FTB Food Technology & Biotechnology

scientific note
ISSN 1330-9862
doi: 10.17113/ftb.56.02.18.5393

# Bioprospecting for Genes Encoding Hydrocarbon-Degrading Enzymes from Metagenomic Samples Isolated from Northern Adriatic Sea Sediments

Ranko Gacesa[1,2,3#], Damir Baranasic[1,4#], Antonio Starcevic[1,5], Janko Diminic[1,5], Marino Korlević[6], Mirjana Najdek[6], Maria Blažina[6], Davor Oršolić[1], Domagoj Kolesarić[1], Paul F. Long[2,3], John Cullum[4], Daslav Hranueli[1,5], Sandi Orlic[7,8] and Jurica Zucko[1,5]*

[1]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, HR-10000 Zagreb, Croatia
[2]Institute of Pharmaceutical Science King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NH, UK
[3]Department of Chemistry, King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NH, UK
[4]Department of Genetics, University of Kaiserslautern, Postfach 3049, DE-67653 Kaiserslautern, Germany
[5]Centre of Research Excellence for Marine Bioprospecting - BioProCro, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia
[6]Centre for Marine Research, Ruđer Bošković Institute, G. Paliaga 5, HR-52210 Rovinj, Croatia
[7]Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia
[8]Center of Excellence for Science and Technology Integrating Mediterranean Region, Microbial Ecology, HR-10000 Zagreb, Croatia

Received: 16 June 2017
Accepted: 12 February 2018

*Corresponding author:
Phone: +38514605151;
Fax: +38514836083;
E-mail: jzucko@pbf.hr

[#]These authors contributed equally to this work

## SUMMARY

Three metagenomic libraries were constructed using surface sediment samples from the northern Adriatic Sea. Two of the samples were taken from a highly polluted and an unpolluted site respectively. The third sample from a polluted site had been enriched using crude oil. The results of the metagenome analyses were incorporated in the REDPET relational database (http://redpet.bioinfo.pbf.hr/REDPET), which was generated using the previously developed MEGGASENSE platform. The database includes taxonomic data to allow the assessment of the biodiversity of metagenomic libraries and a general functional analysis of genes using hidden Markov model (HMM) profiles based on the KEGG database. A set of 22 specialised HMM profiles was developed to detect putative genes for hydrocarbon-degrading enzymes. Use of these profiles showed that the metagenomic library generated after selection on crude oil had enriched genes for aerobic *n*-alkane degradation. The use of this system for bioprospecting was exemplified using potential *alkB* and *almA* genes from this library.

**Key words:** oil pollution, *n*-alkane degradation, database

## INTRODUCTION

There are many industrial sites on the northern Adriatic Sea resulting in decades of hydrocarbon pollution. In order to establish a baseline for future studies, the concentrations of hydrocarbons in surface sediments and the bacterial diversity at seven selected sites were studied (*1*). This paper expands the previous study relying on bioprospecting, process of systemic searching for genetic and biochemical potential of bacterial communities, for hydrocarbon degradation activity. Potential was assessed by searching metagenomic sequences using 22 hidden Markov model (HMM) profiles of most characterised hydrocarbon-degrading enzymes. Many studies have characterised the bacterial aerobic degradation of linear *n*-alkanes, which are the major component of petroleum products. As *n*-alkanes are rather inert chemically, the first step in the degradation pathway is activation, usually involving oxidation to an alcohol by an enzyme using molecular oxygen as a substrate (*2*). Short-chain *n*-alkanes ($C_2$–$C_4$) are oxidised by enzymes related to methane monooxygenases: soluble methane monooxygenases (sMMO) and particulate methane monooxygenases (PMO). Medium-chain ($C_5$–$C_{17}$) *n*-alkanes can be activated by soluble cytochrome P450s (*e.g.* CYP153) or integral membrane non-heme iron monooxygenases, *e.g.* AlkB. Long-chain (>$C_{18}$) *n*-alkanes are hydroxylated by unrelated enzymes including AlmA and LadA (*3*). Anaerobic biodegradation of hydrocarbons also occurs, with the most widely reported mechanism for activation of the substrate being enzymatic addition of the hydrocarbon across the double bond of fumarate. Recently, three phylogenetically related enzymes catalyzing this addition have been identified: alkylsuccinate synthase (AssA) activates *n*-alkanes, benzylsuccinate synthase (BssA) activates toluene and xylene and 2-methylnaphthylsuccinate synthase (Nms) activates 2-methylnaphthalene (*4-6*). The University of Minnesota Biocatalysis/Biodegradation Database (*7*) provides comprehensive information about known hydrocarbon-degrading enzymes.

