UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

DESCRIPCIÓN DE LA DISTRIBUCIÓN FILOGENÉTICA DE ENZIMAS DEL CATABOLISMO DE FOSFONATO EN BACTERIAS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G O

P R E S E N T A:

ROMERO CHORA LUIS GERARDO

DIRECTOR DE TESIS:
DR. LUIS DAVID ALCARAZ PERAZA

Ciudad Universitaria, Cd. Mx., 2022

Agradecimientos académicos

La tesis se realizó en el Laboratorio de Genómica Ambiental, del Departamento de Biología Celular de la Universidad Nacional Autónoma de México. Con apoyo del proyecto DGAPA-PAPIIT UNAM IN221420 bajo responsabilidad del Dr. Luis David Alcaraz Peraza.

Les agradezco su total apoyo y compromiso para realizar este proyecto colaborativo ya que sin su ayuda no se hubiese realizado bajo la dirección de los doctores, entre estas personas agradezco especialmente a las siguientes personas:

Dr. Luis David Alcaraz Peraza, un gran profesor que me enseñó todo lo necesario sobre la biotecnología y siempre me apoyó al no dejarme ir por problemas externos a la academia y me tuvo mucha paciencia.

A los miembros de mi jurado: M. en C. B. Vanessa Vega García, M. en C. Simón Guzmán León.

A la pareja de biólogos la Dra. Irene Sanchez y M. en C. Juan Carlos Flores, por su apoyo incondicional en condiciones difíciles y préstamo de equipo para realizar este escrito.

Dra. Blanca Estela Hernández Baños del taller: Sistemática molecular, filogeografía y genética de la conservación de vertebrados y plantas, de la Facultad de Ciencias, UNAM.

Dedicatoria

A mi familia, que con toda la paciencia del mundo me apoyaron para poder realizar este trabajo en condiciones difíciles de pandemia. A mi madre Gloria y padre Gilberto que no cuestionaron el que pudiera salir con esto a pesar de las vicisitudes. A mis hermanas Gimena y Gina que están en su propia formación y me apoyan en sus respectivos campos para siempre mejorar.

A mis compañeros del Laboratorio de Genómica Ambiental: Gerardo, Miguel, Cristóbal, Ana, Hugo, Angélica y Leslie que me ayudaron siempre que tuve dudas en mis primeros pasos realizando investigación con herramientas bioinformáticas y cuando pudieron me proporcionaron material de primera o a usar el código en primera instancia.

A la familia Reyes Ávila, y especialmente a mi querido amigo Julio, que siempre estuvieron al tanto de mi progreso como biólogo y siempre me dieron de comer cuando iba a su casa.

A mis compañeras de la facultad, quienes me apoyaron y acompañaron a lo largo de mi educación universitaria.

Contenido

1. Resumen

La escasez de fósforo es un problema por ser un recurso vital. Su incorporación a la biósfera inicia gracias al intemperismo de escasas rocas fosfóricas, proceso en el cual intervienen unos pocos microorganismos capaces de degradar este nutriente y sus formas alternativas.

De estos microorganismos se han descrito proteínas involucradas en el metabolismo del fósforo, especialmente de formas fosfonatadas que son energéticamente costosas de catabolizar, y que se clasifican de acuerdo a su mecanismo enzimático: I corte radical del enlace C-P por medio de la PhnJ C-P liasa, II corte hidrolítico del enlace C-P y III corte oxidativo por PhnZ/PnhY*.

Son pocas las especies que se ha reportado que realizan el catabolismo de fosfonato (CPn). Algunos estudios muestran que no hay una relación filogenética entre los organismos con las proteínas encargadas del metabolismo de fosfonatos, y que estas probablemente se hayan distribuido por transferencia horizontal.

Para dilucidar la distribución de estas proteínas en los diferentes phyla de Bacteria se realizó la búsqueda de 22 proteínas involucradas en el catabolismo de fosfonato (CPn) en 10,906 proteomas completos de bacterias. Después de su búsqueda e identificación, se analizó la distribución filogenética, así como se determinó *in silico* el tipo de mecanismo enzimático para estas enzimas de metabolismo del fosfonato.

Se observó gran predominancia del mecanismo de CPn tipo I que involucra las proteínas del *operón phn* (corte radical C-P) a lo largo de muchos miembros de los phyla Proteobacteria y Cyanobacteria no relacionados de forma evolutiva, y que además en el proteoma de estas especies comparten espacio con las proteínas de otro tipo de catabolismos de Pn que involucran la participación de las proteínas PhnW (2-aminoetilfosfonato - piruvato transaminasa) y PhnX (fosfonatasa) del mecanismo enzimático II.

Las especies que destinan mayor porcentaje de su proteoma a este tipo de metabolismo I son los rizobiales *Phyllobacterium sophorae* (0.76%), *Rhizobium miluonense* (0.56%) y *Mesorhizobium tianshanense* (0.49%) y otras 18 proteobacterias junto a las cianobacterias *Rubidibacter lacunae* (0.46%) y *Nodularia spumigena* (0.43%).

Los patrones de distribución de los metabolismos de CPn nos cuentan historias evolutivas para los taxa Burkholderiales, Enterobacterales, Oceanospirillales, Pseudomonadales, Vibrionaceae, Neisseriales y Nostocales donde los ancestros comunes de cada uno contaban con la capacidad de usar este metabolismo.

Se amplió el conocimiento sobre la distribución de las proteínas de CPn en phyla poco descritas además de la cantidad de especies ya mencionadas.

1.1 Lista de abreviaturas

P fósforo

Pad fósforo adsorbido

Pn Fosfonato

Po fósforo orgánico

Poc fósforo inorgánico ocluido

CPn Catabolismo de Fosfonato

CDS Secuencia codificadora

IPG Grupo de Proteínas Idénticas

WGS Secuenciación completa del genoma

COG Clusters de genes ortólogos

Transportador ABC (*ATP-binding cassette* o dominio de unión a ATP)

2. Introducción

2.1. La importancia del fósforo para los organismos

El fósforo (P) es un nutriente limitante en el crecimiento de los organismos, ya que participa en numerosos procesos y funciones dentro de las células, tales como la estructura del material genético, las membranas celulares, procesos como la fotosíntesis, el transporte y almacenamiento de energía, etc. (Buchanan et al., 2015).

Sin embargo, la disponibilidad de este macroelemento está dada por la propia geoquímica del elemento. El ciclo del fósforo en el suelo indica que este elemento químico proviene principalmente del intemperismo de roca fosfórica (compuesto principalmente de apatita) que suministra iones fosfato ($H_2PO_4^-$ y HPO_4^{2-}) a la solución del suelo y posteriormente, las bacterias, hongos y plantas incorporan estos iones fosfato en su biomasa, iniciando con esto la ruta biológica del P. Además, con el paso del tiempo un suelo joven poco intemperizado tendrá una mayor cantidad de fósforo que uno viejo más intemperizado, en el cual dominará la forma de fósforo orgánico (Po) (Tapia-Torres & García-Oliva, 2013).

Además del P asimilado por los microorganismos que queda en forma orgánica (Po), los iones también pueden reaccionar rápidamente quedando adsorbidos en la superficie de partículas órgano-minerales como P adsorbido (Pad) o mineralizado por elementos metálicos como Al, Fe, Ca como P inorgánico ocluido (Poc) quedando no biodisponible. Existen formas de fósforo orgánico (Po) como los ésteres de fosfato, que al igual que los ortofosfatos son altamente reactivos, y son demandados por los microorganismos (Stosiek et al., 2019).

Muchos compuestos en la naturaleza tienen enlaces C-P en sus componentes. Las moléculas más pequeñas de fosfonato son metilfosfonato y 2-aminoetil-fosfonato (fosfonato). Ambos son de naturaleza ubicua pero los

fosfonatos también se encuentran como constituyentes de lípidos, polisacáridos y polipéptidos. En el metabolismo secundario de bacterias funcionan como antibióticos. Por la parte antropogénica, las industrias de detergentes, anticongelantes, pesticidas y herbicidas, como fosfinotricina y glifosato, liberan una gran cantidad de compuestos xenobióticos organofosforados (Hayes et al., 2000).

Se ha cuantificado la presencia de fosfonatos en ecosistemas acuáticos y se sabe que el P contenido en los fosfonatos representa más del 25% del P orgánico disuelto de alto peso molecular en las columnas de agua de los océanos y 23% del P total (Clark et al., 1998). Estos compuestos se encuentran en diversidad de organismos, incluyendo invertebrados y vertebrados. Sin embargo, a pesar de ser producto de la biosíntesis del metabolismo propio de estos organismos, su catálisis es complicada debido a su naturaleza química.

2.1.1 Fósforo y contaminación

Por medio de la industria agrícola es común introducir fertilizantes inorgánicos para sustituir la deficiencia de macro y micronutrientes en el suelo. El fósforo puede ser agregado como apatita que puede contener cantidades de cadmio y otros metales pesados que, con el uso prolongado, puede significar un problema de acumulación de este metal tóxico para plantas y animales (Gilliam et al., 2015).

Investigaciones como la de Hove-Jensen et al., (2014) han considerado la posibilidad de usar contaminantes del suelo como glifosato como fuente de fósforo a través de una ruta metabólica conocida como C-P liasa, una forma particular de llevar a cabo el catabolismo de una gran variedad de moléculas de fosfonato.

Carles et al., (2019) muestran que comunidades de biopelículas bacterianas en ríos pueden disminuir los niveles de glifosato en el agua y aumentar los de fósforo total, y que muy probablemente se trate de microorganismos con algún mecanismo de catabolismo de fosfonato (CPn). Estudios de este tipo de

metabolismos podrían ayudar con el problema de contaminación ambiental y la limitación de fósforo en el futuro (Nowack, 2003).

Se han encontrado varios *operones* de genes en bacterias, codificantes para las enzimas que convierten ciertas formas de fósforo en otras más disponibles, como lo es el *operón phnCDEFGHIJKLMNOP* (C-P liasa), *palRECBA* (Hidrolasa de fosfonopiruvato), *phnWX* (Fosfonacetaldehído Hidrolasa) y, *phnWAY* (Fosfonoacetato Hidrolasa). El *operón phn* C-P liasa está constituido por proteínas de transporte, fosfatasas y la C-P liasa capaz de romper enlaces C-P, los cuales están en un alto porcentaje de P orgánico que no está normalmente accesible para la mayoría de los organismos, por ejemplo, en ambientes marinos pueden representar la única fuente de Po: los fosfatos orgánicos disueltos están compuestos aproximadamente por un 75% de ésteres de fosfato y un 25% de Pn (Bjarne Hove-Jensen et al., 2011b).

2.2. Catabolismo del fosfonato en bacterias

Hay diferentes rutas metabólicas con el propósito de catabolizar fosfonato (Pn) que comparten la forma bioquímica del enlace C-P. Dentro de la bioquímica de los Pn, se han descrito 3 mecanismos enzimáticos para el catabolismo del enlace C-P según Horsman & Zechel (2017):

Mecanismo I. Corte de enlace radical de C-P por carbono-fósforo liasa

También referido simplemente como las proteínas del *operón phn*, este es un mecanismo enzimático generalista, ya que puede hacer uso de una gran variedad de sustratos (como fosfonopiruvato, fosfonacetaldehído, ácido 2-aminoetilfosfónico, ácido 2-hidroxietil fosfónico, ácido 1-hidroxi-2-aminoetilfosfónico, fosfonoalanina, fosfonoacetato y ácido metilfosfónico) pero la transformación que lleva a cabo es compleja (emplea el uso de 5 reacciones

enzimáticas); este mecanismo enzimático fue descrito en *Escherichia coli* K12 y también estudiada en otros organismos como *P. stutzeri* A1501 (Stosiek et al., 2019).

A grandes rasgos, este mecanismo enzimático conlleva la incorporación de sustratos como el ácido 2-aminometilfosfónico que es N-acetilado por aminoalquilfosfonato N-acetiltransferasa (*PhnO*) usando acetil-CoA como donador del grupo acilo (Bjarne Hove-Jensen et al., 2012, **Figura 1**). El siguiente paso es llevado por la subunidad PhnI de α-D-ribosa 1-alquilfosfonato 5-trifosfato sintasa (PhnI) que cataliza la ribosilación del Pn a partir de la hidrólisis de ATP, dejando paso a α-D-ribosa 1-alquilfosfonato 5-trifosfato a ser hidrolizado en su enlace fosforil por la α-D-ribosa 1-alquilfosfonato 5-trifosfato difosfatasa (PhnM) que libera un pirofosfato y α-D-ribosa 1-alquilfosfonato 5-fosfato. En esta parte, se lleva a cabo la escisión del Pn gracias a la acción de la α-ribosa 1-alquilfosfonato 5-fosfato C-P liasa (PhnJ) que corta el enlace C-P haciendo uso de S-adenosil-L-metionina y un donador H^+ y dejando α-D-ribosa 1,2-fosfato cíclico 5-fosfato, el grupo alquilo y el donador H^+ es oxidado.

El producto α-D-ribosa-1,2-fosfato cíclico 5-fosfato es un metabolito sin salida dentro de la célula, por lo que con la acción de la metalo-β-lactamasa fosforribosil 1,2-fosfato cíclico fosfodiesterasa (PhnP) realiza la hidrólisis región-específica del fosfato cíclico para producir α-D-ribosa-1,5-difosfato 5-fosfato, un metabolito que puede ser utilizado por la célula (Hove-Jensen et al, 2011). La enzima ribosa 1,5-bisfosfato fosfoquinasa (PhnN) cataliza la fosforilación (con el uso de ATP) región específica de α-D-ribosa-1,5-difosfato 5-fosfato para formar 5-fosfo-α-D-ribosa-1-difosfato 5-fosfato, la cual puede entrar al metabolismo de guanosina, uridina, NAD^+ histidina o triptófano (**Figura 1**, Horsman y Zechel, 2017).

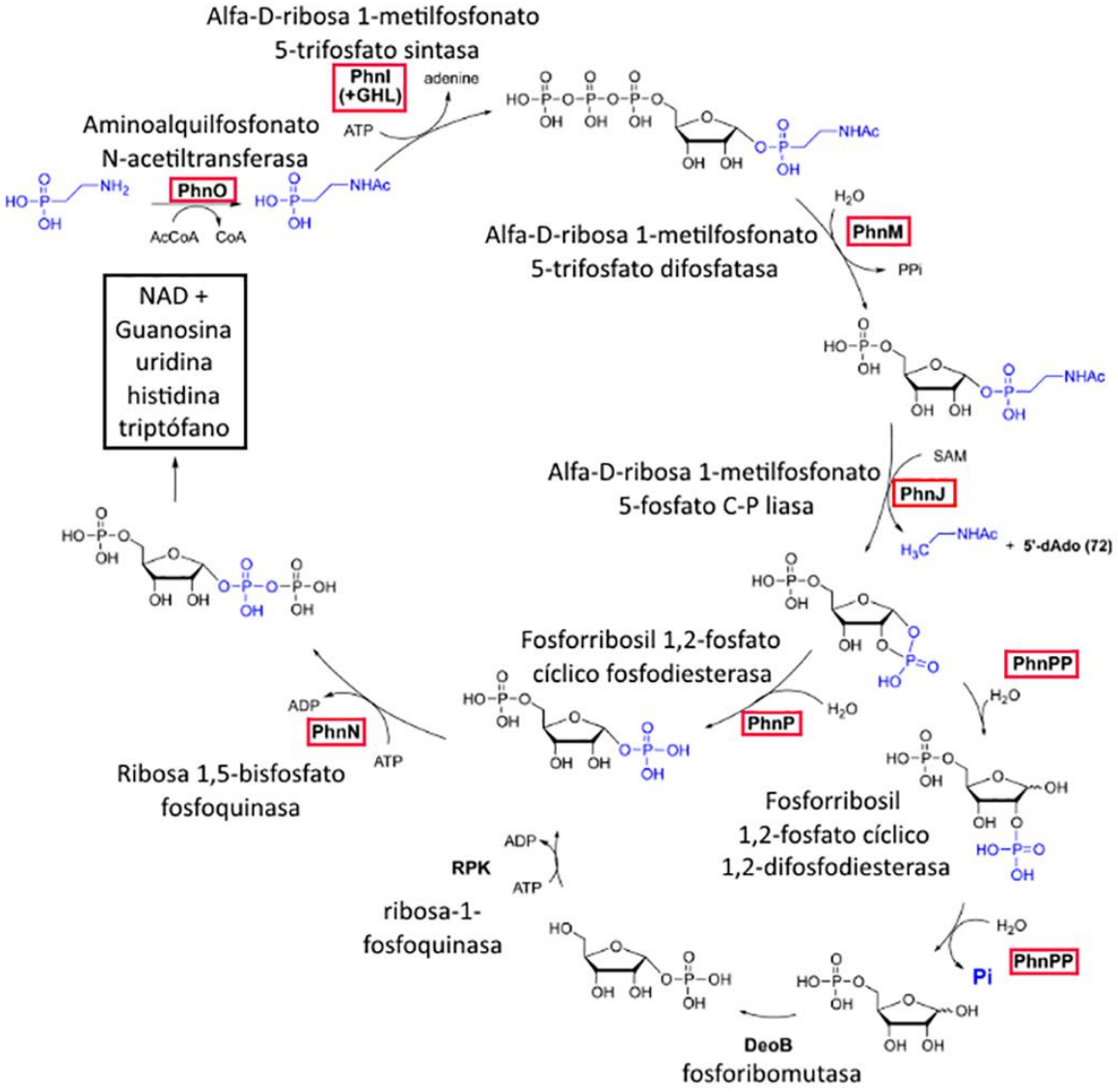

Figura 1. Ruta de catabolismo de fosfonato, con el ácido 2-aminoetil fosfónico como sustrato, modificado de Horsman y Zechel (2017). El sustrato de la reacción de la PhnJ es fosfato de α-D-ribosil-1,2-cíclico que es finalmente convertido en 5-fosfo-α-D-ribosil pirofosfato (PRPP), un donante universal de ribosilo que se utiliza para sintetizar nucleótidos, histidina y triptófano.

Mecanismo II. Cortes hidrolíticos del enlace C-P

Esta ruta catabólica realmente son 3 rutas metabólicas que siguen el mismo mecanismo enzimático: toma de sustrato una molécula aminada de bajo peso molecular que pueda ser usada como primer paso para hidrolizar el enlace C-P,

por lo que el sustrato común para poder entrar a esas pequeñas rutas metabólicas son amino-alquil-fosfonatos como el ácido 2-aminometilfosfónico. En los tres casos, las enzimas se basan en el carbonilo beta para estabilizar el grupo saliente del carbanión que se forma tras el ataque nucleofílico en el centro del fósforo. (Quinn et al., 2007, **Figura 2**).

Uno de los mecanismos de corte hidrolítico es el que se basa en la enzima fosfonopiruvato hidrolasa (PPH), codificada por el gen *palA* en *Variovorax* sp. Pal2 y *Burkholderia cepacia* sp. Pal6, el cual forma parte del *operón PalEDCBA* (PalEDC que es un transportador ABC, PalB que es una transaminasa y PalA fosfonopiruvato hidrolasa), PalB desamina una fosfonoalanina a fosfonopiruvato, el cual es hidrolizado por PalA a piruvato y fosfato. Dicho operón se ha encontrado en registros metagenómicos de hasta 45 sitios de muestreo en agua marina con deficiencia de fósforo inorgánico (Kulakova et al., 2003).

