Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-863X
Doi: 10.4025/actascibiolsci.v47i1.71881

Environmental distribution of polyhydroxyalkanoate (PHA) production in the domain Archaea

Mónica Alejandra Rodríguez Aristizabal[1], Natalia Elizabeth Conde Martínez[2] and Luis Alejandro Acosta González[2]*

[1]Facultad de Ingeniería y Ciencias Básicas, Universidad Central, Bogotá, Colombia. [2]Facultad de Ingeniería, Universidad de La Sabana, Chía, Colombia. *Author for correspondence. E-mail: alejandro.acosta1@unisabana.edu.co

ABSTRACT. Polyhydroxyalkanoates (PHAs) are polyesters synthesised by prokaryotes as a carbon and energy storage strategy. Due to their properties and versatility, these biopolymers are a potential alternative to conventional low-caliber plastics. However, the cost of the process, particularly the carbon sources used as raw materials, limits their large-scale production. The study of new production strains isolated from underexplored sites may be a viable option to improve production capacity. The genes encoding PHA production are organized in the phaCAB operon, with phaC being the key enzyme for polymerization. In this study, we analyzed the environmental distribution of the phaC gene in Archaea using bioinformatic tools to demonstrate the relevance of searching for archaeal strains for PHA production. We searched NCBI for PhaC synthase protein sequences in cultured Archaea and metagenomes. We found 176 sequences of PhaC synthases in cultured Archaea and 66 in metagenomic proteins. Twenty environmental categories were defined based on the associated environmental information. No changes were necessary to ensure grammatical correctness. PhaC genes were found in 41 archaeal genera and 7 possible genus candidates, with *Nitrosopumilus* sp. being the most abundant genus. The distribution of phaC genes was mainly associated with sediments and marine environments, with less presence in soil niches. These results contribute to the knowledge of the taxonomic diversity and habitats where the phaC gene is present in Archaea with potential for polyhydroxyalkanoate production.

Keywords: Archaea; phaC; PhaC synthase; extremophiles; bioinformatics; gene diversity.

Received on April 8, 2024
Accepted on May 9, 2025

Introduction

Polyhydroxyalkanoates (PHAs) are linear polyesters synthesized by various bacteria and Archaea to store carbon as a source of energy. They are synthesized from sugars, fatty acids, simple carbon sources such as methanol, CO_2, and methane, as well as complex substrates including plant biomass, cellulose, lignin, and agro-industrial biowaste (Adeleye et al., 2020; El-malek et al., 2020; Fradinho et al., 2014; Kumar et al., 2020; Li & Wilkins, 2020; Obruča et al., 2022). Based on their chemical structure, PHAs can be classified into three types: short-chain-length monomers (scl-PHA) with 3 to 5 carbon atoms, medium-chain-length monomers (mcl-PHA) with 6 to 14 carbon atoms, and PHAs with a mixture of monomers found in both scl-PHA and mcl-PHA in their polymeric chain (Adeleye et al., 2020; Cruz et al., 2016; Kumar et al., 2020). It is important to note the precise carbon atom range for each type of PHA. In 2016, Kumar et al. found that different physical, chemical, thermal, and mechanical properties are attributed to biopolymers due to the microorganisms that synthesize them and the types of substrates used for this purpose (Obruca et al., 2018; Quillaguaman et al., 2010). Due to their inherent plastic-like properties, PHAs have been given diverse applications, with a focus on packaging materials (bioplastics), pharmaceuticals, and tissue engineering. They offer a potential alternative to petrochemical-based plastics.

The metabolic pathways and genes associated with the production of PHAs have been extensively studied. In general, the carbon source is first converted into coenzyme A thioesters of (R)-hydroxyalkanoic acids. The condensation of two coenzyme A thioester molecules—typically acetyl-CoA and propionyl-CoA-is catalysed by β-ketothiolase. This is followed by a specific reduction of (R) to produce (R)-3-hydroxybutyryl-CoA or (R)-3-hydroxyvaleryl-CoA. These compounds undergo catalysis by an acetoacetyl-CoA reductase and are then converted by a PHA synthase into PHAs. This process has been extensively studied (Obruca et al., 2018; Rehm, 2003; Wang et al., 2019). The phaCAB operon contains the genes encoding the key enzymes involved in this reaction, namely phaA, phaB, and phaC. This operon is present in the genomes of various PHA-producing Archaea and bacteria. Specifically, phaA proteins function as acyl-CoA synthetases, catalyzing the conversion of fatty acids into their corresponding CoA-thioesters—the first step of biosynthesis. PhaB, the β-ketothiolase, catalyzes the condensation of acyl-CoA to form β-ketoacyl-CoA, a key biosynthetic precursor. PhaR regulates gene expression related to PHA synthesis (Laycock et al., 2013; Magagula et al., 2021). PhaC synthase is considered the key enzyme in condensing monomers and polymerizing final products. PhaC synthases are divided into four classes based on their primary structure deduced from amino acid sequences, substrate specificity, and the composition of the PHA synthase subunits (Neoh et al., 2022). Classes I and II comprise a single unit and two domains, primarily producing short-chain PHAs. Classes III and IV Pha synthases have two subunits and produce short- and medium-chain PHAs. The distinguishing subunits are encoded by phaE in the case of class III and by phaR in the case of class IV. These classes are widely distributed among microorganisms (Obruca et al., 2018; Neoh et al., 2022).

Archaea have class III Pha synthases, which require the functionality of both subunits: a catalytic subunit encoded by the phaC gene and another unit encoded by phaE. Furthermore, microorganisms with such synthases prefer to use the CoA thioesters of short-chain hydroxyalkanoates with 3-5 carbon atoms as substrates (Neoh et al., 2022). Class III phaC synthases belong to the PHA_synth_III_C superfamily (cl40647) according to the Clusters of Orthologous Groups (COG) database. This superfamily comprises conserved protein domains such as TIGR01836, which is specifically related to the phaE unit. Although the precise function of this domain is unknown, it is essential for the biosynthesis of ester polymers that accumulate as carbon and energy in the form of inclusions or granules, which polymerize to form short-chain hydroxyalkanoates (Lu et al., 2020; McCool & Cannon, 2001; Wang et al., 2023).

Polyhydroxyalkanoates are synthesized by approximately 20 genera of Archaea and 80 genera of Gram-positive and Gram-negative bacteria. Their biosynthesis involves three major genes: phaA, phaB, and phaC. Genome sequencing of these organisms has enabled phylogenetic and statistical analysis. Around 24 organisms can acquire and adapt genes from taxonomically distant relatives mainly through horizontal gene transfer (HGT) events. This strategy can help organisms adapt to adverse environmental conditions and may result in modified regulatory mechanisms or the evolution of new operons (Alamgeer, 2019; Choi et al., 2020; Kim et al., 2017).

