HADEG: A Curated Database of Hydrocarbon Aerobic Degradation Enzymes and Genes

Jorge Rojas-Vargas[1*], Hugo G. Castelán-Sánchez [2], Liliana Pardo-López[1*]

[1]Departamento de Microbiología Molecular, Instituto de Biotecnología, UNAM, Av. Universidad #2001, Col. Chamilpa, 62210 Cuernavaca, Morelos, México

[2]Programa de Investigadoras e Investigadores por México. Grupo de Genómica y Dinámica Evolutiva de Microorganismos Emergentes. Consejo Nacional de Ciencia y Tecnología. Av. Insurgentes Sur 1582, Crédito Constructor, Benito Juárez, CP 03940, Ciudad de México, México.

* Correspondence:
Liliana Pardo-López, liliana.pardo@ibt.unam.mx
Jorge Rojas-Vargas, jorge.rojas@ibt.unam.mx

ABSTRACT

Databases of genes and enzymes involved in natural and synthetic hydrocarbon degradation have been previously reported. However, these databases specialize on only a specific group of hydrocarbons and/or are constructed partly based on enzyme sequences with putative functions indicated by in silico research, with no experimental evidence. Here, we present a curated database of Hydrocarbon Aerobic Degradation Enzymes and Genes (HADEG) containing enzymes and genes involved in alkane, alkene, aromatic, and polymer aerobic degradation and biosurfactant production based solely on experimental evidence. HADEG includes 259 enzymes for petroleum hydrocarbon degradation, 160 for polymer degradation, and 32 for biosurfactant production. This database will help identify and predict hydrocarbon degradation genes/pathways and biosurfactant production in genomes.

Keywords

Curated database, hydrocarbon degradation, polymer degradation, biosurfactant production.

Impact statement

There is currently no database that integrates enzymes and genes with experimental evidence related to the degradation of different hydrocarbon groups (alkanes, alkenes, aromatics, polymers) or the production of biosurfactants. HADEG is an effort to assemble a single database of enzymes and gene sequences of the aforementioned groups. This database will be helpful for more fully predicting the potential degradation of microorganism genomes (or metagenomes), thus increasing knowledge about the biology of microbes that metabolize pollutants. Our findings from HADEG also shed light on the identification of a set of possible biomarkers for targeted searching of promising hydrocarbon-degrading microorganisms.

Data summary

The HADEG database repository is https://github.com/jarojasva/HADEG.

1 Introduction

Crude oil spills and plastic pollution of anthropogenic origin are two major concerns in the context of environmental equilibrium and conservation (Bhattacharjee & Dutta, 2022; Horton, 2022). Bioremediation technologies remain one of the most promising options for crude oil and plastic cleanup, based on microorganisms that can use hydrocarbons (HCs) and polymers as carbon sources. Interest in bioremediation has led to the identification and characterization of some enzymes involved in degradation processes. Among the features of these enzymes, studies have described their activity conditions, substrates, and products, in addition to their protein structures and active sites in some cases (Brzeszcz & Kaszycki, 2018; Kaushal et al., 2021; Pathak et al., 2022; Xu et al., 2018). Enzymes for biosurfactant production have also been studied; these microbial surface-active compounds increase the bioavailability of HCs and therefore promote higher levels of HC degradation (Banat, 1995).

Advances in sequencing technologies have allowed the nucleotide sequences of hydrocarbon-degrading genes and, thus, the amino acid sequences of the encoded proteins to be obtained. Some of these sequences can be found in existing tailored data repositories and can be used to identify and predict the degradation potential of microbial genomes and metagenomes. Regarding hydrocarbon degradation, the AromaDeg database is focused on the aerobic degradation of aromatic compounds. It was constructed based on a phylogenomic approach considering different oxygenases with known and unknown aromatic substrates (Duarte et al., 2014). A more recent database is that of the Calgary approach to ANnoTating HYDrocarbon degradation genes (CANT-HYD), which contains genes involved in aerobic and anaerobic aliphatic and aromatic hydrocarbon degradation pathways. CANT-HYD uses Hidden Markov Models (HMMs) for the annotation of six monophyletic clades of genes related to aliphatic aerobic degradation (*alkB, almA, ladA, bmo,* CYP153, *prm*), six related to aromatics (*dmp, dsz,* MAH*, ndo, tmo, tom*), and eight related to anaerobic pathways (Khot et al., 2021). The workflow derives 37 HMMs from experimental and in silico-inferred functional enzymes and is designed for the identification and annotation of marker genes, not including some pathways of hydrocarbon degradation.

For polymers, the Plastics Microbial Biodegradation Database (PMBD) (http://pmbd.genome-mining.cn/home/) is an online resource containing 79 experimentally predicted and ~8000 in silico-predicted sequences of enzymes from the literature and the UniProt database (Gan & Zhang, 2019). The Plastics-Active Enzymes Database (PAZy) is an inventory of enzymes that act on only four synthetic fossil fuel-based polymers and three polymers mainly found in renewable resources (Buchholz et al., 2022). Additionally, PlasticDB (Gambarini et al., 2022) includes 174 experimentally tested enzymes (consulted on July 15, 2022) related to the degradation of seven natural and twenty synthetic polymers. Regarding biosurfactants, the only reported database for these compounds is BioSurfDB (www.biosurfdb.org) (Oliveira et al., 2015), which includes experimental and predicted enzymes.

Here, we introduce the manually curated database of Hydrocarbon Aerobic Degradation Enzymes and Genes (HADEG), which is a public repository containing sequences of experimentally characterized enzymes and genes to be used for annotation purposes. HADEG groups these molecules according to their substrates (in the case of alkane, alkene, aromatic, and polymer aerobic degradation) and products (in the case of biosurfactant production). In

this work, we also briefly reviewed the enzymes and pathways involved in petroleum hydrocarbon and polymer degradation.

2 Materials and methods

2.1 Selection of reference sequences

For enzymes involved in alkane, alkene, and aromatic aerobic degradation, we compiled and classified metabolic pathways from the primary literature and two experimentally validated curated databases, MetaCyc (Krieger et al., 2004) and KEGG (Ogata et al., 1999). We searched the sequences of enzymes and genes involved in the aforementioned metabolic pathways in the well-characterized UniProtKB/Swiss-Prot database (Boutet et al., 2007), the NCBI portal, and in the literature if sequences were not available in any database (Supplementary Tables 1, 2, and 3).

To obtain polymer degradation enzymes, we selected nonredundant protein sequences from PlasticDB (Gambarini et al., 2022). To obtain biosurfactant production enzymes, we used the UniProt platform with different biosurfactants as keywords, and the corresponding proteins included in the UniProtKB/Swiss-Prot database were collected. In constructing our polymer and biosurfactant database, we added some new proteins from a search of the primary literature (Supplementary Tables 4 and 5). The nucleotide sequences of all proteins were retrieved from the European Nucleotide Archive (ENA) platform (Leinonen et al., 2011) or the literature.

2.2 Protein annotation and orthology

We used eggNOG-mapper (Huerta-Cepas et al., 2019) online version 2.1.9, run on July 23 (2022), to annotate our protein database, identify orthologs, and study their taxonomic distribution at the phylum level. Protein sequences without orthologs were considered seed sequences (Supplementary Table 6). We used InterProScan v.5.51-85.0 (Jones et al., 2014) for protein domain inference with the default parameters to identify the shared protein domains predicted with Pfam-33.1 (Supplementary Table 7).