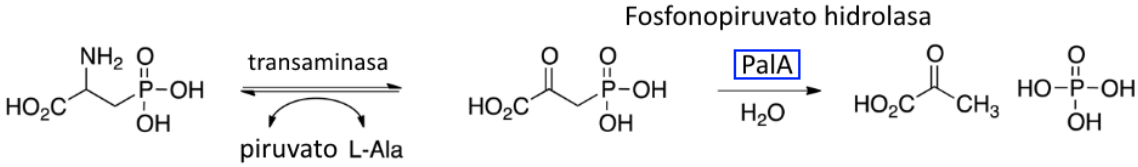

Figura 2. Mecanismo general de catabolismo del ácido 2-aminometilfosfónico el cual es convertido en piruvato por PalA (fosfonopiruvato hidrolasa o PPH) modificado de Horsman y Zechel, 2017.

Otro mecanismo de corte hidrolítico está basado en la acción de la fosfonoacetaldehído hidrolasa (PhnX, también identificada como fosfonatasa) codificada por el gen *phnX* en las bacterias *Bacillus cereus*, *Salmonella typhimurium* LT2129 y *Enterobacter aerogenes*, junto con la transaminasa 2-aminometilfosfonato (codificada en *phnW*) logran la conversión del ácido 2-aminometilfosfónico a fosfonoacetaldehído y su hidrólisis para liberar fosfato y acetaldehído (**Figura 3**, Horsman y Zechel, 2017).

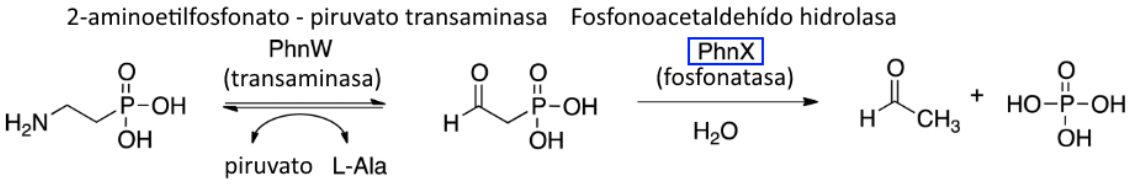

Figura 3. Conversión de ácido 2-aminoetilfosfónico hasta ser hidrolizado por PhnX fosfonatasa, modificado de Horsman y Zechel, 2017.

El tercer mecanismo que involucra el corte hidrolítico del enlace C-P es efectuado por la fosfonoacetato hidrolasa (PhnA), codificada en el gen *phnA* de *Pseudomonas fluorescens* sp. 23F a partir de ácido 2-aminometilfosfónico (si previamente fue transformada por la 2-aminoetilfosfonato piruvato transaminasa, PhnW). La fosfonoacetaldehído deshidrogenasa (PhnY) reduce el fosfonoacetaldehído a fosfonoacetato, y finalmente la PhnA lo hidroliza a ácido acético y fosfato (**Figura 4**).

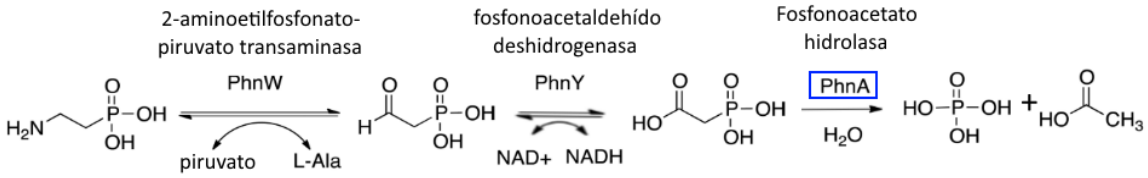

Figura 4. Conversión de ácido 2-aminoetilfosfónico hasta ser hidrolizada por la PhnA fosfonacetato hidrolasa, modificado de Horsman y Zechel, 2017.

Mecanismo III. Corte de enlace oxidativo C-P por PhnYox / PhnZ

Este mecanismo enzimático utiliza oxígeno molecular y iones de hierro para romper el enlace oxidativamente mediante los productos de los genes *phnZ* y *phnY*/phnYox* (encontrados originalmente en bacterias marinas como *Planctomyces maris* DSM8797, *Plesiocystis pacifica* SIR-1, y *Prochlorococcus* sp. MIT9303 y MIT9301), pero también presente en hongos como *Aspergillus niger*. Se ha argumentado que esta distribución entre diferentes organismos, puede hacer que varíe la especificidad de sustrato, entre diferentes organismos (Wörsdörfer et al., 2013).

Al igual que en el mecanismo de corte hidrolítico, el mecanismo III inicia con ácido 2-aminometilfosfónico, pero aquí es sometido a una oxidación en el carbono 1 (C1), pasando a ser (R)-ácido 1-hidroxi-2-aminoetilfosfónico por la 2-aminoetilfosfonato dioxigenasa (PhnY*, ahora escrita como PhnYox que no debe ser confundida con la enzima PhnY, fosfonoacetaldehído deshidrogenasa), para volver a ser oxidado en el C1, y obtener glicina y fosfato, rompiendo el enlace de una forma sencilla por medio de la enzima PhnZ (2-amino-1-hidroxietilfosfonato dioxigenasa, McSorley et al., 2012, **Figura 5**).

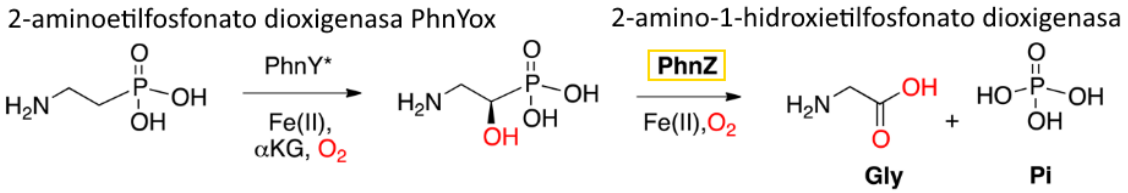

Figura 5. Conversión del ácido 2-aminietilfosfónico hasta ser oxidado por medio de las PhnY 2-aminoetilfosfonato dioxigenasa y PhnZ 2-amino-1-hidroxietilfosfonato dioxigenasa, modificado de Horsman y Zechel (2017).

2.2.1. Estructura del *operón phn* y otras proteínas involucradas en el metabolismo de fosfonatos

El *operón phn* fue descrito por Metcalf & Wanner, (1991), consta de 14 genes denotados como *phnCDEFGHIJKLMNOP* (**Figura 6**). Es muy importante considerar que la actividad del *operón phn* solamente se ha visto expresada en condiciones de inanición de fósforo. En el **Anexo S0** se describen los genes que codifican en este *operón* así como la proteína PhnPP, relacionada y las proteínas que componen los mecanismos enzimáticos II y III anteriormente descritos.

PhnC, PhnD y PhnE: Constituyen un transportador dependiente de proteínas de unión a fosfonato con la capacidad de introducir moléculas de fosfonato, fosfito y fosfato. Responsable del acoplamiento de energía al sistema de transporte. También se observó que PhnC, PhnK, PhnL y PhnN (PhnCKLN)

son cada uno similares a los componentes de la proteína de unión a nucleótidos (Metcalf y Wanner, 1991).

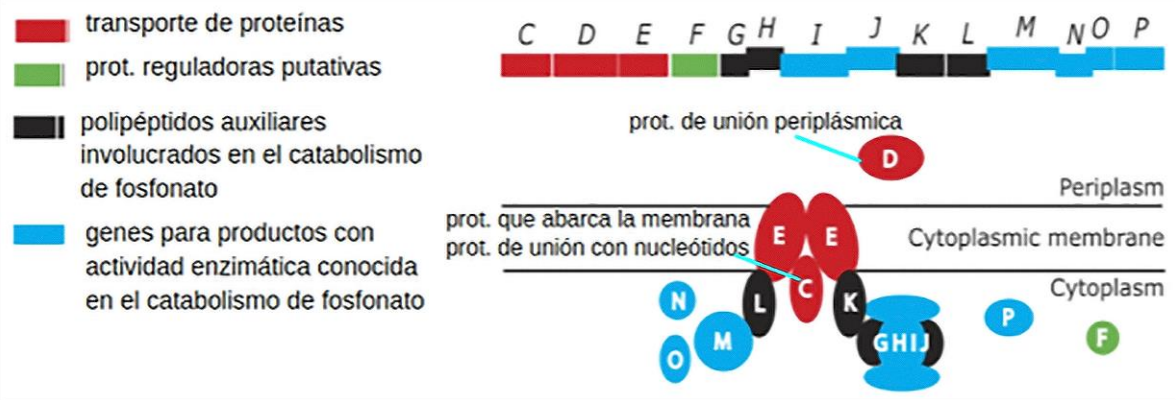

Figura 6. Funciones de los productos del *operón phn*. El *operón phn* consta de 14 genes codificantes, las esferas rojas indican el sistema de transporte ABC, las esferas azules indican polipéptidos con función enzimática asignada, la esfera verde indica el presunto represor de la expresión del *operón phn* y las esferas negras indican polipéptidos auxiliares sin una función bioquímica asignada. Modificado de Hove-Jensen, *et al.* (2014).

White & Metcalf (2007) han mostrado que la estructura del *operón phn* es variable entre diferentes organismos, algo que se puede observar con herramientas como Operon-Mapper (Taboada et al., 2018) y ello podría mostrar una versatilidad de uso del metabolismo CPn o su relación con otras enzimas del metabolismo de fosfonatos (**Figura 7**). La información resumida de las proteínas utilizadas para este estudio se muestra en el **Anexo S0.**

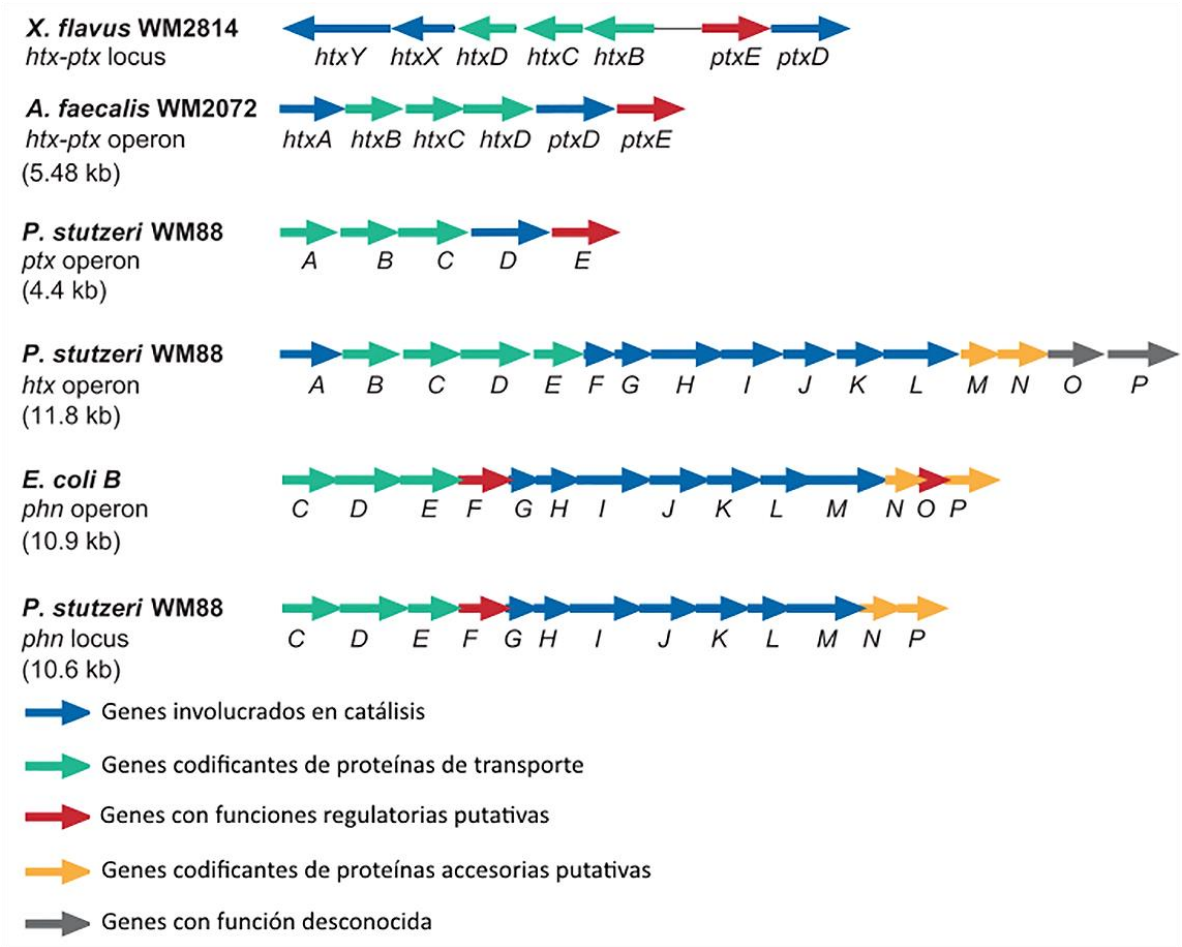

Figura 7. Estructura general del *operón phn* en diferentes especies, se observan diferentes estructuras en la conformación del *operón*. Modificada de White y Metcalf (2007).

2.4. Evolución del catabolismo de fosfonatos

Huang et al. (2005) intentaron descifrar el origen de los genes del mecanismo de corte de enlace radical de C-P por Carbono-Fósforo liasa (el mecanismo I mencionado anteriormente), y encontró que el *operón phn* está distribuido de forma no correlacionada con la filogenia de las bacterias, pues al realizar un árbol filogenético con las proteínas más importantes para el mecanismo de corte del enlace C-P, encontraron que se ha transmitido entre clados por transferencia horizontal (**Figura 8**).

Además hay diferentes versiones del *operón phn* en distintos organismos, conservando solo algunas proteínas de forma universal como son PhnGHIJKL y variando la presencia de las demás proteínas. La razón del por qué existe esta diversidad de estructura del *operón phn* en diferentes organismos es desconocida, ya que esas proteínas que no están tan conservadas son pasos importantes en la vía metabólica de corte radical C-P liasa (I), y que podría haber enzimas de sustrato más inespecífico que lleven a cabo la misma función (Hove-Jensen et al., 2014).

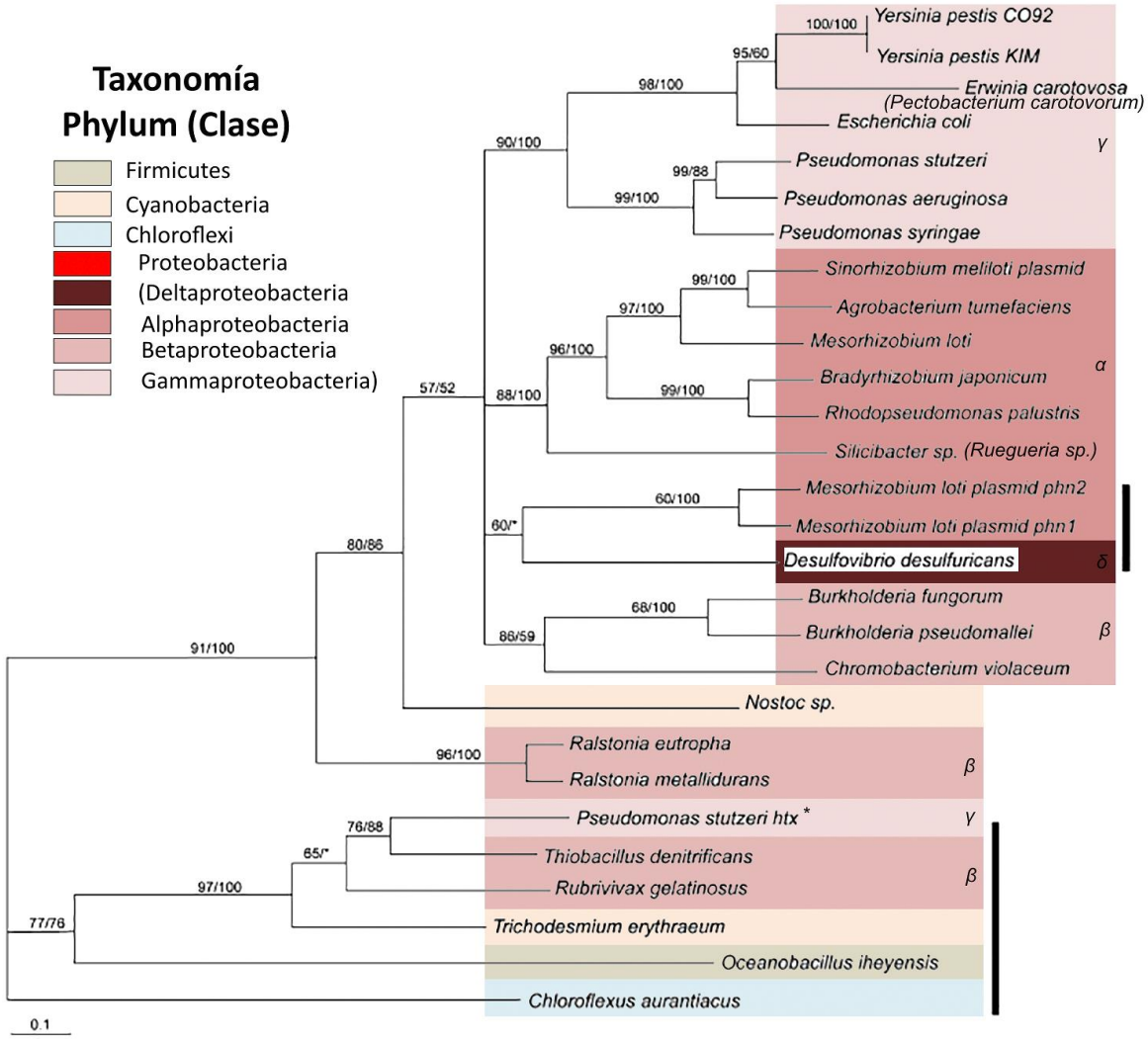

Figura 8. Árbol filogenético generado a partir de las proteínas más importantes del metabolismo I corte de enlace radical de C-P por carbono-fósforo liasa *operón phn* (proteínas PhnGHIJKL), se muestra el grupo taxonómico al que pertenecen en la clasificación de la base de datos del NCBI. Las barras verticales evidencían la transferencia horizontal de genes. Modificada de Huang et al. (2005).

Villarreal-Chiu et al. (2012) analizaron la distribución filogenética de las enzimas de metabolismo de fosfonatos en genomas bacterianos y metagenomas de ambientes marinos. Al no estar bien descritas las diferencias entre los mecanismos y no existir una descripción detallada del mecanismo CPn II (Corte de enlace oxidativo C-P por PhnYox/PhnZ) este trabajo intenta aportar información al respecto.