Hydrocarbon-degrading enzymes are interesting for the development of bioremediation strategies for polluted sites. They are also interesting as potential industrial enzymes and the identification of novel enzymes will increase the armoury of the biotechnologists for the

ORCID IDs: 0000-0003-2119-0539 (Gacesa), 0000-0001-5948-0932 (Baranasic), 0000-0003-2386-2124 (Starcevic), 0000-0001-5104-5813 (Diminic), 0000-0001-5680-3556 (Korlević), 0000-0002-3915-0765 (Najdek), 0000-0002-3756-5860 (Blažina), 0000-0002-5385-1031 (Oršolić), 0000-0003-0831-2685 (Kolesarić), 0000-0001-6410-5803 (Long), 0000-0002-3850-8526 (Cullum) 0000-0001-8336-4384 (Hranueli), 0000-0002-6339-4145 (Orlic), 0000-0001-7782-6503 (Zucko)

development of new processes. Most bacterial species in the environment cannot be grown as pure cultures in the laboratory and a metagenomic approach (*8*) can be used to discover enzymes not present in culturable species. Such bioprospecting of metagenome data presents considerable problems for bioinformatics. Low sequence coverage makes assembly of genes difficult. The probable functions of genes are deduced on the basis of similarity to known genes. This can be achieved using BLAST similarity searches (*9*) of *in silico*-translated DNA sequences. However, BLAST may not be effective for identifying dissimilar to known sequences and, thus, may miss some of the most interesting enzymes. The use of HMM profiles (*10*) has the advantage that amino acid residues are weighted according to their degree of conservation in protein families and is better for identifying dissimilar sequences in a protein family.

In this paper, we describe metagenome sequences from surface sediments in the northern Adriatic Sea. The sequences were incorporated in a newly developed custom database called REDPET (REDucing environmental impact from local PETrochemical industry by novel bioremediation strategies), which was designed for bioprospecting novel hydrocarbon-degrading enzymes. The use of this database is exemplified by the choice of five putative AlkB sequences.

## MATERIALS AND METHODS

Three metagenomes were characterized by shotgun sequencing. The metagenome MET1 was derived from a heavily polluted sediment sampled from the Uljanik shipyard (marked as BN in **Fig. 1**, BN 44.866665°N 13.840400°E) as previously described (*1*). The second metagenome, MET2, was derived from an unpolluted coastal surface sediment sample from Cuvi beach at Rovinj (marked as CU in **Fig. 1**, 45.062290°N 13.652326°E). The third metagenome, MET3, was derived by pooling two moderately polluted sediment samples taken from a tanker berth station (marked as TV in **Fig. 1**, 45.276365°N 14.549654°E) (*1*). These two samples had been enriched for potential hydrocarbon-degrading bacteria by incubating under aerobic or anaerobic conditions in the presence of crude oil (*1*). The sampling procedure for CU sample and the assay for hydrocarbon content were as previously described for the samples BN and TV (*1*).

Total DNA was isolated from 10 g of each sample with the PowerMax Soil DNA Isolation Kit (MO BIO, Carlsbad, CA, USA) according to the manufacturer's instructions. The column was eluted in 5 mL of 10 mM Tris. Sodium acetate (1:10 by volume; Kemika, Zagreb, Croatia) was added and DNA was precipitated for 30 min at -20 °C in one volume of isopropanol (Alkaloid, Skopje, Macedonia). After centrifugation (Eppendorf 5430 R, Leipzig, Germany) for 21 min at 4 °C and 20 000×*g* the DNA pellet was washed with 2 mL of 70 % ethanol solution (Grammol, Zagreb, Croatia). A second centrifugation step lasted for 5 min (4 °C, 20 000×*g*). The pellet was dissolved in 120 µL of 10 mM Tris. Samples were sent to Eurofins MWG GmbH, Ebersberg, Germany for sequencing with the Roche 454 GS FLX+ chemistry. The sequences have been deposited in the European Nucleotide Archive (study accession number: PRJEB13497 and respective sample accession numbers: SAMEA3928486, SAMEA3928487 and SAMEA3928488).

The REDPET database was constructed using the MEGGASENSE platform (*11*). HMM profiles for hydrocarbon-degrading enzymes were generated with HMMER v. 3.0 (*10*) using protein sequences downloaded from the KEGG database (*12*) as a primary source. If less than 10 sequences were present for specific KEGG orthologue (KO), all protein sequences of respective KO were used to search UniRef50 database (*13*). Identified unique UniRef50 clusters were used to build HMMs. Metagenomic read was assigned to specific HMM profile if E-value was lower than $10^{-5}$ and the length of alignment was greater than a third of the HMM profile length. Taxonomic profiling of metagenomes was done using Kaiju (*14*) on entire metagenomic dataset. Default settings were used and RefSeq Genomes (*15*) was used as a reference database.