Therefore, investigating the distribution of the phaC gene in Archaea and the diversity of their habitats is important for understanding their potential in PHA production. This can be achieved through bioinformatic analyses of phaC, phaA, phaB, phaP, phaR, and phaZ gene records (Wang et al., 2019). However, studies aiming to develop or compare mutant Pha synthases with varied substrate specificities to overcome challenges in biopolymer production have been found to cause loss of biosynthesis activity due to the lack of metabolic pathways in the mutant microorganisms and the highly conserved residues in Pha synthases. Therefore, it is more promising to search for microorganisms with a natural potential for polymer production (Lee & Kim, 2015; McCool & Cannon, 2001; Obulisamy & Mehariya, 2021).

Studies on PHA production have often focused on extreme environments, particularly those that are halophilic. It has been suggested that taxa containing extremophiles may have developed physiological strategies that involve PHA production (Vuong et al., 2021). However, recent research has shown that microorganisms from other environments, including clinical samples and hosts such as humans and animals, can also accumulate these polymers (Ji et al., 2021; Lee et al., 1994). To better understand the microbial genera associated with PHA production and their environmental diversity, we can rely on the results of genetic diversity studies based on sequencing data. Since the first metagenome-assembled genomes (MAG) provided evidence of the important functions of bacteria and Archaea in the world's ecosystems, these studies have become the basis to find specific information about the genetics and functions of microorganisms (Singleton et al., 2021). However, it is important to note that not all information derived from metagenomic studies is of good quality (Ji et al., 2021; Lee et al., 1994; Singleton et al., 2021).

The aim of this study was to analyze the environmental distribution of the phaC gene for PHA production in Archaea. Bioinformatic and statistical tools were used to compare and establish phylogenetic and environmental relationships. This contributes to the knowledge about the diversity of habitats where these microorganisms are found. This study also discusses the potential of biotechnological approaches to address current challenges in bioplastics production. Specifically, it highlights the advantages of using PHA produced by Archaea, which are more resistant to heat, pressure, and salinity. Additionally, their large-scale production reduces the risk of contamination, and they can use a variety of substrates as raw materials, making them highly suitable for bioprocessing (Chen et al., 2020; Obulisamy & Mehariya, 2021).

Material and methods

Search for culturable Archaea

The National Center for Biotechnology Information (NCBI) database was searched for sequences of PhaC protein synthases from culturable Archaea. The search was filtered using the following criteria: Identical Protein Group + phaC + Protein name and GENE + phaC + Find related data + Protein. The results were limited to Archaea only by including a filter for types of microorganisms. For this search, we downloaded data to create working matrices and identify clusters of similar sequences for future metagenome searches. These clusters were determined from the PhaC protein sequence alignment distance matrix in Archaea, using Mega X version 10.2.6. The inclusion criterion was a value < 0.5, indicating the similarity between sequences. The closer the value was to zero, the more similar the sequences were. The process of curating the data involved reviewing each record and excluding those that lacked information about the sample's origin environment.

Metagenome search

The KAUST Metagenomic Analysis Platform (available at https://www.cbrc.kaust.edu.sa/aamg/kmap.start/) was used to gather information from 47 databases distributed by biome. The search was performed by taking into account the clusters defined for the protein sequences of culturable Archaea. A representative sequence was selected from each cluster and analyzed using BLAST within the KAUST platform in the 47 databases to download metagenome sequences. To approximate the archaeal genera compatible with the metagenome sequences, BLASTp was performed on the NCBI platform with the following parameters: Database: Non-redundant protein sequences, Max number of sequences: 100 E-value threshold: 0.05, Matrix: BLOSUM62 Conditional compositional score matrix adjustment: Enabled To verify the relationship between the metagenome sequences and those found in the same environments, a second BLASTp analysis was performed using metagenomic proteins as the database parameters. During data curation, sequences with < 80% coverage or < 50% identity were excluded. Additionally, 100% identical sequences were removed to avoid redundancy.

Protein sequence alignments

Multiple sequence alignments were performed using Mega X (version 10.2.6), with the sequences of PhaC synthase proteins from cultivable Archaea and selected metagenomes. The representative sequences of class I-IV PhaC synthases in bacteria, as well as sequences obtained from the KAUST BLAST results, were included. The alignments were refined with Bioedit (version 7.2.6) to eliminate unaligned regions. Distance matrices were calculated and phylogenetic trees were constructed using Mega X (version 10.2.6).

Phylogenetic trees and distance matrices

Phylogenetic trees were constructed using the Neighbor-Joining algorithm in Mega X (version 10.2.6). The bootstrap consensus method was inferred from 10,000 replicates for greater precision. The sequence WP_004060157.1 (corresponding to *Haloferax mediterranei*) was used as the root for all trees. Three phylogenetic trees were created by comparing the aligned and edited sequences of culturable Archaea, representative bacterial PhaC synthases I-V, and metagenomic proteins. To analyze the sequences and establish their possible relationships, pairwise distance matrices were performed using Mega X. These matrices show the number of amino acid differences per site between sequences, which were used to define clusters and their environmental relationships.

Principal coordinates analysis

Principal Coordinates Analysis (PCoA) was performed using the R package *vegan* (version 4.3.3). The analysis was conducted with the vegdist function, which calculates distance matrices between samples—here representing sequences of class III PhaC synthase proteins from Archaea and metagenomes. The analysis enables comparison of sequence clustering and establishment of environmental relationships based on sequence origin data.

Results and discussion

PhaC sequences in cultured Archaea

For PhaC synthases, the archaeal sequence data were listed under 10 different names. A UniProt search revealed that these names are synonymous with PHA synthase enzymes (UniProt, 2023). However, only one of them named the PHA synthase class as "class III poly(R)-hydroxyalkanoic acid synthase PhaC subunit". This presented the first challenge in organizing the information. Additionally, much of the data associated with the identified sequences contained redundancies and gaps: several entries were repeated multiple times, while others were left unspecified. This made the search more exhaustive for sequences with missing data. Furthermore, downloading the databases was problematic due to the inclusion of irrelevant and repeated information. The initial stage of data collection involved curation or cleaning of data, which can be challenging. This required reviewing all records to detect errors, complete missing data, and remove duplicates to organize arrays with data in a consistent format for further analysis. Therefore, it is crucial to evaluate the quality of metadata found in public genetic repositories where sequences are archived. Access to these resources has allowed sharing research results with the scientific community, resulting in invaluable collections of microbial ecology sequences. However, gaps and redundancies in reported information make it difficult to use the data in specific applications and contexts (Huttenhower et al., 2023; Jurburg et al., 2020).