2.5. Búsqueda de proteínas homólogas en el proteoma bacteriano y proteomas representativos

En los genomas bacterianos, los genes relacionados funcionalmente que corresponden a rutas metabólicas o complejos proteicos comúnmente están codificados por *loci* vecinos co-regulados y co-transcritos (unidades *cistrónicas*), denominados *operones*. En la naturaleza, la constitución de los *operones* puede ser muy variada, debido a la forma en la que evolucionan estos en el genoma bacteriano, ya que además de mutaciones puede haber duplicaciones, inserciones o deleciones de cistrones dentro de los *operones*, así como rearreglos dentro del genoma y transferencia horizontal que pueden mover genes y tener una constitución muy diferente a la original (Sobecky, y Hazen, 2009).

Estos mecanismos de modificación del material genético pueden resultar en una diversificación funcional, aumentando la evolución potencial de nuevas rutas metabólicas y permite que las rutas preexistentes se adapten a los requisitos de entornos genómicos particulares. Las proteínas homólogas son distintas versiones de una proteína pero en diferentes especies (ortólogos), o bien pueden tener diferentes funciones (parálogos) (Bundalovic-Torma et al., 2019).

Para la siguiente tesis se usaron proteomas anotados como "**representativos**" de la base de datos de Refseq (O'leary et al., 2015). Los

genomas anotados son producto de un proceso de ensamblaje *de novo* a partir de la secuencia de DNA de un organismo, iniciando con la identificación de genes de RNA y codificantes para identificar ORF (Open Reading Frames en inglés, o MAL Marcos Abiertos de Lectura) en la secuencia del genoma.

Con esos ORFs se buscan homólogos con el programa BLAST contra bases de datos como GenBank y UniProt para identificar funciones putativas y evidencia de proteínas. Los ORF se asignan a las vías metabólicas utilizando una base de datos KEGG. Los dominios de proteínas se identifican mediante búsquedas de InterProScan. Los ORF se buscan en la base de datos del dominio conservado que incluye los COG para identificar los ortólogos correspondientes (Christoffels & Van Heusden, 2019).

Los genomas representativos se calculan y seleccionan computacionalmente para representar diversas especies, por lo que en el presente trabajo se utilizan proteomas completos (total de proteínas codificadas por un genoma) de bacterias (Wilkins, 2009).

3. Antecedentes y Justificación

Los microorganismos llevan a cabo funciones muy diversas en el balance y funcionamiento de los ciclos biogeoquímicos llevados a cabo por grupos funcionales de microorganismos con metabolismos especializados en la solubilización de nutrientes como el nitrógeno y el fósforo en el suelo con organismos de la familia Nitrososphaeraceae (Wu et al., 2021).

Específicamente el fósforo se encuentra en gran diversidad de composiciones químicas y organismos especializados pueden usar las formas menos accesibles de fósforo como compuestos organofosforados en condiciones de estrés (McMullan & Quinn, 1994) para introducir este recurso a la red trófica.

Ciertos productos de degradación de vías metabólicas relacionadas con el fosfonato han causado un gran misterio en la comunidad científica por muchos años con la llamada “Paradoja de la sobresaturación de metano”, la cual consiste en la exagerada producción de metano en ambientes acuáticos que no cumplen con las condiciones comunes de producción de este gas sencillo, tales como la estricta anaerobiosis.

Gracias al grupo de Repeta et al. (2016) y Wang et al. (2017) queda claro que este problema está relacionado con el metabolismo del metilfosfonato y otros compuestos aminoetil-fosfonatados, pues su biodegradación puede verse exacerbada por la actividad de los microorganismos fijadores de nitrógeno y en condiciones de deficiencia de nutrientes en el medio marino que probablemente sean el resultado del calentamiento de los océanos causado por los gases de efecto invernadero. Los microorganismos capaces de realizar este metabolismo tienen proteínas componentes de la C-P liasa.

Los avances en la secuenciación a partir de 2012 han permitido conocer que el 10% de los genomas bacterianos secuenciados contenían genes para la biosíntesis de Pn mientras que el 40% de los genomas bacterianos contenían

genes que codifican a proteínas del CPn por uno o más de tres mecanismos catabólicos. Inclusive la ampliamente utilizada *E. coli*, que evolucionó en el ambiente rico en Pi del intestino de los mamíferos, ha retenido los genes que codifican la C-P liasa (Horsman y Zechel, 2017).

En el caso del mecanismo catabólico I por medio de la C-P liasa, se encontró que es capaz de degradar un amplio espectro de organofosfonatos, incluyendo el glifosato (Hove-Jensen et al, 2014), por lo que detectar bacterias que posean este metabolismo, potencialmente podrían utilizarse para disminuir el contenido de organofosfatos contaminantes en el suelo y en el agua.

Morales, et al. (2020) crean los *primers* de las proteínas de las enzimas más importantes de la metabólica de corte de enlace radical por medio de la C-P liasa con el objetivo de en un futuro utilizar la biorremediación con bacterias modificadas genéticamente, mostrando el gran interés en este tipo de tecnología.

4. Objetivos

4.1. General

Identificar y describir la distribución filogenética de los genes codificantes y proteínas predichas del catabolismo de fosfonato en proteomas representativos del dominio Bacteria.

4.2. Particulares

1. Identificar las proteínas relevantes en el catabolismo de fosfonatos.
2. Construir modelos ocultos de Márkov (HMMs) de cada una de las enzimas identificadas en el catabolismo de fosfonatos.
3. Utilizar los HMMs, para buscar e identificar secuencias homólogas a proteínas ya descritas de los 3 mecanismos enzimáticos de catabolismo de fosfonato en los proteomas bacterianos disponibles en www.ncbi.nlm.nih.gov/refseq/
4. Describir la abundancia y distribución filogenética en Bacteria de las secuencias homólogas de aminoácidos de las proteínas del catabolismo de fosfonatos (CPn) y su distribución filogenética.

5. Metodología

Procedimientos bioinformáticos

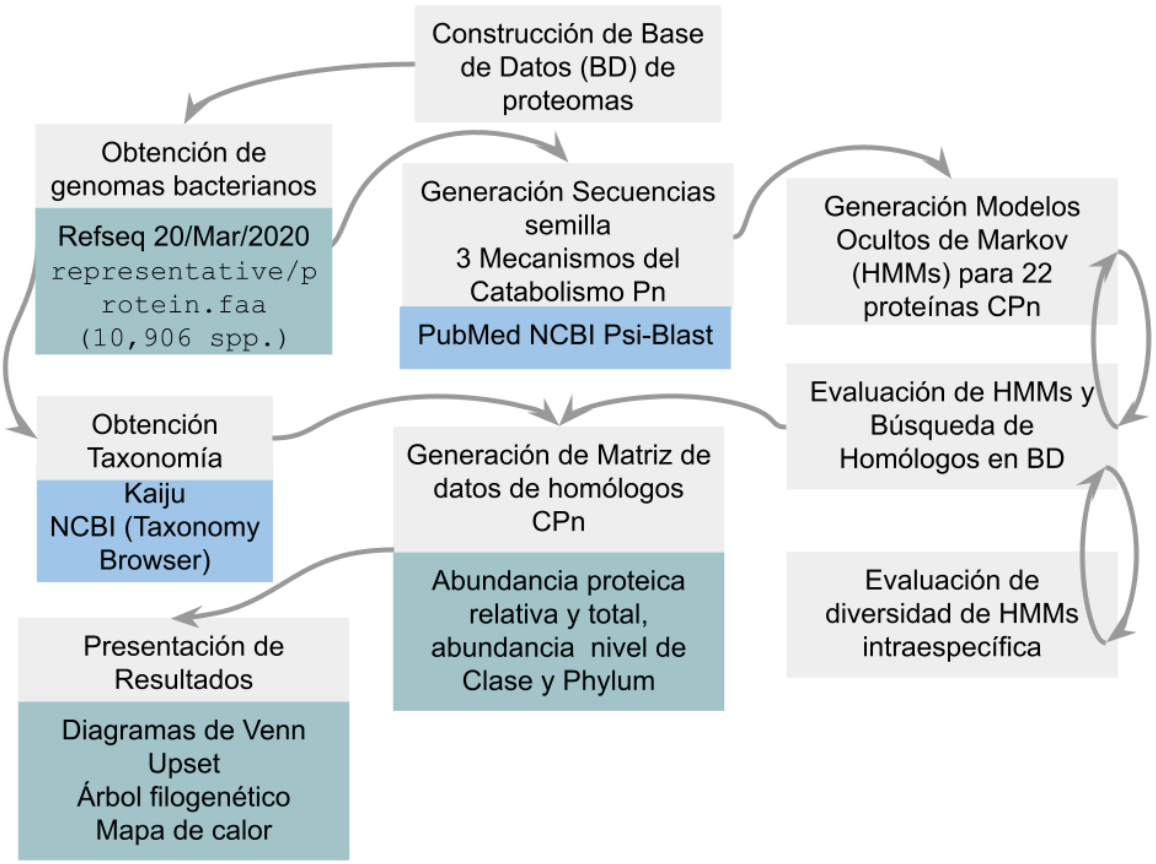

Diagrama de flujo. Resumen de los procedimientos realizados para este trabajo de tesis.

5.1. Obtención de proteomas bacterianos

Para la creación de la base de datos local de proteomas bacterianos se utilizó el siguiente comando en la terminal de bash:

```
$ rsync -ahvmR --copy-links rsync://ftp.ncbi.nih.gov/genomes/refseq/bacteria/./*/representative/*/*protein.faa.gz Documents/Bacteria/
```

rsync descarga los archivos desde una base de datos dada una dirección web (por ejemplo Refseq). En este caso los proteomas de bacterias en formato *.faa* que muestra proteínas anotadas en formato fasta (aminoácidos) del grupo de proteomas representativos (el conjunto de los proteomas representativos contiene a las especies que tienen al menos 10 presentaciones de genomas, es decir que fueron unidos de diferentes trabajos de secuenciación y anotación, O'leary et al., 2015). A la fecha (marzo 2020), se descargaron 10,906 proteomas representativos de Bacteria (**Anexo S1**).

Luego se concatenaron los archivos que pertenecieran a la misma especie para evitar problemas al manejar los archivos y se borraron los demás, se usó el comando:

```
     $      cat      GCF_000175255.2_ASM17525v2_protein.faa
GCF_000218875.1_ASM21887v1_protein.faa                    >
```

5.1.1 Taxonomía de los proteomas bacterianos

Se usó el método de Ciccarelli et al. (2006) que usa Clusters de Genes ortólogos (COGs) que están inocuamente distribuidos en los organismos.
El método consistió en buscar los COGs de referencia que se usaron en dicho artículo (COG0016, COG0052, COG0081, COG0092, COG0172, Tabla suplementaria S1 del artículo de Ciccarelli):

Copiar los identificadores de los ortólogos en un archivo, se buscaron las secuencias COGs en la página ftp://ftp.ncbi.nih.gov/pub/COG/COG/, (Anexo github.com/LChora/LChora/blob/main/COGs_bacteria.bz2) usando el archivo whog, el cual tiene una lista de nombres de ortólogos de proteínas en diferentes organismos. Se hizo una copia local para cada COG y se modificó para que quedara dentro del archivo lista_COGnum con el formato de *myva:nombre_ortólogo*. Para que se realice una búsqueda del archivo myva las secuencias que tengan los nombres de las proteínas que están en la

lista_COGnum se usó por medio del programa EMBOSS:6.6.0.0 (https://www.bioinformatics.nl/cgi-bin/emboss/help/xmltext) con el comando:

```
$ seqret -auto list:lista_COG0num" >
```

seqret realiza una búsqueda de la secuencia a partir de los nombres de identificadores que estén en la lista.

A partir de éstas secuencias, se utilizó el programa MUSCLE v3.8.31 (Edgar, 2004) para realizar una alineación múltiple con el programa de clustalW, con la función:

```
$ muscle -in secuencias_COG0num.faa -out COG0num.aln
```

donde a *muscle* se le da la instrucción de qué usar como archivo de entrada *-in* y como archivo de salida *-out*.

A partir del alineamiento múltiple *COG0num.aln*, con la intención de encontrar proteínas homólogas en los genomas de la base de datos local para este trabajo, se realizó un Modelo Oculto de Márkov (HMM) con HMMER 3.1b2 (Eddy, 1998):

```
$ hmmbuild COG0num.hmm COG0num.aln
```

Para buscar una secuencia a partir de un perfil HMM y dar un resultado en forma de tabla en el archivo *genoma_protein.faa.out* (para ver el archivo de salida ir a **Código 1**):

```
$ hmmsearch --tblout ruta/del/genoma_protein.faa.out
ruta/del/COG/COGnum.hmm ruta/del/genoma_protein.faa
```

Una vez obtenido el perfil HMM, se usó el parámetro *bit-Score*, que es una medida para observar el grado de confiabilidad del alineamiento y que es independiente del tamaño de la base de datos de la referencia (diferente a la medida de e-value). Para hacer esta búsqueda se usó el comando *awk*, el cual busca un cierto valor (en la columna *$6* del archivo de texto), que con la

condicional de mayor o igual un cierto valor de bit-Score especificado imprime el nombre de la proteína buscada y ese valor y lo guarda en *identificador_obtenido_de_hmmsearch.fin*:

```
$ awk '{if($6 >= 285) print $6, $1}' ruta/del/genoma_protein.faa.out > ruta/del/identificador_obtenido_de_hmmsearch.fin
```

En el caso de obtener más de un identificador, se unificó, seleccionando la proteína que tuviese la puntuación más alta de bit-Score para posteriormente buscar en cada proteoma la secuencia de aminoácidos de la proteína encontrada con el comando:

```
$ seqret -auto -stdout ruta/del/genoma_protein.faa:identificador_obtenido_de_hmmsearch > ruta/representative.faa
```

El cual extrae la secuencia en el proteoma *ruta/del/genoma_protein.faa* para posteriormente ser usado en Kaiju (Menzel et al., 2016a).

Como antecedente, se utilizó un script local del laboratorio (https://github.com/genomica-fciencias-unam/SOP/blob/master/scripts/header.fasta.numbers.pl) para renombrar el contenido de una secuencia al nombre del archivo y eliminar los saltos de línea (es decir *n*) que pueden complicar el análisis de Kaiju. Para ello se usó:

```
$ perl header.fasta.numbers.pl Nombre_Especie_COG00num.faa
```

Posteriormente se concatenaron todos los proteomas en un solo archivo usando:

```
$ cat */representative/*.numbered.fas > COGs bacteria
```

Este último archivo es el que se usó en el programa Kaiju, como lo establece su misma página de uso (Menzel et al., 2016).

El resultado del análisis se vació en una hoja de cálculo (**Anexo S1**), y se curó usando la información del taxonomy browser del NCBI (Federhen, 2012).

5.2. Obtención de secuencias semilla de proteínas de catabolismo de fosfonato (CPn)

Se llevó a cabo una revisión bibliográfica en el servidor PubMed (http://www.ncbi.nlm.nih.gov/pubmed) y con ayuda de herramientas como STRING (Szklarczyk et al., 2021) se enlistaron inicialmente 22 proteínas que participan en el catabolismo de fosfonato en bacterias, que participan directamente en la actividad de corte de la C-P liasa, corte hidrolítico y oxidativo del enlace carbono-fósforo (C-P) de bacterias. Estas se resumen en la **Tabla S1**.

Con el objetivo de buscar proteínas homólogas necesarias para realizar un alineamiento y poder construir un HMM se usaron las proteínas que se encuentran en tres mecanismos enzimáticos de catabolismo de fosfonato como base para hacer un PSI-BLAST (Position-Specific Iterated BLAST https://blast.ncbi.nlm.nih.gov/). Los parámetros para esta búsqueda con BLAST fueron los siguientes: matriz de sustitución BLOSUM45 con un límite de 0.005 con 4 iteraciones y un rango de 70-100% de identidad.

Se dividieron las proteínas en la siguiente clasificación de acuerdo al mecanismo enzimático al que pertenecen (en total 22 proteínas: 15 del mecanismo I, 5 del mecanismo II y 2 del mecanismo III):

Corte de enlace radical de C-P por carbono-fósforo liasa: *phnCDEFGHIJKLNOP* de *Escherichia coli* str. K-12 substr. MG1655.

Corte hidrolítico del enlace C-P: *phnA* de *Pseudomonas fluorescens* 23F / *Sinorhizobium meliloti* 1021, PalA de *Burkholderia* / *Variovorax* sp., PhnX de

Bacillus cereus / *Salmonella typhimurium*, *phnW* de *Salmonella typhimurium*, PhnY de *Sinorhizobium meliloti* sp. 1021.

Corte de enlace oxidativo C-P por PhnYox / PhnZ: PhnYox y PhnZ de *Planctomyces maris* DSM8797.

El resultado de esto se guardó en un archivo multifasta y en https://github.com/LChora/LChora/blob/main/HMMs.zip se anotan todas las proteínas utilizadas para generar los alineamientos a partir de los cuales se hicieron los perfiles de modelos ocultos de Márkov (HMMs por sus siglas en inglés *Hidden Márkov Model*).

Las secuencias semilla que se utilizaron pertenecen a las especies *Acidithiobacillus ferrooxidans, Agrobacterium tumefaciens, Bacillus cereus, Bacillus megaterium, Bacillus terrae, Bradyrhizobium japonicum, Burkholderia pseudomallei, Chloroflexus aurantiacus, Cupriavidus necator, Desulfotignum phosphitoxidans, Escherichia coli, Klebsiella aerogenes, Klebsiella oxytoca, Klebsiella pneumoniae, Kluyvera ascorbata, Kluyvera cryocrescens, Mesorhizobium loti, Mycolicibacterium smegmatis, Nostoc sp. PCC 7120, Ochrobactrum anthropi, Prochlorococcus marinus, Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas stutzeri, Rhodobacter capsulatus, Ruegeria pomeroyi, Salmonella enterica, Sinorhizobium meliloti, Trichodesmium erythraeum* (Horsman y Zechel, 2017; Stosiek *et al*., 2019), **Anexo S2**.