2.3 Searching for hydrocarbon degradation enzymes in genomes using HADEG

We selected 60 genomes (Supplementary Table 8) of recognized HC- and plastic-degrading microorganisms and biosurfactant producers for validation purposes. The protein FASTA files were retrieved from the NCBI web portal (consulted on July 1, 2022). BLAST software (McGinnis & Madden, 2004) was used to predict enzymes by applying the following cutoff values: 50 for the percentage of identity, $1x10^{-60}$ for the e-value, ±10% of the variation in the query sequences, and the higher bit score, using the protein FASTA files from our database.

3 Results and discussion

3.1 Hydrocarbon aerobic degradation

HCs are relatively stable molecules but can serve as a source of carbon and energy for any organism able to activate them. This activation by the aerobic insertion of oxygen or the

anaerobic addition of succinate distinguishes hydrocarbon-degrading organisms from their counterparts (Prince et al., 2010). In HADEG, we included enzymes related to aerobic HC degradation, such as alkanes, alkenes, aromatics, and polymers (Supplementary Tables 1-4).

3.1.1 Alkanes

Alkanes are the principal components of crude oil. Four microbial pathways of alkane degradation have been identified (Fig. 1) (Ji et al., 2013; Moreno & Rojo, 2019; Wentzel et al., 2007). All of them begin with the activation of the alkane chain through its oxidation to a primary or secondary alcohol, as observed in terminal (TO), biterminal (BO), and subterminal oxidation (SO), or to an *n*-alkyl hydroperoxide via the Finnerty pathway (FP). Among the activation enzymes included in HADEG, we report the extensively studied alkane monooxygenase (*alkB*), cytochrome P153 (CYP153), and P450 (CYP450), flavin-binding protein monooxygenase (*almA*), long-chain alkane monooxygenase (*ladA*), and terminal alkane-hydroxylase (*alkM*). The rubredoxins (*rubA*, *rubB*), which play a crucial role in electron transfer to *alkB*, are also included. Some other included enzymes belong to a particular pathway, such as Baeyer-Villiger monooxygenase (BVMO) in SO and alkyl hydroperoxide reductase (*ahpCF*) in FP. Enzymes with specific substrates, such as methane monooxygenase (*pmoA*), propane 2-monooxygenase (*prm*), and butane monooxygenase hydroxylase (*bmo*), are also included in HADEG.

Some other enzymes included in HADEG are alcohol and aldehyde dehydrogenases (*adh*, *ald*), cyclohexanone 1,2-monooxygenase (CHMO), alkanesulfonate monooxygenase (*ssuD*), the helix-turn-helix transcriptional regulator of *alkB* (*alkS*), and enzymes related to the uptake of alkanes, such as an outer membrane protein (*alkL*), an outer membrane lipoprotein (*blc*), long-chain fatty acid transport enzyme (*fadL*), and a methyl-accepting chemotaxis protein (*alkN* or *mcp*).

3.1.2 Alkenes

Alkenes are transformed into corresponding epoxides by an alkene monooxygenase. Epoxides are somewhat toxic to cells but can be processed using epoxide hydrolases, such as the enzyme epoxyalkane CoM transferase, and their products ultimately enter the central metabolism. HADEG includes the alkene monooxygenase complex (*amoABCD*) identified in *Nocardia corallina* B-276, isoprene monooxygenase (*isoABCDEHI*) from *Rhodococcus* sp. AD45, the propene monooxygenase system (*xamoABCDE*) and epoxyalkane CoM system (*xecADC*) from *Xanthobacter autotrophicus* Py2, the epoxyalkane CoM (*etnE*) from *Mycolicibacterium rhodesiae*, and the dehydrogenase *mpdBC* from *M. austroafricanum*.

3.1.3 Aromatics

Aromatic compounds are commonly known as recalcitrant pollutants because they show relatively low susceptibility to microbial degradation (Ghosal et al., 2016). Some also represent mutagenic and carcinogenic risks to human health. In contrast to alkanes and alkenes, aromatic compound degradation requires complex metabolic machinery. In general, aerobic organisms take advantage of the availability of molecular oxygen and use oxygenases that introduce hydroxyl groups and cleave aromatic rings. These pathways usually involved a few central intermediates, most commonly including catechol, protocatechuate, and gentisate (Phale et al., 2020).

The goal of enzymatic degradation is the activation and cleavage of aromatic rings to obtain products that can subsequently be integrated into central pathways. Our research included pathways with characterized enzymes involved in the metabolic processes of poly- and monoaromatic degradation. Regarding polyaromatic degradation pathways, HADEG includes enzymes involved in the degradation of anthracene, biphenyl, dibenzothiophene, fluorene, naphthalene, and phenanthrene. Pyrene degradation involves pyrene dioxygenases, but none of these enzymes had been experimentally characterized when this manuscript was written, although their activity has been inferred in *Pseudomonas* sp. Jpyr-1 and *Mycolicibacterium* sp. 16F. Nevertheless, cytochrome P450 monooxygenase (CYP450), which is involved in alkane degradation, seems to participate in pyrene degradation in *Roseobacter* (Zhou et al., 2020). Regarding monoaromatic degradation pathways, we included enzymes involved in the degradation of benzene, toluene, xylene (BTX), anthranilate, benzoate, 2-nitrobenzoate, catechol (ortho- and meta-cleavage pathways), gentisate, phenol, phenylacetate, phthalate, protocatechuate (ortho- and meta-cleavage pathways), salicylate, styrene, and terephthalate.

3.1.4 Polymers

Enzymes involved in the degradation of polymers were also added to the HADEG database, since polymers are important environmental contaminants, and it has been reported that some microorganisms can degrade them (Kaushal et al., 2021). Polymers can be classified according to their origin into natural or synthetic polymers. Natural polymers, also known as biopolymers, are derived from plants, animals, and microorganisms, are degradable and environmentally sensitive, and have been studied for potential medical and engineering applications (Kumbar et al., 2014; Shrivastava, 2018; Swain & Jawaid, 2019; Wagner et al., 2020); (Garcia-Gareta, 2019). Some natural polymers are essential for life, such as polysaccharides (e.g., cellulose, pectin, chitin, and alginic acid), polypeptides (e.g., proteins), and polynucleotides (e.g., DNA). On the other hand, synthetic or man-made polymers are obtained in the laboratory (Saldívar-Guerra & Vivaldo-Lima, 2013), and they can have an organic or inorganic backbone (Zainudin et al., 2020). They are commonly derived from petroleum (e.g., polyethylene terephthalate-PET, polyurethane-PU, polyvinyl chloride-PVC, polystyrene-PS, and polypropylene-PP) and have many applications (Shrivastava, 2018). Most of them are very difficult to degrade due to inherent characteristics such as high molecular weights, crystallinity, and hydrophobicity (Amobonye et al., 2021; Geyer, 2020). In this context, microorganisms with the capacity for polymer degradation remain one of the most promising options for the reduction or recycling of synthetic polymers with long lifetimes (Muriel-Millán et al., 2021).

The polymer degradation process begins with the superficial deterioration and subsequent cleavage of polymers into intermediates, which can be used as energy and carbon sources. The depolymerization process mainly involves the hydrolytic cleavage of glycosidic, ester, and peptide bonds, resulting in constituent oligomers or monomers (Amobonye et al., 2021). Implementing physicochemical pretreatments such as the application of chemical additives or UV irradiation can facilitate microbial polymer degradation (Ru et al., 2020).

