELEX Data Analysis

Nikita Aleksandrovich Skrylnik, Stepan Petrovich Chumakov, Natalia Mikhailovna Ratnikova, Yulia Evgenyevna Kravchenko, Elena Ivanovna Frolova

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, 16/10, Ulitsa Miklukho-Maklaya, GSP-7, 117997, Moscow, Russian Federation

Abstract

Aim: Implementing deep sequencing for analysis of DNA aptamer selection results requires for specialized bioinformatic software. Analysis steps include search for homologous sequences, clustering, and comparing cluster enrichment across different samples. These procedures allow deeper characterization of selected sequences by target affinity, non-specific amplification, and off-target binding, thus highlighting most promising variants or motifs. **Materials and Methods:** Sequencing results of systematic evolution of ligands by exponential enrichment for 40 nucleotide aptamers against extracellular CD47 protein were used as datasets for comparative clustering. Modified fast clustering script was developed based on FASTAptamer-Cluster and adapted as a galaxy tool. The algorithm was modified to terminate calculations after achieving the threshold value, and an exceeding edit distance was then assigned to non-matching pair of sequences. **Results and Discussion:** We have developed a set of galaxy compatible applications for rapid clustering of sequencing results and further comparative analysis of clusters. Our clustering algorithm is specifically optimized for searching for highly homologous sequences that usually form aptamer clusters and provides an average 8.4-fold increase in speed. **Conclusion:** Our modified clustering algorithm substantially surpasses existing alternatives in speed, thus simplifying analysis of large data sets, while its Galaxy version allows easy integration in standard workflows for preprocessing and analysis of the deep sequencing results.

Key words: HT-SELEX, levenshtein edit distance, aptamer clustering

INTRODUCTION

With the development of deep-sequencing technologies, the cost of conducting a research which involves this method decreases, gradually making its implementation in the new areas of study more accessible. One of such areas is combinatorial selection, with particular respect to the selection of high-affinity protein or nucleic acid ligands from extensive libraries. Nucleic acid ligands - aptamers - are usually selected from the libraries comprising random DNA or RNA fragments with the systematic evolution of ligands by exponential enrichment (SELEX) procedure [Figure 1].[1] Following the adsorption to the immobilized selection target, the pool of DNA molecules is exposed to serial elution to remove weakly bound sequences, after which the molecules that are still bound to the target are polymerase chain reaction amplified and subjected to another selection. After several round of selection, the library is cloned into a plasmid, and the most widely represented aptamers are identified with Sanger sequencing. Deep sequencing allows one to omit the cloning stage; moreover, its implementation in

Address for correspondence:
Nikita Aleksandrovich Skrylnik, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, 16/10, Ulitsa Miklukho-Maklaya, GSP-7, 117997, Moscow, Russian Federation
E-mail: stepan@chumakov.email

Received: 18-02-2017
Revised: 24-05-2017
Accepted: 01-06-2017

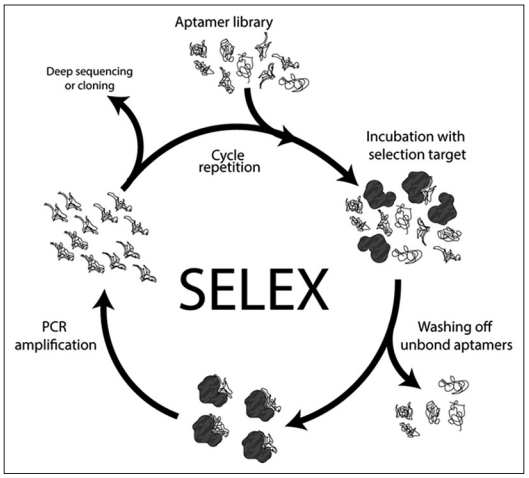

Figure 1: Flow chart of the aptamer selection process

the analysis of the selection results - HT-SELEX, comprising sequencing of hundreds of thousands or millions of individual sequences, makes it possible to decrease the number of the selection rounds from several dozens to 5-10, substantially accelerating generation of the results.[2]