## RESULTS AND DISCUSSION

### *Metagenome samples*

A metagenomic library MET1 was constructed from a heavily polluted sample from the Uljanik shipyard (BN, **Fig. 1**). In order to understand the effects of pollution, it was also necessary to have a sample from an uncontaminated area in the northern Adriatic Sea. A surface sediment sample was collected from Cuvi beach in Rovinj (CU, **Fig. 1**). This had low levels on dry mass basis of aliphatic hydrocarbons (resolved *n*-alkanes 4.13 µg/g, unresolved complex mixture 22.48 µg/g) and polycyclic aromatic hydrocarbons (PAHs; 0.08 µg/g) compared to the two polluted sites, BN and TV (resolved *n*-alkanes 38.07 and 6.59 µg/g, unresolved complex mixture 518.19 and 10.75 µg/g, PAHs 73.53 and 0.40 µg/g) previously reported (*1*). This sample was used to construct the metagenomic library MET2. The third library, MET3, was derived from a moderately polluted sample from a tanker berth station (TV, **Fig. 1**), which was grown under the crude oil selection pressure. Samples from aerobic and anaerobic selection were pooled to construct the library.

The three libraries were sequenced using pyrosequencing and each yielded similar amounts of sequence data (**Table 1**) with similar read lengths. However, while over 60 % of the reads in MET3 could be assembled into contigs, less than 3 % of the reads in the other two libraries could be assembled.

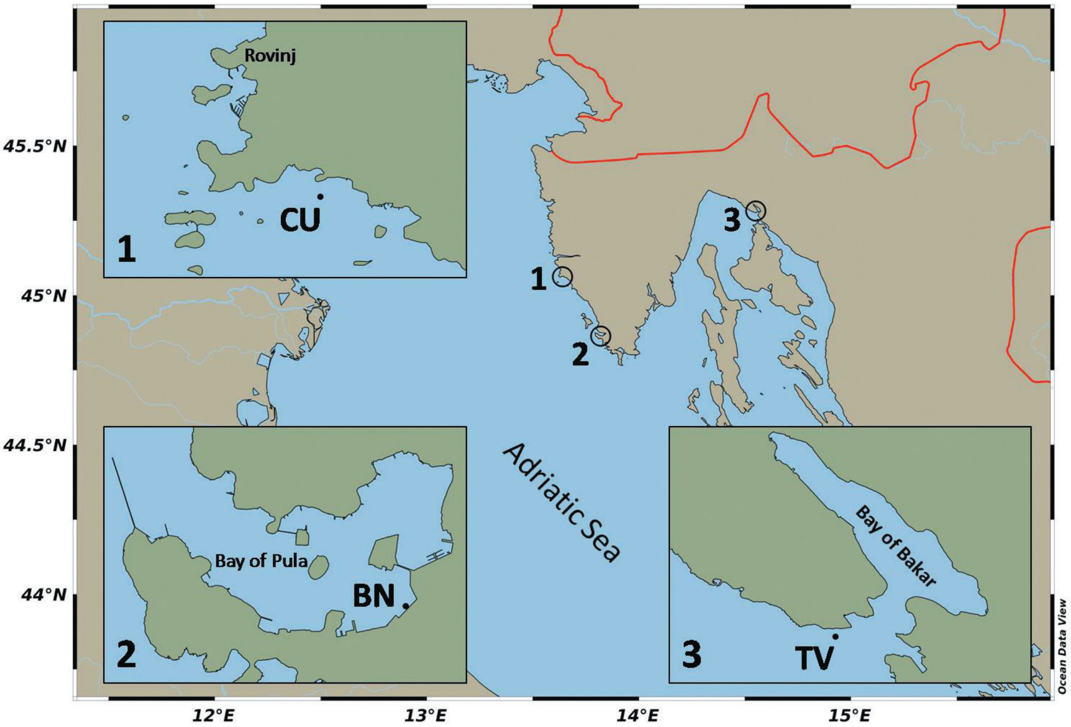

**Fig. 1.** Sampling sites used for construction of metagenomic libraries. Cuvi beach (CU), repair shipyard Uljanik (BN) and tanker berth (TV)

**Table 1.** Metagenome sequencing and assembly statistics

| Metagenome | MET1 (contaminated) | MET2 (uncontaminated) | MET3 (enriched on crude oil) |
|---|---|---|---|
| Number of reads | 1329237 | 1373730 | 1244899 |
| Number of bases/Mb | 879 | 838 | 796 |
| Range of read length | 21-1459 | 26-1424 | 19-1764 |
| Mean read length | 640 | 630 | 640 |
| Mode read length | 782 | 765 | 762 |
| Reads assembled/% | 2.5 | 1.9 | 62.0 |
| Number of contigs | 13442 | 7329 | 28426 |
| Bases in contigs/Mb | 9.54 | 5.47 | 63.84 |
| Average contig size/bp | 710 | 747 | 2246 |