The first search filter in NCBI yielded 108 records, on average associated with more than 10,000 sequence entries. The second filter produced 7,465 records. After the initial data cleansing, 291 sequences of PhaC synthase proteins were obtained for Archaea. These sequences were then reviewed individually to select 176 sequences for further analysis. The selected sequences were downloaded for phylogenetic analysis and for constructing the metadata matrix (Table 1, Supplementary Material available at Zenodo: https://doi.org/10.5281/zenodo.15866018). The selected records consist of 120 reference sequences from the RefSeq database and 56 INSCD sequences, representing 48 different genera of Archaea and 6 potential candidate genera. Of the 176 registries, 152 corresponded to class III PhaC synthases, and 24 corresponded to unclassified PhaC enzymes. Using these data, a phylogenetic tree and a distance matrix (with a threshold of < 0.5) were created to identify clusters for the Archaea found. The results revealed three clusters (Figure 1): the first cluster contained 22 Archaea sequences, the second contained 140, and the third contained 9. This study suggests that there may be significant genetic differences among the sequences, despite the limited number of clusters identified. Additionally, five archaeal sequences were not grouped in any of the clusters, indicating values > 0.6 in the similarity matrix. These correspond to the *Ferroglobus placidus* and *Geoglobus ahangari* genera, as well as three possible candidates genera: *Candidatus Nitrosocosmicus oleophilus, Candidatus Nitrosocaldus cavascurensis,* and *Candidatus Methanoliparum*. It is important to consider that genetic differences among Archaea may be related to specific adaptations to different ecological niches, particularly those associated with extreme environments (Pinhassi et al., 2016; Vuong et al., 2021; Wang et al., 2019). Further research on the grouping of archaeal sequences in these clusters is suggested to better understand these microorganisms' ecology.

Clusters of cultivated archaeal sequences and their environmental distribution

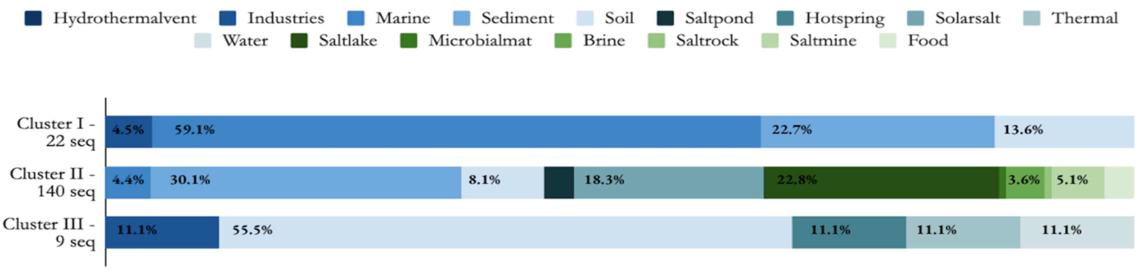


Figure 1. Distribution of the protein sequences of PhaC synthases from cultured Archaea. Sequences were grouped using a distance and similarity matrix analysis performed in Mega X (version 10.2.6). The model of evolutionary divergence estimates between sequences was determined using the p-distance method in pairs.

Sequences in metagenomes

A search was conducted on the KAUST platform using the reference sequences selected for each cluster. The search yielded 6,754 data points for cluster I, 6,886 for cluster II, and 6,742 for cluster III. These data were individually curated to remove redundancy and errors, resulting in a final set of 66 different sequences derived from 20,382 metagenomic records. The curated dataset comprised 64 records of class III PhaC synthases and 2 records corresponding to unclassified PhaC. To classify at the genus or species level, a BLASTp analysis was performed. The results showed 3 genera, 2 potential candidates, and 4 sequences classified at the family level. All sequences had 100% coverage and >90% identity, indicating a high degree of homology (Table 2, Supplementary Material). The taxonomic data of the sequences found correspond to *Natronobeatus* sp*.*, *Nitrosopumilus* sp*.*, *Nitrososphaera* sp*.*, *Candidatus Nitrosopelagicus* sp*.*, *Candidatus Nitrosotenuis acuario*, *Nitrosophaeraceae archaeon, Nitrosopumilaceae archaeon, Thaumarchaeota archaeon,* and *Thermoproteota archaeon.*

The analysis of archaeal sequences indicated low genetic diversity. The similarity matrix showed values below 0.5, where 0 indicates identical sequences and 1 indicates completely different sequences in all compared positions (Kumar et al., 2018). This is illustrated by the prevalence of specific genera found in the metagenomic proteins, which have values below 0.5. Records with 100% homology in metagenomic protein sequences were obtained, despite being classified with different codes or retrieved from databases of different habitats at KAUST.

Type of synthases in PhaC sequences from cultured and metagenomic Archaea

Although most sequences were classified by the type of synthase, we verified and completed the data for those that were not. As previously mentioned, there were 24 sequences of Pha synthase from Archaea and 2 from metagenomic proteins. It is known that some microorganisms have genes that encode multiple Pha synthases. However, current records show that all PhaC synthase enzymes in Archaea belong to class III. It has been reported that the haloarchaeal PhaC synthase and the bacterial type III synthase share some conserved residues (Chek et al., 2017; Neoh et al., 2022).

All 242 sequences, including both Archaea and metagenomes, were verified for the classification of the superfamily and the presence of the PhaE_TIGR01834 domain. After curating the sequences and reviewing the conserved domains reported by NCBI, the 24 archaeal sequences were classified, along with the 2 metagenomic sequences that did not specify the synthase class (Table 1, Supplementary Material available at Zenodo: https://doi.org/10.5281/zenodo.15866018). The classification of PHA_synth_III_C also relies on the highly conserved motif <rmekwifdspd> (amino acid residues 245 to 254), which is typical of PHA_synth_III_C. This motif is mainly found in bacteria but also in some Archaea (Hai et al., 2004), however, for the sequences analyzed in this study it was not present. The distribution of PHA_synth_III_C in microorganisms that are not closely related phylogenetically is also noteworthy. It has been suggested that the genes for these enzymes are acquired from a common source through horizontal DNA transfer during evolution, indicating that they did not evolve independently (Hai et al., 2004). This explanation sheds light on some of the phylogenetic relationships observed in this study as *Ferroglobus placidus, Geoglobus ahangari, Candidatus Nitrosocosmicus oleophilus, Candidatus Nitrosocaldus cavascurensis,* and *Candidatus Metanoliparum* were not grouped with any of the clusters, they were not closely related to the other genera found. Furthermore, they were found in unusual, non-halophilic environments.

Archaeal and metagenomic sequences were found in diverse environments, leading to the establishment of 20 categories (Table 1, Supplementary Material available at Zenodo: https://doi.org/10.5281/zenodo.15866018). Figure 2 displays the distribution of the origin environment for the sequence, along with the taxonomic and environmental distribution of the data associated with the cultured archaeal protein sequences and metagenomes. The environmental distribution for cultivated Archaea was identified as follows: sediments (26.70%), followed by salt lakes (17.61%), salt mines (14.77%), and solar salt (14.77%). Marine and soil samples each account for 10.79%, while brine, food, industries, and water each accounted for less than 3%. The remaining categories, including hot springs, hydrothermal vents, microbial mats, salt rocks, and thermal springs, each accounted for less than 2%. In terms of metagenomes, marine environments accounted for 41.66%, followed by soil (31.48%), and sediments (9.25%). Hydrothermal vents, anaerobic digesters, and water samples each accounted for less than 7%, while biogas fermenters, hydrocarbons, microbial mats, and activated sludge each accounted for less than 2% (Figure 2).

