The compost metagenome as a source of T4 bacteriophage pyrimidine dimer glycosylase homologues

Alexandra N. Karmanova^{1.2,*}, Yinhua Lu³, and Andrei A. Zimin¹

¹G.K. Scriabin Institute of Biochemistry and Physiology of Microorganisms RAS, Pushchino, Russia ²Pushchino State Institute of Natural Science, Pushchino, Russia ³Callage of Life Spinnee, Shenghei Nermal University, Shenghei, Ching

³College of Life Sciences, Shanghai Normal University, Shanghai, China

Abstract. Compost is a promising source of thermotolerant enzymes for their application in biotechnology. Homologues of bacteriophage T4 DNA glycosylase can find their application in pharmaceuticals and perfumery. Five homologues of glycosylase of pyrimidine dimers of bacteriophage T4, a product of the *denV* gene, were found by comparing using the DELTA-BLAST algorithm with the compost metagenome proteins. Phylogenetic analysis of the found sequences of enzyme homologues was carried out using the Maximum Likelihood algorithm in the MegaX software package. Thus, an interesting spectrum of promising proteins, homologues of the repair enzyme, DNA glycosylase of pyrimidine dimers of bacteriophage T4, was found. After structural modeling, they can be tested for their thermal stability and tested as a basis for therapeutic and prophylactic drugs.

1 Introduction

Many enterprises and farms, whose activities involve the presence of biowaste, use various methods of composting to process certain types of organic matter. This is justified both by the ecological purity and safety of the method, and by its economic value. Composting implies a process of aerobic decomposition of waste as a result of the vital activity of microorganisms to water, carbon dioxide, heat and the final product - compost [1]. The latter, as a rule, is subsequently used as an organic fertilizer for the soil. In addition, it can act as a source of microorganisms with enzymes, which can be further used in various industries. The most promising is the selection of biochemical catalysts for the production of biofuels, the processing of cellulose, plastic and other hard-to-decompose substances, as well as the search for biomolecules that can form the basis of pharmaceuticals.

One of the methods that will help to more clearly determine the prospects for research in these areas is the metagenomic analysis of microbial consortia from preselected and sequenced samples of various compost ecosystems. For example, a group of researchers analyzed a library of pyrotags 16S rRNA metagenomic sequence of microbial consortia

^{*} Corresponding author: Firetiger2011@yandex.ru

[©] The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

from compost enriched with rice straw [2]. They showed that the predominant group among bacteria were representatives of *Actinobacteria*, *Proteobacteria*, *Firmicutes*, *Chloroflexi* and *Bacteroidetes*. Among these communities, many previously unknown species of the taxon Actinobacteria were also found, which made it possible to conclude that there is a greater ecological diversity of thermophilic actinobacteria than previously assumed [2]. Thus, this approach to metagenomes will help determine the presence or identify the presence of sequences of necessary enzymes and their hosts in the selected community for a wide variety of purposes.

One of the most popular trends in cosmetology and pharmacology is the search and creation of drugs to protect the skin from the effects of UV radiation. They can be used to combat diseases, such as xeroderma pigmentosa, or used to create daily care products. DNA glycosylases of pyrimidine dimers have significant potential in this area and are already beginning to recommend themselves as pharmaceutical and cosmetic protectors and reparations. Quite a lot of studies have been devoted to the product of the denV gene of Escherichia virus T4. It is a multifunctional enzyme capable of performing excisional repair of pyrimidine dimers due to N-glycosylase and AP (apurinic/apyrimidinic) lyase activities [3]. A number of articles are devoted to the study of its activity and effects in eukaryotic and prokaryotic cells [4-5]; there are also several patents associated with it [6-7]. For example, Korean inventors have created a cosmetic composition containing a component that can prevent signs of aging, and promote skin recovery from sun stress [6]. Inventors from the University of Oregon of Health received a Canadian patent for polypeptides of glycosylase specific to pyrimidine dimers and methods of their use for repairing damaged DNA [7]. It was the product of the denV gene and its specially engineered mutants. Moreover, in addition to the forms of the bacteriophage T4 enzyme, its homologue, pyrimidine dimer specific glycosylase (CV-PDG) of Paramecium bursaria chlorella virus-1 (PBCV-1) was also used [8].