### *Taxonomic classification*

All three metagenomic libraries were used for taxonomic classification using Kaiju (*14*). Each read was translated in all six possible reading frames and searched against RefSeq Genome database (*15*) to identify originating taxon. For MET1 and MET2, 48 % of reads were unclassified while for MET3 only 24 % of reads were unclassified. The most abundant in all three metagenomes is phylum Proteobacteria with 56, 52 and 75 % of all classified reads. Second most abundant is the phylum Bacteroidetes with 23, 23 and 11 % of all classified reads, followed by Firmicutes (4, 4 and 5 %) and Actinobacteria (4, 5 and 1 %). Most abundant species in MET1 is *Woeseia oceani* with 3 % of all classified reads, followed by *Halioglobus pacificus* (2 %) and *Halioglobus japonicus* (2 %). Most abundant species in MET2 is also *Woeseia oceani* with 3 % of all classified reads, followed by *Maribacter* sp. HTCC2170 (2 %). MET3 is dominated by *Immundisolibacter cernigliae*, corresponding to 16 % of all classified reads, followed by several *Marinobacter* species (15 %). All three metagenomes have high diversity with identified 3280, 3349 and 3279 bacterial species reflecting on high Shannon index (6.8, 6.7 and 5.5). The reason for such high diversity might be due to the method used for taxonomic classification that tries to assign taxa to every metagenomic read and retention of taxa present in low numbers, as singletons were kept for analysis. When taking into account only bacterial species present with more than 0.1 % abundance, number of identified species falls to 193, 182 and 111. Enrichment with crude oil reflected on the bacterial composition on species level with dominant bacterial species in MET3 being *Immundisolibacter cernigliae,* making up

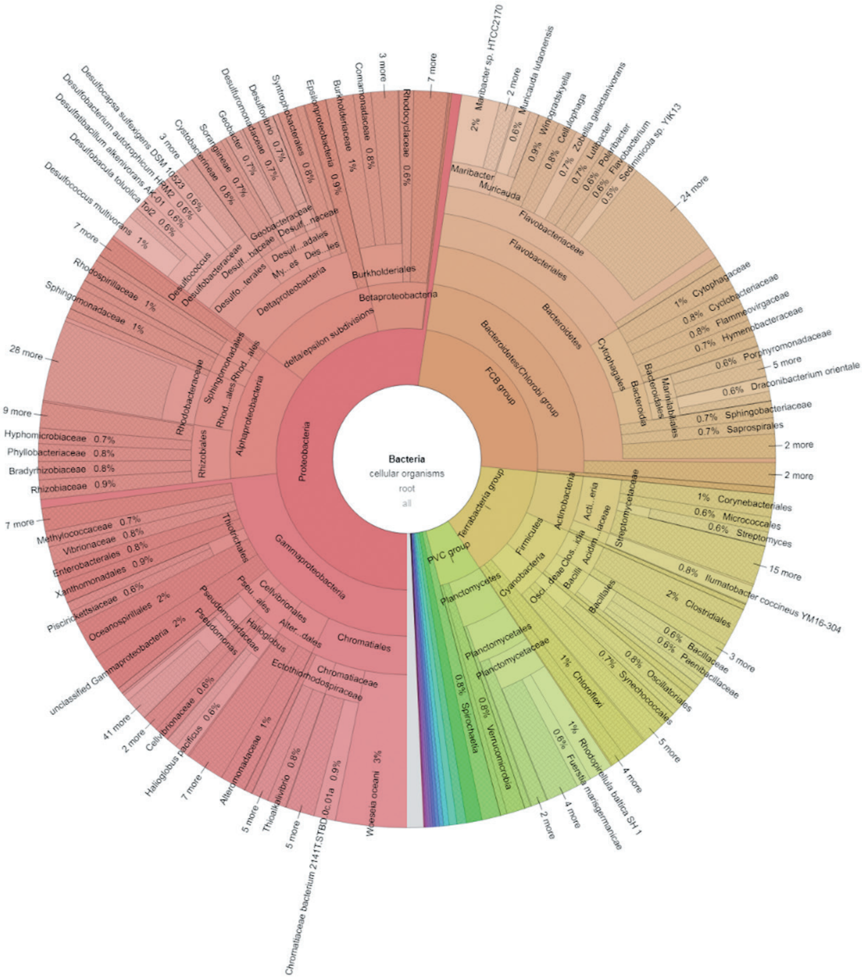

**Fig. 2.** Phylogeny of the metagenome MET2 at the level of phylum

16 % of all classified reads. With several *Marinobacter* species, they represent 31 % of all classified reads in MET3 and both genera are involved in hydrocarbon degradation (*4,6,16,17*). Dominance of several bacteria in MET3 metagenome is also reflected on greater number of reads assembled into contigs.