Environmental and taxonomic distribution of phaC genes in Archaea and metagenomes

Protein sequences for cultivable Archaea were most commonly found in sediments, salt lakes, solar salt environments, soil, and marine habitats. Meanwhile, the categories with the highest number of sequences from metagenomes were marine, soil, sediments, hydrothermal vents, and anaerobic digesters. Regarding this matter, sediments, salt lakes, soil, and marine samples had the highest number of associated PhaC synthase sequences from Archaea and metagenomes (Figure 2). Archaeal isolation is common in saline habitats, particularly marine environments. It has been demonstrated that there is an association between the PHA production gene and genes regulating adaptation to halophilic environments (Wang et al., 2019). Therefore, the presence of PHA-producing Archaea in high-salinity environments offers a chance to expand the bioprocess while reducing costs and pollution associated with the nutritional requirements of these microorganisms (Wang et al., 2019).

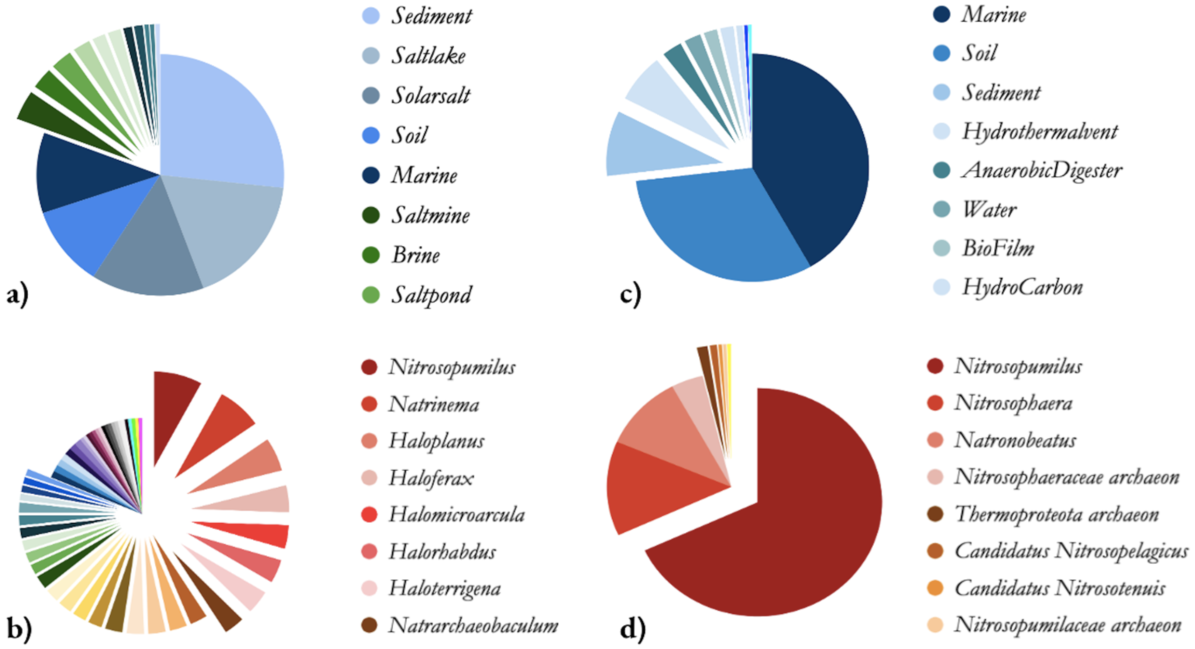
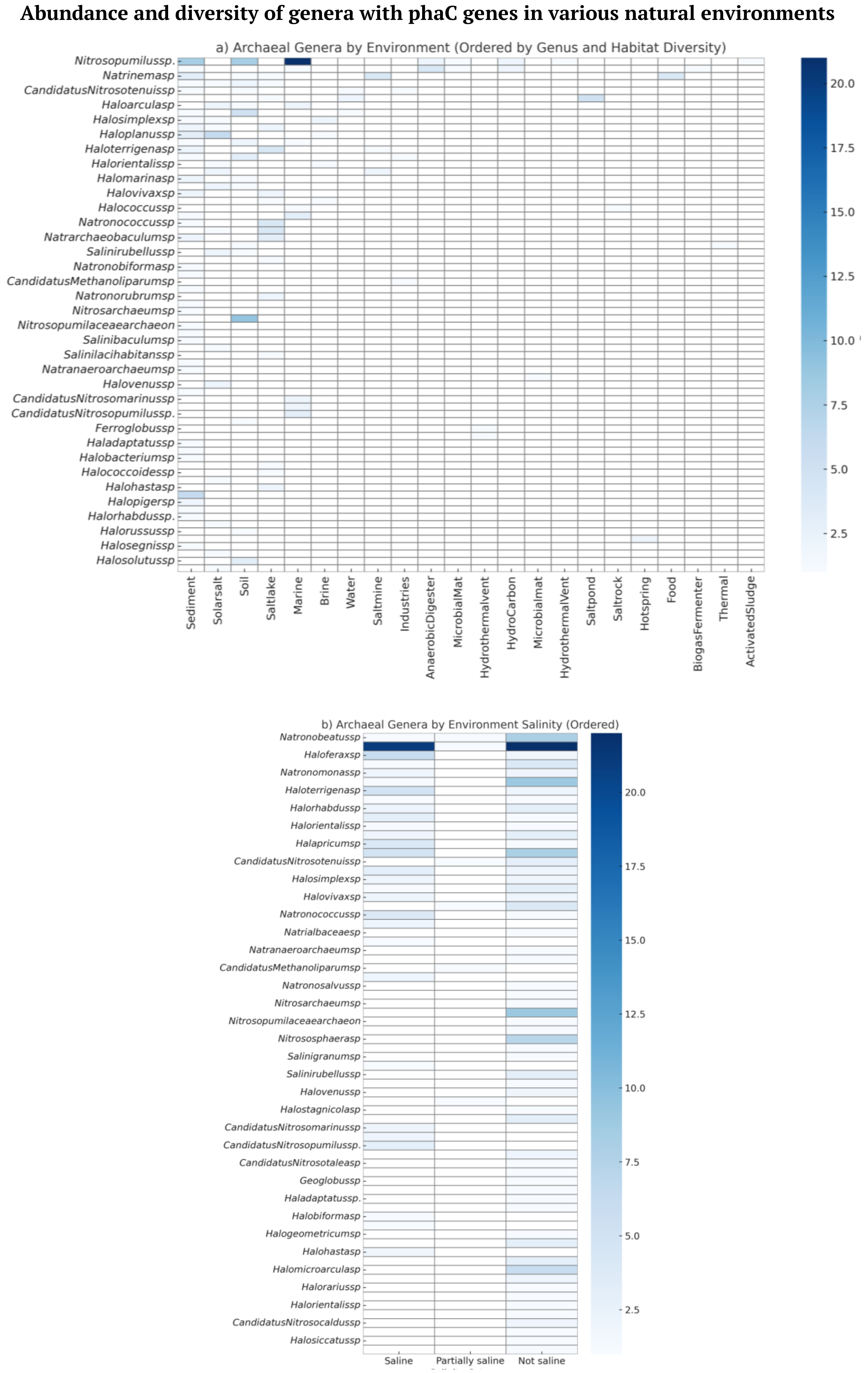