It follows that the search for DNA glycosylases, in particular homologs of the already proven DenV enzyme of bacteriophage T4, can be of great value for the preparatory stages of the selection and development of UV protectors for the needs of medicine, veterinary medicine and cosmetology. As mentioned earlier, composts are a fairly specific ecosystem in which a large number of microorganisms exist. In theory, they can possess more thermally stable and valuable homologs of DNA glycosylase. Thus, the aim of this work was the search in the compost metagenomes for new amino acid sequences of DNA glycosylases of pyrimidine dimers and their taxonomic analysis

2 Materials and methods

To search for homologous, we took the amino acid sequence of the denV gene product - DNA glycosylase DenV of Escherichia virus T4:

>NP_049733.1 DenV endonuclease V, N-glycosylase UV repair enzyme [Escherichia virus T4]:

MTRINLTLVSELADQHLMAEYRELPRVFGAVRKHVANGKRVRDFKISPTFILGAGH VTFFYDKLEFLRKRQIELIAECLKRGFNIKDTTVQDISDIPQEFRGDYIPHEASIAISQ ARLDEKIAQRPTWYKYYGKAIYA

2.1 Search for homologues among the compost metagenomes

To find homologues among metagenomes, the DELTA-BLAST algorithm was used with the following parameters: matrix: BLOSUM62, gap costs: existence: 9 extension: 1. The second iteration was carried out using the PSI-BLAST algorithm, parameters: incl. Threshold: 0.005. The search was performed in the database env_nr, taxid: 702656.

2.2 Reverse search for compost DNA glycosylases

For the reliability of bioinformatics analysis, it was decided to introduce controls - homologues of glycosylases found in metagenomes. For each found glycosylase, a search was carried out using the PSI-BLAST algorithm, parameters: incl. Threshold: 0.001 in several iterations. The parameters for each are presented in table 1:

Glycosylase from metagenome (GenBank number)	Allowed E-value for homolog	Allowed% identity for homolog	Number of iterations
MNQ25265.1	<e-25< td=""><td>47-100%</td><td>5</td></e-25<>	47-100%	5
MNL43486.1	<e-33< td=""><td>47-100%</td><td>4</td></e-33<>	47-100%	4
MMZ46843.1	<e-70< td=""><td>80-100%</td><td>5</td></e-70<>	80-100%	5
MNW40567.1	<e-70< td=""><td>80-100%</td><td>5</td></e-70<>	80-100%	5
MNS97894.1	<e-25< td=""><td>58-100%</td><td>5</td></e-25<>	58-100%	5

Table 1. Parameters of the reverse search for glycosylase homologs from metagenomes

2.3 Search for homologues among the taxa Tequatrovirus and Bacteria

For bioinformatics analysis, homologous sequences of DNA glycosylases of other members of the *Tequatrovirus* genus, to which Escherichia virus T4 belongs, were required, as well as homologous proteins among bacteria. The search was carried out using the PSI-BLAST algorithm, in several iterations, incl. Threshold: 0.005. Sequences with E-value $<3e^{-29}$ were selected. For bacteria, an additional criterion was their belonging to ecological niches: soil, human and animal microbiome.

2.4 Phylogenetic analysis

The resulting file with amino acid sequences of DenV T4 homologs in FASTA format were combined into one file and used for processing in the MEGA X software package [9]. Alignment was performed using the MUSCLE program [10]. The phylogenetic tree was constructed using the Maximum Likelihood algorithm [11], bootstrap of 1000 repetitions [12] was chosen as a statistical method.

3 Result

As a result of searching for homologues among compost metagenomes, we were able to find five glycosylase sequences, their GenBank numbers: MMZ46843.1, MNW40567.1, MNS97894.1, MNQ25265.1, MNL43486.1 [13-14]. Metagenomes containing these sequences were extracted from bacterial cultures obtained from the compost soil of the Experimental Botanical Garden Goettingen, Germany, by the research team Egelkamp, R., Zimmermann, T., Hertel, R. and Daniel, R [13-14]. These cultures were additionally enriched with nitriles and their corresponding carboxylic acids; they were subsequently sequenced and composed in metagenome. Thus, the found glycosylases are quite unique, since they were found in an artificially created unique habitat, which makes them more valuable.