Among the polymer enzymes included in HADEG, there are some hydrolases for PET degradation, such as the recently well-characterized polyester hydrolase of *Halopseudomonas aestusnigri* VXGO4 (6SCD PDB), one hydrolase from a marine metagenome (Hajighasemi et al., 2018), a lacasse of *Rhodococcus ruber* C208, and some

cutinases. These enzymes also include PVA dehydrogenase (*pvaA, pvaB*) for polyvinyl alcohol (PVA) degradation (Shimao, 2001); exo-cleaving and endo-cleaving rubber dioxygenases (*roxA, roxB*) for natural rubber degradation from *Xanthomonas* sp. 35Y; polyurethanases (*pueA, pueB*) for PU degradation from *Pseudomonas choloraphis* and *Pseudomonas protegens* Pf-5; manganese peroxidase 1 (*mnp1*) for polyethylene degradation from *Phanerodontia chrysosporium*; and a cutinase for poly(epsilon-caprolactone) (PCL) and poly(1,3-propylene adipate) (PPA) degradation from *Kineococcus radiotolerans* DSM 14245.

3.2 Biosurfactant production

The first step in HC degradation is accessing the molecules. Short-chain alkanes of less than 9 carbons are soluble enough to be transported into cells (Moreno & Rojo, 2017; Wang & Shao, 2013) Alkanes of greater length require cell adhesion to crude oil droplets, as is the case in *Acinetobacter*, which uses adhesion protein filaments (fimbriae) or the action of biosurfactants (Baldi et al., 1999). Biosurfactants increase the bioavailability of HCs (Patowary et al., 2016), and it is speculated that once they are excreted, they cannot enter cells (Wang & Shao, 2013). Some microorganisms capable of degrading HCs secrete biosurfactants of a different nature, such as lipoproteins or rhamnolipids, to emulsify HCs. Biosurfactants have also been associated with viral factors (Sharma et al., 2021).

The HADEG database includes the enzymes involved in the production of fengycin (*ffp*), plipastatin (*ppsABCDE*), surfactin (*srf*), iturin (*itu*), and mycosubtilin (*mycABC*) in *Bacillus subtilis*. Additionally, enzymes related to rhamnolipids (*rhl*) from *Pseudomonas aeruginosa*; emulsan from *Acinetobacter lwoffii* RAG-1 (*wzb, wzc*); arthrofactin (*arfABC*) from *Pseudomonas* sp. MIS38; amphisin (*amsY*) from *Pseudomonas* sp. DSS73; serrawetin (*pswP*) in *Serratia marcescens*; and lactonic sophorose lipid (*sble*), and sophorolipids (*at*) from *Starmerella bombicola* are also included. Two fungal biosurfactants are also included; mannosylerythritol lipids (*emt1*) from *Ustilago maydis* and hydrophobin (*hfb1, hfb2*) from *Hypocrea jecorina*.

3.3 Taxonomic distribution of orthologs in the HADEG database

Among the proteins present in HADEG, a search for orthologs was carried out to determine their distribution in different microorganisms and their taxonomic assignment. We used eggNOG-mapper to identify 19,749 orthologs of 445 protein sequences among the 451 sequences included in HADEG (Supplementary Table 6). The six proteins without identified orthologs were the PBAT esterase of an uncultured bacterium (accession no. A0A1C9T7G6), the PVA hydrolase of *Sphingomonas sp.* (accession no. Q588Z2, *oph*), the butane monooxygenase hydroxylase gamma subunit of *Thauera butanivorans* (a.n. Q8KQE7, *bmoZ*), the fluoren-9-ol dehydrogenase of *Terrabacter* sp. DBF63 (a.n. Q93UV4, *flnB*), the oxidized PVA hydrolase of *Pseudomonas* sp. (a.n. Q9LCQ7, *pvaB*), and the anthranilate-CoA ligase of *Rhodococcus erythropolis* (a.n. S6BVH5, *rauF*).

The taxonomic analysis of orthologs at the phylum level revealed that Proteobacteria was the most abundant phylum related to HC degradation pathways and biosurfactant production (forest-green color in Fig. 2A). Many studies have confirmed the degradation capacities of species in genera such as *Alcanivorax* and *Marinobacter* (hydrocarbonoclastics), *Acinetobacter* (versatile HC degraders), and ubiquitous *Pseudomonas*. Biosurfactants such

as amphisin, arthrobactin, emulsan and serrawetin have only been described in Proteobacteria (Soberón-Chávez, 2010), and rhamnolipid production mechanisms have been well described in *Pseudomonas* (Soberón-Chávez et al., 2021) (Fig. 2B). Actinobacteria was the second most abundant phylum related to HC degradation pathways (goldenrod color in Fig. 2), and it includes the hydrocarbon-degrading genera *Rhodococcus, Mycobacterium, Norcadia, Gordonia*, and *Streptomyces*, among others. Firmicutes was the third most abundant phylum related to HC pathways and the second most abundant related to biosurfactant production (*Bacillus, Geobacillus,* and *Paenibacillus*) (royal blue color in Fig. 2), within which surfactin and iturin production by *Bacillus* genus members has been well studied (Soberón-Chávez, 2010). In the Bacteria domain, the fourth most abundant phylum was Bacteroidetes (*Aequorivita, Algoriphagus, Cytophaga*), particularly related to alkene degradation (orange color in Fig. 2). Ascomycota was the most abundant phylum from the fungal domain and was mainly associated with alkane and polymer degradation and biosurfactant production (brick color in Fig. 2). Fungi have received interest in relation to bioremediation applications because they do not rely solely on soluble organic compounds for nutrition; they also show diverse enzymatic mechanisms and distinct mechanisms of physiological adaptability to environmental conditions (Prenafeta-Boldú et al., 2018).

3.4 HADEG Pfams network

A network analysis was performed to determine whether the HADEG proteins shared domains, which may indicate that a protein could participate in the degradation of more than one type of HC (Fig. 3). Identifying the shared and nonshared protein domains between hydrocarbon degradation groups will be helpful to establish a set of biomarkers for the targeted searching of potential hydrocarbon-degrading microorganisms. A total of 166 Pfam domains were identified in HADEG (Supplementary Table 7). Network analyses reveal that the polymer and aromatic degradation groups contained the largest numbers of protein domains, at 75 and 48, respectively. These results are congruent with the diversity and complexity of aromatic and polymeric chemical structures. On the other hand, we predicted 30 Pfams among biosurfactants, 28 among alkanes, and 18 among alkenes.