### *Development of the REDPET database*

The results of the analyses of the metagenomes were used to construct the REDPET relational database (*18*), which was based on the MEGGASENSE platform (*11*) and incorporates the generic functionality of this platform. This includes assembly of reads, taxonomic analysis and an intelligent search engine. General functional analysis of the metagenomes uses a set of profiles based on KEGG orthologues (KO), which are ordered according to the BRITE hierarchical functional scheme (*12*). The taxonomic analysis was carried out using the Kaiju program (*14*) and the database incorporates the Krona viewer (*19*) for display of the phylogenetic composition of the metagenome. An example is shown for the MET2 library (**Fig. 2**) at a phylum level. As Krona is an interactive viewer, it is possible to view the taxonomy results at different hierarchical levels down to the individual species level.

As the main aim of the project was bioprospecting for potential hydrocarbon-degrading enzymes, it was necessary to add some custom features to the REDPET database. A collection of HMM profiles was developed to recognise these enzymes (**Table 2**). Some of the enzyme activities were already included in the KEGG orthology database, so that the MEGGASENSE profiles corresponding to the KEGG KOs could be used (*e.g.* AlkB, PmoA). If there were only a small number of sequences present in the KO (less than 10 protein sequences), UniRef50 clusters of all members belonging to KO were obtained and distinct UniRef50 clusters were used to construct the profile (*e.g.* MmoB, MmoX). If the enzyme did not have corresponding KO entry (*e.g.* AlmA), its sequence was identified in the UniProt database (*13*) based on a literature search and the UniRef50 cluster of the identified protein was then used to construct the HMM profile. The collection of 22 HMM profiles (**Table 2**) was used to mine the assembled contigs of the metagenomic libraries. The HMM profiles are available from the REDPET database (*18*).

**Table 2.** HMM profiles of hydrocarbon-degrading enzymes

| Peptide | Substrate | Source* |
|---|---|---|
| MmoB | short-chain *n*-alkanes | D2K2D0+P18797 |
| MmoD | | B8EPN0+G8RLG0+G0A1L1+P22867 |
| MmoC | | B8EPN1+G8RLG1+P22868 |
| MmoX | | B2HJD2+D2K2E0+P22869 |
| MmoY | | D2K2B9+G8RLF8+G8RJD6+P18798 |
| MmoZ | | G8RLG2+P11987+P27355 |
| PmoA | short-chain *n*-alkanes | K10944 |
| PmoB | | K10945 |
| PmoC | | K10946 |
| AlkB | medium-chain *n*-alkanes | K00496 |
| CYP153 | medium-chain *n*-alkanes | H8WCT8 |
| AlmA | long-chain *n*-alkanes | D4XTS7 |
| LadA | long-chain *n*-alkanes | F0KLN9 |
| AssA | *n*-alkanes, anaerobic | B8FEM4 |
| BssA | toluene, xylene | O87943 |
| BssB | | O87944 |
| BssC | | O87942 |
| BssD | | O87941 |
| NmsA | 2-methylnaphthalene | B0CLU5 |
| NmsB | | D8F4E4 |
| NmsC | | R4KAJ4 |
| NmsD | | D8F4E3 |

*Entries with an initial K are derived from Kegg orthologues. The other entries are derived from UniRef protein sequences (*12*,*13*)

**Table 3.** Genes for putative hydrocarbon-degrading enzymes in the metagenomic libraries

| Peptide | MET1 | MET2 | MET3 | |
|---|---|---|---|---|
| | Hits | Hits | Hits | Complete genes |
| MmoB | 21 | 31 | 64 | 0 |
| MmoD | 19 | 28 | 10 | 0 |
| MmoC | 1018 | 1093 | 2002 | 13 |
| MmoX | 44 | 66 | 165 | 4 |
| MmoY | 26 | 33 | 215 | 1 |
| MmoZ | 16 | 11 | 10 | 0 |
| PmoA | 39 | 25 | 68 | 2 |
| PmoB | 37 | 54 | 110 | 2 |
| PmoC | 31 | 37 | 128 | 3 |
| AlkB | 458 | 422 | 1750 | 19 |
| CYP153 | 948 | 920 | 952 | 8 |
| AlmA | 828 | 735 | 1241 | 6 |
| LadA | 397 | 483 | 626 | 2 |
| AssA | 417 | 402 | 263 | 1 |
| BssA | 446 | 429 | 245 | 1 |
| BssB | 92 | 87 | 167 | 0 |
| BssC | 39 | 36 | 38 | 0 |
| BssD | 4807 | 4421 | 3752 | 2 |
| NmsA | 392 | 373 | 211 | 1 |
| NmsB | 62 | 55 | 72 | 0 |
| NmsC | 34 | 18 | 11 | 0 |
| NmsD | 4152 | 3818 | 3334 | 2 |
| Total | 14323 | 13577 | 15434 | 67 |

### *Hydrocarbon-degrading enzymes*

The number of reads corresponding to putative hydrocarbon-degrading genes found in each of the metagenomic libraries is shown in Table 3. In addition to scanning the reads, the assembled contigs were also analysed to find complete genes. As expected for the small contig sizes (Table 1), no complete genes were found in MET1 and MET2, but 67 complete genes were found in MET3 (Table 3).

