Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 269–275

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Using of bioinformatic software methods for determining PAM sequences by analyzing of CRISPR sites

Elena Maslovskaya[a*], Dmitriy Sosin[b], Aleksey Frolov[a], Ivan Pavlenko[a], Vladislav Zhadaev[a], Mihail Lebedev[a], Yaroslav Leonov[a], Nikita Ivashchenko[a], Viktoriya Sapozhnikova[a] and Yulia Shaltaeva[a]

[a]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute, Russia
[b]Obninsk Institute for Nuclear Power Engineering, Russia

Abstract

Currently, CRISPR/Cas technology is one of the most actively developing in molecular biology as an instrument for genomic DNA editing and controlling gene activity at various levels. For today this technology has a number of limitations that makes it difficult to use. One of the main limitation is the conditions under which the Cas type proteins or their analogues cut DNA: 1) the complementary between sequence of protospacer and guiding RNA; 2)the special sequences flanking this protospacer on one or both ends. This special sequences usually unique for each enzyme and named PAM sites. However, if the ability to change guiding RNA leads to the emergence of an accurate genomic editing tool, the presence of an immutable PAM site significantly limits the areas of the genome in which editing can be performed. At the moment, more than 3.5 thousand of Cas proteins using different PAM sites have been identified in various bacteria with the help of bioinfomatic tools, but in practice less than ten PAM are used. To be able to use new proteins, it is necessary to obtain their characteristics. In particular the optimal physical and chemical conditions for enzyme, the structure of guiding RNA and PAM-sequence, characteristic for new protein. Currently, several programs have been created to determine PAM sequences for previously unexplored Cas protein analogues, but all of them prone to mistakes and require data confirmation from classical molecular biology.

The purpose of this work was to create a new efficient algorithm and software for idetification of PAM sequences characteristic of any microorganism. To test the program, we used CRISPR bacterial cassettes, formed by Cas9 enzyme, as well as current homology search programs for the bacterial genomes.

As a result of the study, a modular program was created that allows identifying the sequence of the PAM site based on the analysis of CRISPR cassettes with an efficiency not inferior to world analogues.

* Corresponding author. Tel.: +7-926-601-95-40/
E-mail address: ev_maslovskaya@1511.ru

1877-0509
10.1016/j.procs.2020.02.176

Keywords: bionformatics, CRISPR, CAS9, PAM

Introduction.

The CRISPR-Cas bacteria “adaptive immunity” system (clustered regularly alternating short palindromic repeats and CRISPR-associated proteins) was discovered in 1987 [1]. However, the immune function of CRISPR systems was only established in 2005 [2-4]. The system is a prokaryotic analogue of the vertebrate immune system, allowing to protect single prokaryotic cells from destruction by phages. It is known that this system works in the cells of 90% of archaea and 50% of bacteria [5].

Crispr cassettes and Cas proteins are the main components of the CRISPR system [6]. Each functional cassette contains three types of elements: lead sequences, spacers, and repeats. Repeats within one cassette, as a rule, identical to each other in sequence and length, less often — may differ in one or two, often terminal, nucleotides. Spacers are unrelated, repetitive, short sequences located between repeats and originating from fragments of alien genetic elements that enter the prokaryotic cell and are called protospacers [7]. Длина спейсеров внутри кассеты приблизительно равна длине повторов. Набор спейсеров у штаммов одного и того же вида обычно очень разный [8].

When a virus penetrates a bacterium or archaea equipped with a CRISPR system, the adaptive functional module of the system is activated: specific Cas proteins cut protospacers from the foreign genome. Proteins select sites near a particular sequence of pam (protospacer adjacent motif) - only a few nucleotides that have been identified near one end of protospacers but are not the same for different CRISPR systems [9,10]. These same adaptive proteins then embed the fragment in the CRISPR cassette in the leader sequence. So a new spacer is formed, and along with it — and a new repeat. The whole process is called adaptation, or acquisition, and in fact it is — remembering the virus. Information about all of the memorable viruses receive when they tick all offspring cells. Thus, the mechanism of bacterial immunity is to form a database of viruses that the bacterium has previously encountered. If necessary, the system "gets" a fragment from the database and reads the RNA copy, which forms a ribonucleic complex, joining the Cas9 protein. In turn, the complex binds to the foreign target gene, which corresponds to the original fragment, and the Cas9 protein conducts cutting of the foreign DNA chain [11].