Among the protein domains without connectivity or nonshared protein domains, the alkane set included 16 Pfams such as the related to rubredoxin proteins (PF00301 and PF18113) and a flavin-binding monooxygenase-like protein (PF00743) in *almA* and BVMO. The alkene set included four Pfams with no connectivity: cobalamin-independent synthase (PF01717) in *etnE*, a glutathione S-transferase C-terminal domain (PF17171 and PF17172) in *isoI*, and a pyridine nucleotide-disulfide oxidoreductase dimerization domain (PF02852) in *xecC*. The aromatic set included 57 nonshared Pfams like the ring-hydroxylating alpha and beta subunits (PF00848 and PF00866) present in aromatic-ring dioxygenase systems, the dioxygenase superfamily (PF00903), and the catechol dioxygenase N-terminus (PF04444). The polymer set included 42 nonshared protein domains like the RTX calcium-binding nonapeptide repeat 4 copies (PF00353) present in the polyester polyurethanases *pueA* and *pueB*, the alpha/beta hydrolase family (PF12695) involved in PET degradation, the esterase PHB depolymerase (PF10503) involved in PHA/PHB degradation and a cutinase (PF01083) that acts on carboxylic ester bonds. Finally, the biosurfactant set included 26 nonshared Pfams like a condensation domain (PF00668) present in *Bacillus* biosurfactants synthesis proteins (iturin, mycosubtilin, surfactin, and plipastatin), the 4'-phosphopantetheinyl transferase superfamily (PF01648) in *ffp, pswP*

and *sfp*, and helicase domains (PF00270 and PF00271) in *rhlB* and (PF13641) in *rhlC*, both involved in rhamnolipids production.

Concerning the protein domains that were connected in the network or shared, only the aldehyde dehydrogenase family (PF00171) participates in the degradation of the four HC groups. This domain is responsible for the oxidation of aldehydes, which is considered a detoxification reaction (*Pfam: Family: Aldedh (PF00171)*, n.d.). The largest set of shared domains was found between alkenes and aromatics, which shared 19 Pfams. These domains included the Rieske [2Fe-2S] domain (PF00355), which is involved in the electron transfer chain. PF00355 was present in enzymes such as *andA, benA, doxAB, ndoAB, tmoC, todB, bedB, isoC*, and *xamoC*. The alkane, alkene and aromatic groups shared seven Pfam domains, including the oxidoreductase NAD-binding and FAD-binding domains (PF00175 and PF00970), present in *prmB, isoF, xamoF, benC,* and *tmoF*; the methane/phenol/toluene hydroxylase domains (PF02332), found in *bmoX, xamoA*, and *tmoA*; and the MmoB/DmpM family (PF02406), found in *prmD, xamoD*, and *tmoD*. The alkane and aromatic groups shared a fatty acid desaturase domain (PF00487) that catalyzes the insertion of a double bond at the delta position in fatty acids (*Pfam: Family: FA_desaturase (PF00487)*, n.d.). This domain is present in alkane monooxygenases (*alkB*), alkane hydroxylases (*alkMa, alkMb*), and xylene monooxygenase subunit 1 (*xylM*).

3.5 Hydrocarbon degradation enzymes in different genomes

One way to use this database could be to identify hydrocarbon degradation genes in different genomes based on homology through BLAST searches, although other methods can be used for this purpose. To this end, a BLAST comparison was conducted using the amino acid sequences of HADEG against 60 genomes with previously proven hydrocarbon degradation and/or biosurfactant production activity. The results of this comparison showed that 53 of 60 genomes presented experimentally verified hydrocarbon degradation genes; thus, the data were consistent with previous reports (Fig. 4). Among the remaining 7 genomes, three exhibited experimental evidence hits with a percentage of identity between 44% and 49%, which was below the cutoff (50%). This was the case for the BVMO gene of *Thalassolituus oleivorans* MIL-1 involved in the SO pathway, *pegAC* of *Sphingopyxis macrogoltabida* 203 involved in polyethylene glycol degradation, and *rhlABC* of *Burkholderia thailandensis* E264 involved in rhamnolipid synthesis. Although genomes (11, 28, 42, and 50) showed some hits against the HADEG database, they did not show matches with any genes or substrates with previously reported experimental results. These genomes likely include other potential pathways that are not covered in the HADEG database.

Additionally, it was possible to identify unreported hydrocarbon degradation and biosurfactant production capabilities in 59 genomes with HADEG. Many of these genomes included genes for the degradation of aliphatic and aromatic HCs and polymers and the production of multiple biosurfactants on the same chromosome. The occurrence of these numerous pathways might be due to microbial environmental adaptations and be widely observed within the bacterial world (Brzeszcz & Kaszycki, 2018). However, empirical evidence indicates that microbes show metabolic pathway preferences (Brzeszcz & Kaszycki, 2018). The concordance with experimental evidence and the predicted pathway repertoire show that the HADEG database is a reliable resource for identifying and predicting hydrocarbon degradation pathways and biosurfactant production in genomes.

4 Conclusions

The HADEG database is a bioinformatic resource that can be used to identify potential microorganisms for developing HC-bioremediation methods. To the best of our knowledge, HADEG, which contains 451 experimentally verified proteins, is the first database to integrate different types of HCs (alkane, alkenes, aromatics, and polymers) and biosurfactants in a single resource. Through ortholog analysis, we found that the HADEG proteins are distributed in bacteria and fungi and that they are most highly represented in Proteobacteria. Additionally, Pfam analysis showed that the domains of the HC degradation proteins correspond primarily to polymers and aromatics. Most domains are not shared among HC groups, indicating that their functions are HC-degradation-specific and that they can probably be used as biomarkers for the identification of microbes with degradation abilities. Finally, we show that HADEG is a reliable tool for annotating HC degradation and biosurfactant production proteins.

Author contributions

JR-V conceived the initial study. JR-V and HGC-S designed the experiments and analyzed the data. JR-V, HGC-S, and LP-L contributed to the interpretation and discussion of the results. LP-L provided financial support. All authors contributed to the writing of the manuscript and approved the final version.

Funding

Jorge Rojas-Vargas is a doctoral student from Programa de Doctorado en Ciencias Bioquímicas, Universidad Nacional Autónoma de México (UNAM), and received fellowship 965003 from the Consejo Nacional de Ciencia y Tecnología (CONACYT). Liliana Pardo-Lopez received financial support from PASPA DGAPA-UNAM for a sabbatical year.

Acknowledgments

We thank the Unidad de Secuenciación Masiva y Bioinformática del Instituto de Biotecnología-UNAM for giving us access to its computer cluster.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Amobonye, A., Bhagwat, P., Singh, S., & Pillai, S. (2021). Plastic biodegradation: Frontline microbes and their enzymes. *The Science of the Total Environment*, *759*, 143536.

Baldi, F., Ivošević, N., Minacci, A., Pepi, M., Fani, R., Svetličić, V., & Žutić, V. (1999). Adhesion of Acinetobacter venetianus to diesel fuel droplets studied with in situ electrochemical and molecular probes. *Applied and Environmental Microbiology*, *65*(5), 2041–2048.

Banat, I. M. (1995). Biosurfactants production and possible uses in microbial enhanced oil recovery and oil pollution remediation: A review. *Bioresource Technology*, *51*(1), 1–12.