5.3. Generación de perfiles de Modelos Ocultos de Márkov (MOM) para las proteínas CPn

A partir de los archivos multifasta de cada proteína, se realizó un alineamiento múltiple que utiliza probabilidades p_i y p_{ij} derivadas de la matriz 240 PAM VTML (Edgar, 2004), usando estos archivos:

```
$ muscle -in phnJ.faa -out phnJ.aln -clw
```

con la secuencias alineadas, se generó un perfil HMM:

```
$ hmmbuild --amino phnJ.hmm phnJ.aln
```

La búsqueda de perfiles fue realizada con ayuda del programa HMMER. El comando utilizado para buscar secuencias homólogas en los proteomas fue usado de la siguiente manera:

```
$ hmmsearch --tblout phnJ.out phnJ.hmm phnJ.faa
```

donde *hmmsearch*: es la instrucción para buscar un perfil HMM en una base de datos (en este caso la base de datos local)

--tblout es una opción del comando hmmsearch que permite obtener resultados en formato tabular

phnJ.out define el nombre del archivo de texto con los principales resultados de la búsqueda en phn J.faa

phnJ.hmm es el perfil HMM que será buscado, en este caso el de la proteína C-P liasa.

en *phnJ.out* aparecen los valores de la búsqueda de homólogos al perfil HMM.

```
#                                                               --- full sequence ---- --- best 1 domain ---- --- domain number estimation ----
# target name     query name  E-value  score  bias   description of target
WP_000002282.1 phnJ       7.2e-162  535.1    0.1    alpha-D-ribose 1-methylphosphonate 5-phosphate C-P-lyase PhnJ [Escherichia coli]
#
# Program:          hmmsearch
# Version:          3.1b2 (February 2015)
# Pipeline mode:    SEARCH
# Query file:       Structure/phn/phnJ.hmm
#                                                                 Target                                                                 file:
Bacteria/Escherichia_coli/representative/GCF_003697165.2_ASM369716v2/GCF_003697165.2_ASM369716v2_protein.faa
#    Option    settings:    hmmsearch    --tblout    Bacteria/Escherichia_coli/representative/phnJ_Escherichia_coli.out
```

Código 1. Salida de archivo *.out* cual muestra la alpha-D-ribose 1-methylphosphonate 5-phosphate C-P-lyase PhnJ (alfa-D-ribosa 1-metilfosfonato 5-fosfato C-P-liasa) con sus valores correspondientes de *7.2e-162* e-value y *535.1* bit-score.

5.3.1 Evaluación de modelos ocultos de Márkov (HMM)

Antes de realizar las búsquedas de los perfiles HMM en los proteomas bacterianos de la base de datos local, se realizaron diversas pruebas.
Para evaluar la capacidad de los perfiles HMM en los proteomas de bacterias, la primera prueba consistió en evaluar la capacidad de los perfiles para encontrar proteínas homólogas en algunos proteomas control, que previamente hayan sido reportados en la literatura, o aparezcan como homólogos en el programa de Gene Context Tool (Martinez-Guerrero et al., 2008). Estos organismos control se muestran en el **Anexo S2**. La búsqueda se realizó usando *hmmsearch*.

Se definió un rango de selección de *bit-Score* para discriminar entre secuencias homólogas y no homólogas. Este parámetro fue elegido debido a que su cálculo es independiente de la base de datos de referencia sobre la que se está trabajando para construir el perfil de HMM (Eddy, 2010), lo que permite comparar similitud de secuencias de los diferentes proteomas (González, 2016).

El rango de selección fue definido con el *bit score* máximo y mínimo posible, que pudieran tener las proteínas homólogas conocidas para dicha familia de proteínas. El bit score máximo y mínimo se obtuvo al comparar el perfil HMM de cada proteína contra cada una de los proteomas usados en el paso anterior.

Para analizar detenidamente los datos que definieron al rango de selección, se decidió graficar la distribución de los bit score, en un diagrama de caja, el cual fue generado en el programa R (versión 3.0.2) (R Core Team, 2013a). Los diagramas de caja se muestran en https://github.com/LChora/LChora/blob/main/boxplot-distr-bitscore-proteina.png así como el rango de selección, la cantidad de secuencias con las que fue calibrado cada perfil y la longitud de cada perfil.

Cada perfil HMM se representó gráficamente como un HMM logo usando el programa Skylign en línea (Wheeler et al., 2014). Un HMM logo permite visualizar la información contenida y la distribución de proteínas de un alineamiento múltiple

de secuencias (Demarchi et al., 2006), estos se incluyen en https://github.com/LChora/LChora/blob/main/HMMsLogos.zip.

5.3.2 Búsqueda de homólogos con perfiles HMM

Al igual que en el caso de la evaluación de los perfiles, se usó el programa HMMER:

```
$ hmmsearch --tblout ruta/del/genoma_phnJ.out phnJ.hmm ruta/del/genoma.faa
```

Se usó el valor del bit-score para calcular la media y la desviación estándar para definir un valor de corte para incluir a las proteínas que sí entran en la búsqueda de homólogos, para el cual, en cada caso específico se usó su propio valor de corte para buscar en la base de datos local de proteomas bacterianos:

```
$ awk '{if($6 >= 285) print $6, $1, $3, $19, $20, $21, $22, $23, $24, $25, $26, $27, $28}' ruta/del/genoma_phnJ.out | grep -e '^[0-9][0-9][0-9].' >
```

En la columna $6 de la salida .out aparece el valor de bit-score, con la condicional *if* se seleccionan sólo los valores mayores al valor de corte previamente establecidos para cada proteína, el valor $1 pertenece al identificador de la proteína, $3 al nombre de la proteína buscada, y de $19-28 es la descripción con el que está anotada esa proteína.

Este método se corrió para cada una de las proteínas en los 10,906 proteomas de nuestra base de datos. Los resultados de presencia de secuencias homólogas de proteínas CPn se encuentran en el **Anexo S3**.

Las secuencias utilizadas de Refseq se encuentran anotadas como 'WP' lo cual significa que una secuencia de proteína procariota no redundante, que corresponde a traducciones idénticas de al menos una, pero a veces miles de regiones de secuencia codificadora (CDS por sus siglas en inglés) anotadas en

proteomas de RefSeq. Cada proteína no redundante de RefSeq representa un grupo de proteínas idénticas (IPG), así como traducciones de ensamblajes Ref-Seq de secuenciación completa del genoma (WGS), de modo que es esperable que haya secuencias de proteínas que no se encuentren en el genoma utilizado (Haft et al., 2018).

5.3.3 Evaluación de la diversidad de la presencia de proteínas entre bases de datos

Debido al carácter histórico de la investigación del *operón phn* y la variante *htx* en la especie Pseudomonas stutzeri (White y Metcalf, 2004; Kononova y Nesmeyanova, 2002), y dado que los HMMs de este trabajo no fueron capaces de encontrar las respectivas proteínas homólogas para esta especie, se decidió analizar la diversidad de perfiles de presencia de proteínas CPn en la misma especie pero en diferentes genomas de las bases de datos de Refseq y Genbank.

Dado que se usaron proteomas representativos de la base de datos de RefSeq, (O'leary et al., 2015, https://ftp.ncbi.nih.gov/genomes/refseq/bacteria/), se compararon los que se encuentran en GenBank (ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/). La base de datos de GenBank tiene una cantidad mayor de secuencias al ser una base redundante, es decir que hay muchas secuencias que pertenecen a la misma especie pero diferentes cepas.

Se utilizaron varias cepas de *P. stutzeri*, se conoce que hay diversidad de cepas como *P. stutzeri WM88* que tienen ambos *operones phn* y *htx*, sin embargo, al realizar la búsqueda en el proteoma de la base de datos de proteomas representativos de Refseq, no se encontraron los genes buscados, por lo que se buscaron en GenBank. Para la mejor visualización de esta información, se realizó un mapa de calor (heatmap) como en la **Figura 16**.

5.4. Generación de matriz de datos con los homólogos del catabolismo de fosfonatos (CPn)

Para analizar la abundancia y la distribución de las proteínas homólogas encontradas con los perfiles HMM, se generó una matriz de datos (**Tabla S3**). En las columnas se ordenaron las proteínas analizadas y en las filas se colocaron los nombres de los proteomas organizados por phylum y por clase. En las celdas intermedias se agregó el número de las copias de proteínas homólogas halladas para cada proteoma.

Al observar que existían proteomas con más de una copia para una misma proteína, se tomaron solo los que tuvieran mayor puntuación de bit-score, en posteriores análisis se contaron únicamente como presencia o ausencia de los genes estudiados. Estos casos de proteínas con "exceso de copias" son el caso de algunas proteínas que se encuentran dentro de familias génicas grandes como los transportadores acoplados a ATP (proteínas PhnC, PhnD y PhnE), por lo que los HMMs pueden estar haciendo falso positivo al encontrar proteínas homólogas que no estén relacionadas con el catabolismo del Pn.

5.4.1 Abundancia a nivel de proteoma y total de proteínas

Para contar el total de proteínas por proteoma se usó el comando:

```
$ grep -o -Ee '>' ruta/del/genoma.faa | echo `wc -l` ruta/del/genoma.faa
```

Luego se sumaron el total de proteínas homólogas de cada proteína dentro del archivo que se generó usando el archivo de la sección anterior:

```
$ wc -l ruta/del/genoma phnJ.out.fin
```

La abundancia total de proteínas se obtuvo al contar todos los homólogos encontrados con los HMMs. Adicionalmente, se reportó la cantidad de homólogos encontrados por cada una de las proteínas de CPn analizadas. Los resultados se

vaciaron en una hoja de cálculo, en la cual se sumó lo encontrado para el mecanismo de corte de enlace por C-P liasa, y en conjunto los de corte de enlace oxidativo y corte hidrolítico. La abundancia total fue graficada en un histograma (**Figura 10**). A partir de esta abundancia se obtuvo el porcentaje de cada proteína, tomando como el 100% el total de proteínas encontradas en todo el taxón.

5.4.2 Abundancia a nivel de clase y phylum

En la abundancia a nivel de phylum se obtuvieron tres datos. El primer dato hace referencia a la abundancia de proteínas por phylum (**Tabla 1**). El segundo dato deriva de la abundancia de cada phylum, normalizada entre el número de organismos analizados en el respectivo phylum. El tercer dato, muestra la proporción en la que fueron encontradas las proteínas CPn en cada phylum. Por ejemplo, en el caso del phylum Proteobacteria, el cual tiene una abundancia muy predominante (**Figura 9**), se analizó la abundancia relativa de la proteína PhnJ respecto al total de proteínas encontradas en cianobacterias. El resumen por phylum y clase se muestra en el **Anexo S3**.

A partir de estos datos en conjunto con los datos obtenidos de la sección de taxonomía, se graficaron las categorías taxonómicas de clase y phylum y la imagen fue editada para mejor comprensión con el programa paint.NET 4.2.13 (https://www.getpaint.net/) (**Figura 9**).

5.5. Diagramas de Venn, comparación de conjuntos de proteínas CPn

También se realizaron diagramas de Venn para visualizar más fácilmente la presencia de las proteínas en los proteomas bacterianos, para esto se usaron 2 paquetes diferentes de R, el primero fue con la función venn 1.9 (2020) y upset 1.4.0 (2019). El primero (venn) con la idea de observar conjuntos particulares y el

segundo (upset) de observar la diversidad de conjuntos que pueden encontrarse en los proteomas bacterianos seleccionando a los que estén reportados en la literatura y que sean significativos biológicamente (que exista la posibilidad de interacción metabólica entre las proteínas), en qué organismos, qué tipo de interacciones son preguntas que se abordarán en la parte de discusión. Se usaron los siguientes scripts:

Para los diagramas de Venn (**Figura 12**) se usó :

```
# Cargar Librerías que se necesitan para que se genere el diagrama de venn
library(VennDiagram); library(venn); library(ggplot2)
#Cargar datos de Tabla
resultS<- read.csv("~/Yayo/R/Taxonomy_Kaiju_Resultados_s.csv", row.names=1)
venn_prot <-read.csv("~/Yayo/R/Venn_prot.csv", header=T, sep=",")


likes <- function(dpn) {
  ppl <- venn_prot
  names(ppl) <- c("Phylum", "Clase","phnC", "phnD", "phnE", "phnF", "phnG",
"phnH", "phnI", "phnJ", "phnK", "phnL", "phnM", "phnN", "phnO", "phnP", "phnPP",
"palA", "phnA", "phnW", "phnX", "phnY", "phnZ", "phnYox")
  for (i in 1:length(dpn)) {
    ppl <- subset(ppl, ppl[dpn[i]] > 0)
  }
  nrow(ppl)
}
#dibujar hasta 5 conjuntos
plotProt <- function(a, ...) {
  grid.newpage()
  if (length(a) == 1) {
    out <- draw.single.venn(likes(a), ...)
  }
  if (length(a) == 2) {
    out <- draw.pairwise.venn(likes(a[1]), likes(a[2]), likes(a[1:2]), ...)
  }
  if (length(a) == 3) {
    out <- draw.triple.venn(likes(a[1]), likes(a[2]), likes(a[3]),
                            likes(a[1:2]),   likes(a[2:3]),   likes(a[c(1,
3)]), likes(a), ...)
  }
  if (length(a) == 4) {
    out   <-   draw.quad.venn(likes(a[1]),   likes(a[2]),   likes(a[3]),
likes(a[4]),
                            likes(a[1:2]),  likes(a[c(1,  3)]),  likes(a[c(1,
4)]), likes(a[2:3]),
                            likes(a[c(2,  4)]), likes(a[3:4]), likes(a[1:3]),
likes(a[c(1, 2,

4)]), likes(a[c(1, 3, 4)]), likes(a[2:4]), likes(a), ...)
  }
  if (length(a) == 5) {
    out  <-  draw.quintuple.venn(likes(a[1]),  likes(a[2]),  likes(a[3]),
likes(a[4]), likes(a[5]),
                            likes(a[1:2]),  likes(a[c(1,  3)]),  likes(a[c(1,
4)]),likes(a[c(1, 5)]),
                            likes(a[2:3]),  likes(a[c(2,  4)]),  likes(a[c(2,
5)]),
```

```
                                likes(a[3:4]),  likes(a[c(3,  5)]),  likes(a[c(4,
5)]),
                                likes(a[1:3]), likes(a[c(1, 2, 4)]), likes(a[c(1,
2, 5)]), likes(a[c(1, 3, 4)]), likes(a[c(1, 3, 5)]),
                                likes(a[c(1,    4,    5)]),        likes(a[2:4]),
likes(a[c(2, 3, 5)]),  likes(a[c(2, 4, 5)]), likes(a[c(3, 4, 5)]),
                                likes(a[1:4]),    likes(a[c(1,    2,    3,    5)]),
likes(a[c(1, 2, 4, 5)]),  likes(a[c(1, 3, 4, 5)]), likes(a[2:5]),
                                likes(a[1:5]), likes(a), ...
                                )
  }
  if (!exists("out"))
    out <- "Oops"
  return(out)
}

plotProt(c("phnC", "phnD", "phnE"),
         category =c("phnC", "phnD", "phnE"),
         lty = "blank", fill = c("skyblue", "pink1", "mediumorchid")
)
… # y así dibujar los conjuntos de interés
```

Código 2. Script para la creación de diagramas de Venn.

Y para los upset y observar mayor número de conjuntos (**Figura 13**):

```
# Cargar Librerías que se necesitan para que se genere el upset
install.packages("UpSetR");  library(UpSetR);  install.packages("turner");
library(turner); install.packages("ComplexHeatmap"); library(ComplexHeatmap)
#Cargar datos a partir de una hoja de cálculo en formato csv y
visualización en forma de Tabla
data    <-    read.csv("C:/Users/Dell/Documents/Yayo/R/venn/Taxonomy_Kaiju-
Resultados_s_m.csv")
View(data)

#Diseño de Upset, carga de datos y selección de conjuntos principales
graf <- upset(
  (data),
  sets = c(
    "palA", "phnA", "phnW", "phnX", "phnY"
    ,"phnZ", "phnYox"
    ,"phnC",  "phnD",  "phnE",  "phnF",  "phnG",  "phnH",  "phnI",  "phnJ",
"phnK", "phnL", "phnM", "phnN", "phnO", "phnP"
    ),
  number.angles = 0, point.size = 3, line.size = 1,
  mainbar.y.label = "Intersecciones de genes",
  sets.x.label = "No. de proteomas por proteínas",
  text.scale = c(1.7, 1.5, 1.5, 1, 1.5, 1.2),
  mb.ratio = c(0.55, 0.45), order.by = c("degree"),
  keep.order = TRUE,
  nintersects = 1000,
  group.by = "sets",
  cutoff = 3,
  sets.bar.color = c(
    "blue", "blue", "blue"
    , "blue", "blue"
    , "orange", "orange"
    ,
"red","red","red","firebrick","indianred","indianred","firebrick2","firebrick2","
indianred","indianred","firebrick2","firebrick2","firebrick2","firebrick2"
    )
)
graf
```

Código 3. Script para la creación de Upset.

Los datos utilizados para la creación de este upset y las correspondientes se encuentran en el **Anexo S4**.

5.6. Determinación de la distribución filogenética de los homólogos de las proteínas del Catabolismo de fosfonatos (CPn)

Para analizar los patrones de distribución de las proteínas del CPn, se graficó un heatmap asociado a una filogenia con la taxonomía a nivel de phylum y clase donde se observaron los emparentamientos filogenéticos (**Figura 16**).

5.6.1 Construcción del árbol filogenético

El árbol filogenético que se realizó se basó en el de Ciccarelli et al., (2006), haciendo coincidir el orden de la tabla con el de la filogenia, en la filogenia base se eliminaron los taxa que pertenecieran a más de una especie (como diferentes cepas) y se agregaron al menos una especie de cada phylum, respetando en todo momento la topología de la filogenia base. La filogenia propuesta por Ciccarelli fue modificada y obtenida del servidor iTOL (http://itol.embl.de/itol.cgi), en la sección de "TREE OF LIFE", y se modificó con el programa de edición de imágenes *Paint.net* en donde también se incluyó el heatmap de presencia de proteínas CPn.

5.6.2 Construcción del mapa de calor (heatmap)

El heatmap fue construido a partir de la matriz de la presencia y ausencia de proteínas del CPn generada previamente, en el programa R con la función *heatmaply* de la paquetería heatmap 3.6.3 (Galili et al., 2018) con el siguiente script:

```
       # Instalar paquetes y Cargar Librerias que se necesitan para que se genere el
heatmap
       install.packages('heatmaply');                          install.packages("broom");
install.packages('d3heatmap');                                 install.packages('ggplot2')
library(tidyverse);library(hrbrthemes);        library(viridis);        library(plotly);
library(heatmaply); library(ggplot2)


       #  Cargar datos desde una hoja de calculo en formato csv
       data                                                                              <-
read.table(file="C:/Users/Usuario/Documents/Luis/Tesis/Rscripts/Taxonomy_Kaiju-
Resultados_g_m.csv", header=T, sep=",")
       colnames(data) <- gsub("\\.", "", colnames(data))


       # Seleccionar ciertos organismos, aqui incluimos los organismos que queremos que
sean representados en el heatmap, esto se hace si tienes muchos datos en la hoja de
calculo
       org <- c("Clostridium acetobutylicum"...) #lista de organismos usados
       data <- data %>%
         filter(Especie %in% org) %>%
         mutate(Especie = factor(Especie, Especie))


       # Formateo de los datos a forma de matriz
       matz <- data
       rownames(matz) <- matz[,1]
       matz  <-  matz  %>%  dplyr::select(-Especie,...)  #quitar  todas  las  columnas  no
numericas
       matz <- as.matrix(matz)


       #  Dibujar  el  Heatmap    d3heatmap(matx,  scale="column",  dendrogram  =  "none",
width="800px", height="80Opx", colors = "Blues")
       #png("C:/Users/Usuario/Documents/Luis/Tesis/Rscripts/heatmap_filogenia.png")
```

Código 4. Script usado en la construcción del heatmap.