The big breakthrough in crispr biology came with the realization that Cas9 can be reprogrammed to cleave not only viral DNA but also other DNA sequences by altering the filament of the guide RNA associated with Cas9. The enzyme is able to remove any unwanted DNA fragment and leave in its place a specially created fragment [12]. Thus, the technology can be actively used in molecular biotechnology, but it has its limitations. The RAM motif is strictly specific to each protein. In places of combination of these nucleotides editing capabilities will be limited.

The definition of frame motifs can expand the use of alternative Cas proteins, which will greatly increase the editing capabilities.

Currently, a relatively new direction of research is widely known, which uses mathematical and algorithmic methods to solve molecular biological problems. This direction was called bioinformatics [13].

Genome research using bioinformatics programs opens a new path in the study of molecular and search for solutions to many medical problems.

The aim of the study was to develop an application to determine the pam sequences of Cas effectors, to determine the General structure of the pam sequence search algorithm, to write software modules for the implementation of individual steps of the algorithm.

Materials and methods.

The material was 17 genome sequences of bacteria of the genus Streptococcus from the international online library NCBI. Phage are searched through BLAST, in which the program, using the application programming interface, finds all bacteriophages and viruses containing a spacer. Using database queries, the program receives all the DNA of the necessary viruses found using BLAST [14]. Weblogo libraries that generate sequence logos were used to visualize the output. The full algorithm of the program is presented as a plan:

1. A CRISPR cassette is fed to the program input.
2. The program divides it into spacers and repeats.

3. Using BLAST, the program with usage the API finds all bacteriophages and viruses containing one of the spacers.

4. Using queries to the NCBI databases, the program receives all the DNA of the required viruses found using BLAST.

5. The program processes all DNA sequences and finds all parts which are next to protospacers that are complementary or equal to spacers

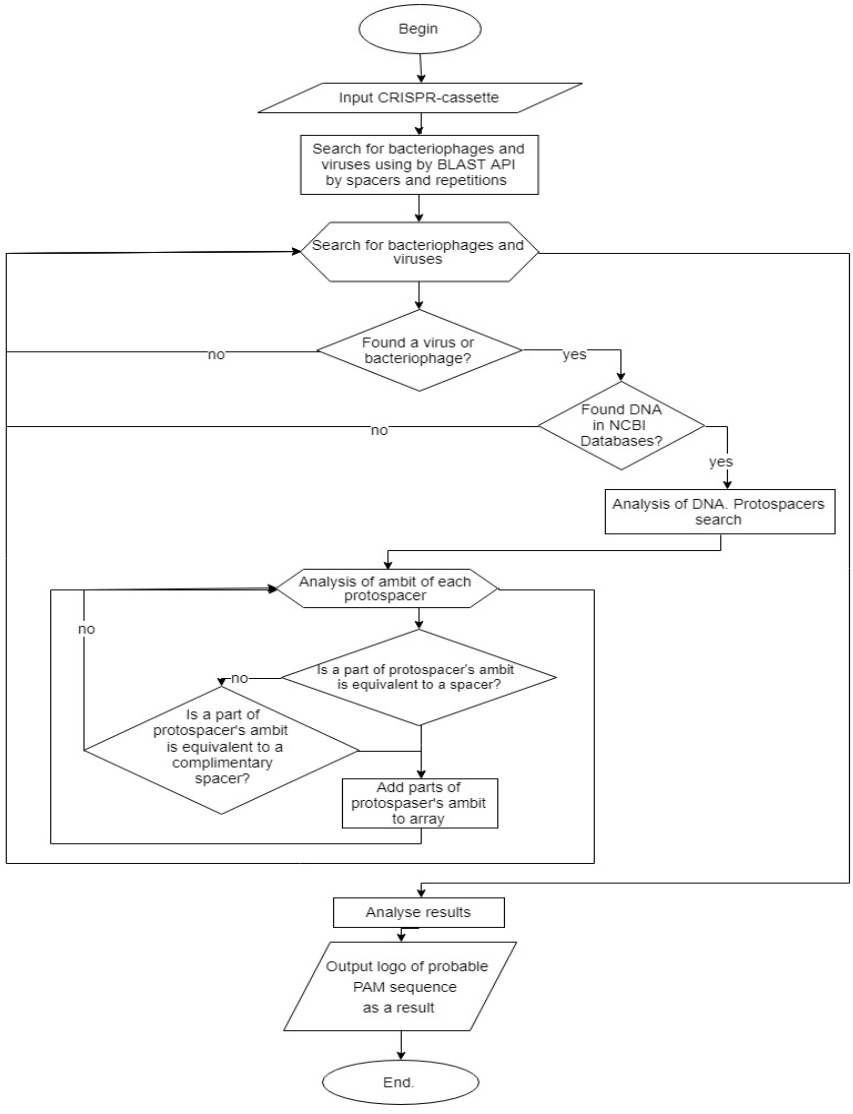
6. The sections obtained in the previous paragraph are sorted and the probability of hitting all adjacent sequences adjacent to the spacer is calculated.