As the three metagenomic libraries have similar numbers of reads and total sequence lengths (Table 1), it is possible to compare the numbers of hits directly. It can be seen (Table 3) that the three libraries have comparable total numbers of hits. There is little difference in the sample from the highly polluted shipyard (MET1) and the sample from the unpolluted site (MET2). However, the samples selected on crude oil (MET3) have an increased proportion of genes for aerobic *n*-alkane degradation. Genes for short- (MmoC), medium- (AlkB) and long-chain (AlmA, LadA) *n*-alkane degradation are all present in an increased proportion. However, there was a lower proportion of the anaerobic *n*-alkane-degrading enzyme AssA and the anaerobic cyclic hydrocarbon-degrading enzymes (Bss and NMS).

An important aim of this research was to identify novel hydrocarbon-degrading enzymes with initial interest focusing on the 19 complete *alkB* genes (Table 3) encoded in the MET3 metagenomic library. BLAST searches (*9*) were carried out with the translated sequences and all best hits showed at least 85 % coverage of the query sequence. Most of the hits were annotated as alkane 1-monooxygenase (Table 4). Five of the hits were chosen for further work, including cloning by synthesis. Table 5 shows the six putative *almA* genes. None of them are annotated as *n*-alkane monooxygenase.

Three metagenomic libraries were constructed from samples collected in the northern Adriatic Sea (Fig. 1). All the data from the metagenomes are presented in the REDPET database (*18*). The main analysis tool used to assign gene function in the REDPET database is the use of HMM profiles (*10*). The profiles are derived from sequence families and match residues in a weighted manner depending on their degree of conservation in the family. This gives better recognition of function than BLAST searches (*9*), which use no weighting. This advantage is particularly important when short sequences or sequences evolutionarily distant from most of those in the databases are being analysed; both these cases occur with metagenomic data. Use of BLAST searches (*9*) with uncurated databases such as GenBank (*20*) is particularly problematic, because there is no fixed standard of annotation and there are a lot of mistakes. For general analysis of gene function, HMM profiles derived from KEGG orthologues are used (*12*), which are presented using the hierarchical BRITE classification, thus simplifying functional analysis of the genes present. In order to detect putative hydrocarbon-degrading enzymes and potential enzymatic functions of interest, 22 HMM profiles were developed.

**Table 4.** Selection of *alkB* genes for synthesis. The results of BLAST analyses are shown with the annotation of the best hits and the percentage of identity (ID) with the hits. All hits shown had at least 85 % query coverage

| Gene | Size/bp | BLAST hit | ID/% | Accession number |
|---|---|---|---|---|
| | | Genes selected for synthesis | | |
| *M3-AlkB1* | 1146 | alkane 1-monooxygenase | 51 | WP_020931882 |
| *M3-AlkB2* | 1152 | hypothetical protein | 44 | WP_020162762 |
| *M3-AlkB3* | 1212 | fatty acid desaturase | 47 | WP_013832824 |
| *M3-AlkB4* | 1272 | alkane 1-monooxygenase | 93 | WP_008173998 |
| *M3-AlkB5* | 1275 | alkane 1-monooxygenase | 41 | KEF32438 |
| | | Genes not selected for synthesis | | |
| *M3-AlkB6* | 1104 | hypothetical protein | 67 | WP_029889235 |
| *M3-AlkB7* | 1116 | xylene monooxygenase | 40 | WP_025548206 |
| *M3-AlkB8* | 1116 | xylene monooxygenase subunit 1 | 42 | WP_016390893 |
| *M3-AlkB9* | 1119 | xylene monooxygenase hydroxylase subunit | 93 | WP_008828148 |
| *M3-AlkB10* | 1134 | alkane 1-monooxygenase | 99 | WP_008175258 |
| *M3-AlkB11* | 1137 | alkane 1-monooxygenase | 94 | WP_008175258 |
| *M3-AlkB12* | 1200 | fatty acid desaturase | 43 | KDA01186 |
| *M3-AlkB13* | 1212 | alkane 1-monooxygenase | 92 | WP_009506508 |
| *M3-AlkB14* | 1218 | hypothetical protein | 45 | WP_022988941 |
| *M3-AlkB15* | 1218 | alkane 1-monooxygenase | 100 | WP_014422298 |
| *M3-AlkB16* | 1236 | hypothetical protein | 32 | WP_022988941 |
| *M3-AlkB17* | 1254 | alkanal monooxygenase | 37 | EUC69885 |
| *M3-AlkB18* | 1290 | fatty acid desaturase | 40 | WP_023446487 |
| *M3-AlkB19* | 1291 | alkane 1-monooxygenase | 83 | WP_008939065 |