6. Resultados

6.1. Descripción taxonómica de Bacteria en la base de datos local

La clasificación completa para cada especie se encuentra en el **Anexo S1**. En la base de datos de proteomas descargada de Refseq (conocida como base de datos local) se analizaron 10,906 especies bacterianas, distribuidas en 36 phyla y 80 clases. De los cuales se encontraron 4,413 proteomas (40.46% del total) con al menos una secuencia homóloga de proteína de catabolismo de fosfonatos (CPn) calculada por medio de los modelos ocultos de Márkov (HMM), distribuidas en 24 phyla y 50 clases.

La distribución de estos proteomas se encuentra en el **Anexo S5** y la **Figura 9** (el restante 59.54% se excluye en las siguientes descripciones debido a que las especies no se les detectó ninguna proteína del catabolismo de Pn).

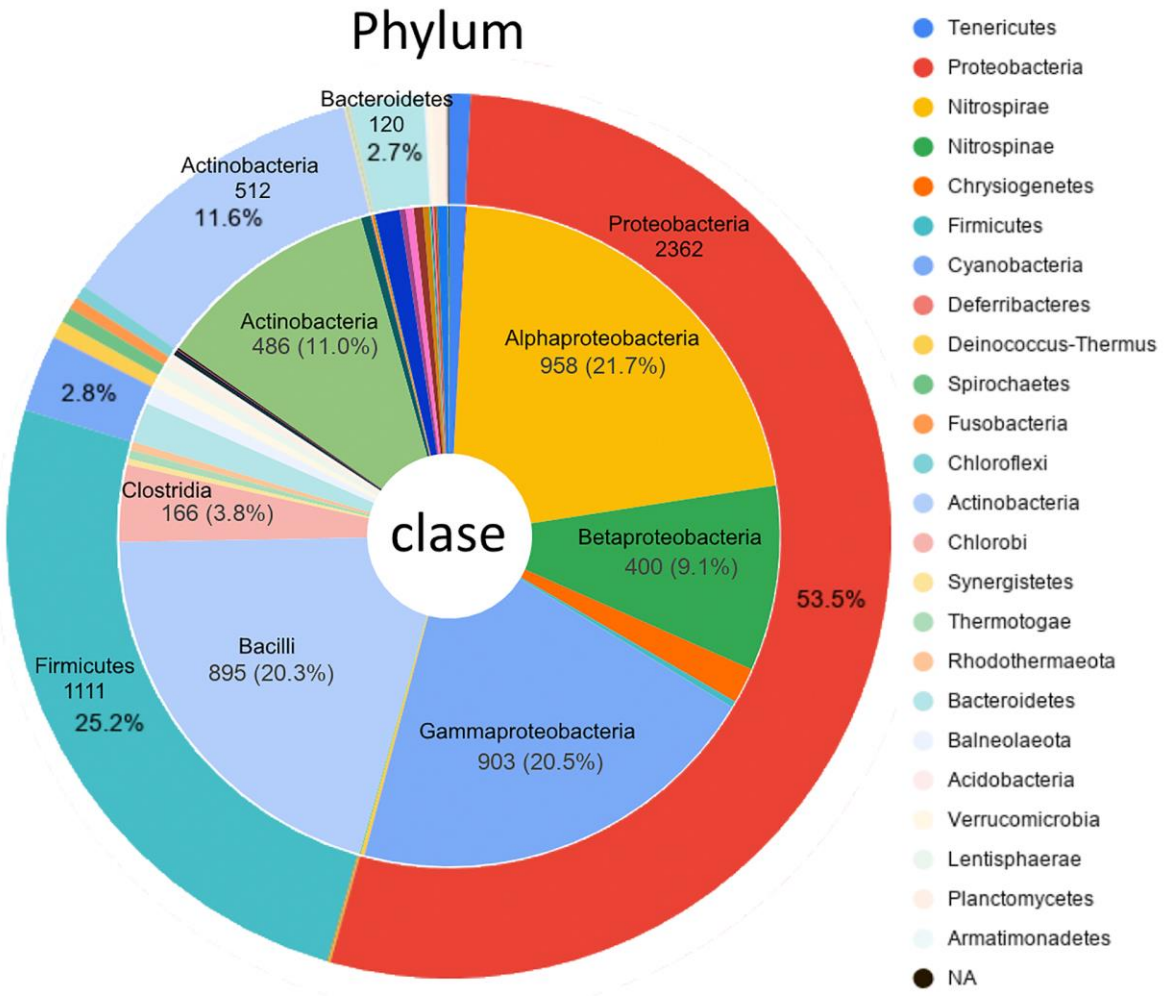

Figura 9. Distribución por phylum y clase de proteínas CPn. En el círculo externo se encuentran los phyla bacterianos y en el círculo interno las clases, con la cantidad de proteomas totales para cada uno y el porcentaje que representan del total de bacterias que poseen al menos una proteína CPn. NA = Sin clasificación taxonómica disponible.

Para algunas de las siguientes especies no se ha probado o reportado algún tipo de actividad de este tipo de metabolismo de Pn. Se puede apreciar que los phyla dominante de las especies que tienen al menos una proteína del metabolismo de CPn son el phylum Proteobacteria (53.5%), con las clases:

Alphaproteobacteria 21.7% con especies tales como: *Phyllobacterium sophorae, Mesorhizobium tianshanense, Rhizobium miluonense,* bacterium M00.F.Ca.ET.156.01.1.1, *Mesorhizobium delmotii*;

Gammaproteobacteria 20.5%: *Pectobacterium carotovorum, Pectobacterium polaris, Pseudomonas abietaniphila, Pseudomonas silesiensis, Enterobacter soli*); y

Betaproteobacteria 9.1%: *Paraburkholderia aspalathi, Paraburkholderia dipogonis, Burkholderia agricolaris, Paraburkholderia diazotrophica, Paraburkholderia hospita*).

Seguidos en abundancia por el phylum Firmicutes (25.2%) con las clases:

Bacilli 20.3%, siendo las especies más abundantes del phymun: *Oceanobacillus damuensis, Paenibacillus ferrarius, Paenibacillus terrigena, Paenibacillus guangzhouensis, Paenibacillus harenae*; y

Clostridia 3.8%: *Clostridium acidisoli, Clostridiisalibacter paucivorans, Alkaliphilus metalliredigens, Caldanaerovirga acetigignens, Oxobacter pfennigii*.

Finalmente los phyla Actinobacteria 11.6%: *Gordonibacter urolithinfaciens, Curtobacterium plantarum, Microbacterium sediminis, Arthrobacter alpinus, Arthrobacter glacialis*; y

Bacteroidetes 2.7%: *Rikenella microfusus, Flavihumibacter solisilvae, Ilyomonas limi, Segetibacter koreensis* y *Chitinophaga rhizosphaerae*.

6.2. Abundancia relativa de proteínas CPn

Una vez que se buscaron los perfiles de modelos ocultos de Márkov (HMM) en la base de datos local de los proteomas bacterianos, se encontraron en total 24,023 secuencias de proteínas homólogas de todas las proteínas del catabolismo de fosfonato (CPn). Se calculó la cantidad de proteínas homólogas totales y el porcentaje con respecto al total de copias de genes de proteínas del proteoma para cada phylum, en la **Figura 10** y **Tabla 1** se muestran las 3 especies con mayor porcentaje en su proteoma, con respecto a los demás miembros de su mismo phylum.

Tabla 1. Se muestran las 3 especies de cada phylum con mayor abundancia relativa en número de copias de secuencias de proteínas homólogas de catabolismo de fosfonato.

Especie	Phylum	Clase	Orden	Familia	Género	No. Proteínas	Abundancia total	Abundancia relativa en el proteoma
Phyllobacterium sophorae	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Phyllobacterium*	5804	44	0.76%
Rhizobium miluonense	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Rhizobium*	6129	35	0.57%
Mesorhizobium tianshanense	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Mesorhizobium*	7124	35	0.49%
Rubidibacter lacunae	Cyanobacteria	Oscillatoriophycideae	Chroococcales	Aphanothecaceae	*Rubidibacter*	3287	15	0.46%
Nodularia spumigena	Cyanobacteria	-Cyanobacteria	Nostocales	Aphanizomenonaceae	*Nodularia*	4443	19	0.43%
Microbacterium sediminis	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	*Microbacterium*	2712	11	0.41%
Phormidium tenue	Cyanobacteria	Oscillatoriophycideae	Oscillatoriales	Oscillatoriaceae	*Phormidium*	4992	16	0.32%
Oceanobacillus damuensis	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Oceanobacillus*	3790	12	0.32%
Entomoplasma somnilux	Tenericutes	Mollicutes	Entomoplasmatales	Entomoplasmataceae	*Entomoplasma*	713	2	0.28%
Curtobacterium plantarum	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	*Curtobacterium*	4487	12	0.27%
Spiroplasma litorale	Tenericutes	Mollicutes	Entomoplasmatales	Spiroplasmataceae	*Spiroplasma*	1105	3	0.27%
Alkalispirochaeta americana	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Alkalispirochaeta*	2815	6	0.21%
Hydrogenibacillus schlegelii	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	*Thermodesulfovibrio*	2328	5	0.21%
Mycoplasma salivarium	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	*Mycoplasma*	1409	3	0.21%
Chloroflexus aurantiacus	Chloroflexi	Chloroflexia	Chloroflexales	Chloroflexaceae	*Chloroflexus*	3939	8	0.20%
Paenibacillus guangzhouensis	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	*Paenibacillus*	6014	11	0.18%
Paenibacillus ferrarius	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	*Paenibacillus*	7834	12	0.15%
Chloroflexus aggregans	Chloroflexi	Chloroflexia	Chloroflexales	Chloroflexaceae	*Chloroflexus*	3648	5	0.14%
Rikenella microfusus	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	*Rikenella*	2212	3	0.14%
Rubritalea squalenifaciens	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Rubritaleaceae	*Rubritalea*	3715	5	0.13%
Fusobacterium russii	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	*Fusobacterium*	1690	2	0.12%
Cetobacterium somerae	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	*Cetobacterium*	2826	3	0.11%
Nitrospina gracilis	Nitrospinae	Nitrospinia	Nitrospinales	Nitrospinaceae	*Nitrospina*	2782	3	0.11%
Alkalispirochaeta sphaeroplastigenens	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Alkalispirochaeta*	2759	3	0.11%
Chloroflexus islandicus	Chloroflexi	Chloroflexia	Chloroflexales	Chloroflexaceae	*Chloroflexus*	3861	4	0.10%

Deinococcus koreensis	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	3978	4	0.10%
Haloplasma contractile	NA	NA	Haloplasmatales	Haloplasmataceae	*Haloplasma*	2935	3	0.10%
Alkalispirochaeta odontotermitis	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Alkalispirochaeta*	3124	3	0.10%
Chrysiogenes arsenatis	Chrysiogenetes	Chrysiogenetes	Chrysiogenales	Chrysiogenaceae	*Chrysiogenes*	2495	2	0.08%
Desulfurispirillum indicum	Chrysiogenetes	Chrysiogenetes	Chrysiogenales	Chrysiogenaceae	*Desulfurispirillum*	2538	2	0.08%
Akkermansia glycaniphila	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Akkermansiaceae	*Akkermansia*	2480	2	0.08%
Gordonibacter urolithinfaciens	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	*Gordonibacter*	21162	14	0.07%
Geovibrio thiophilus	Deferribacteres	Deferribacteres	Deferribacterales	Deferribacteraceae	*Geovibrio*	2693	2	0.07%
Isachenkonia alkalipeptolytica	NA	NA	NA	NA	*Isachenkonia*	2870	2	0.07%
Chitinophaga barathri	Bacteroidetes	Chitinophagia	Chitinophagales	Chitinophagaceae	*Chitinophaga*	5410	3	0.06%
Fusobacterium ulcerans	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	*Fusobacterium*	3102	2	0.06%
Isosphaera pallida	Planctomycetes	Planctomycetia	Planctomycetales	Isosphaeraceae	*Isosphaera*	3575	2	0.06%
Cloacibacillus porcorum	Synergistetes	Synergistia	Synergistales	Synergistaceae	*Cloacibacillus*	3078	2	0.06%
Lacunisphaera limnophila	Verrucomicrobia	Opitutae	Opitutales	Opitutaceae	*Lacunisphaera*	3491	2	0.06%
Edaphobacter dinghuensis	Acidobacteria	Acidobacteriia	Acidobacteriales	Acidobacteriaceae	*Edaphobacter*	3655	2	0.05%
Chitinophaga niabensis	Bacteroidetes	Chitinophagia	Chitinophagales	Chitinophagaceae	*Chitinophaga*	5730	3	0.05%
Deinococcus alpinitundrae	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	4389	2	0.05%
Deinococcus peraridilitoris	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	4203	2	0.05%
Rubrivirga marina	Rhodothermaeota	Rhodothermia	Rhodothermales	Rubricoccaceae	*Rubrivirga*	4115	2	0.05%
Prosthecochloris vibrioformis	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Prosthecochloris*	2056	1	0.05%
Thermodesulfovibrio thiophilus	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	*Thermodesulfovibrio*	1874	1	0.05%
Petrotoga sibirica	Thermotogae	Thermotogae	Petrotogales	Petrotogaceae	*Petrotoga*	1834	1	0.05%
Pseudothermotoga thermarum	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Pseudothermotoga*	1951	1	0.05%
Aquisphaera giovannonii	Planctomycetes	Planctomycetia	Planctomycetales	Isosphaeraceae	*Aquisphaera*	7414	3	0.04%
Rhodohalobacter halophilus	Balneolaeota	Balneolia	Balneolales	Balneolaceae	*Rhodohalobacter*	2679	1	0.04%
Oceanotoga teriensis	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Oceanotoga*	2604	1	0.04%
Gimesia maris	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	*Gimesia*	5745	2	0.03%
Gracilimonas tropica	Balneolaeota	Balneolia	Balneolales	Balneolaceae	*Gracilimonas*	3271	1	0.03%
Victivallis vadensis	Lentisphaerae	Lentisphaeria	Victivallales	Victivallaceae	*Victivallis*	3957	1	0.03%
Natronospirillum operossus	NA	NA	NA	NA	*Natronospirillum*	3946	1	0.03%

Rubricoccus marinus	Rhodothermaeota	Rhodothermia	Rhodothermales	Rubricoccaceae	*Rubricoccus*	3648	1	0.03%
Cloacibacillus evryensis	Synergistetes	Synergistia	Synergistales	Synergistaceae	*Cloacibacillus*	3018	1	0.03%
Terriglobus saanensis	Acidobacteria	Acidobacteriia	Acidobacteriales	Acidobacteriaceae	*Terriglobus*	4102	1	0.02%
Granulicella mallensis	Acidobacteria	Acidobacteriia	Acidobacteriales	Acidobacteriaceae	*Granulicella*	4674	1	0.02%
Capsulimonas corticalis	Armatimonadetes	Armatimonadia	Capsulimonadales	Capsulimonadaceae	*Capsulimonas*	6401	1	0.02%

De forma congruente con lo mostrado en la **Figura 9** anterior, el grupo que contiene más proteínas CPn son las Proteobacteria y además son las que poseen mayor número de copias (abundancia total) y mayor abundancia relativa en su proteoma, aunque el segundo grupo en abundancia relativa es el de Cyanobacteria. El resto de las especies pertenecientes a los demás phyla no se encuentran agrupadas y poseen menos copias de proteínas CPn.

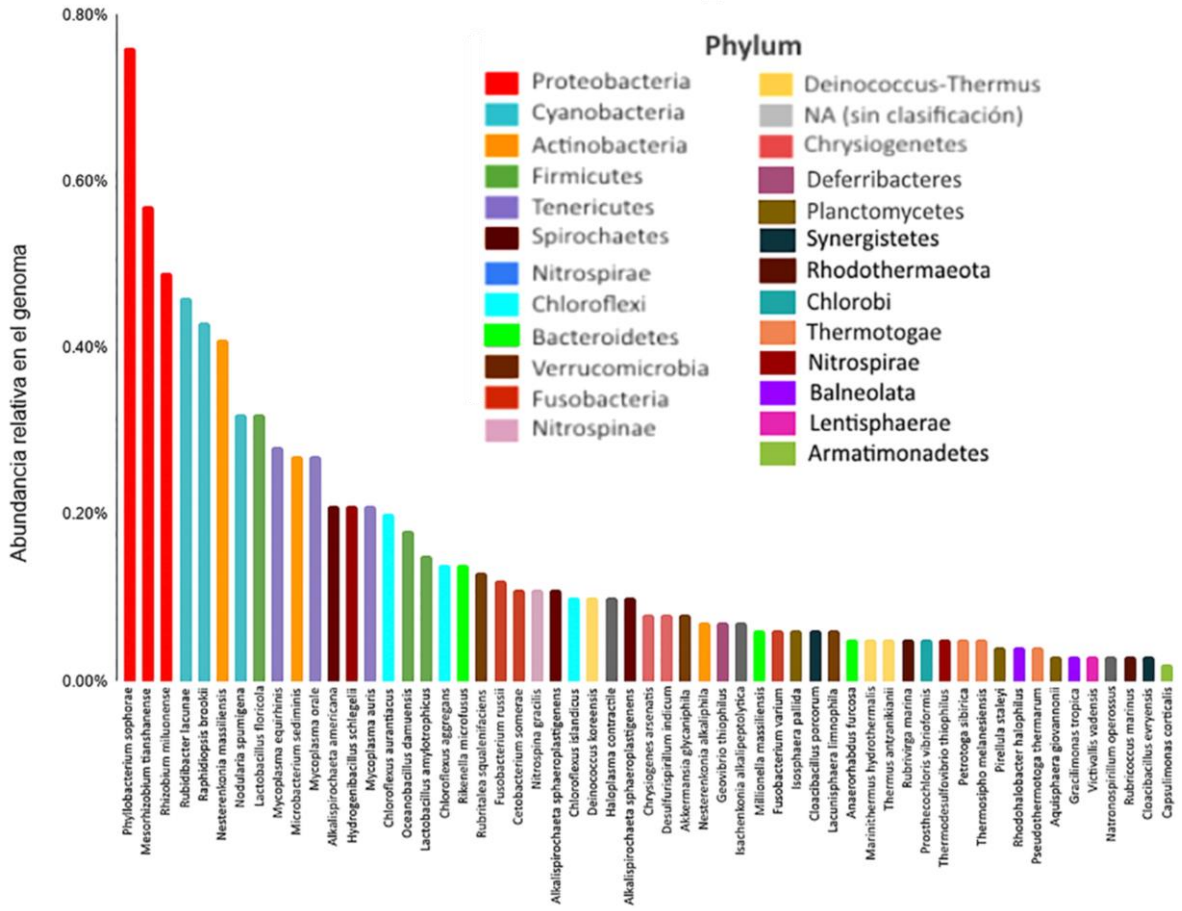

Figura 10. Abundancia relativa de proteínas CPn por phylum, calculada como la abundancia de proteínas por phylum entre proteínas totales por phylum.

6.3. Distribución las proteínas CPn por mecanismo enzimático

Podemos dividir las proteínas encontradas en la base de datos local en los tres mecanismos metabólicos descritos anteriormente. En la **Figura 11** se aprecia que las proteínas pertenecientes al mecanismo de corte de enlace radical de C-P por C-P liasa (I) son las predominantes con 86.65%, seguidas por 12.55% del mecanismo de corte hidrolítico del enlace C-P (II) y por último las de corte de enlace oxidativo C-P por PhnYox / PhnZ (III) con 0.79% de las secuencias.