Bhattacharjee, S., & Dutta, T. (2022). Chapter 1 - An overview of oil pollution and oil-spilling incidents. In P. Das, S. Manna, & J. K. Pandey (Eds.), *Advances in Oil-Water*

Separation (pp. 3–15). Elsevier.
Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. In D. Edwards (Ed.), *Plant Bioinformatics: Methods and Protocols* (pp. 89–112). Humana Press.
Brzeszcz, J., & Kaszycki, P. (2018). Aerobic bacteria degrading both n-alkanes and aromatic hydrocarbons: an undervalued strategy for metabolic diversity and flexibility. *Biodegradation*, *29*(4), 359–407.
Buchholz, P. C. F., Feuerriegel, G., Zhang, H., Perez-Garcia, P., Nover, L.-L., Chow, J., Streit, W. R., & Pleiss, J. (2022). Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database-PAZy. *Proteins*, *90*(7), 1443–1456.
Duarte, M., Jauregui, R., Vilchez-Vargas, R., Junca, H., & Pieper, D. H. (2014). AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database: The Journal of Biological Databases and Curation*, *2014*, bau118.
Gambarini, V., Pantos, O., Kingsbury, J. M., Weaver, L., Handley, K. M., & Lear, G. (2022). PlasticDB: a database of microorganisms and proteins linked to plastic biodegradation. *Database: The Journal of Biological Databases and Curation*, *2022*. https://doi.org/10.1093/database/baac008
Gan, Z., & Zhang, H. (2019). PMBD: a Comprehensive Plastics Microbial Biodegradation Database. *Database: The Journal of Biological Databases and Curation*, *2019*. https://doi.org/10.1093/database/baz119
Garcia-Gareta, E. (2019). *Biomaterials for Skin Repair and Regeneration*. Elsevier Science.
Geyer, R. (2020). Chapter 2 - Production, use, and fate of synthetic polymers. In T. M. Letcher (Ed.), *Plastic Waste and Recycling* (pp. 13–32). Academic Press.
Ghosal, D., Ghosh, S., Dutta, T. K., & Ahn, Y. (2016). Current State of Knowledge in Microbial Degradation of Polycyclic Aromatic Hydrocarbons (PAHs): A Review. *Frontiers in Microbiology*, *7*, 1369.
Hajighasemi, M., Tchigvintsev, A., Nocek, B., Flick, R., Popovic, A., Hai, T., Khusnutdinova, A. N., Brown, G., Xu, X., Cui, H., Anstett, J., Chernikova, T. N., Brüls, T., Le Paslier, D., Yakimov, M. M., Joachimiak, A., Golyshina, O. V., Savchenko, A., Golyshin, P. N., … Yakunin, A. F. (2018). Screening and Characterization of Novel Polyesterases from Environmental Metagenomes with High Hydrolytic Activity against Synthetic Polyesters. *Environmental Science & Technology*, *52*(21), 12388–12401.
Horton, A. A. (2022). Plastic pollution: When do we know enough? *Journal of Hazardous Materials*, *422*, 126885.
Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314.
Ji, Y., Mao, G., Wang, Y., & Bartlam, M. (2013). Structural insights into diversity and n-alkane biodegradation mechanisms of alkane hydroxylases. *Frontiers in Microbiology*, *4*, 58.
Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* , *30*(9), 1236–1240.
Kaushal, J., Khatri, M., & Arya, S. K. (2021). Recent insight into enzymatic degradation of plastics prevalent in the environment: A mini - review. *Cleaner Engineering and*

Technology, *2*, 100083.
Khot, V., Zorz, J., Gittins, D. A., Chakraborty, A., Bell, E., Bautista, M. A., Paquette, A. J., Hawley, A. K., Novotnik, B., Hubert, C. R. J., Strous, M., & Bhatnagar, S. (2021). CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Frontiers in Microbiology*, *12*, 764058.
Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y., & Karp, P. D. (2004). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, *32*(Database issue), D438-42.
Kumbar, S., Laurencin, C., & Deng, M. (2014). *Natural and Synthetic Biomedical Polymers*. Elsevier Science.
Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(Database issue), D28-31.
McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(Web Server issue), W20-5.
Moreno, R., & Rojo, F. (2017). Enzymes for aerobic degradation of alkanes in bacteria. In *Aerobic Utilization of Hydrocarbons, Oils and Lipids* (pp. 1–25). Springer International Publishing.
Moreno, R., & Rojo, F. (2019). Enzymes for Aerobic Degradation of Alkanes in Bacteria. In F. Rojo (Ed.), *Aerobic Utilization of Hydrocarbons, Oils, and Lipids* (pp. 117–142). Springer International Publishing.
Muriel-Millán, L. F., Millán-López, S., & Pardo-López, L. (2021). Biotechnological applications of marine bacteria in bioremediation of environments polluted with hydrocarbons and plastics. *Applied Microbiology and Biotechnology*, *105*(19), 7171–7185.
Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34.
Oliveira, J. S., Araújo, W., Lopes Sales, A. I., Brito Guerra, A. de, Silva Araújo, S. C. da, de Vasconcelos, A. T. R., Agnez-Lima, L. F., & Freitas, A. T. (2015). BioSurfDB: knowledge and algorithms to support biosurfactants and biodegradation studies. *Database: The Journal of Biological Databases and Curation*, *2015*. https://doi.org/10.1093/database/bav033
Pathak, U., Jhunjhunwala, A., Singh, S., Bajaj, N., & Mandal, T. (2022). Chapter 19 - Potentiality of enzymes as a green tool in degradation of petroleum hydrocarbons. In P. Das, S. Manna, & J. K. Pandey (Eds.), *Advances in Oil-Water Separation* (pp. 337–351). Elsevier.
Patowary, K., Patowary, R., Kalita, M. C., & Deka, S. (2016). Development of an Efficient Bacterial Consortium for the Potential Remediation of Hydrocarbons from Contaminated Sites. *Frontiers in Microbiology*, *7*, 1092.
Pfam: Family: Aldedh (PF00171). (n.d.). Retrieved July 31, 2022, from http://pfam.xfam.org/family/pf00171
Pfam: Family: FA_desaturase (PF00487). (n.d.). Retrieved July 31, 2022, from http://pfam.xfam.org/family/PF00487
Phale, P. S., Malhotra, H., & Shah, B. A. (2020). Chapter One - Degradation strategies and associated regulatory mechanisms/features for aromatic compound metabolism in

bacteria. In G. M. Gadd & S. Sariaslani (Eds.), *Advances in Applied Microbiology* (Vol. 112, pp. 1–65). Academic Press.

Prenafeta-Boldú, F. X., de Hoog, G. S., & Summerbell, R. C. (2018). Fungal Communities in Hydrocarbon Degradation. In T. J. McGenity (Ed.), *Microbial Communities Utilizing Hydrocarbons and Lipids: Members, Metagenomics and Ecophysiology* (pp. 1–36). Springer International Publishing.

Prince, R. C., Gramain, A., & McGenity, T. J. (2010). Prokaryotic Hydrocarbon Degraders. In K. N. Timmis (Ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (pp. 1669–1692). Springer Berlin Heidelberg.

Ru, J., Huo, Y., & Yang, Y. (2020). Microbial Degradation and Valorization of Plastic Wastes. *Frontiers in Microbiology*, *11*, 442.

Saldívar-Guerra, E., & Vivaldo-Lima, E. (2013). Introduction to polymers and polymer types. In *Handbook of Polymer Synthesis, Characterization, and Processing* (pp. 1–14). John Wiley & Sons, Inc.

Sharma, J., Sundar, D., & Srivastava, P. (2021). Biosurfactants: Potential Agents for Controlling Cellular Communication, Motility, and Antagonism. *Frontiers in Molecular Biosciences*, *8*, 727070.

Shimao, M. (2001). Biodegradation of plastics. *Current Opinion in Biotechnology*, *12*(3), 242–247.

Shrivastava, A. (2018). *Introduction to Plastics Engineering*. Elsevier Science.

Soberón-Chávez, G. (2010). *Biosurfactants: From Genes to Applications*. Springer Berlin Heidelberg.