Figure 2. a) Representative environments associated with PHA_synth_III_C protein sequences in Archaea (NCBI), b) Representative genera associated with PHA_synth_III_C protein sequences in Archaea (NCBI), c) Representative environments associated with PHA_synth_III_C protein sequences in metagenomes (KAUST), and d) Representative genera associated with metagenomic protein sequences found in KAUST, whose association was performed using BLASTp.

Studies have reported that PHA_synth_III_C genotypes are present in various ecosystems, with a higher abundance in terrestrial and aquatic environments, including freshwater and marine habitats (Obulisamy & Mehariya, 2021; Vuong et al., 2021), as well as non-marine saline and alkaline habitats, thermal springs, sediments, agricultural fields, geological sites, and caves. In this study, it was found that Archaea have been isolated with the presence of the phaC gene in uncommon environments, such as rhizospheric soils, agricultural fields, environments contaminated with hydrocarbons, anaerobic digesters, and samples of industrial and thermal waste. This suggests that these environments could be a potential source of microorganisms for the production of PHA, despite saline environments being the most common for these genes (Lehtovirta-Morley et al., 2014; Obulisamy & Mehariya, 2021; Vuong et al., 2021).

Regarding the identified genera, the abundances for cultivated Archaea were as follows: *Nitrosopumilus* sp. (7.95%)*, Natrinema* sp. (7.39%)*, Haloplanus* sp. (5.68%)*, Haloferax* sp. (4.55%)*, Halomicroarcula* sp. (3.98%)*, Halorhabdus* sp. (3.98%)*, Haloterrigena* sp. (3.98%)*,* and *Natrarchaeobaculum* sp. (3.41%) (Figure 2). For metagenomic proteins, the associated genera were present in varying abundances, as follows: *Nitrosopumilus* sp. (68.05%)*, Nitrosophaera* sp. (13.42%), *Natronobeatus* sp. (10.64%), *Nitrosophaeraceae archaeon* (4.16%), *Thermoproteota archaeon* (1.38%), *Candidatus Nitrosopelagicus* (0.92%), *Candidatus Nitrosotenuis* (0.46%), *Nitrosopumilaceae archaeon* (0.46%), and *Thaumarchaeota archaeon* (0.46%) (Figure 2).

Based on the information provided, *Nitrosopumilus* sp. sequences were the most abundant, found in 7 out of 20 categories for both Archaea and metagenomes. However, it is important to note that when considering only the abundance in Archaea, this percentage is not significantly different from that of *Natrinema* sp., which was present in 7.39% of the analyzed sequences. Furthermore, the analysis of the presence and absence of genera across environments revealed that although *Natronobeatus* sp. was not the most abundant genus in terms of the number of sequences found, it was present in the highest number of environments, specifically 9 out of the 20 defined categories (Figure 3).



Abundance and diversity of genera with phaC genes in various natural environments

a) Archaeal Genera by Environment (Ordered by Genus and Habitat Diversity)

b) Archaeal Genera by Environment Salinity (Ordered)

Figure 3. a) The most abundant genera in different environments and the environments with the highest diversity of archaeal genera with the presence of the phaC gene for PHA production are observed. b) The abundance and diversity of genera in saline environments are shown.

It is important to note that both PHA_synth_III_C sequences of Archaea and PHA_synth_III_C sequences of metagenomic proteins contain records of microorganisms that have not yet been classified but are potential candidates (Ji et al., 2021; Obulisamy & Mehariya, 2021; Vuong et al., 2021). Many of the genera found in these sequences have been reported in other studies (El-malek et al., 2020; Han et al., 2010; Lehtovirta-Morley et al., 2014; Wang et al., 2019). Hydrothermal vents contain a variety of microorganisms, including *Ferroglobus* sp. and *Geoglobus* sp., which are less commonly reported (Anderson et al., 2011; Salwan & Sharma, 2020). Metagenomic analysis has identified *Natronobeatus* sp. in samples such as activated sludge, anaerobic digesters, and biogas fermenters. These findings suggest that *Nitrosopumilus* sp., *Natrinema* sp., and *Natronobeatus* sp. could be considered potential candidates for PHA production, even though no previous reviews have explored their use in biopolymer production.

Several trees were constructed to establish phylogenetic and environmental relationships. The first tree was constructed using archaeal protein sequences, and three clusters were defined according to the similarity matrix (Figure 1). Figure 4a shows the second tree, which was constructed using archaeal protein sequences and representative protein sequences in bacteria for each type of PhaC synthase. The tree reveals that the sequences of class I and class II PhaC synthases in bacteria cluster together, while the sequences of reference class III PhaC synthases cluster closer to the class III PhaC synthases found in Archaea. For class IV synthases, both the tree and the NCBI search show that these sequences are now classified as PHA_synth_III_C. Furthermore, sequence WP_013055939.1, previously associated with the species *Bacillus megaterium*, has been reclassified as *Priestia megaterium*. This sequence is consistently referred to in the literature as a class IV PhaC synthase (Gupta et al., 2020; Neoh et al., 2022). The independent classification of this member of the PHA_synt_III_C family is likely due to its requirement of PhaR, unlike other members that require PhaE (McCool & Cannon, 2001).