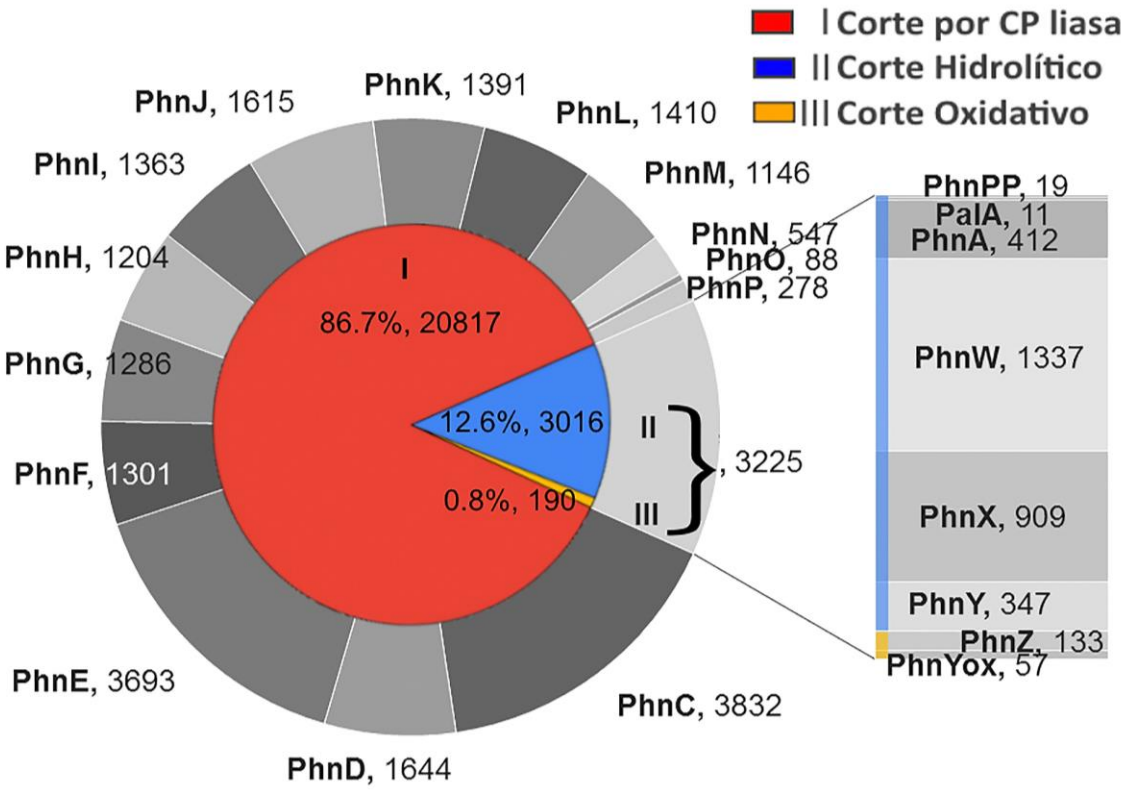

Figura 11. Distribución de las proteínas CPn en Bacteria por proporción. Se grafica el número de copias de proteínas homólogas CPn totales encontradas por medio de los HMMs en la Base de datos local . En el círculo exterior se muestra el total de proteínas homólogas encontradas para cada proteína de la literatura, en el círculo interior se muestra la proporción de los que pertenecen a ciertos mecanismos enzimáticos, es decir los mecanismos I (corte radical por medio de la C-P liasa), II (corte hidrolítico) y III (corte oxidativo).

6.4. Mecanismos enzimáticos CPn y su predominancia en Bacteria

6.4.1. Conteo de conjuntos de proteínas en Bacteria

Las 24,023 secuencias de proteínas CPn homólogas están distribuídas en 4,413 proteomas completos. Se calcula que las 4,413 especies con al menos una proteína CPn se encuentran con la siguiente composición: 3,810 especies (86.34%) tienen al menos una proteína perteneciente al mecanismo I (Corte de enlace radical de C-P por carbono-fósforo liasa), 1,507 (34.15%) al mecanismo II (Cortes hidrolíticos del enlace C-P) y 188 (4.26%) al mecanismo III (Corte de enlace oxidativo C-P por PhnYox / PhnZ).

La suma de porcentajes no deben ser interpretados como el 100% de las especies de bacterias, pues como se verá más adelante, hay especies que poseen proteínas pertenecientes a más de un mecanismo enzimático CPn (a diferencia de la **Figura 11** donde se muestra la presencia del total de proteínas en toda la base de datos local).

Algunos proteomas únicamente poseen una copia de un gen de una proteína y otros proteomas poseen hasta 44 copias de la misma proteína (**Anexo S0**). Por ello se realizaron diagramas de Venn contabilizando únicamente la presencia de las proteínas, y se calcularon las intersecciones por grupos de proteínas según lo reportado en la literatura (en conjuntos biológicamente significativos según la literatura, conjuntos como el de las proteínas PhnCDEFGHIJKLMOP que son componentes del operón phn del mecanismo I).

En la **Figura 12** se muestra el número de proteomas que contienen las proteínas más relevantes para el mecanismo enzimático I del corte de enlace radical de C-P liasa.

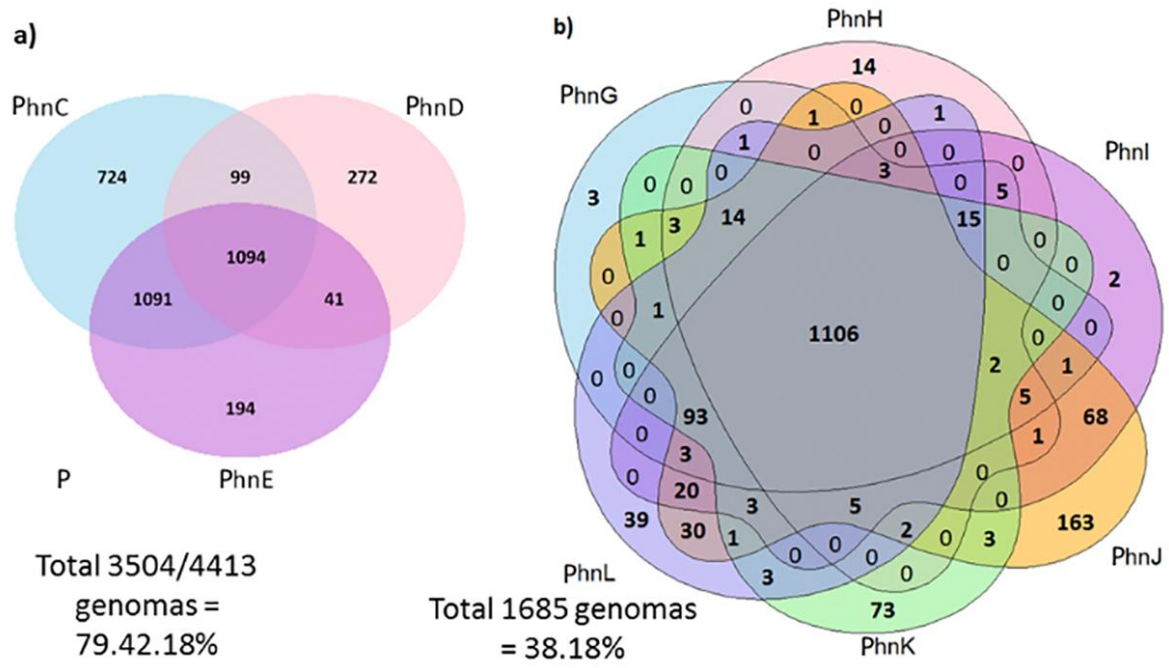

Figura 12. Conjunto de especies que contienen proteínas del mecanismo I de corte radical por medio de C-P liasa. Se muestra la cantidad de proteomas que contienen a las proteínas del transportador ABC de fosfonatos PhnCDE (**a**) y las proteínas componentes de la enzima C-P liasa PhnGHIJKL(**b)**.

En este mecanismo enzimático I hay al menos 1,094 especies de bacteria con las proteínas componentes del transportador ABC de fosfatos/fosfonatos (25% del total de bacterias con al menos una proteína CPn y 10.1% de todas las 10,906 bacterias analizadas de la base de datos local).

Observando los datos del **Anexo S3** podemos determinar lo siguiente:

Hay al menos 1,106 especies (25%) con las 6 proteínas que conforman el complejo $PhnG_2H_2I_2J_2K$ de la C-P liasa que parecen ser capaces de llevar a cabo las reacciones más importantes de la ruta metabólica del mecanismo I, dadas las condiciones ambientales de inanición de fósforo. Es decir, al menos estas bacterias tienen la mayoría de las proteínas del *operón phn* funcional y en las condiciones de inanición de fósforo llevarían a cabo las funciones más importantes para catabolizar el fosfonato por este mecanismo enzimático.

Otro conjunto de proteínas, conformado por PhnNOP, también participan en esa ruta metabólica (**Figura 1**), estas proteínas se encuentran menos distribuídas, únicamente 33 especies tienen estas 3 secuencias de proteínas en el mismo proteoma.

Los otros dos mecanismos enzimáticos II y III están menos distribuidos, ya que solo 1,506 proteomas (el 34%) cuentan con alguna proteína con el mecanismo de corte hidrolítico del enlace C-P y únicamente 227 (5.1%) que cuentan con las 3 proteínas de la ruta PhnWAY que realiza la ruta de fosfonoacetato hidrolasa (mecanismo II de corte hidrolítico).

La mayor cantidad de proteomas con un conjunto biológicamente significativo de proteínas corresponden al conjunto PhnCDE (1,094 proteomas con las proteínas del sistema de transporte ABC) y PhnGHIJKLM (862 proteomas), que son proteínas que componen a la C-P liasa y otras enzimas de esa misma ruta metabólica.

En segundo lugar se encuentra el conjunto de proteínas PhnCX (854 proteomas 19.3%, con un solo componente del sistema ABC y las enzima fosfonacetaldehído hidrolasa respectivamente) y PhnXW (652 proteomas 14.7%, con la fosfonacetaldehído hidrolasa y la 2-aminometil fosfonato transaminasa), componentes del metabolismo II de corte hidrolítico.

6.4.2. La distribución filogenética de las proteínas del metabolismo de fosfonatos es dispersa

Debido a las limitaciones de visualización por medio de diagramas de Venn, para visualizar los posibles conjuntos de intersecciones de proteínas se muestran en forma de diagrama de UpSetR en la **Figura 13**.

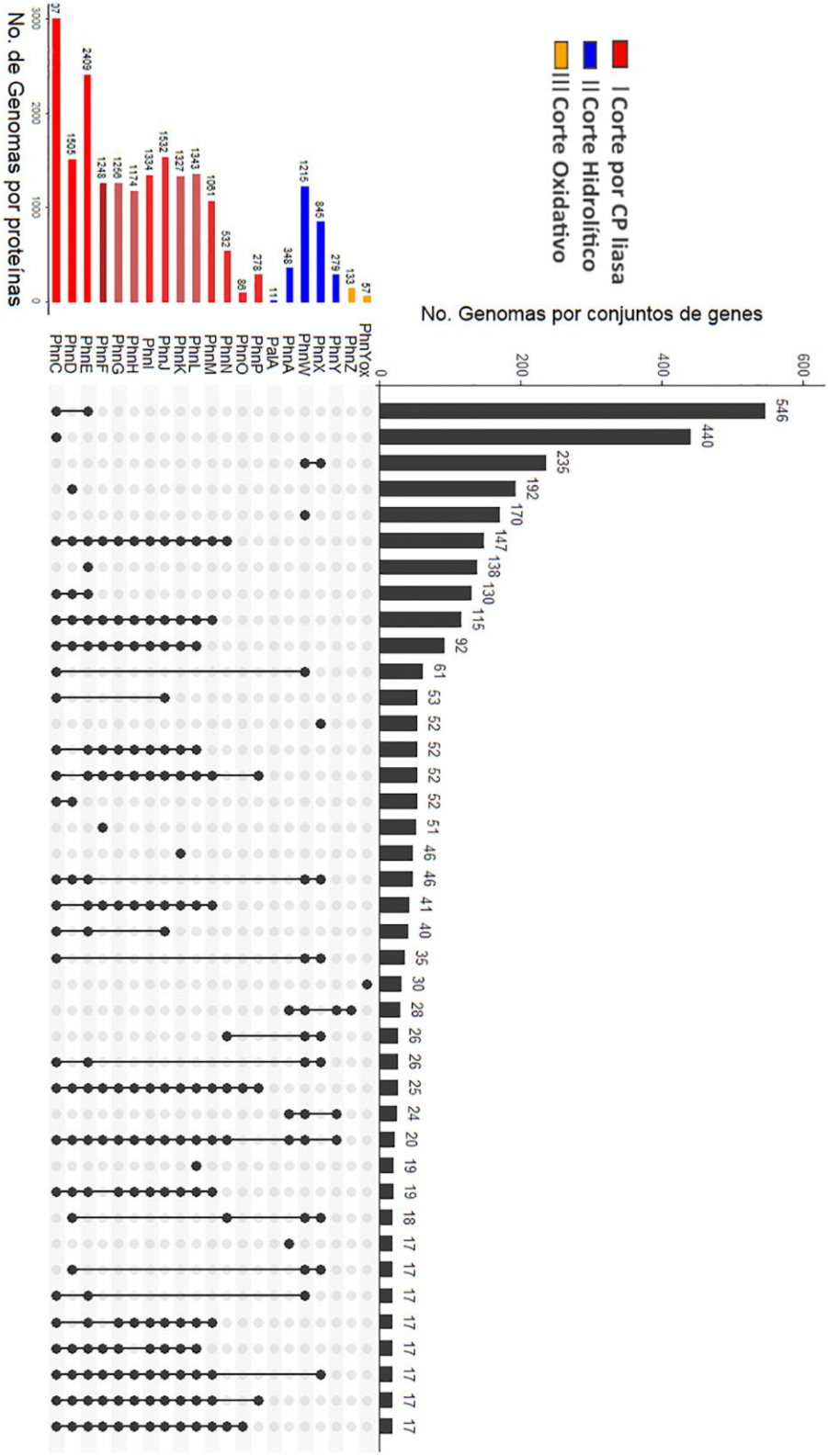

Figura 13. Diagrama UpSetR de proteínas que muestra los conjuntos intersección de los diferentes genes de proteínas CPn que se encuentran en los proteomas bacterianos. Se muestran en la parte inferior izquierda los tamaños de las intersecciones

de conjuntos de proteínas CPn en los 3 mecanismos enzimáticos, en la parte superior se muestra el número de proteomas donde se encuentran los conjuntos.

A partir de la **Figura 13,** el primer aspecto notable es que no hay organismos que tengan todas las proteínas analizadas en el presente estudio.

Los conjuntos más abundantes son los pertenecientes al transportador ABC de fosfatos/fosfonatos (PhnCDE) del mecanismo I, aunque de forma incompleta ya que se requieren de la presencia de las 3 proteínas para construir el transportador.

A continuación están las proteínas del mecanismo II, PhnXW (fosfonacetaldehído hidrolasa y 2-aminometilfosfonato transaminasa) descritas en las especies *Bacillus cereus*, *Salmonella typhimurium* LT2129 y *Enterobacter aerogenes*, 235 proteomas, de las cuales:

112 son Gammaproteobacteria: 31 Alteromonadales, 12 Pseudomonadales, 8 Aeromonadales, 33 Vibrionales, 8 Enterobacterales, 14 Oceanospirillales;

37 Firmicutes: 30 Bacillales familias Planococcaceae, Bacillaceae, Paenibacillaceae y 6 Clostridiales géneros *Clostridium*, *Anaerofustis*, *Lawsonibacter* y *Negativibacillus*;

27 Bacteroidetes: Bacteroidales familias Bacteroidaceae, Rikenellaceae, Tannerellaceae y Prevotellacea;

18 Actinobacteria: Bifidobacteriales género *Bifidobacterium*;

15 Betaproteobacteria: Burkholderiales, Rhodocyclales y Neisseriales;

10 Alphaproteobacteria: 10 Rhodospirillaceae géneros *Azospirillum, Komagataeibacter, Pararhodospirillum, Sinirhodobacter* y *Tropicimonas*;

6 Deltaproteobacteria: 5 Desulfobacterales familia Desulfobacteraceae y 1 Syntrophobacterales género *Syntrophobacter*;

3 Fusobacteria: Fusobacteriales géneros *Fusobacterium* y *Sebaldella*;

3 Planctomycetes: Planctomycetales géneros *Gimesia, Isosphaera* y *Pirellula*;

1 Synergistetes familia Synergistaceae género *Cloacibacillus*;

1 Verrucomicrobia Akkermansiaceae género *Akkermansia*;

1 Chloroflexi: Caldilineae género *Caldilinea*; 1 Spirochaetes familia Spirochaetaceae género Sediminispirochaeta.

Otro conjunto muy abundante es el de PhnW*,* con 170 especies, de las cuales:

84 son Firmicutes: 43 Clostridia 16 Clostridiaceae, 13 Lachnospiraceae, 5 Eubacteriaceae, 3 Ruminococcaceae, 1 Christensenellaceae, 1 Peptostreptococcaceae, 6 Negativicutes, 3 Erysipelotrichia, 32 Bacilli, 23 Bacillales 9 Lactobacillales;

30 Bacteroidetes: 21 Bacteroidia, 7 Cytophagia, 1 Flavobacteriia;

17 Actinobacteria: 11 Bifidobacteriaceae 3 Streptomycetaceae, 2 Actinomycetaceae, 1 Actinopolysporaceae;

13 Gammaproteobacteria: 5 Xanthomonadales, 2 Alteromonadales, 1 Chromatiales, 1 Nevskiales, 1 Chromatiales, 1 Oceanospirillales, 1 Thiotrichales, 1 Legionellales, 1 Aeromonadales;

12 Spirochaetes: 10 Spirochaetaceae géneros *Treponema*, *Sphaerochaeta* y *Sediminispirochaeta* y 2 Brachyspiraceae *Brachyspira*;

4 Epsilonproteobacteria: 3 Helicobacteraceae, 1 Campylobacteraceae;

3 Betaproteobacteria: Burkholderiaceae, Oxalobacteraceae y Neisseriaceae;

2 Alphaproteobacteria: Rhizobiales y Rhodospirillales;

1 Deltaproteobacteria: Syntrophaceae género Desulfobacca;

1 Verrucomicrobia: *incertae sedis*;

1 Balneolaeota Balneolaceae género *Gracilimonas*;

1 Acidobacteria: Acidobacteriaceae género *Granulicella*.

Con un total de 147 especies se encuentra el conjunto PhnCDEFGHIJKLMN, de las cuales:

146 son Alphaproteobacteria: 131 Rhizobiales familias 62 Rhizobiaceae, 25 Phyllobacteriaceae, 18 Bradyrhizobiaceae, 10 Brucellaceae, 9 Rhodobacteraceae, 7 Hyphomicrobiaceae, 2 Cohaesibacteraceae, 1 Aurantimonadaceae, 1 Rhodobiaceae, 1 Xanthobacteraceae;

1 Deltaproteobacteria: Desulfomicrobiaceae (*Desulfomicrobium*).

Muy parecido a esto se encuentra el conjunto PhnCDEFGHIJKLM, el cual es muy parecido en su composición taxonómica al conjunto de proteínas PhnCDEFGHIJKLMN, pero incluye cianobacterias. Está distribuído en 115 proteomas de los cuales:

82 son Alphaproteobacteria: 47 Rhizobiales, 23 Rhodobacterales, 11 Rhodospirillales, 1 Minwuiales;

21 Betaproteobacteria: 19 Burkholderiales, 2 Neisseriales;

6 Cyanobacteria: 3 Oscillatoriales, 2 Nostocales, 1 Synechococcales;

4 Deltaproteobacteria: Desulfomicrobiaceae géneros *Desulfovibrio, Halodesulfovibrio, Pseudodesulfovibrio*;

2 Gammaproteobacteria: géneros *Marinobacter* y *Erwinia*.

Existen 25 especies con el *operón phn* completo (mecanismo I). Éstas bacterias son Gammaproteobacteria de la familia Enterobactericeae (14 *Enterobacter,* 6 *Citrobacter* , 3 *Lelliottia,* 1 *Leclercia,* 1 *Escherichia*).