Apart from faster selection process, deep sequencing of the aptamer libraries provides the means to overcome the major drawback of this method - the influence of bidirectional selection, causing the enrichment of the mixture, along with the highly affine aptamers, with the non-specific molecules (full-sized aptamers or primer dimers), whose selective advantage is increased amplification capacity rather than high affinity. With a greater number of the selection cycles, easily amplifiable sequences might outrank the highly affine ones, resulting in the total absence of aptamers matching the criteria of the researcher after simple sequencing of several clones following the selection. Moreover, highly affine sequences are also selected by their amplification capacity; therefore, even when there are no unwanted primer dimers, more preferable candidates may become less represented and be excluded from the resulting mixture by the less affine aptamers, which are easier to amplify. Two major factors facilitating this scenario are the excessive number of the amplification cycles following the selection and the excessive number of the selection rounds. The application of deep sequencing allows one to analyze the enrichment of multiple aptamers after several consecutive selection cycles, which enables the researcher to reveal poorly amplified variants, which, nevertheless, meet the necessary affinity requirements.

The analysis of HT-SELEX results is usually conducted as follows: Several (2-5) pre-selected pools of aptamers, which have been subject to several selection rounds, are sequenced, after which the enrichment of each identified sequence in the mixture is determined from the number of individual readings during sequencing.[3] Comparing the enrichment of the next pool with a certain sequence, as compared to the previous round, one can assess its ability for selective binding to the target [Figure 2].[4] Parallel sequencing of the aptamer libraries, conducted before each selection round, and the analysis of the enrichment of individual sequences as opposed to their enrichment after selection allows the researchers to determine, whether a certain sequence is highly affine or easily amplifiable. This approach, combined with *in silico* modeling of the spatial structure of the candidate aptamers,[5] allows to select the molecules which most likely possess the required qualities for further individual testing.

A specific feature of the aptamer selection process is an extremely high complexity of the initial libraries of random sequences. As a result, after several selection cycles and deep sequencing, the coverage might not be deep enough to provide statistically sufficient amount of readings of the individual sequences. In this case, a major proportion of the identified aptamers will appear only once in the sequencing results; furthermore, the enrichment of the same molecules in the results of the next selection round might be weakly correlated with the previous round or with the properties of the molecule due to statistical errors. Extensive amplification after each round also leads to the aptamer replication errors and creates similar yet non-identical molecules, which undergo the selection process in parallel.

Assuming that highly homologous aptamers either descend from a common precursor in the mixture or have similar affinity and are independently selected, the issue of an insufficient number of individual readings after the initial rounds of selection might be overcome, if similar sequences are clustered and changes in the enrichment of all the cluster members as a whole in the mixture are analyzed.[6] The most commonly used algorithm of sequence clustering is calculation of Levenshtein edit distance; in case, if its value is lower than the threshold, the compared sequences are clustered.[7] Aptamer clustering usually implements low thresholds, ranging from 1 to 5, which corresponds to the number of permissible substitutions, deletions, or insertions that constitute the difference between the aptamers. Clustering during the HT-SELEX analysis with the help of the FASTAptamer set of scripts[8] proceeds as following: First, the number of individual readings for each aptamer is determined, then the normalized value (reads per million [RPM]) is calculated from this number and from the total number of readings, and the aptamers are ranked by their RPM in descending order. Clustering starts from the first, the most enriched aptamer. It is consecutively compared to all the rest sequences whose RPM is above the threshold set by the operator. Levenshtein edit distance is calculated for each pair and the first cluster is formed. Then, the procedure is reiterated for the aptamer with the next highest RPM, which has not been included in the first cluster, and so on, until the required number of clusters is created or all aptamers are sorted.

According to the description, full clustering requires pair-wise comparisons of all the sequences; therefore, the number of necessary calculations increases as the square of their total amount. In case of the mixtures with a low number of clusters, each comprising a large number of members, the total amount of calculations turns out to be slightly smaller.