7. A pattern is derived from these nucleotides and possible PAM sequences are determined.

8. Align of output sequences

9. Creating the logo of the PAM sequence

10. Output logo as a result

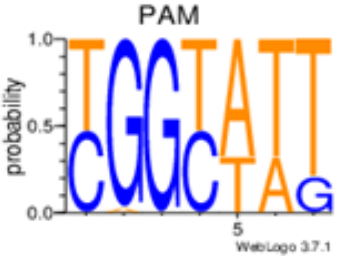
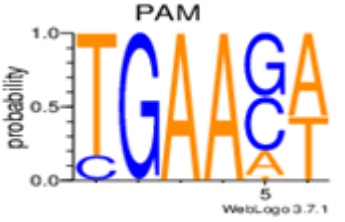
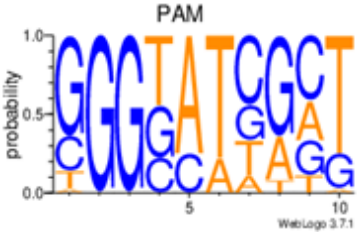
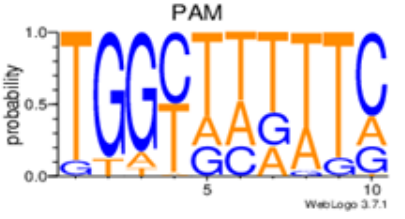
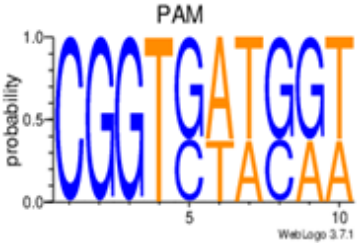
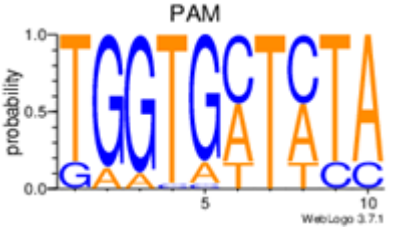