Phylogenetic distribution of phaC in Archaea and metagenomes

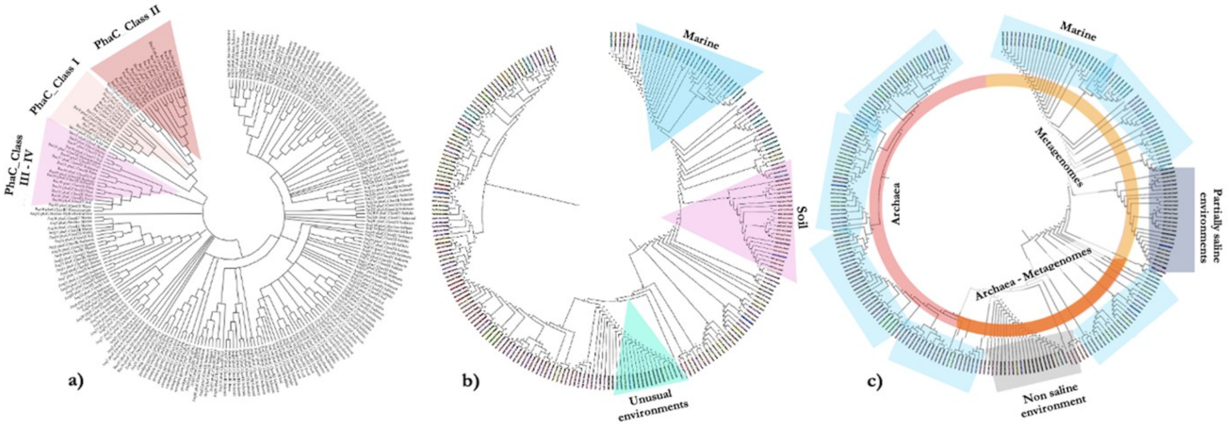


Figure 4. Phylogenetic tree constructed with the 242 sequences of cultivated Archaea and metagenomes (Mega X, version 10.2.6, neighbor-joining algorithm) a) Archaeal PhaC synthases and representative PhaC I-V from bacteria, showing the following clustering: PhaC class II, PhaC class I, PhaC synthases class III-IV. b) Archaeal and metagenomic sequences, three regions where sequences are grouped by environments: blue region = marine habitats, pink region = soil habitats, green region = activated sludge habitats, hydrocarbon-contaminated environments, anaerobic digesters, and fermenters; c) Archaeal Pha synthase sequences are shown in red, metagenomic protein sequences in yellow, and the two together in orange; the areas marked in blue correspond to saline environments, dark gray partially saline environments, and dark blue and light gray non-saline environments.

A phylogenetic tree was constructed using all archaeal PhaC synthase sequences and metagenomic protein sequences (Figure 4). No clustering based on environmental categories was observed. However, three regions collect metagenomic protein sequences from marine and soil environments (Figure 4b). Additionally, there is a region with sequences originating from samples isolated from activated sludge, hydrocarbon-contaminated environments, anaerobic digesters, and fermenters. This region contains both archaeal and metagenomic protein sequences (Figure 4b). The results obtained in the similarity matrix (Mega X, version 10.2.6) are compared with those presented here. The similarity matrix shows the number of amino acid differences per site between sequences. All 242 amino acid sequences were included, and all ambiguous positions were removed for each pair of sequences. A total of 301 positions in the dataset were compared to demonstrate environmental relationships using the Principal Coordinates Analysis (PCoA) (Figure 5).

Phylogenetic and environmental distribution of the phaC gene in Archaea and metagenomes

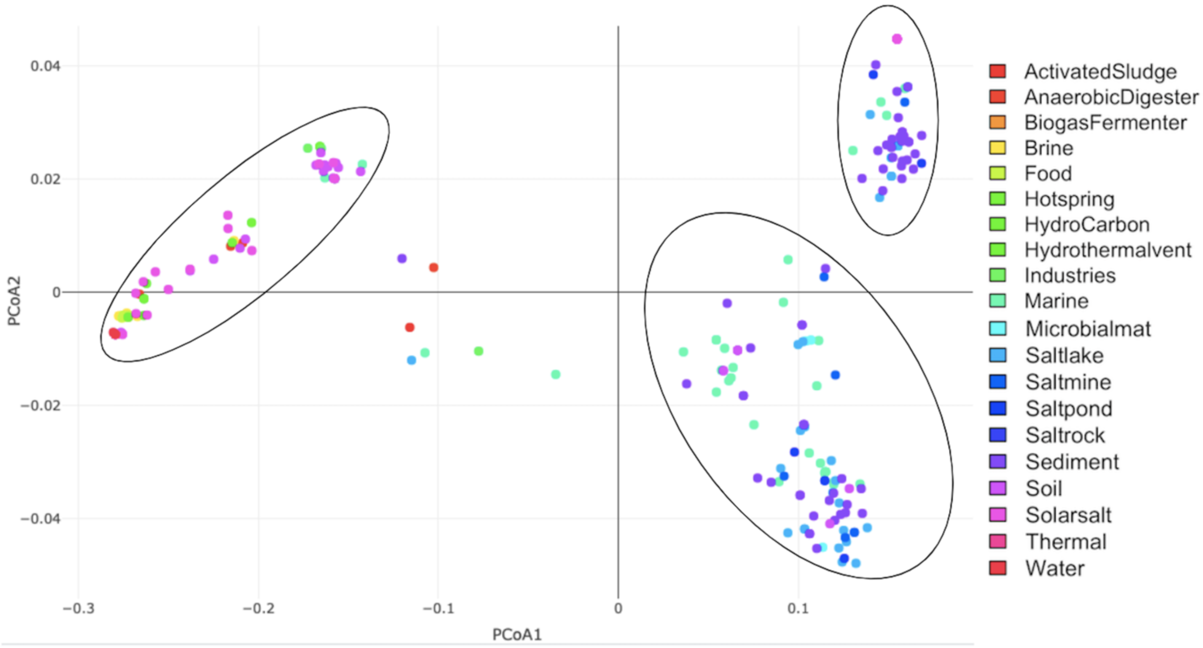


Figure 5. Principal Coordinates Analysis (PCoA) conducted using the vegdist function in the R vegan package (R version 4.3.3). The distribution of the 176 Archaea (NCBI) and 66 metagenomic proteins (KAUST) is observed based on the comparison of their sequences performed with Mega X (version 10.2.6), through a similarity matrix, considering the sample origin and classification into the 20 categories defined for environments.

The principal coordinates analysis shows that the presence of the phaC gene, and consequently the PHA_synth_III_C protein, is not exclusively determined by the origin of the Archaea's environment. Although microorganisms in saline environments tend to exhibit extremophilic characteristics, including the presence of Archaea and metagenomes, similar characteristics have also been observed in other challenging environments such as thermal areas, hydrocarbon-contaminated soils, anaerobic digesters, fermenters, and activated sludge (Lehtovirta-Morley et al., 2014; Obulisamy & Mehariya, 2021). These environments are not necessarily saline but still present challenging conditions for microbial survival. Archaea are highly adapted and metabolically active in unusual or challenging environments, where other microorganisms may struggle to survive. This is related to their ability to produce PHAs, which accumulate carbon and energy in response to nutrient deficits or stressful conditions. Some studies suggest that PHA accumulation can protect against oxidative, saline, and thermal stress (Adeleye et al., 2020; Cruz et al., 2016; Obulisamy & Mehariya, 2021; Pu et al., 2020; Vuong et al., 2021).