Soberón-Chávez, G., González-Valdez, A., Soto-Aceves, M. P., & Cocotl-Yañez, M. (2021). Rhamnolipids produced by Pseudomonas: from molecular genetics to the market. *Microbial Biotechnology*, *14*(1), 136–146.

Swain, S. K., & Jawaid, M. (2019). *Nanostructured Polymer Composites for Biomedical Applications*. Elsevier Science.

Wagner, W., Sakiyama-Elbert, S., Yaszemski, M., & Zhang, G. (2020). *Biomaterials Science: An Introduction to Materials in Medicine*. Elsevier Science & Technology.

Wang, W., & Shao, Z. (2013). Enzymes and genes involved in aerobic alkane degradation. *Frontiers in Microbiology*, *4*, 116.

Wentzel, A., Ellingsen, T. E., Kotlar, H.-K., Zotchev, S. B., & Throne-Holst, M. (2007). Bacterial metabolism of long-chain n-alkanes. *Applied Microbiology and Biotechnology*, *76*(6), 1209–1221.

Xu, X., Liu, W., Tian, S., Wang, W., Qi, Q., Jiang, P., Gao, X., Li, F., Li, H., & Yu, H. (2018). Petroleum Hydrocarbon-Degrading Bacteria for the Remediation of Oil Pollution Under Aerobic Conditions: A Perspective Analysis. *Frontiers in Microbiology*, *9*, 2885.

Zainudin, B. H., Wong, T. W., & Hamdan, H. (2020). Chapter 8 - Pectin as oral colon-specific nano- and microparticulate drug carriers. In M. A. A. AlMaadeed, D. Ponnamma, & M. A. Carignano (Eds.), *Polymer Science and Innovative Applications* (pp. 257–286). Elsevier.

Zhou, H., Zhang, S., Xie, J., Wei, H., Hu, Z., & Wang, H. (2020). Pyrene biodegradation and its potential pathway involving Roseobacter clade bacteria. *International Biodeterioration & Biodegradation*, *150*, 104961.

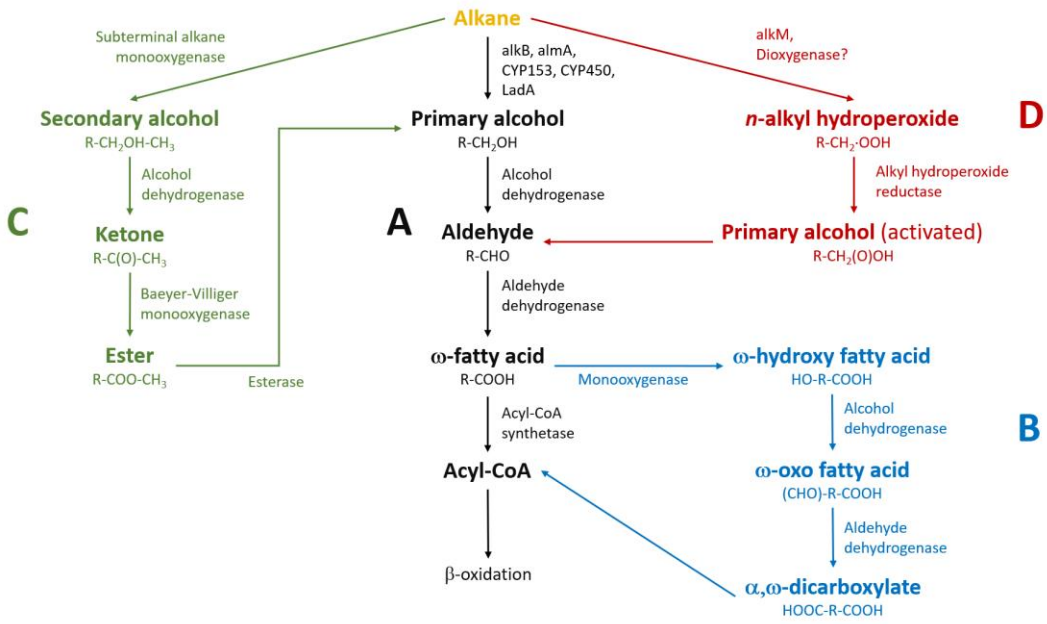
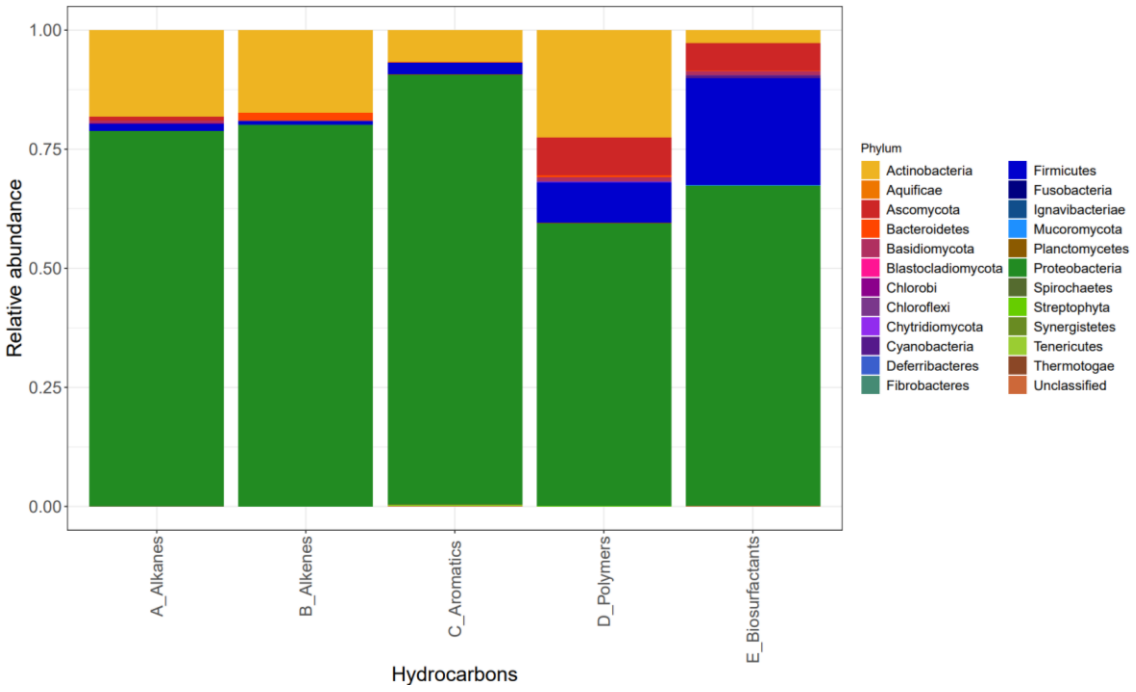


Fig. 1. Alkane aerobic degradation pathways. **A.** Terminal oxidation (black color). **B.** Biterminal oxidation (blue). **C.** Subterminal oxidation (green). **D.** Finnerty pathway (red).