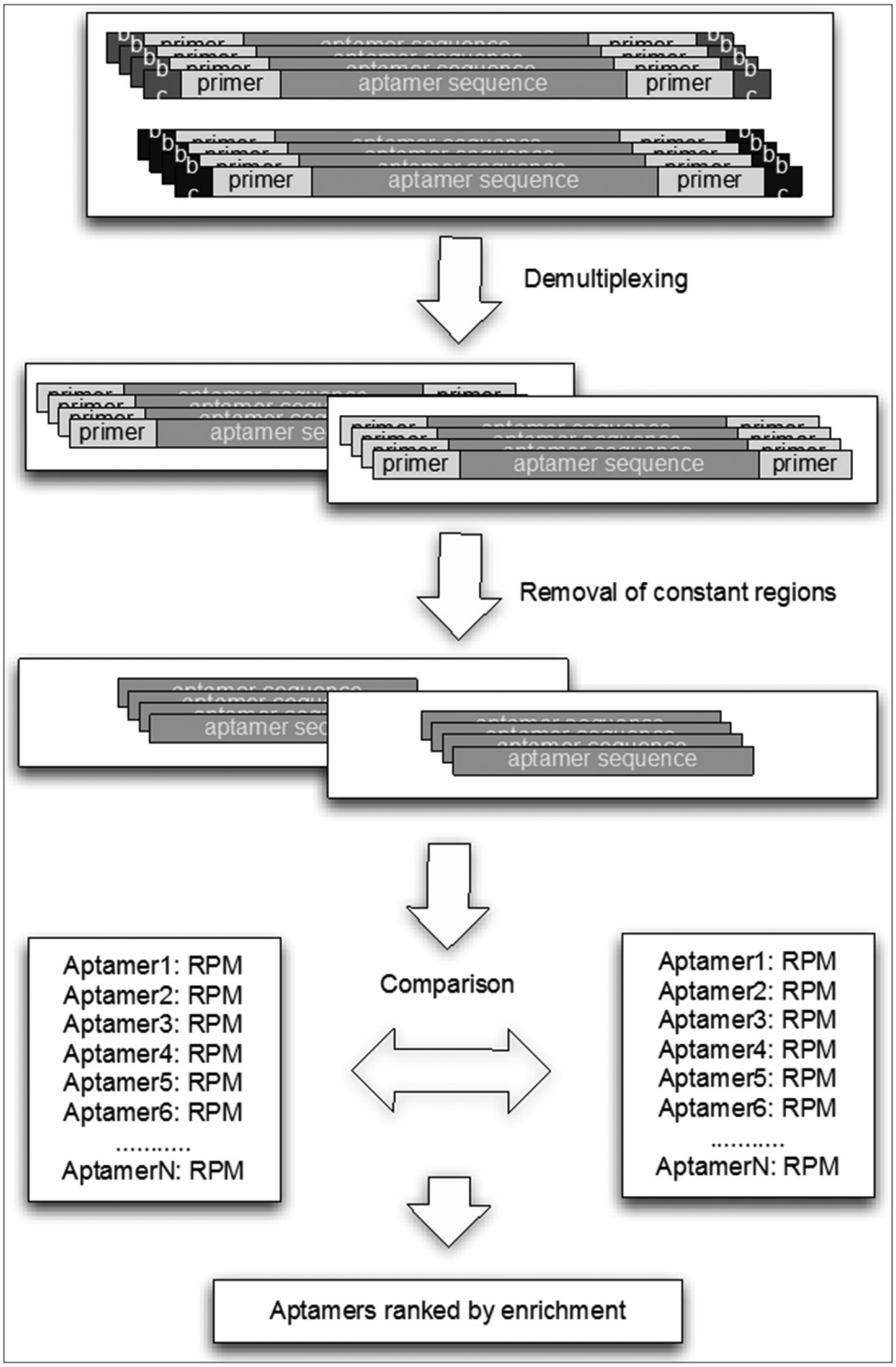

Figure 2: Flow chart of HT-systematic evolution of ligands by exponential enrichment analysis

In actual practice, full clustering may not be necessary, as the initial 100-1000 clusters usually prove to be enough for the analysis and search for the optimal candidates. Nevertheless, even this mode of clustering is often impossible due to the high demands for the computational capacity.