Conclusion

The study confirms that Archaea have class III PhaC synthases, which include another catalytic subunit encoded by phaE. This was confirmed in all sequences analyzed. The presence of the phaC gene in Archaea appears to be closely linked to saline environments and the production of polyhydroxyalkanoates. However, the diverse range of environments studied in the present research suggests the possibility of isolating Archaea capable of producing these biopolymers in atypical habitats. The identification of genera such as *Nitrosopumilus* sp., *Natrinema* sp., and *Natronobeatus* sp. harboring the phaC in metagenomic data is a significant finding, indicating their potential as candidates for the production of polyhydroxyalkanoate. These microorganisms have not been previously studied for this bioprocess, either *in vitro* or at scale for the production of bioplastics. The data analysis showed that specific PHA_synth_III_C proteins found in Archaea and metagenomes were grouped into categories of unusual environments, such as hydrocarbon-contaminated soils, anaerobic digesters, fermenters, and activated sludge. However, it is not possible to conclude that there is a trend in the clustering of the found protein sequences regarding environments. The abundance and diversity of PhaC synthase protein sequences found in Archaea and metagenomes may enable the production of new monomeric polymer compositions, providing an alternative to conventional plastics and addressing the challenges of this bioprocess. Additionally, it is important to note that manual curation or data cleaning is still required to ensure data quality. This study highlights the usefulness of bioinformatics tools in managing information. Organizing the information into a consistent format suitable for analysis was identified as one of the main challenges in this study. It is important to note that the gap between the amount of generated data and the available analysis tools still presents a challenge, particularly regarding the necessary skills for their use and the technological capacity related to the equipment and servers that enable data analysis.

References

Adeleye, A. T., Odoh, C. K., Enudi, O. C., Banjoko, O. O., Osiboye, O. O., Odediran, E. T., & Louis, H. (2020). Sustainable synthesis and applications of polyhydroxyalkanoates (PHAs) from biomass. *Process Biochemistry*, *96*, 174-193. https://doi.org/10.1016/j.procbio.2020.05.032

Alamgeer, M. (2019). Polyhydroxyalkanoates (PHA) genes database. *Bioinformation*, *15*(1), 36-39. https://doi.org/10.6026/97320630015036

Anderson, I., Risso, C., Holmes, D., Lucas, S., Copeland, A., Lapidus, A., Cheng, J.-F., Bruce, D., Goodwin,L., Pitluck, S., Saunders,E., Brettin, T., Detter, J. C., Han, C., Tapia, R., Larimer, F., Land, M., Hauser, L., Woyke, T., … Ivanova, N. (2011). Complete genome sequence of *Ferroglobus placidus* AEDII12DO. *Standards in Genomic Sciences*, *5*(1), 50-60. https://doi.org/10.4056/sigs.2225018

Chek, M. F., Kim, S.-Y, Mori, T., Arsad, H., Samian, M. R., Sudesh, K., & Hakoshima, T. (2017). Structure of polyhydroxyalkanoate (PHA) synthase PhaC from *Chromobacterium* sp. USM2, producing biodegradable plastics. *Scientific Reports*, *7*(5312), 1-15. https://doi.org/10.1038/s41598-017-05509-4

Chen, G.-Q., Chen, X.-Y., Wu, F.-Q, & Chen, J.-C. (2020). Polyhydroxyalkanoates (PHA) toward cost competitiveness and functionality. *Advanced Industrial and Engineering Polymer Research*, *3*(1), 1-7. https://doi.org/10.1016/j.aiepr.2019.11.001

Choi, S. Y., Cho, I. J., Lee, Y., Kim, Y.-J., Kim, K.-J, & Lee, S. Y. (2020). Microbial polyhydroxyalkanoates and nonnatural polyesters. *Advanced Materials*, *32*(35), 1907138. https://doi.org/10.1002/adma.201907138

Cruz, M. V., Freitas, F., Paiva, A., Mano, F., Dionísio, M., Ramos, A. M., & Reis, M. A. M. (2016). Valorization of fatty acids-containing wastes and byproducts into short- and medium-chain length polyhydroxyalkanoates. *New Biotechnology, 33*(1), 206-215. https://doi.org/10.1016/j.nbt.2015.05.005

El-malek, F. A., Farag, A., Omar, S., & Khairy, H. (2020). Polyhydroxyalkanoates (PHA) from halomonas pacifica ASL10 and *Halomonas salifodiane* ASL11 isolated from mariout salt lakes. *International Journal of Biological Macromolecules, 161*, 1318-1328. https://doi.org/10.1016/j.ijbiomac.2020.07.258

Gupta, R. S., Patel, S., Saini, N., & Chen, S. (2020). Robust demarcation of 17 distinct *Bacillus* species clades, proposed as novel *Bacillaceae* genera, by phylogenomics and comparative genomic analyses: description of *Robertmurraya kyonggiensis* sp. nov. and proposal for an emended genus *Bacillus* limiting it only to the members of the Subtilis and Cereus clades of species. *International Journal of Systematic and Evolutionary Microbiology*, *70*(11), 5753-5798. https://doi.org/10.1099/ijsem.0.004475

Fradinho, J.C., Oehmen, A., & Reis, M. A. M. (2014). Photosynthetic mixed culture polyhydroxyalkanoate (PHA) production from individual and mixed volatile fatty acids (VFAs): Substrate preferences and co-substrate uptake. *Journal of Biotechnology*, *185*, 19–27. https://doi.org/10.1016/j.jbiotec.2014.05.035

Hai, T., Lange, D., Rabus, R., & Steinbüchel, A. (2004). Polyhydroxyalkanoate (PHA) accumulation in sulfate-reducing bacteria and identification of a class III PHA synthase (PhaEC) in *Desulfococcus multivorans*. *Applied and Environmental Microbiology, 70*(8), 4440-4448. https://doi.org/10.1128/AEM.70.8.4440-4448.2004

Han, J., Li, M., Hou, J., Wu, L., Zhou, J., & Xiang, H. (2010). Comparison of four phaC genes from *Haloferax mediterranei* and their function in different PHBV copolymer biosyntheses in *Haloarcula hispanica. Aquatic Biosystems, 6*(9), 1-10. https://doi.org/10.1186/1746-1448-6-9

Huttenhower, C., Finn, R. D., & McHardy, A. C. (2023). Challenges and opportunities in sharing microbiome data and analyses. *Nature Microbiology, 8*(11), 1960-1970. https://doi.org/10.1038/s41564-023-01484-x

Ji, M., Williams, T. J., Montgomery, K., Wong, H. L., Zaugg, J., Berengut, J. F., Bissett, A., Chuvochina, M., Hugenholtz, P., & Ferrari, B. C. (2021). *Candidatus eremiobacterota*, a metabolically and phylogenetically diverse terrestrial phylum with acid-tolerant adaptations. *The ISME Journal, 15*(9), 2692-2707. https://doi.org/10.1038/s41396-021-00944-8

Jurburg, S. D., Konzack, M., Eisenhauer, N., & Heintz-Buschart, A. (2020). The archives are half-empty: An assessment of the availability of microbial community sequencing data. *Communications Biology, 3*(474), 1-8. https://doi.org/10.1038/s42003-020-01204-9