A



B

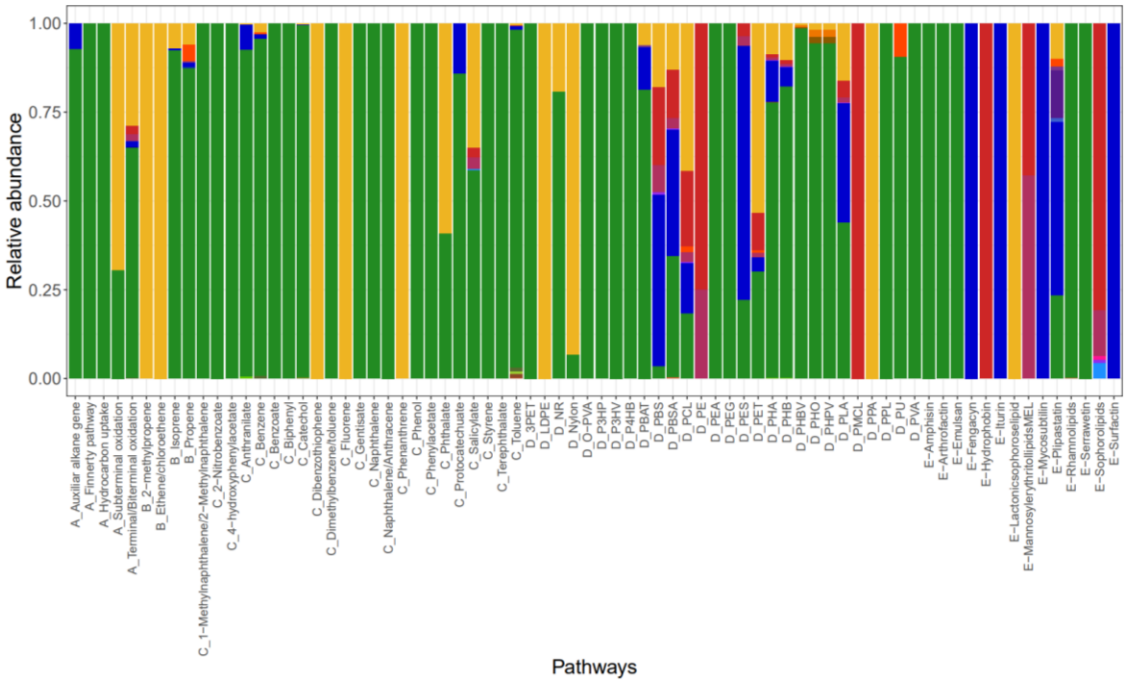


Fig. 2. Taxonomic distribution of enzyme orthologs obtained with eggNOG-Mapper. **A.** Hydrocarbon groups. **B.** Degradation pathways. A_alkane groups, B_alkenes, C_aromatics, D_polymers, E_biosrufactants.

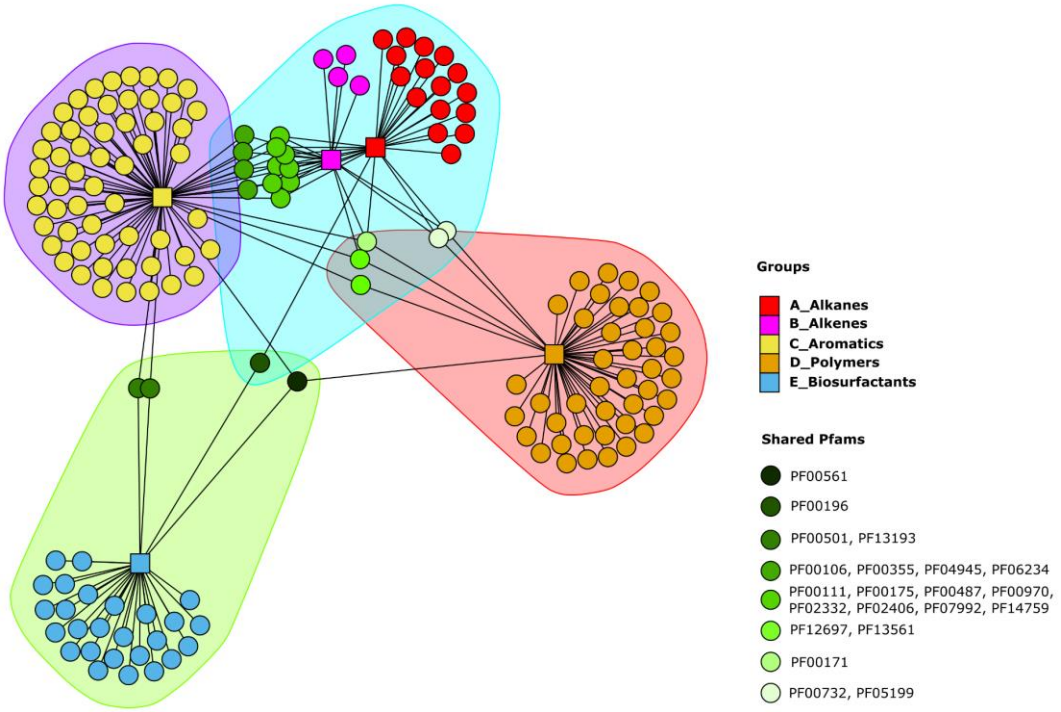


Fig. 3. HADEG Pfam network. Hydrocarbon degradation groups and biosurfactant production are represented by squared: alkanes (red), alkenes (magenta), aromatics (yellow), polymers (orange), and biosurfactants (blue). Pfams are represented by circles, and the color convention is as described above. The shared protein domains among groups are shown on green scale. The list of the Pfams is shown in Supplementary Table 7.