También se encuentran combinados los componentes de los mecanismos I y II, como el caso de PhnWAYZ en 28 especies (de las cuales 27 son Betaproteobacteria, 25 *Burkholderia,* 2 *Caballeronia,* 1 *Trinickia*; 1 Gammaproteobacteria: género *Pseudomonas*).

En el presente trabajo sólo se encontraron homólogos de PalA en *Bordetella bronchiseptica*.

6.5. Presencia de secuencias homólogas de proteínas CPn en un contexto filogenético

En la **Figura 14** se observa un mapa de calor (heatmap) donde se presentan las especies de Bacteria acopladas con el árbol filogenético de Ciccarelli et al., (2006). En este árbol se incluyó al menos un representante de cada phylum según la clasificación del *Taxonomy Browser* de NCBI, y que tuvieran al menos una proteína homóloga CPn predicha por los modelos HMM con el objetivo de observar organismos con este metabolismo. Es importante mencionar que hay phyla enteros que no tienen una sola proteína CPn, el árbol ayuda a visualizar estos casos.

Los grupos en los que mayor se aprecia la presencia de metabolismo del Pn, especialmente el metabolismo I de corte radical por medio de C-P liasa (al menos una proteína perteneciente a este mecanismo) es en el grupo de Proteobacteria (2,362 especies 53.5%), congruente con la información mostrada en las **Anexo S3, Figuras 9**, **10** y **11**. Mientras que la baja presencia de las proteínas de los metabolismos de corte hidrolítico (mecanismo II, 1,507 especies 34.1%) y corte oxidativo (mecanismo III, solo 188 especies 4.2%) no ayudan a visualizar un patrón importante en otros grupos de bacterias.

En este trabajo fue poco común encontrar especies con proteínas CPn aisladas (1,133 25.6%), usualmente se encuentran con otras proteínas del mismo mecanismo enzimático. Es interesante observar (**Figura 14**) que muchas especies que tienen las proteínas del mecanismo I también cuentan con proteínas del mecanismo II, específicamente PhnW (2-aminoetilfosfonato - piruvato transaminasa) y PhnX (fosfonoacetaldehído hidrolasa), y en menor medida PhnA (fosfonoacetato hidrolasa) y PhnY (fosfonoacetaldehído deshidrogenasa).

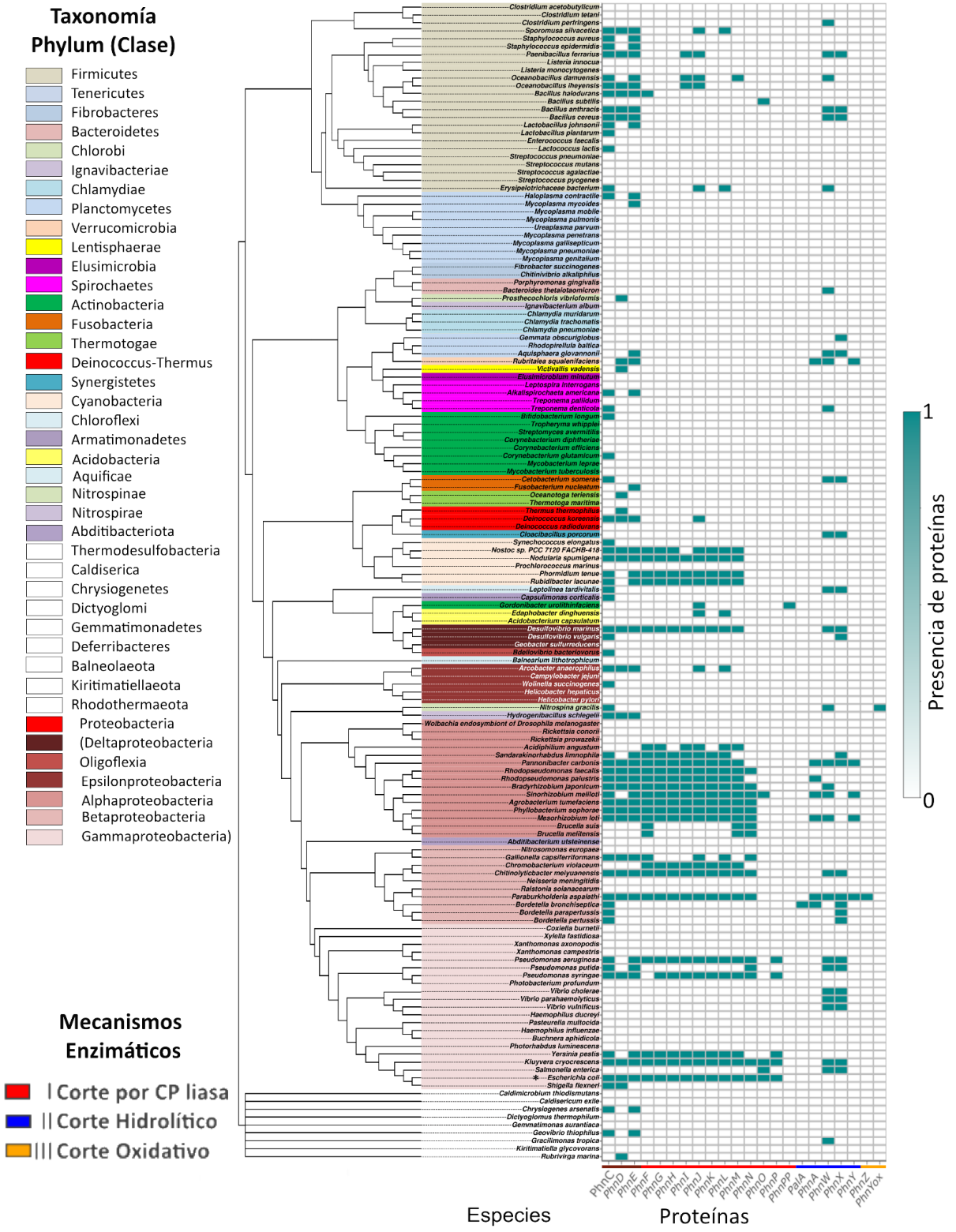

Figura 14. Mapa de calor acoplado a una filogenia basada en la de Ciccarelli et al. (2006) (se basa en la concatenación de 31 ortólogos que ocurren en 191 especies con genomas secuenciados). Este mapa muestra la presencia de las proteínas CPn en especies control, en el eje x se muestran las proteínas usadas en nuestra búsqueda, la

escala de color azul en el heatmap indica presencia de las proteínas CPn (como ausencia y presencia), se agregaron los phyla a los que pertenece cada especie para ilustrar relaciones taxonómicas entre esas comparaciones de especies de bacterias, al ser la más predominante el phylum Proteobacteria se incluyó la clase (para observar mejor resolución visitar https://github.com/LChora/LChora).

Si realizamos un enfoque más detallado en el grupo más predominante que es Proteobacteria (53.5%, **Figura 15**), observamos que aún hay bastantes diferencias entre la constitución de estos metabolismos de catabolismo de fosfonato.

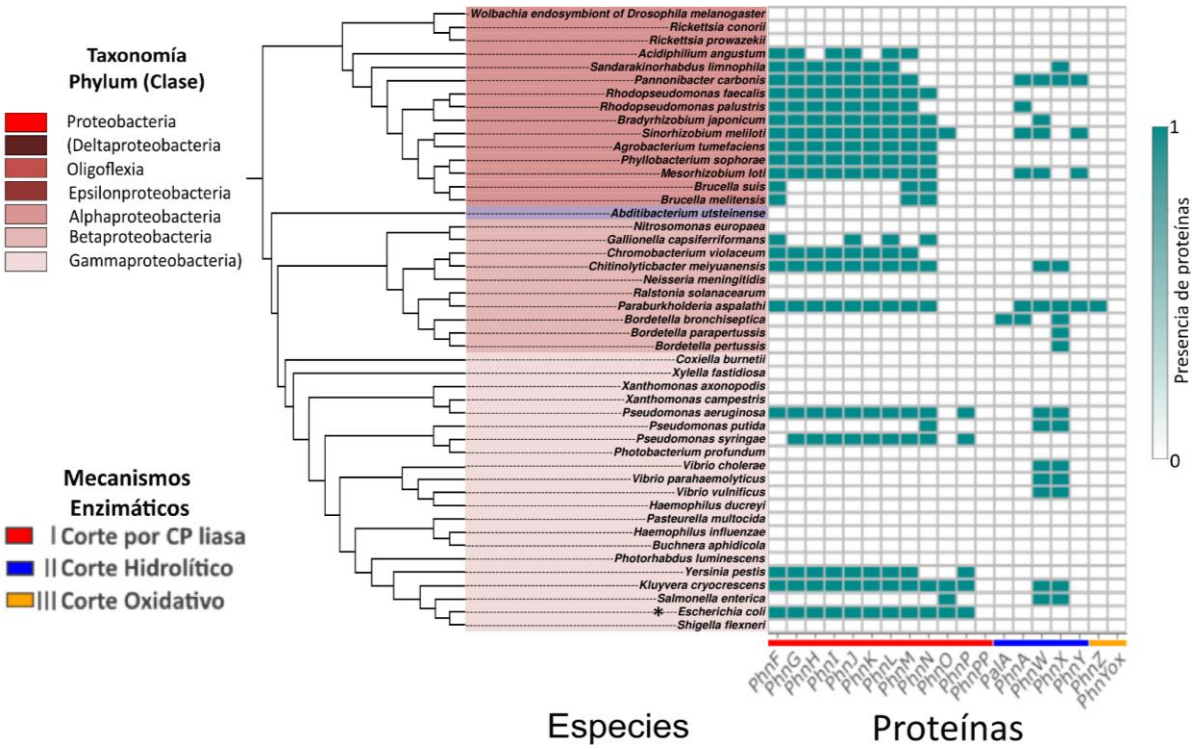

Figura 15. Mapa de calor acoplado a una filogenia que muestra el patrón de presencia de proteínas CPn en el grupo de Proteobacteria.

En el grupo de Proteobacteria (**Figura 15**), especialmente en Alphaproteobacteria hay 18/46 (39%) representantes que tienen la mayoría (8/11) de las proteínas del mecanismo I (sin contar el transportador ABC de las proteínas PhnCDE) y algunas proteínas del mecanismo II como PhnA (5/46) y PhnW (13/46) así como PhnY (14/46).

7. Discusión

Hay especies de bacterias que están reportadas en la literatura que poseen la presencia de proteínas del catabolismo de fosfonatos (CPn) tales como *Acidithiobacillus ferrooxidans, Agrobacterium tumefaciens, Bacillus cereus, Bacillus megaterium, Bacillus terrae, Bradyrhizobium japonicum, Burkholderia pseudomallei, Chloroflexus aurantiacus, Cupriavidus necator, Desulfotignum phosphitoxidans, Escherichia coli, Klebsiella aerogenes, Klebsiella oxytoca, Klebsiella pneumoniae, Kluyvera ascorbata, Kluyvera cryocrescens, Mesorhizobium loti, Mycolicibacterium smegmatis, Nostoc sp. PCC 7120, Ochrobactrum anthropi, Prochlorococcus marinus, Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas stutzeri, Rhodobacter capsulatus, Ruegeria pomeroyi, Salmonella enterica, Sinorhizobium meliloti* y *Trichodesmium erythraeum*, en este trabajo se trataron como “especies control” para la evaluación de modelos ocultos de Márkov (HMMs).

7.1. Evaluación de los Modelos Ocultos de Márkov usando especies control

Es importante recordar que las secuencias de proteínas utilizadas para estos modelos fueron descargados de las bases de datos de Uniprot y NCBI (**Anexo S0**), y los genomas anotados (proteomas completos) se descargaron de Refseq (con formato .faa, notación WP, **Anexo S2**).

En este trabajo se corrieron los HMMs en los genomas anotados de nuestra base de datos local con los genomas de las especies control (como el caso de *E. coli* donde la literatura reporta que está el *operón* completo *phnCDEFGHIJKLMOP,* Makino et al., 1991). Esto se hizo con la finalidad de observar si los modelos eran capaces de encontrarse “a sí mismos” en los proteomas de las especies control, como se observa en la **Figura 14**, **Anexo S4** y

https://github.com/LChora/LChora/blob/main/Presencia%20de%20prote%C3%ADnas%20CPn%20en%20controles.png.

Como se muestra en este trabajo (**Figura 14**) los modelos HMM fueron capaces de encontrarse a “sí mismos” en las especies control, pero no todas las especies control mostraron los mismos perfiles de presencia de proteínas CPn (con algunas proteínas ausentes) como se esperaba en la literatura (White y Metcalf, 2004; Kononova y Nesmeyanova, 2002), algunas de estas especies que no cumplieron con lo esperado fueron: *Pseudomonas stutzeri, Salmonella enterica, Salmonella bongori, Acidovorax citrulli, Legionella anisa, Bacillus megaterium, , Rubneribacter badeniensis, Eggerthella sinensis, Grodonibacter pamelaeae, Eggerthella lenta* y *Trinickia soli*.

Dada la incongruencia de la no-presencia de los genes del *operón phn* en éstas especies control se realizó el mismo ejercicio del mapa de calor para 25 ensambles de secuencia de genoma de *Pseudomonas stutzeri* WM88 (ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/Pseudomonas_stutzeri/) con el objetivo de conocer la robustez de trabajar con los genomas representativos de Refseq. No se encontró el mismo perfil de presencia de proteínas CPn en el genoma representativo de Refseq de *P. stutzeri*, por lo que en la **Figura 16** se muestra un mapa de calor con todos los genomas de *P. stutzeri* en GenBank, donde se logró observar diferencias entre los ensambles de secuencias de los genomas de esta bacteria.

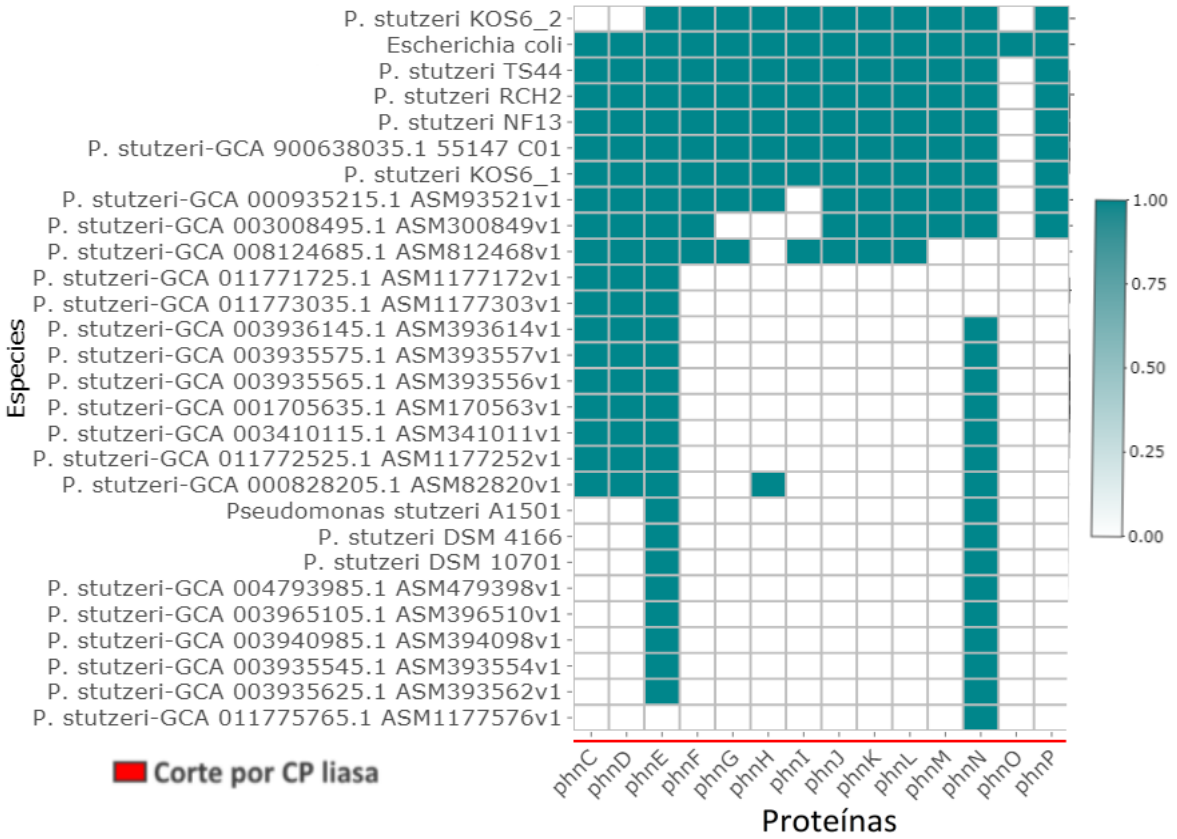

Figura 16. Mapa de calor que muestra la presencia de los genes CPn en los proteomas de varios ensambles de secuencia de genoma de *Pseudomonas stutzeri*, se observa diversidad entre los patrones de presencia del *operón phn* dentro de esta especie.

7.2. Consistencia de la proporcionalidad de la abundancia y distribución filogenética de las proteínas CPn

Hay inconsistencias en la proporcionalidad de las proteínas CPn (**Figura 11**). Por ejemplo, las proteínas PhnCDE: componentes del transportador de fosfonatos ABC (ATP-binding cassette o dominio de unión a ATP) no tienen el mismo número de copias totales. La proteína PhnC es la más abundante de todas las proteínas CPn (3,832) y supera por 100 copias a la PhnE (3,693) y por más del doble a la PhnD (1,644). Cabe aclarar que el número de copias totales

visualizadas en la **Figura 11** considera varias copias de genes de una misma proteína en un mismo genoma.

Esto último podría resultar en la conclusión de que la inconsistencia de proporcionalidad de las proteínas CPn es un artefacto del método que se utilizó con los modelos ocultos de Márkov.