As Levenshtein edit distance is determined through the calculation of the whole matrix of two strings' prefixes, the final value can only be obtained after the end of the computation.[9] However, during the matrix calculation, the intermediate result increases, eventually forming the Levenshtein edit distance. As the aptamer clustering is usually conducted under the low threshold values of Levenshtein edit distance and most of the sequences of the processed set considerably differ, the intermediate Levenshtein edit distance starts to exceed the threshold as early as at the initial

steps of the full matrix calculation for most of the compared pairs of sequences. We suggest that simple filtering by Levenshtein edit distance instead of calculation of the precise value could massively reduce the computation time, allowing for more effective implementation of the aptamer clustering in the analysis of the HT-SELEX results [Figure 3].

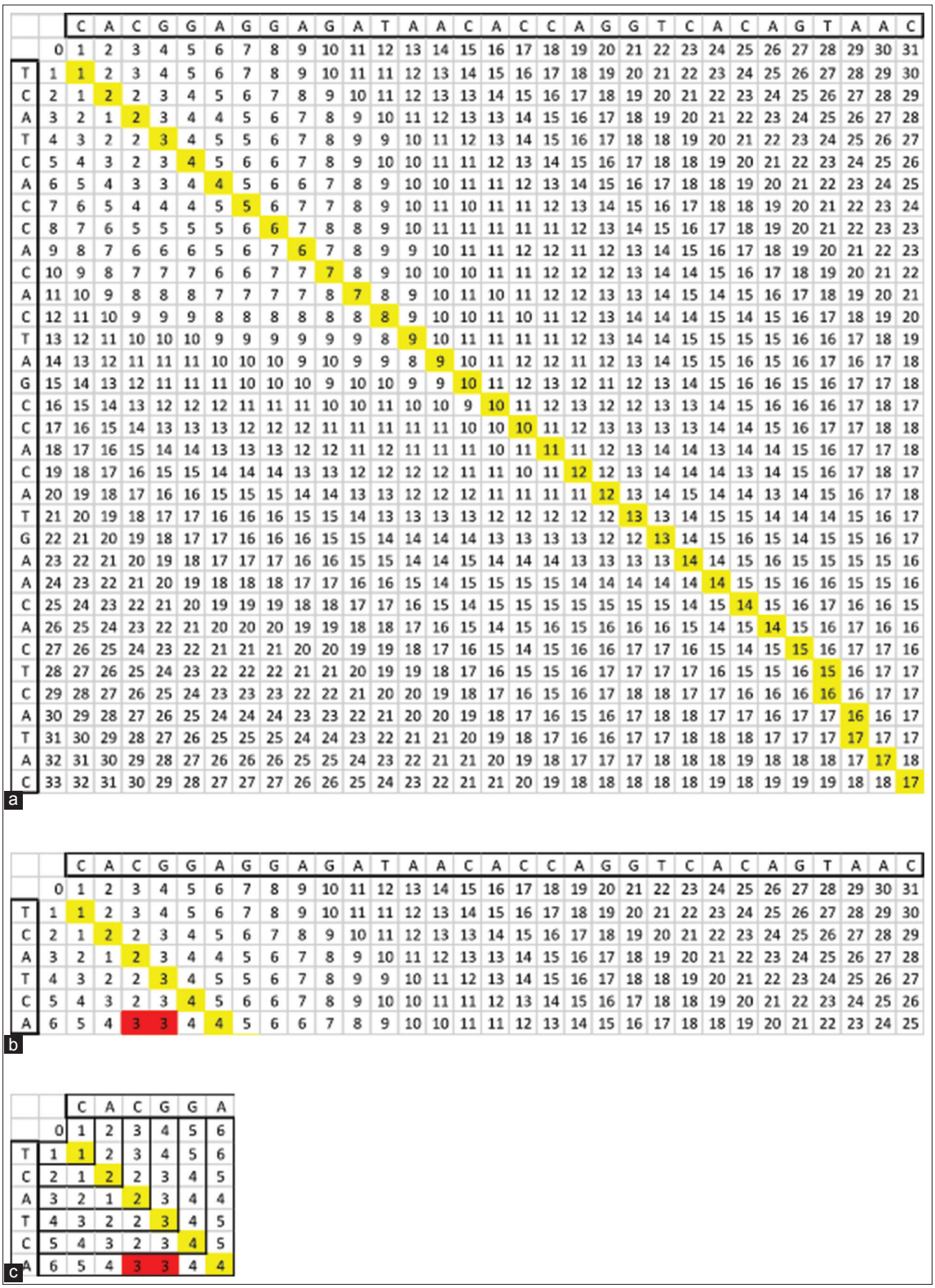

Figure 3: Application of different clustering algorithms for a pair of sequences, CACGGAGGAGATAACACCAGGTCACAGTAAC and TCATCACCACACTAGCCACATGAACACTCATAC (a) full matrix processing, 1023 calculations in total, (b) matrix processing terminates, as soon as the algorithm finds the string with a minimal value (3), exceeding the threshold (2). 186 calculations were required, (c) the matrix is processed layer by layer, starting from the top-left corner. The processing terminates, as soon as the algorithm finds the layer with a minimal value (3), exceeding the threshold (2). 36 calculations were required

MATERIALS AND METHODS

The sequencing results of the rounds 5, 6, and 7 of the SELEX procedure for the 40 nucleotide aptamers against the extracellular CD47 protein fragment immobilized on the Ni-NTA-coated magnetic beads (Cube Biotech) were used as the test data set for the comparative clustering. The selection was conducted on the sequences which contained the terminal primers PO dir (TAGGGAAGAGAAGGACATATGAT) and PO rev (TCAAGTGGTCATGTACTAGTCAA). In course of preparation for sequencing, the libraries were amplified with the elongated primers of seq dir (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNTAGGGAAGAGAAGGACATATGAT) and seq rev (CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCNNNNNNTCAAGTGGTCATGTACTAGTCAA) types with 6-nucleotide barcodes (designated as NNNNNN) on 5′- and 3′-terminals of the sequenced site. The raw data were uploaded to galaxy cloud cluster before clustering,[10] the multiplexed data were separated with the barcode splitter from FASTX-Toolkit, the constant sites were removed with the primer clip sequences script,[11] and the resulting FASTQ files were processed by the FASTAptamer-Count script to obtain the FASTA files containing only the unique sequences with the data on the number of readings and rpm. The standard FASTAptamer script (FASTAptamer-Cluster), which was installed as a standard galaxy tool with modifications allowing to set the maximum required number of clusters (FASTAptamer-Cluster_limit repository in galaxy test toolshed), was used as a control for clustering.

The modified fast-clustering script has been developed on the basis of FASTAptamer-Cluster and adapted for implementation as a galaxy tool, and it is located at http://toolshed.g2.bx.psu.edu/repos/hathkul/rapidcluster. The Levenshtein edit distance calculating algorithm was modified to terminate the calculations after achieving the threshold value more than 2, and an arbitrary edit distance, which is one more than the threshold value, was then assigned to the pair of sequences.

The set of 15,000 unique sequences was used to compare the calculation speed. Values ranging from 1 to 10 were used as threshold, and the script was used to form one cluster.

RESULTS

As Figure 4 shows, the optimized algorithm, implementing Levenshtein edit distance filtering instead of full calculation, allows to perform clustering in a substantially shorter time than the conventional FASTAptamer-count. Moreover, the speed of the modified algorithm considerably varies, depending on the threshold value.

The most common range of the threshold values (1-4) provides an average 8.4-fold increase in speed, allowing significant reduction of the total computational duration [Figure 4].