**Table 5.** Properties of the *alm*A genes found in metagenome MET3. The results of BLAST analyses are shown with the annotation of the best hits and the percentage of identity (ID) with the hits. All hits shown had at least 98 % query coverage

| Gene | Size/bp | BLAST hit | ID/% | Accession number |
|---|---|---|---|---|
| *M3-AlmA1* | 1467 | hypothetical protein | 97 | KCZ62165 |
| *M3-AlmA2* | 1485 | flavin-containing monooxygenase | 71 | WP_012137381 |
| *M3-AlmA3* | 1488 | 4-hydroxyacetophenone monooxygenase | 86 | WP_014577005 |
| *M3-AlmA4* | 1527 | monooxygenase, flavin-binding family protein | 90 | WP_014578764 |
| *M3-AlmA5* | 1539 | cyclohexanone monooxygenase | 84 | WP_012138025 |
| *M3-AlmA6* | 1575 | 4-hydroxyacetophenone monooxygenase | 85 | WP_027830175 |

Although similar amounts of sequence data were gathered from each sample, MET3 allowed a much greater degree of sequence assembly than the other two libraries (Table 1). This is because the selection on crude oil has resulted in a large reduction in the number of species present (*1*) so that the depth of sequencing for MET1 and MET2 was lower. Use of an alternative method such as Illumina sequencing would increase the depth of sequencing, but might cause problems due to the shorter length of reads (*21*).

The REDPET database includes a taxonomic browser (Fig. 2) so that the biodiversity of samples can be assessed. The metabolic activities of the microbial consortia represented in the metagenomes can be assessed using the results of the analyses with the HMM profiles. General metabolic activities are organised using the BRITE hierarchical classification of the KEGG database (*12*) allowing systematic analyses. The hydrocarbon-degrading enzymes (Table 3) in the different metagenomes were analysed using the specially constructed HMM profiles (Table 2). It is intended to undertake a more detailed analysis of the metabolic activities and establish base lines to assess the future development of pollution and remediation in the northern Adriatic Sea. Metagenomic analyses have already proven useful for the analysis of the consequences of oil spills (*4*).

For bioprospecting, it is necessary to analyse the putative genes of interest in more detail. The REDPET database allows easy downloading of genes for external analysis. The graphical user interface also allows BLAST searches (*9*) to be launched. Although useful results may also be extracted from partial gene sequences, it is much easier to assess complete assembled genes. When putative *almA* genes were analysed (Table 5), none of the best hits were annotated as long-chain ($>C_{18}$) *n*-alkane monooxygenases despite high sequence identity. This reflects the fact that such enzymes are less well studied than enzymes degrading shorter-chain *n*-alkanes (*3*) and illustrates the advantage of using HMM profiles to identify putative genes. In contrast, for the well-studied *alkB* genes, many of the putative genes were annotated as *n*-alkane monooxygenases (Table 4). The sequence identity in the BLAST searches, combined with prediction of active sites, can be used as a criterion for the potential novelty of the genes. Sequences for five potentially novel *alkB* genes will next be synthesised and putative degradation of medium-chain ($C_5$–$C_{17}$) *n*-alkanes then assessed following expression in a heterologous host.

## CONCLUSION

Metagenomics is a powerful tool for dealing with environmental problems and for bioprospecting for novel enzymes. The volume of generated data and the quality make analyses difficult. Standard analysis pipelines and databases do not provide the tools needed for specific projects. The REDPET database was constructed with metagenome sequences from sediments in the Adriatic Sea and was designed to facilitate analysis of genes involved in hydrocarbon degradation. This allowed easy identification of interesting target genes for bioprospecting. The construction of specialized databases using the MEGGASENSE platform is a general strategy, which can be applied to any metagenomic datasets.

## ACKNOWLEDGEMENTS

This work was supported by the grant 09/5 (to D.H.) from the Croatian Science Foundation, Republic of Croatia and by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (to D.H. and J.C.). It was also supported by King's College London (to P.F.L.) and BioProCro Centre of Excellence (to A.S., D.H., J.D. and J.Z.), a project co-financed by the Croatian Government and the European Union through the European Regional Development Fund – the Competitiveness and Cohesion Operational Programme (KK.01.1.1.01).