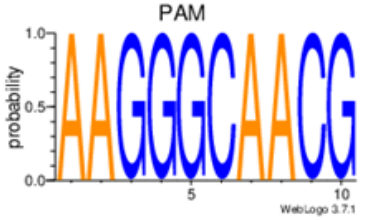
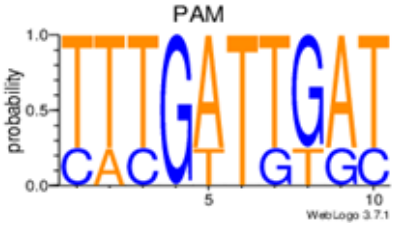
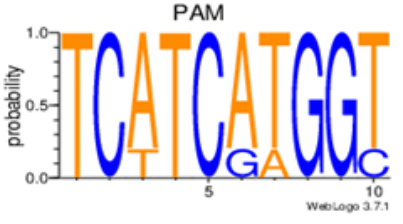
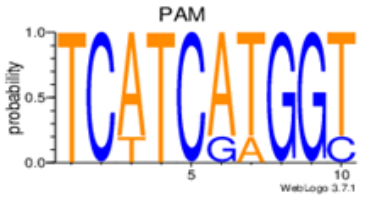
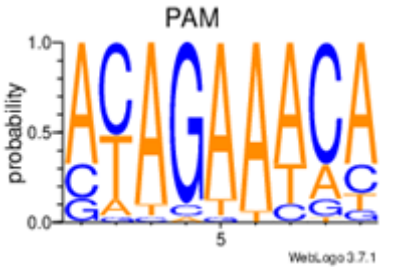
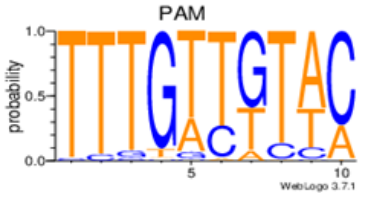
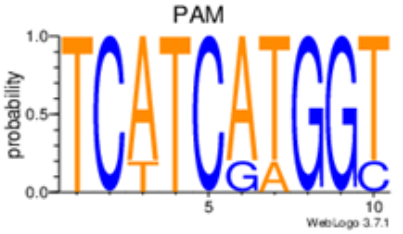
Flowchart 1. Detailed algorithm of program

Results.

In most cases, the program we developed reveals the correct PAM sequence, or close to it.

In the process of processing the request, the program prepares a graphical representation of the logo of the WebLogo sequence, an example is shown in Table. 1.

I) Genus:Streptococcus, Species: pyogenes					
№	Cassette/Result	PAM	№	Cassette/Result	PAM
1	A20 (NC_018936_1) 	NGG	4	GCF_001635975(NZ_CP013840_4) 	NGG
2	GCF_000767505 (NZ_CP007240_1) 	NGG	5	M1 GAS (NC_002737_1) 	NGG
3	GCF_001635975 (NZ_CP013840_2) 	NGG	6	NZ131 (NC_011375_1) 	NGG
II) Genus:Streptococcus, Species: mutans					
№	Cassette/Result	PAM	№	Cassette/Result	PAM

7	UA159-FR GCF_000817065 (NZ_CP007016_1) 	NGG	8	GS-5 (NC_018089_1) 	NGG
III) Genus:Streptococcus, Species: thermophilus					
№	Cassette/Result	PAM	№	Cassette/Result	PAM
9	LMD-9 (NC_008532_2) 	NAAAAW	14	GCF_000971665(NZ_CP013939_2) 	NGGNG
10	LMD-9 (NC_008532_4) 	NGGNG	15	ASCC 1275 GCF_000698885 (NZ_CP013939_2) 	NNAGAAW
11	GCF_001663795 (NZ_CP016026_2) 	NNAGAAW	16	ASCC 1275 GCF_000698885 (NZ_CP006819_5) 	NNAGAAW

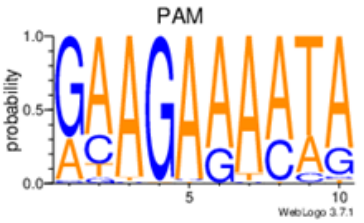
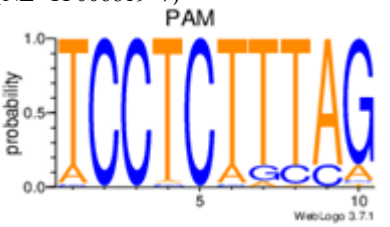
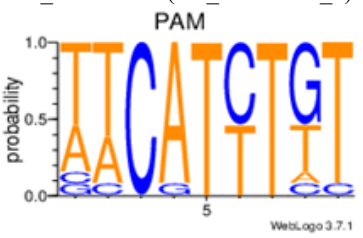
1 2	GCF_001663795 (NZ_CP016026_8) 	NNAAAW	17	ASCC 1275 GCF_000698885 (NZ CP006819 7) 	NGGNG
1 3	GCF_001514435 (NZ_CP013939_2) 	NNAGAAW			

Table 1. Results of program testing

Conclusion.