Rapid clustering of the HT-SELEX results opens new opportunities to use this method for the analysis of the initial rounds of selection when a majority of data comprises unique sequences. Merging the homologous readings into clusters might result in more precise ranking and help to determine the rank of each particular cluster. Thereupon, the properties of the sequences from the most promising clusters can be compared to determine the final candidates for individual testing.

For this purpose, we have developed two FASTAptamer-Compatible modified scripts, allowing to compare the clusters of sequences.

The first script, seed select, uses the clustering data obtained from either FASTAptamer-Cluster or previously described modified Rapidcluster script as an input. All the sequences other than the initial cluster seeds are removed from the clustering results; the numbers of readings and RPMs for all sequences of the cluster are added together and put in the cluster seeds descriptor. The total number of the cluster members is also used as an input.

The second script, cluster compare, allows to compare the cluster enrichment in the selection results after a different number of rounds or in the selection results and in the control group. As the most abundant sequences are chosen as cluster seeds during the primary clustering, the homologous clusters can be formed on the basis of different seeds. It restricts a direct comparison of the sequence features, implemented in the FASTAptamer-Compare script. We have also developed a modified script which also uses the Levenshtein edit distance calculation to compare the cluster parameters. For each sequence from the seed select output for the selection results, the edit distance to each sequence from the seed select output for the control

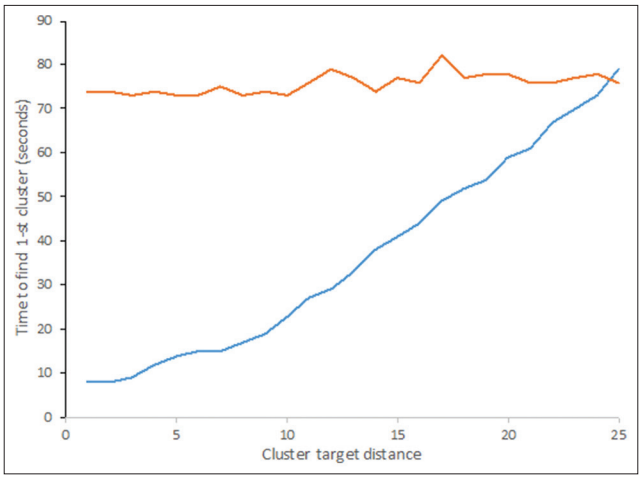

Figure 4: Speed of the first cluster calculation through FASTAptamer-Cluster (orange) and rapidcluster (blue), depending on the preset Levenshtein edit distance

sample. The threshold is set at the value which exceeds the threshold edit distance used for the initial clustering by two. All clusters form the control group, satisfying this condition, are clustered around the seed with the highest RPM, and the sum of their values is used for comparing with the cluster of the selection results. In case, if more than one cluster meets the criteria, an additional round of searching for homologous clusters is run on the selection results, and their RPM values are added to the values of the initial seed.

These results allow the researchers to compare the enrichment of the whole clusters in different rounds of selection or before and after a certain round and to determine the ones with the most suitable features [Figure 5].

DISCUSSION

The application of HT-SELEX allows to obtain better control over the high-affinity aptamers selection process and avoid losses of the target sequences which are underrepresented in the terminal selective pool. However, this method also has some blind spots, such as the lack of opportunity to precisely determine the most suitable number of the selection rounds. An excessive number can result in waste of time and resources as well as reduction in the abundance of rare and potentially promising aptamer sequences or contamination of the results with the primer dimers. An insufficient number of cycles results in a prevalence of the unique sequences which cannot be analyzed with traditional methods.