Fig. 1. Phylogenetic tree of DNA glycosylase homologues among compost metagenomes, constructed by the Maximum Likelihood method. Legend: colored squares - glycosylases from compost metagenomes; figures of the same color as the squares - homologues from the reverse search for a given metagenomic sequence: circles - bacterial sequences (red circles - bacterial homologues for

MMZ6843.1 and MNW40567.1), triangles - sequences from archaea, rhombus - from phages, yellow diamonds - Tequatrovirus glycosylases, black circles - glycosylases from human and animal bacteria, white circles with a black border - glycosylases from ubiquitous bacteria. The numbers mean the bootstrap values for each branch. Abbreviations denote proteins of the following bacteriophages, bacteria, arhaea and compost metagenomes (GenBank accession numbers are given in brackets): MMZ46843.1 (compost metagenome), WP 023986688.1 (MULTISPECIES: Paenibacillus), WP 019685919.1 (Paenibacillus polymyxa), WP 104497419.1 (Paenibacillus peoriae), WP 044647753.1 (Paenibacillus terrae), MNW40567.1 (compost metagenome), KAF6631859.1 (Paenibacillus sp. EKM208P), MNS97894.1 (compost metagenome), WP_019378734.1 (Virgibacillus halodenitrificans), WP 089182766.1 (Campylobacter sputorum), WP 066235590.1 (Metabacillus fastidiosus), EOB22415.1 (Streptococcus mitis 11/5), KRN44031.1 (Pediococcus damnosus), KRL37874.1 (Lactobacillus uvarum DSM 19971), MNQ25265.1 (compost metagenome). EEQ93896.1 (Ochrobactrum intermedium LMG 3301), RRD67058.1 (Comamonadaceae bacterium OH2310 COT-174), QDP53519.1 (Prokaryotic dsDNA virus sp.), RKW20205.1 (Candidatus Gracilibacteria bacterium), MNL43486.1 (compost metagenome), MBA4371370.1 (Thermodesulfovibrio sp.), HBE44341.1 (Deltaproteobacteria bacterium), KYK30128.1 (Thermoplasmatales archaeon SG8-52-1), WP 012870710.1 (Sphaerobacter thermophilus), WP 012463983.1 (Methylacidiphilum infernorum), TAK56910.1 (Bacteriodetes bacterium), NP 049733.1 (Escherichia virus T4), YP_009290393.1 (Escherichia phage vB_EcoM-UFV13), QBO60867.1 (Escherichia phage vB EcoM G2133), QCQ57104.1 (Escherichia phage EcNP1), QBQ79735.1 (Escherichia phage vB EcoM R5505), YP 009210313.1 (Escherichia phage wV7), VEV88971.1 (Yersinia phage fPS-90), YP 002854083.1 (Enterobacteria phage RB51), YP 009110946.1 (Shigella phage pSs-1), ANZ51756.1 (Enterobacteria phage Kha5h), YP 009277499.1 (Shigella phage SHFML-11), YP 009618934.1 (Shigella phage Sf21), YP 009149369.1 (Yersinia phage phiD1), QEG05166.1 (Shigella phage JK23), QBO63640.1 vB EcoM G10400), ANZ51556.1 (Enterobacteria phage GiZh), (Escherichia phage YP 009197330.1 (Escherichia phage slur07), YP 009180626.1 (Escherichia phage slur14), YP 009148567.1 (Escherichia phage HY01), QBO61133.1 (Escherichia phage D5505), YP 007004503.1 (Escherichia phage ime09), YP 009279118.1 (Shigella phage SHFML-26), ANH49739.1 (Escherichia phage PE37), CAA84453.1 (Enterobacteria phage RB70), QBO63362.1 (Escherichia phage vB EcoM G2540), EAA6275583.1 (Salmonella enterica subsp. enterica), EBK7480192.1 (Salmonella enterica), WP 088209736.1 (Escherichia coli Strain: E38), WP 100771269.1 (Bacillus cereus), OTE92210.1 (Escherichia coli Strain: 34VL), PID34675.1 (candidate division SR1 bacterium), WP 140460916.1 (Cellulomonas oligotrophica), RYD66589.1 (Verrucomicrobiaceae bacterium), EEW87445.1 (Brucella melitensis bv. 1 str. 16M), RYD63982.1 (Verrucomicrobiaceae WP 044287779.1 (Brucella *bacterium*), abortus), WP 006073748.1 (MULTISPECIES: Brucella), WP 082912581.1 (Sinorhizobium americanum), WP 006199911.1 (Brucella suis), WP 111303198.1 (Aggregatibacter aphrophilus), WP 065295504.1 WP 038320634.1 (Aggregatibacter aphrophilus), (Kingella kingae), WP 069315562.1 WP 111296191.1 (Aggregatibacter segnis), (Xenorhabdus hominickii), WP 115180151.1 (Haemophilus parainfluenzae), WP 109079219.1 (Aggregatibacter kilianii), WP 126371097.1 (Avibacterium volantium), WP 026822735.1 (Arsenophonus nasoniae), RTK96651.1 (Neisseriaceae bacterium), EGV07391.1 (Haemophilus pittmaniae HK 85), ENT03393.1 (Brucella sp. 63/311), TMJ15802.1 (Alphaproteobacteria bacterium), WP 006906677.1 (Shuttleworthia satelles), WP 043487880.1 (Streptomyces bingchenggensis). Clostridioides difficile (MTRINLTLVSELTDQHLMAEYRELPRVFGAVRKHVQNGKRVKDFKISPTFILGTGHVTFFYD KLEFLRLRQIELIAECLKRGFKIKDTTVQDISDIPAEFRNNYVPSEASIAISQARLDEKIAQRPT WYKYYGKSIY).