The program allows you to automate the search, so it will be useful for biologists and geneticists without specialized training to work with databases. It is planned to improve the developed program in the following components of the work:

Improving the performance when searching for sequences, by solving the problem of accelerated loading of large genomes from the site (at the moment, loading genomes is the largest part of the program execution time.

The program will be modified to detect "sophisticated" cassettes of different organisms, as the first version of the program was initially tested on more unified cassettes Streptococcus pyogenes. In future versions of the program repeats in cassettes should be the same length, and differ by no more than 40%. It is worth paying attention to the fact that the results of our tests, even taking into account the error satisfy the literature data [15, 16].

References

[1] Ishino Y., Shinagawa H., Makino K., Amemura M., Nakata A. (1988) "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product" J Bacteriol. **169(12)**:5429-5433.

[2] Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. (2005) "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin". Microbiology **151(8):**2551-2561.

[3] Pourcel C., Salvignol G., Vergnaud G. (2005) "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies". Microbiology **151(3)**:653-663

[4] Mojica F. J. M., Díez-Villaseñor C., García-Martínez J., Almendros C. (2005) "Short motif sequences determine the targets of the prokaryotic CRISPR defence system". Microbiology **155(3)**: 733–740.

[5] De Dong., Minghui Guo., Sihan Wang., Yuwei Zhu., Shuo Wang., Zhi Xiong., Jianzheng Yang., Zengliang Xu & Zhiwei Huang.(2017) "Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein". Nature **546:** 436–439.

[6] Barrangou R., J. van der Oost (Eds.) "CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea, Springer Press, Heidelburg (2013), pp. 1-129

[7] Shah SA, Erdmann S, Mojica FJ, Garrett RA. (2013) "Protospacer recognition motifs: mixed identities and functional diversity." RNA Biol.**10(5)**:891-899.

[8] Sorek R, Kunin V, Hugenholtz P (2008) "CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea" Nature Reviews Microbiology **6(3)**: 181

[9] Mojica F. J. M., Díez-Villaseñor C., García-Martínez J., Almendros C. (2005) "Short motif sequences determine the targets of the prokaryotic

CRISPR defence system". Microbiology **155(3):** 733–740.

[10] Leenay RT, Beisel CL. Deciphering (2017) "Communicating, and Engineering the CRISPR PAM" J Mol Biol. **429(2):** 177-191.

[11] Jiang F, Doudna JA (2017) "CRISPR-Cas9 Structures and Mechanisms" Annu Rev Biophys **46:** 505-529.

[12] Palermo G., Clarisse G. Ricci, J. Andrew McCammon, Joseph E. (2019) "The invisible dance of CRISPR-Cas9. Simulations unveil the molecular side of the gene-editing revolution" Phys Today 72(4): 30–36.

[13] Khlebnikov V. V., Averkov V. A. (2008) "Development of software for bioinformatics research" Tomsk State University Journal 13(1):113.

[14] Johnson, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL.(2008) "NCBI BLAST: a better web interface." Nucleic Acids Res. 36 : 5-9.

[15]Mülle M., Lee C.M., Gasiunas G., Davis T. H., CradickT.J., Siksnys V., Bao G., Cathomen T. (2016) "Mussolino C.Streptococcus thermophilus CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome" Mol Ther. 24(3): 636–644.

[16] Chatterjee P., Jakimo N. , Jacobson J. M. (2018) "Minimal PAM specificity of a highly similar SpCas9 ortholog" Sci Adv. 4(10).