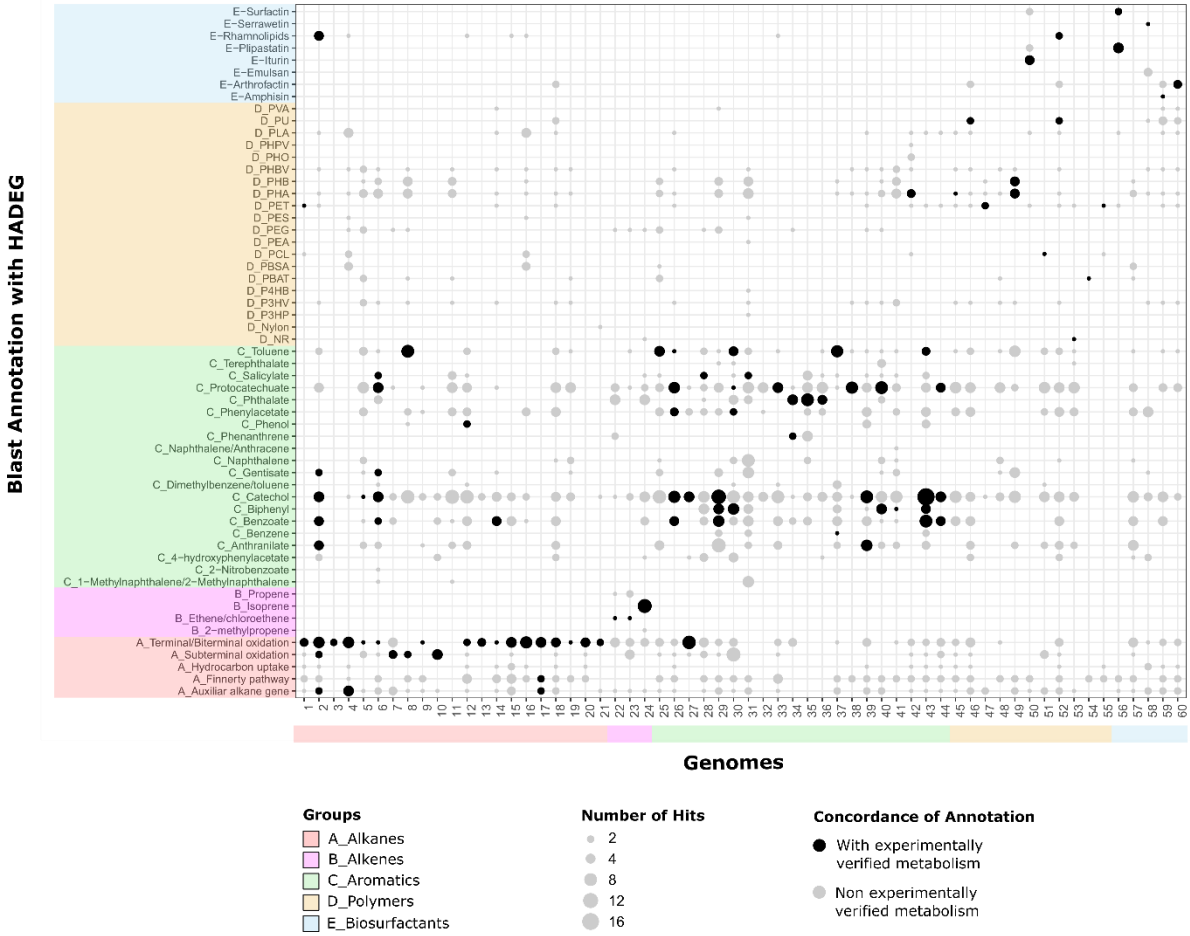
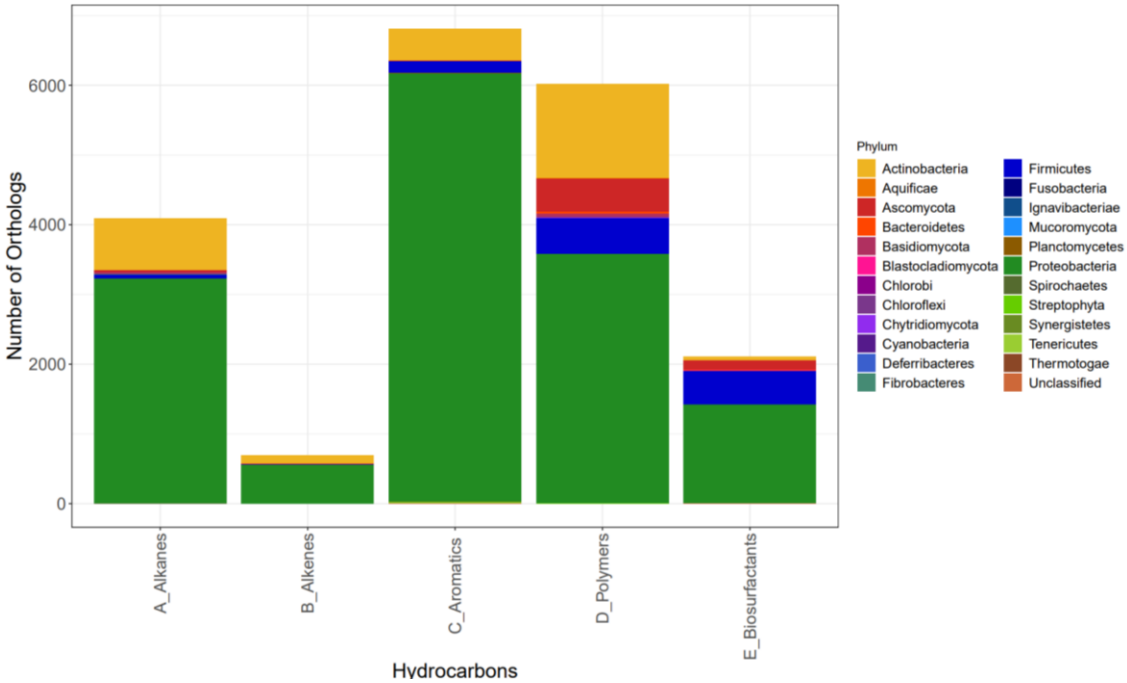
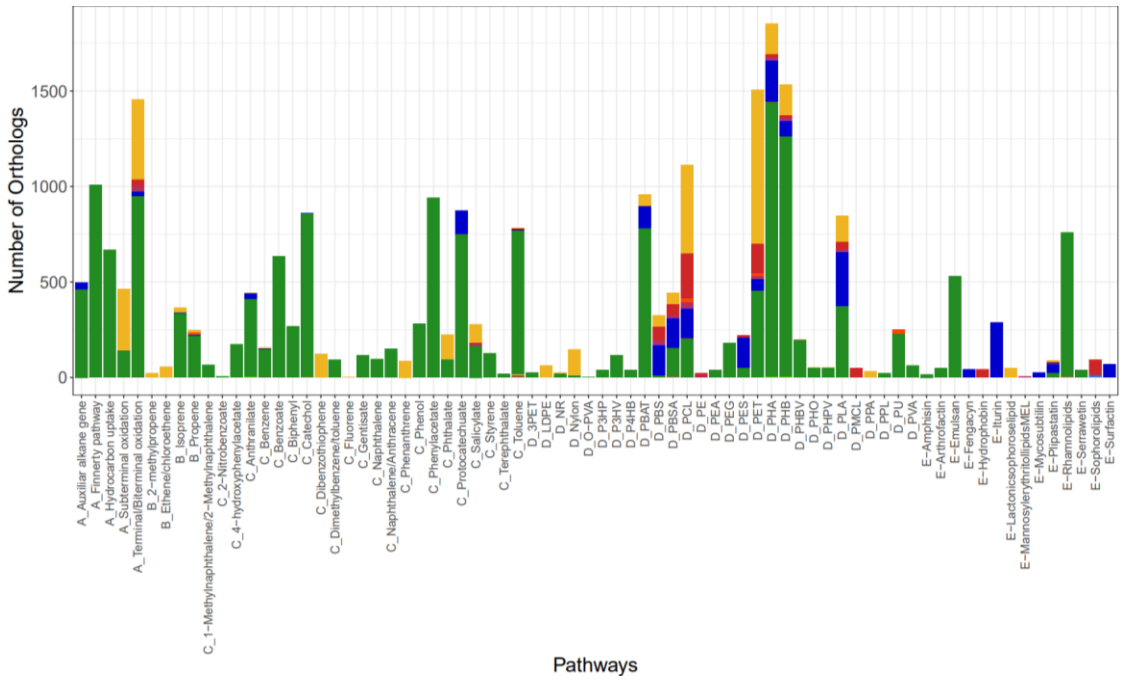


Fig. 4. BLAST annotation of genomes with the HADEG database and experimentally validated activities. The corresponding hydrocarbon groups are used to organize both the genomes (x-axis) and the HADEG pathways (y-axis). Bubble indicates hits above the confidence threshold. The hits with experimental evidence are shown in black, and the predicted hits are shown in gray. The number of hits in the genome is indicated by the size of the bubble. Supplementary Table 8 contains a complete list of isolated genomes and their accession numbers.

A



B



Supplementary Fig. 1. Orthologous proteins obtained with eggNOG-mapper. **A.** Hydrocarbon groups. **B.** Degradation pathways. A_alkane groups, B_alkenes, C_aromatics, D_polymers, E_biosrufactants.