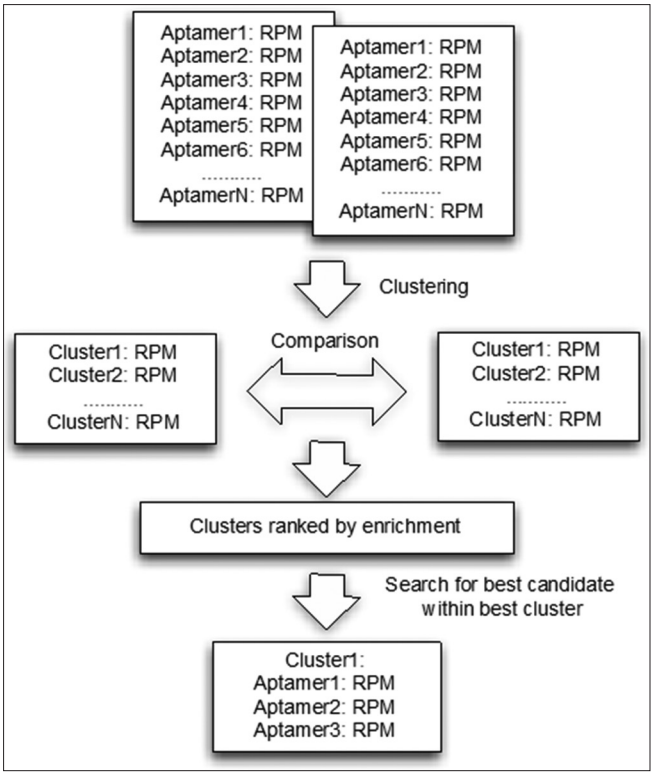

Figure 5: Flow chart of the HT-systematic evolution of ligands by exponential enrichment cluster analysis

The developed tools for rapid clustering of the results of the aptamer pool sequencing and further analysis of the enrichment of the sequence clusters allow to obtain the relevant data from the samples which have not been subject to a sufficient number of the selection cycles and to determine the promising sequences without additional selection and resequencing.

All the developed scripts can be used in galaxy cloud,[12] which can be crucial for clustering, as an elastic cloud platform properly provides the computational capacity, necessary for clustering, without installation and adjustment of the software on a local computer.

CONCLUSIONS

We have developed a set of programs for rapid clustering of the homologous sequences from the results of the HT-SELEX deep sequencing and subsequent comparison of the individual samples in terms of the cluster enrichment. The software allows to conduct the analysis of under selected SELEX results, which cannot be analyzed by other means.

ACKNOWLEDGMENTS

The work was funded by MESR contract No. 14.604.21.0110 (RFMEFI60414X0110).

REFERENCES

1. Zhu G, Ye M, Donovan MJ, Song E, Zhao Z, Tan W. Nucleic acid aptamers: An emerging frontier in cancer therapy. Chem Commun (Camb) 2012;48:10472-80.
2. Beier R, Boschke E, Labudde D. New strategies for evaluation and analysis of SELEX experiments. Biomed Res Int 2014;2014:849743.
3. Blind M, Blank M. Aptamer selection technology and recent advances. Mol Ther Nucleic Acids 2015;4:e223.
4. Dupont DM, Larsen N, Jensen JK, Andreasen PA, Kjems J. Characterisation of aptamer-target interactions by branched selection and high-throughput sequencing of SELEX pools. Nucleic Acids Res 2015;43:e139.
5. Caroli J, Taccioli C, De La Fuente A, Serafini P, Bicciato S. APTANI: A computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. Bioinformatics 2016;32:161-4.
6. Hoinka J, Berezhnoy A, Dao P, Sauna ZE, Gilboa E, Przytycka TM. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. Nucleic Acids Res 2015;43:5699-707.
7. Hoinka J, Berezhnoy A, Sauna ZE, Gilboa E, Przytycka TM. AptaCluster - A method to cluster HT-SELEX aptamer pools and lessons from its application. Res Comput Mol Biol 2014;8394:115-28.

8. Alam KK, Chang JL, Burke DH. FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. Mol Ther Nucleic Acids 2015;4:e230.
9. Navarro G. A guided tour to approximate string matching. ACM Comput Surv 2001;33:31-88.
10. Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, *et al.* Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol 2011;29:972-4.
11. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422-3.
12. Thiel WH, Giangrande PH. Analyzing HT-SELEX data with the Galaxy Project tools - A web based bioinformatics platform for biomedical research. Methods 2016;97:3-10.

Source of Support: Nil. **Conflict of Interest:** None declared.