## CONFLICT OF INTEREST

We acknowledge support of SemGen Ltd. for use of portions of MEGGASENSE pipeline developed under FP7-funded project "Amylomics". Antonio Starcevic, Janko Diminic and Jurica Zucko have a business interest in SemGen Ltd., which can be contracted to provide third party services. The publication of this paper will serve as an advertisement for some of these services. Antonio Starcevic, Janko Diminic and Jurica Zucko declare conflict of interest.

## REFERENCES

*1.* Korlević M, Zucko J, Dragić MN, Blažina M, Pustijanac E, Zeljko TV, et al. Bacterial diversity of polluted surface sediments in the northern Adriatic Sea. Syst Appl Microbiol. 2015;38(3):189-97.
https://doi.org/10.1016/j.syapm.2015.03.001

*2.* van Beilen JB, Funhoff EG. Alkane hydroxylases involved in microbial alkane degradation. Appl Microbiol Biotechnol. 2007;74(1):13–21.
https://doi.org/10.1007/s00253-006-0748-0

*3.* Wang W, Shao Z. Enzymes and genes involved in aerobic alkane degradation. Front Microbiol. 2013;4:116.
https://doi.org/10.3389/fmicb.2013.00116

*4.* Kimes NE, Callaghan AV, Aktas DF, Smith WL, Sunner J, Golding B, et al. Metagenomic analysis and metabolite profiling of deep-sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill. Front Microbiol. 2013;4:50.
https://doi.org/10.3389/fmicb.2013.00050

*5.* Tan B, Dong X, Sensen CW, Foght J. Metagenomic analysis of an anaerobic alkane-degrading microbial culture: Potential hydrocarbon-activating pathways and inferred roles of community members. Genome. 2013;56(10):599-611.
https://doi.org/10.1139/gen-2013-0069

*6.* Gao W, Cui Z, Li Q, Xu G, Jia X, Zheng L. Marinobacter nanhaiticus sp. nov., polycyclic aromatic hydrocarbon-degrading bacterium isolated from the sediment of the South China Sea. Antonie van Leeuwenhoek. 2013;103(3):485-91.
https://doi.org/10.1007/s10482-012-9830-z

*7.* Gao J, Ellis LBM, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. Nucleic Acids Res. 2010;38(Suppl. 1):D488-91.
https://doi.org/10.1093/nar/gkp771

*8.* Vilchez-Vargas R, Junca H, Pieper DH. Metabolic networks, microbial ecology and 'omics' technologies: Towards understanding in situ biodegradation processes. Environ Microbiol. 2010;12(12):3089-104.
https://doi.org/10.1111/j.1462-2920.2010.02340.x

*9.* Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
https://doi.org/10.1093/nar/25.17.3389

*10.* Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205-11.
https://doi.org/10.1142/9781848165632_0019

*11.* Gacesa R, Zucko J, Petursdottir SK, Gudmundsdottir EE, Fridjonsson OH, Diminic J, et al. MEGGASENSE - The metagenome/genome annotated sequence natural language search engine: A platform for the construction of sequence data warehouses. Food Technol Biotechnol. 2017;55(2):251–7.
https://doi.org/10.17113/ftb.55.02.17.4749

*12.* Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(D1):D109-14.
https://doi.org/10.1093/nar/gkr988

*13.* Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926-32.
https://doi.org/10.1093/bioinformatics/btu739

*14.* Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.
https://doi.org/10.1038/ncomms11257

*15.* O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-45.
https://doi.org/10.1093/nar/gkv1189

*16.* Corteselli EM, Aitken MD, Singleton DR. Description of Immundisolibacter cernigliae gen. nov., sp. nov., a high-molecular-weight polycyclic aromatic hydrocarbon-degrading bacterium within the class Gammaproteobacteria, and proposal of Immundisolibacterales ord. nov. and Immundisolibacteraceae fam. nov. Int J Syst Evol Microbiol. 2017;67(4):925-931.
https://doi.org/10.1099/ijsem.0.001714

*17.* King GM, Kostka JE, Hazen TC, Sobecky PA. Microbial responses to the Deepwater Horizon oil spill: From coastal wetlands to the deep sea. Ann Rev Mar Sci. 2015;7:377-401.
https://doi.org/10.1146/annurev-marine-010814-015543

*18.* The REDPET database. Zagreb, Croatia; 2017. Available from: http://redpet.bioinfo.pbf.hr/REDPET.

*19.* Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.
https://doi.org/10.1186/1471-2105-12-385

*20.* Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2017;45(D1):D37-42.
https://doi.org/10.1093/nar/gkw1070

*21.* Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(10):315-27.
https://doi.org/10.1016/j.ygeno.2010.03.001