Kim, J., Kim, Y.-J., Choi, S. Y., Lee, S. Y., & Kim, K.-J (2017). Crystal structure of *Ralstonia eutropha* polyhydroxyalkanoate synthase C-terminal domain and reaction mechanisms. *Biotechnology Journal*, *12*(1), e1600648. https://doi.org/10.1002/biot.201600648

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547-1549. https://doi.org/10.1093/molbev/msy096

Kumar, V., Kumar, S., & Singh, D. (2020). Microbial polyhydroxyalkanoates from extreme niches: Bioprospection status, opportunities and challenges. *International Journal of Biological Macromolecules*, *147*, 1255-1267. https://doi.org/10.1016/j.ijbiomac.2019.09.253

Laycock, B., Halley, P., Pratt, S., Werker, A., & Lant, P. (2013). The chemomechanical properties of microbial polyhydroxyalkanoates. *Progress in Polymer Science*, *38*(3-4), 536-583. https://doi.org/10.1016/j.progpolymsci.2012.06.003

Lee, S. Y., & Kim, H. U. (2015). Systems strategies for developing industrial microbial strains. *Nature Biotechnology*, *33*(10), 1061-1072. https://doi.org/10.1038/nbt.3365

Lee, S. Y., Lee, K. M., Chan, H. N., & Steinbüchel, A. (1994). Comparison of recombinant *Escherichia coli* strains for synthesis and accumulation of poly-(3-hydroxybutyric acid) and morphological changes. *Biotechnology and Bioengineering, 44*(11), 1337-1347. https://doi.org/10.1002/bit.260441110

Lehtovirta-Morley, L. E., Ge, C., Ross, J., Yao, H., Nicol, G. W., & Prosser, J. I. (2014). Characterisation of terrestrial acidophilic archaeal ammonia oxidisers and their inhibition and stimulation by organic compounds. *FEMS Microbiology Ecology*, *89*(3), 542-552. https://doi.org/10.1111/1574-6941.12353

Li, M., & Wilkins, M. R. (2020). Recent advances in polyhydroxyalkanoate production: Feedstocks, strains and process developments. *International Journal of Biological Macromolecules*, *156*, 691–703. https://doi.org/10.1016/j.ijbiomac.2020.04.082

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Research, 48*(1), 265-268. https://doi.org/10.1093/nar/gkz991

Magagula, S. I., Mohapi, M., Sefadi, J. S., & Mochane, M. J. (2021). The production and applications of microbial-derived polyhydroxybutyrates. In A. Vaishnav, & D. K. Choudhary (Eds), *Microbial polymers* (pp. 3-43). Springer. https://doi.org/10.1007/978-981-16-0045-6

McCool, G. J., & Cannon, M. C. (2001). PhaC and PhaR are required for polyhydroxyalkanoic acid synthase activity in *Bacillus megaterium*. *Journal of Bacteriology, 183*(14), 4235-4243. https://doi.org/10.1128/JB.183.14.4235-4243.2001

Neoh, S. Z., Chek, M. F., Tan, H. T., Linares-Pastén, J. A., Nandakumar, A., Hakoshima, T., & Sudesh, K. (2022). Polyhydroxyalkanoate synthase (PhaC): The key enzyme for biopolyester synthesis. *Current Research in Biotechnology*, *4*, 87-101. https://doi.org/10.1016/j.crbiot.2022.01.002

Obruča, S., Dvořák, P., Sedláček, P., Koller, M., Sedlář, K., Pernicová, I., & Šafránek, D. (2022). Polyhydroxyalkanoates synthesis by halophiles and thermophiles: Towards sustainable production of microbial bioplastics. *Biotechnology Advances, 58*, 107906. https://doi.org/10.1016/j.biotechadv.2022.107906

Obruca, S., Sedlacek, P., Koller, M., Kucera, D., & Pernicova, I. (2018). Involvement of polyhydroxyalkanoates in stress resistance of microbial cells: Biotechnological consequences and applications. *Biotechnology Advances, 36*(3), 856-870. https://doi.org/10.1016/j.biotechadv.2017.12.006

Obulisamy, P. K., & Mehariya, S. (2021). Polyhydroxyalkanoates from extremophiles: A review. *Bioresource Technology, 325*, 124653. https://doi.org/10.1016/j.biortech.2020.124653

Pinhassi, J., DeLong, E. F., Béjà, O., González, J. M., & Pedrós-Alió, C. (2016). Marine bacterial and archaeal ion-pumping rhodopsins: Genetic diversity, physiology, and ecology. *Microbiology and Molecular Biology Reviews*, *80*(4), 929-954. https://doi.org/10.1128/MMBR.00003-16

Pu, N., Wang, M.-R., & Li, Z.-J. (2020). Characterization of polyhydroxyalkanoate synthases from the marine bacterium *Neptunomonas concharum* JCM17730. *Journal of Biotechnology, 319*, 69-73. https://doi.org/10.1016/j.jbiotec.2020.06.002

Quillaguaman, J., Guzman, H., Van-thuoc, D., & Hatti-kaul, R. (2010). Synthesis and production of polyhydroxyalkanoates by halophiles: Current potential and future prospects. *Applied Microbiology and Biotechnology, 85*(6), 1687-1696. https://doi.org/10.1007/s00253-009-2397-6

Rehm, B. H. A. (2003). Polyester synthases: Natural catalysts for plastics. *Biochemical Journal, 376*(Pt 1), 15-33. https://doi.org/10.1042/BJ20031254

Salwan, R., & Sharma, V. (2020). *Physiological and biotechnological aspects of extremophiles*. Academic Press. https://doi.org/10.1016/B978-0-12-818322-9.00002-2

Singleton, C. M., Petriglieri, F., Kristensen, J. M., Kirkegaard, R. H., Michaelsen, T. Y., Andersen, M. H., Kondrotaite, Z., Karst, S. M., Dueholm, M. S., Nielsen, P. H., & Albertsen, M. (2021). Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications, 12*(1), 1-13. https://doi.org/10.1038/s41467-021-22203-2

UniProt. (2023). *UniProtKB – Protein knowledgebase*. https://www.uniprot.org

Vuong, P., Lim, D. J., Murphy, D. V., Wise, M. J., Whiteley, A. S., & Kaur, P. (2021). Developing bioprospecting strategies for bioplastics through the large-scale mining of microbial genomes. *Frontiers in Microbiology, 12*(697309), 1-12. https://doi.org/10.3389/fmicb.2021.697309

Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2023). The conserved domain database in 2023. *Nucleic Acids Research, 51*(1), 384-388. https://doi.org/10.1093/nar/gkac1096

Wang, L., Liu, Q., Wu, X., Huang, Y., Wise, M. J., Liu, Z., Wang, W., Hu, J., & Wang, C. (2019). Bioinformatics analysis of metabolism pathways of archaeal energy reserves. *Scientific Reports, 9*(1034), 1-12. https://doi.org/10.1038/s41598-018-37768-0