UPGMA - analysis of type II CRISPR RNA-guided endonuclease Cas9 homologues from the compost metagenome

Andrei A. Zimin[1*], *Alexandra* N. Karmanova[1,2] and *Yinhua* Lu [3]

[1]G.K. Scriabin Institute of Biochemistry and Physiology of Microorganisms RAS, Pushchino, Russia
[2]Pushchino State Institute of Natural Science, Pushchino, Russia
[3]College of Life Sciences, Shanghai Normal University, Shanghai, China

Abstract. Metagenomic approaches provide access to the genetic diversity of the environment for biotechnological applications, allowing the discovery of new enzymes and new pathways for numerous catalytic processes. Five new putative type II CRISPR-Cas9 DNA endonucleases were identified from the compost community using the DELTA-BLAST algorithm. It was determined using phylogenetic UPGMA analysis that four of these potential enzymes are similar to those of the Bacteroidetes. Protein structural modeling confirmed the data of DELTA-BLAST and UPGMA analysis. These new five proteins found may be promising for genome editing in termoresistant *Actinomyces*.

1 Introduction

Extreme environmental conditions such as hot springs, deep-sea hydrothermal vents and organic composts are reservoirs of unique microbial diversity, providing the potential for the release of new enzymes with desirable properties. The adaptation of microbial communities to these environmental conditions explains their high genomic and metabolic flexibility, and they often encode enzymes with novel properties suitable for many applications [1].

The aim of this work was to search for homologues of CRISPR-Cas9 DNA nuclease from the compost metagenome. Such homologues may be interesting for the development of a system for editing the genes of various bacteria inhabiting this artificial biotope. These enzymes must be thermotolerant, because temperatures during the incubation of compost rise to 90 degrees Celsius or more. Thermotolerant enzymes can also be used to edit the genome of bacteria isolated from other extreme biotopes. An additional bonus of using such sequences can be the use of a thermostable in vitro DNA editing system. An interesting fundamental study of the found TR (thermoresistant) homologues of type II CRISPR-Cas9 DNA endonucleases can be a structural study of these enzymes for the subsequent production of biotechnologically significant mutants based on amino acid sequences extracted from the compost metagenome.

* Corresponding author: dr.zimin8@yandex.ru

2 Materials and methods

The main details of the methodological approaches were published by us earlier [1]. The reference point for the search for homologues was the amino acid sequence of the second type II CRISPR-Cas9 gene product. The search for homologues in the compost metagenome (env_nr, taxid: 702656 [3-4] was performed using the DELTA-BLAST algorithm with the following change in the default parameters: gap costs: existence: 9 extension: 1. Sequences MNK39233.1, MNQ43276.1, MNF63500.1, MNX56837.1, and MNU15097.1 were found with statistical confidence. Sequences of CRISPR-Cas9 homologues from various bacterial taxa were taken as controls. Phylogenetic analysis was performed using the UPGMA algorithm with 2000 repetitions of the bootstrap statistical analysis [7] in the MEGAX software package [5].

Building the model by SWISS-MODEL [11]. Templates searches were done using BLAST [9] and HHblits [10] against the SWISS-MODEL Matrix Library (SMTL [12], last updated: 2021-03-03, last included PDB release: 2021-02-26). The model was built by aligning the target sequence with the template using the ProMod3 software tool [13].

3 Results

Amino acid sequences found using DELTA-BLAST in the metagenome of the compost from the Experimental Botanical Garden Goettingen, Germany [3-4] were used for phylogenetic analysis (Fig.1).

The result of the taxonomic identification of homologues was: MNK39233.1 - clustered with enzymes from *Chryseobacterium*; MNQ43276.1, MNF63500.1, MNX56837.1 - with enzymes of the taxa *Cytophagales* and *Flavobacterium*.

The evolutionary relationships of the homologue MNU15097.1 remain unclear. It is interesting that attempts to compare it with databases in GenBank also did not give an unambiguous result. It is possible that a new rare sequence of the second type CRISPR-Cas9 endonuclease has been found.

Building the MNK39233.1 model by SWISS-MODEL: homology modelling of protein structures [11]. The search for patterns in the SMTL database [12] for structural modeling confirmed the data of DELTA-BLAST and UPGMA analysis.

Protein structural modeling was performed for all five found sequences: MNK39233.1, MNQ43276.1, MNF63500.1, MNX56837.1, and MNU15097.1 (Tab.1). As an example, we present the best result obtained for the MNK39233.1 sequence. A total of 19 templates for MNK39233.1 found. Model on the structure from SMTL DB ID: 6jdv.1 (Crystal structure of Nme1Cas9 in complex with sgRNA and target DNA (ATATGATT PAM) in catalytic state) [14] for sequence MNK39233.1 was found (Tab.1). The model was obtained by superimposing the studied sequence on the structure of the protein ID: 6jdv.1 from SMTL (Fig.2).