On the resulting phylogenetic tree (Fig. 1), we obtained an interesting distribution among glycosylases from the compost metagenome, DenV phage glycosylases homologs from the *Tequatrovirus* genus, and sequences found in soil bacteria and bacteria that can be opportunistic and pathogenic for humans and animals. Also, as a result of the reverse search, a homolog was found among the Archaea domain.

Each glycosylase from the compost formed its own branch, which included its bacterial and archaeal homologues found by reverse search. Exceptions were glucosylases numbered MNL43486.1 and MNQ25265.1 In addition to their own homologues, most of their

branches were proteins from bacteria from the human oral cavity and pathogens that cause respiratory infections.

In the reverse search, it was found that glycosylases MMZ46843.1 and MNW40567.1 have common homologues from *Paenibacillus* sp. - important producers of antibiotics

The sequences from bacteriophages formed a separate branch of their own, taking with them a pair of homologs from bacilli and enterobacteria.

4 Findings

Thus, the found glycosylases from the compost metagenomes are more similar to bacterial sequences than to phage ones. The proteins MNL43486.1 and MNQ25265.1 may be the closest to phage proteins. It should be noted that these sequences had the greatest similarity to the DenV of bacteriophage T4 at the time of the search for homologues. Their close similarity with homologues from bacterial pathogens is also alarming, since the studied enzyme affects the resistance to UV-radiation.

The fact that these proteins are located in a rather peculiar ecosystem can influence the functions and characteristics, therefore, their further study may turn out to be very promising for both industrial and fundamental studies. The latter may raise the question of the pathways of origin of this protein both among bacteria and among phages. The authors have already carried out some studies on oceanic homologues of this glycosylase [15], some of the results turned out to be similar - enzymes from *Tequatrovirus* phages and proteins from enterobacteria and bacilli were closely related.

After structural modeling, the found glycosylases can be tested for their thermal stability and tested as a basis for therapeutic and prophylactic drugs.

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

References

- 1. Wen Yi Chia et al., Environmental pollution, 267, 115662 (2020)
- 2. Cheng Wang et al., Biotechnology for biofuels, 9, 22 (2016)
- 3. https://www.uniprot.org/uniprot/P04418
- 4. K. Valerie, J.K. de Riel, and E.E. Henderson, Proceedings of the National Academy of Sciences of the United States of America, **82(22)**, 7656-7660 (1985)
- 5. J.T. Kibitel et al., Photochemistry and photobiology **54(5**), 753-760 (1991)
- 6. KR20060108665A South Korea (US51030703P, USA) Composition comprising a rosmarinus officinalis plant extract, a centella, echinacea or alpinia plant extract and a dna repair enzyme. Worldwide applications 2004 JP WO KR TW 2006 US
- CA2712900A1 Inventors: Amanda K. McculloughR. Stephen Lloyd Dna repair polypeptides and methods of delivery and use. Canada, Oregon Health Science University Worldwide applications 2009 WO US AU CA
- 8. A.K. McCullough et al., J. Biol. Chem. 273(21), 13136-13142 (1998)
- 9. S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, Molecular Biology and Evolution, **35(6)**, 1547-1549 (2018).
- 10. R. C. Edgar, Nucleic Acids Research, 32(5), 1792-1797 (2004)
- 11. J. Felsenstein J. Evolution 39, 783-791 (1985)
- 12. D.T, Jones, W.R, Taylor, and J.M. Thornton, Computer Applications in the Biosciences 8, 275-282 (1992)

- 13. NCBI:taxid702656:https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id =702656
- 14. R. Egelkamp et al, Front. Environ. Sci., 7(JUL), 103 (2019)
- 15. A.N. Karmanova, A.A. Zimin, J. Physics: Conf. Series, 1701, 012022 (2020)