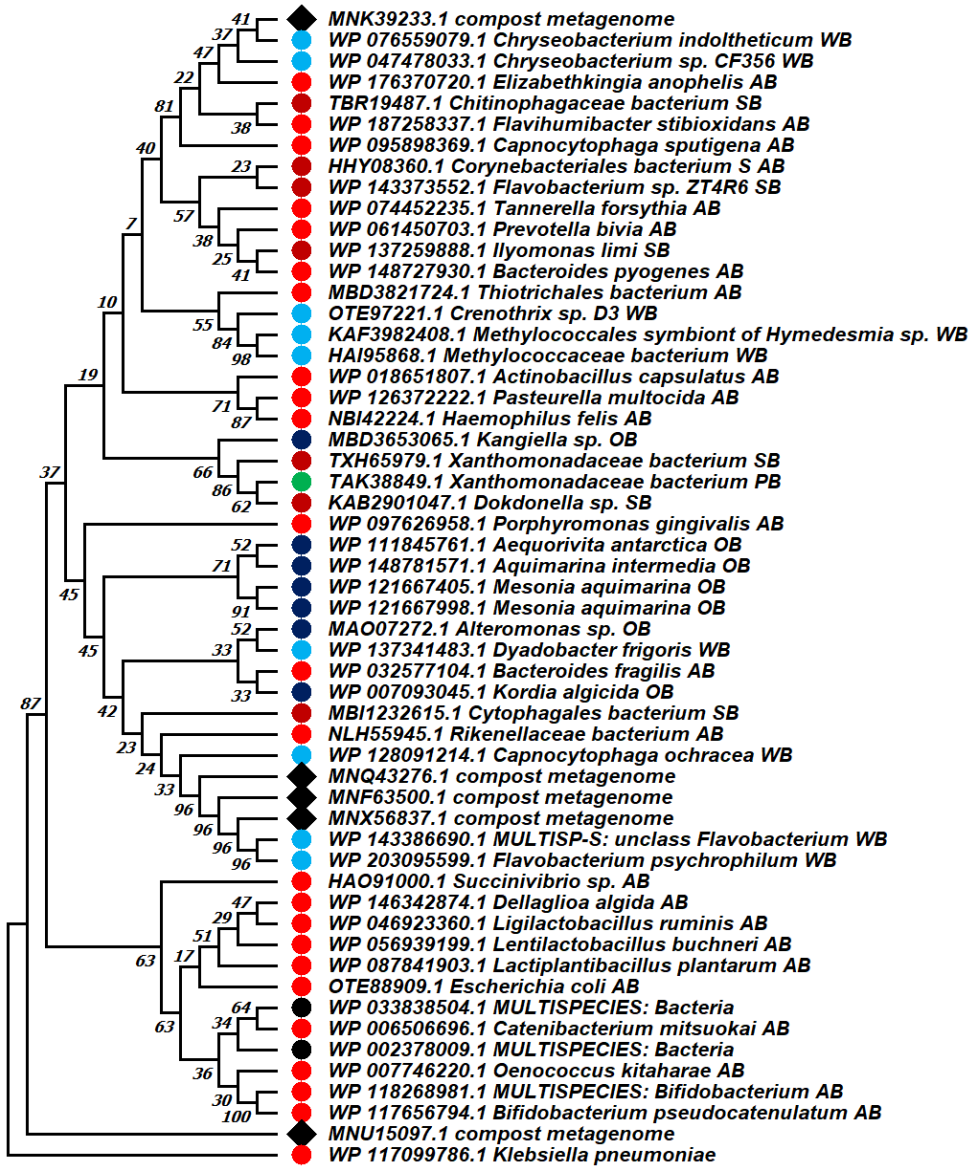

Fig. 1. Phylogeny of amino acid sequences of type II CRISPR-Cas9 homologues - endonucleases from the compost metagenome. Evolutionary distances were calculated using the method based on the JTT matrix [8] in terms of the number of amino acid substitutions per site. This analysis involved 55 amino acid sequences. Black diamonds denote sequences from the compost metagenome. The colors of the circles indicate different source biotopes of sequences: green - plants, brown - soil, light blue - fresh water, dark blue - ocean, red - animals, black - taxon *Bacteria.*

Table 1. The best results of template search for MNK39233.1 sequence from the compost metagenome in the SWISS-MODEL Matrix Library (SMTL).

Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Coverage	Description
6je4.1.A	23.18	monomer	-	HHblits	X-ray	3.07Å	0.31	0.77	CRISPR-associated endonuclease Cas9
6jdq.1.A	23.42	monomer	-	HHblits	X-ray	2.95Å	0.32	0.77	CRISPR-associated endonuclease Cas9
6jdv.1.A	23.27	monomer	-	HHblits	X-ray	3.10Å	0.31	0.77	CRISPR-associated endonuclease Cas9
6je9.1.A	23.42	monomer	-	HHblits	X-ray	3.46Å	0.32	0.77	CRISPR-associated endonuclease Cas9
6kc8.1.A	23.45	monomer	-	HHblits	X-ray	2.90Å	0.32	0.76	CRISPR-associated endonuclease Cas9
6je9.1.C	23.42	monomer	-	HHblits	X-ray	3.46Å	0.32	0.77	CRISPR-associated endonuclease Cas9
6kc7.1.A	23.27	monomer	-	HHblits	X-ray	3.30Å	0.31	0.77	CRISPR-associated endonuclease Cas9

4 Findings

Five new putative type II CRISPR-Cas9 endonucleases were identified from the compost microbial community using the DELTA-BLAST algorithm. It was determined using phylogenetic UPGMA analysis that four of these potential enzymes are similar to those of the *Bacteroidetes* from the taxa *Cytophagales*, *Chryseobacterium* and *Flavobacterium*. Protein structural modeling confirmed the data of DELTA-BLAST and UPGMA analysis. Structural modeling of proteins showed that potential secondary structures similar to the known endonuclease Cas9 are distributed along the entire length of the paired sequence alignment. This distribution is typical for all studied paired elements. This indicates the possibility of using the data of the found sequences for the synthesis of artificial genes and for obtaining real endonucleases for their experimental verification and subsequent use. The potential thermal stability may determine the scope of application of these new proteins - Cas9 homologues in practice. These new five proteins found may be promising for genome editing in thermoresistant microbe.

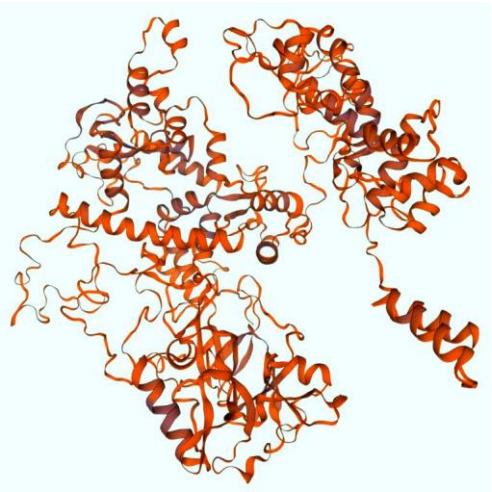

Fig. 2. Protein structural model for the MNK39233.1 sequence from the compost metagenome. The structures with the highest similarity in pair structural alignment are shown in purple.

The study was funded by RFBR and NSFC, project number 20-54-53018.

References

1. C. Wang, D. Dong, H. Wang, K. Müller, Y. Qin, H. Wang, W. Wu, Biotechnol Biofuels, **9**, 22 (2016)
2. A.N. Karmanova, A.A. Zimin, *J. of Physics: Conference Series*, **1701** (2020) ,
3. NCBI:taxid702656:https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=702656
4. R. Egelkamp, T. Zimmermann, D. Schneider, R. Hertel, R. Daniel, Frontiers in Environmental Science, **7,** 103 (2019)
5. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, Molecular Biology and Evolution, **35** .(2018).
6. R. C. Edgar, Nucleic Acids Research**, 32(5)**, 1792-1797 (2004)
7. Felsenstein J. *Evolution* **39**, 783-791 (1985)
8. Jones D.T., Taylor W.R., Thornton J.M. Computer Applications in the Biosciences **8,** 275-282 (1992).
9. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BMC Bioinformatics, 10, 421-430 (2009).
10. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S.J. Haunsberger, J. Söding, BMC Bioinformatics, 20, 473 (2019).
11. A.Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F.T. Heer, T.A.P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, Nucleic Acids Res., **46(W1)**, W296-W303 (2018).
12. S. Bienert, A. Waterhouse, T.A.P. de Beer, G., Tauriello, G. Studer, L. Bordoli, T. Schwede, Nucleic Acids Res., **45**, D313-D319 (2017).
13. G. Studer, G. Tauriello, S. Bienert, M. Biasini, N. Johner, T. Schwede, PLOS Comp. Biol., **17(1)**, e1008667 (2021).
14. W. Sun, J. Yang, Z. Cheng, N. Amrani, C. Liu, K. Wang, R. Ibraheim, A. Edraki, X. Huang, M. Wang, J. Wang, L. Liu, G. Sheng, Y. Yang, J. Lou, E.J. Sontheimer, Y. Wang, Mol. Cell., **76(6),** 938-952.e5 (2019)