Hay grupos filogenéticos congruentes con la presencia de las proteínas CPn (**Figura 14**) especialmente las proteínas pertenecientes al mecanismo I (de corte radical por C-P liasa) como algunos miembros de Cyanobacteria y Alphaproteobacteria, aunque dentro de Proteobacteria parecieran formarse grupos no naturales al haber convergencias dentro de algunas especies de Betaproteobacteria y Gammaproteobacteria. Desde el trabajo de Huang et al. (2005) se le atribuía la transferencia horizontal de genes a la distribución dispersa entre distintos organismos, especialmente en Proteobacteria, jugando así, un papel importante en la evolución de la degradación del fosfonato en bacterias.

Las relaciones filogenéticas de los grupos de Bacteria pueden ser descritas de diferentes formas y por lo tanto la topología de los grupos no es la misma que con el trabajo de Huang et al. (2005). Observando el árbol filogenético de la **Figura 15** se puede resaltar que los grupos de Proteobacteria están relacionados cercanamente y es plausible pensar en un modelo parsimonioso, en el que el ancestro común de Alpha-, Beta- y Gamma-proteobacteria tenía el *operón phn* y se fueron perdiendo genes o incluso el operón completo.

Los datos muestran que los mecanismos CPn I y II usualmente se encuentran conjugados con componentes del transportador ABC de fosfonatos (PhnCDE), el cual fue descrito inicialmente en el *operón phn* del mecanismo I. Esto podría indicar que este tipo de metabolismo cuenta con un componente específico para internalizar las moléculas de fosfonatos.

El *operón phn* se encuentra distribuido en diversas variantes del *operón* en varios grupos de Bacteria (**Figura 14**), es decir, grupos de genes componentes del *operón phn* de constitución variable (como ocurre en la **Figura 7**, de White y Metcalf, 2007). Según Gebhard et al. (2006), una posible interpretación es que los fragmentos del *operón phn* han sido cooptados por bacterias (que resulta en estructuras que no habían existido en los antepasados antes mediante el uso de sistemas genéticos existentes, como grupos de genes y sus rasgos subyacentes y programas de desarrollo, en otros lugares o en diferentes momentos) para funciones alternativas o desconocidas.

Por ejemplo, un locus que contiene los genes *phnCDEF* se encuentra en *Mycobacterium smegmatis,* aunque esta bacteria no puede usar Pn como fuente de Pi (ya que en este organismo es específico para transporte de Pi, Gebhard et al., 2006). En este caso, los genes *phnCDEF* parecen haber sido adoptados para codificar el transporte de Pi, en lugar de Pn, al interior de la célula cuando Pi es limitante (Horsman y Zechel, 2017). También *phnO* se encuentra en *Salmonella enterica* adyacente a los genes que codifican las vías de PhnX y PhnA para el catabolismo de ácido 2-aminoetil fosfónico y fosfonoacetato, respectivamente.

Según Horsman y Zechel (2017) *phnO* codifica una N-acetil transferasa, una química que no otorga un propósito obvio a PhnX y vías PhnA, es decir son reacciones enzimáticas que no están relacionadas. También se plantea la implementación de nuevas rutas metabólicas o hasta redundantes, por ejemplo, *Enterobacter aerogenes* IFO 12010 tiene operones que codifican vías de C-P liasa y PhnX para la conversión de ácido 2-aminoetil fosfónico a Pi. En este trabajo se encuentran situaciones similares cuando se observan los datos (**Figura 14**).

Hay poca representatividad en el árbol filogenético (**Figura 14**) de los mecanismos CPn II (35/147, 23.8%) y III (2/147, 1.4%) y en la base de datos (12.6%, **Anexo S3**). El otro gran conjunto de proteínas representadas es el del mecanismo de corte hidrolítico, representado por las proteínas PhnWX que está ampliamente distribuido a lo largo de todo el árbol (18/147).

En el trabajo de Quinn et al. (2007) encontraron homólogos putativos de PhnA presentes en representantes de los subgrupos Alpha-, Beta- (cepas de *Burkholderia*), Gamma- y Delta-proteobacterias, y en un representante de las Acidobacterias. De manera similar, en el presente trabajo se encontraron 348 especies con homólogos de PhnA, 294 en Proteobacteria, distribuídos en las clases 135 Apha-, 144 Beta-, 2 Delta-, 13 Gamma-, 2 Acidithiobacillia,); 24 en Actinobacterias (clases Actinobacteria); además de 27 en Bacteroidetes (clases 10 Chitinophagia, 10 Flavobacteriia y 7 Sphingobacteriia), y 3 en Verrucomicrobia.

Además el grupo de Quinn et al. (2007) encontraron un homólogo putativo de la proteína estructural de fosfonopiruvato hidrolasa PalA presentes en 10 proteomas, todos de miembros de Alpha-, Beta- o Gamma-proteobacteria; en siete casos (*Bordetella bronchiseptica* RB50, *Pseudomonas fluorescens* Pf-5, *Acidovorax avenae* subsp. *citrulli* AAC00-1, *Pseudomonas entomophila* str. L48, *Mesorhizobium loti* MAFF303099. En el presente trabajo se encontró presente en 11 especies, sólo de Betaproteobacteria, siendo incapaces de encontrar en los mismos grupos que Quinn et al. (2007), siendo éste un resultado menos alentador.

La distribución de estas proteínas CPn tiende a ser menor en ciertos grupos (como en Tenericutes (35/170), Spirochaetes (25/141), Deinococcus-Thermus (27/72), Chloroflexi (22/40), Planctomycetes (20/40), Fusobacteria (20/38), Thermotogae (4/38), Verrucomicrobia (8/32), Acidobacteria (3/27), Aquificae (0/20), Synergistetes (2/18), Chlorobi (1/14), Thermodesulfobacteria (0/11), Nitrospirae (2/10), Rhodothermaeota (2/3), Elusimicrobia (0/3), Gemmatimonadetes (0/3), Kiritimatiellaeota (0/3), Armatimonadetes (1/2), Chrysiogenetes (2/2), Coprothermobacterota (0/2), Ignavibacteriae (0/2), Nitrospinae (1/1), Abditibacteriota (0/1), Caldiserica (0/1), Calditrichaeota (0/1), Dictyoglomi (0/1), ver **Anexo S3**), por lo que su adaptabilidad en diferentes especies y ambientes podría facilitar su transferencia horizontal de una bacteria a otra.

7.3. Taxa notables en la distribución filogenética de los CPn

En la **Figura 13** se observa la predominancia de ciertos conjuntos de proteínas (detallados más adelante), que podrían contar una historia diferente a la escrita en la literatura. Se puede especular que la composición específica de dichos conjuntos de proteínas CPn nos indicará la funcionalidad metabólica de éstas proteínas en las especies que pertenecen. Gracias a esto podemos llegar a discutir los siguientes puntos observados:

PhnCDEFGHIJKLM y PhnCDEFGHIJKLMN: Es interesante recalcar que hay conjuntos proteicos donde la diferencia de una sola proteína marca grupos taxonómicos enteros. Este es el caso del conjunto PhnCDEFGHIJKLM y PhnCDEFGHIJKLMN incluye Rhizobiales (Alphaproteobacteria) de 10 familias diferentes y algunas Deltaproteobacteria, pero el que carece de la PhnN incluye a Cianobacterias y Gammaproteobacterias. Además de que el *operón phn* completo solo está en miembros de la familia Enterobacteriaceae.

PhnJ-PhnPP: En la literatura se encuentra reportado que la proteína PhnPP (fosforribosil 1,2-fosfato cíclico 1,2-difosfodiesterasa) se puede usar como sustituto de la PhnP (fosforribosil 1,2-fosfato cíclico fosfodiesterasa). En el trabajo de Ghodge et al. (2013) mencionan que es interesante encontrar esta proteína en *Clostridium difficile,* un patógeno multiresistente altamente virulento, que tiene ribosa-5-fosfato y Pi como productos terminales de la vía C-P liasa en lugar de la enzima 5-fosforibosil-1-pirofosfato (PRPP) sintasa PhnP. En este trabajo, la presencia de la PhnPP (en solo 13 especies) se encontró estrictamente en presencia de la PhnJ (Alfa-D-ribosa 1-metilfosfonato 5-fosfato C-P liasa), condiciones que se cumplieron en especies pertenecientes al phylum Actinobacteria familia Eggerthellaceae (que incluye a la especie *Eggerthella lenta* que es donde se describió esta proteína, y a *C. difficile*, pero en este trabajo no se trabajó con esta).

En la **Tabla 1** se muestran tres especies de cada phylum que destinan mayor porcentaje de su genoma a proteínas del catabolismo de fosfonato.

En el **Anexo S3** las 22 especies con mayor abundancia relativa son Alphaproteobacteria y Gammaproteobacteria entre Rhizobiales, Rhodospirillales y Rhodobacterales con porcentajes que van desde 0.76% al 0.46% del proteoma, con constituciones del *operón phn* de las proteínas PhnCDEFGHIJKLM y PhnCDEFGHIJKLMN con algunos miembros con proteínas como la PhnA, PhnW y PhnX (estas especies son *Phyllobacterium sophorae, Rhizobium miluonense, Mesorhizobium tianshanense, Elioraea thermophila, Desulfovibrio marinus, Microvirga tunisiensis, Haematobacter massiliensis, Pseudochrobactrum saccharolyticum, Pannonibacter carbonis, Maritalea mobilis, Rhodovulum bhavnagarense, Devosia limi, Aliiroseovarius halocynthiae, Actibacterium mucosum, Halomonas jeotgali, Rhodovulum robiginosum, Roseibaca calidilacus, Bosea thiooxidans, Marivita cryptomonadis, Fodinicurvata fenggangensis* y *Pseudorhodobacter ferrugineus*).

El segundo grupo en mayor porcentaje del proteoma ocupado por proteínas CPn está constituido por las cianobacterias *Rubidibacter lacunae* y *Nodularia spumigena* con 0.46% y 0.43% respectivamente, ambas con la presencia de las proteínas PhnCDEFGHIJKLM, en ambos casos con varias copias de cada proteína, por lo que estas especies podrían ser estudiadas más a fondo y determinar por qué tienen tantas copias de este metabolismo. Si realmente se encuentran en situaciones de inanición de fósforo (P) para poder producir la cantidad necesaria de proteína, realizar el catabolismo de fosfonatos y obtener el P del ambiente.

A partir de generar mapas de calor de taxa específicos se puede observar la predominancia de muchas especies con proteínas CPn del mecanismo I, especialmente los grupos de Proteobacteria: Burkholderiales, Enterobacterales, Oceanospirillales, Pseudomonadales, Vibrionaceae, Neisseriales y de

Cyanobacteria destaca Nostocales. Y algunos grupos como Rhizobiales, Lactobacillales destacan con PhnW.

7.4. Repercusiones ambientales de los microorganismos con capacidad de metabolismo CPn y biotecnología

Los mecanismos hidrolíticos (estrategia metabólica II), que se encuentran en las vías de la fosfonatasa (PhnX), la fosfonoacetato hidrolasa (PhnY) y la fosfonopiruvato hidrolasa (PalA), requieren sustratos con un carbonilo beta en el núcleo de fósforo, para estabilizar un grupo saliente de carbanión tras la ruptura del enlace C-P (Horsman y Zechel, 2017). Por esta razón, estas vías son incapaces de catabolizar ácido metil fosfónico. Por el contrario, la vía de las enzimas C-P liasa (estrategia metabólica I) es notable por su capacidad para convertir una amplia variedad de Pn inactivados, incluido ácido metil fosfónico, en los correspondientes hidrocarburos y Pi (Villarreal-Chiu et al., 2012), por lo que la presencia de estos mecanismos puede estar más localizada en las especies de Bacteria dependiendo del ambiente donde estos organismos se desarrollan.

Hay algunos organismos cuya relevancia ecológica podría estar marcada por la presencia de los metabolismos del CPn, como el caso de las especies pertenecientes al género *Pseudomonas*. En trabajos como el de Wang, Dore, y McDermott (2017) mencionan las características de posibles cepas de *Pseudomonas* sp. con proteínas del metabolismo I con la PhnJ está implicada la producción de metano como desecho de la vía metabólica de la C-P liasa (**Figura 1**) en un contexto ambiental donde ocurre la “Paradoja de la sobresaturación de metano”, que implica que la sobresaturación de metano (Edwards y Owens, 1991), un gas de efecto invernadero que puede modificar drásticamente el ambiente a partir de la necesidad de los microorganismos por obtener recursos escasos como el P.

Se había planteado la posibilidad de usar contaminantes como el glifosato (N-fosfonometilglicina) como fuente de P con Hove-Jensen et. al. (2014) que consideran todas las posibles interacciones que puedan tener las proteínas componentes de los distintos mecanismos de catabolismo de fosfonatos, cuando los esfuerzos se han enfocado más en el estudio del mecanismo (I) de la C-P liasa por la versatilidad que tiene de usar diferentes sustratos.

Por ejemplo en el trabajo de Drzyzga y Lipok (2018), usaron especies de Cyanobacteria para crecer en la presencia de este compuesto fosfonatado y fósforo inorgánico, creciendo en condiciones de inanición y en medio rico de fósforo. En ese trabajo mencionan que la participación de las proteínas de la ruta del corte de enlace radical por medio de la C-P liasa es una posibilidad, y que se requiere tener los genomas completos de estas especies para hacer un análisis más detallado. Las especies con las que trabajaron fueron *Nostoc* cf. *muscorum*, *Anabaena variabilis*, *Chroococcus minutus* y *Fischerella* cf. *maior*, de las cuales este trabajo solo se puede comparar del símil *Nostoc* sp. PCC 7120 FACHB-418 que tiene las proteínas principales del mecanismo I PhnCDEFGHJKLM y potencialmente puede realizar el catabolismo de este tipo de fosfonatos contaminantes, la carencia de información acerca de los genomas de las otras tres especies de Cyanobacteria sigue siendo un limitante en el análisis.

7.5. Transferencia horizontal de los genes de las proteínas CPn

La presencia de más de una copia de un gen se puede explicar por medio de la duplicación génica y la transferencia horizontal, siendo esta última la causante del 88% de la expansión del genoma de las bacterias (Treangen y Rocha, 2011).

Lo que podría ocurrir en el caso de los genes que codifican a las proteínas del mecanismo II (PhnX, PhnW y PhnA) que no pertenecen a un *operón* es que se distribuyan por medio de la transferencia horizontal de genes, lo que explicaría su mayor distribución.

La hipótesis de la transferencia horizontal se ha sustentado por el trabajo de Raoult et al. (2004), en el que se encontró un homólogo del gen para la proteína PhnZ (del mecanismo III) en el genoma de un mimivirus lo que sugiere que los virus podrían ser un vehículo para la HGT, como se ha visto en otros genes del huésped (Sobecky y Hazen, 2009). Lamentablemente para este trabajo los HMMs para las proteínas PhnZ y PhnYox del mecanismo III encontraron muy pocos homólogos, por lo que no se pueden comparar los resultados a esta escala.

El fenómeno de transferencia horizontal parece poco factible con el mecanismo más representado (mecanismo I corte por C-P liasa), este podría darse debido a lo difícil que es pasar un operón de 14 genes que codifican a 14 proteínas.

Con la evidencia planteada en este trabajo acerca de la gran distribución del *operón phn* en Proteobacteria (53.5% de todo el phylum, **Figura 13** y **Figura 15**), se puede hipotetizar que este grupo comparte un ancestro en común que contaba con la capacidad de catabolizar moléculas de fosfonatos por medio del *operón phn*, pero al ser un metabolismo solamente activo en condiciones de inanición del fósforo, se fueron perdiendo algunos de los genes que lo componen o todo el operón. Con muchos casos donde se conservan los genes de transporte de fosfonatos PhnCDE que también se utilizan para transporte de fosfatos, en éstas condiciones de poseer fragmentos del *operón phn*, es posible que se hayan distribuido por transferencia horizontal a miembros del mismo grupo de Proteobacteria (según el análisis de Huang et al. 2005) y en otros phyla como Cyanobacteria, Firmicutes, etc.

8. Conclusiones

Se logró identificar y describir la distribución filogenética de las proteínas predichas del catabolismo de fosfonato en los phyla del dominio Bacteria donde:

- Se calcula que de las 4,413 especies con al menos una proteína CPn predominan las que poseen el mecanismo de corte de enlace radical de C-P por carbono-fósforo liasa con 3,810 especies (86.34%), seguidas de 1507 (34.15%) que poseen el mecanismo de Cortes hidrolíticos del enlace C-P) y 188 (4.26%) con el mecanismo de Corte de enlace oxidativo C-P por PhnYox / PhnZ tal y como se sabe de la literatura.
- La predominancia de especies con metabolismo CPn en taxa como Burkholderiales, Enterobacterales, Oceanospirillales, Pseudomonadales, Vibrionaceae, Neisseriales y Nostocales parece indicar que en la base de sus respectivos árboles evolutivos había ancestros comunes con la capacidad de usar el metabolismo CPn.
- Las especies de Alphaproteobacteria *Phyllobacterium sophorae* (0.76%), *Rhizobium miluonense* (0.57%) y *Mesorhizobium tianshanense* (0.49%) y otras 18 proteobacterias, son las que destinan mayor porcentaje de su proteoma al metabolismo de catabolismo de fosfonatos utilizando las proteínas PhnCDEFGHIJKLMN.
- Se llega a la hipótesis evolutiva de que el phylum Proteobacteria tenía un ancestro común con las proteínas de catabolismo de fosfonato de corte de enlace radical por medio de C-P liasa y algunos organismos han perdido y vuelto a adquirir el metabolismo por medio de transferencia horizontal.

9. Perspectivas

La tesis aquí presentada cuenta con las bases suficientes para iniciar un trabajo de carácter científico publicable en una revista indexada. Para lograr tal cometido es necesario considerar los siguientes puntos:

1. Algunas de las especies encontradas como (Phyllobacterium sophorae, Rhizobium miluonense y Mesorhizobium tianshanense) tienen el potencial metabólico de acceder a alternativas poco convencionales de fósforo, por lo que podrían aplicarse en proyectos biotecnológicos para la obtención de fosfonatos.
2. Los proteomas siempre se están actualizando, previo a publicar se necesitaría reconstruir los modelos ocultos de Márkov en una base de datos de proteomas más amplia para tener muestras representativas en distintos phyla y taxa.
3. Realizar un árbol filogenético en rigor con un número mayor de taxa.
4. Buscar el contexto genómico y evolución de cada proteína para entender mejor la historia evolutiva de estos metabolismos.

10. Bibliografía

1. Agarwal, V., Peck, S. C., Chen, J. H., Borisova, S. A., Chekan, J. R., Van Der Donk, W. A., & Nair, S. K. (2014). Structure and function of phosphonoacetaldehyde dehydrogenase: The missing link in phosphonoacetate formation. *Chemistry and Biology*, *21*(1), 125–135. https://doi.org/10.1016/j.chembiol.2013.11.006
2. Baker, A. S., Ciocci, M. J., Metcalf, W. W., Kim, J., Babbitt, P. C., Wanner, B. L., Martin, B. M., & Dunaway-Mariano, D. (1998). Insights into the mechanism of catalysis by the P-C bond-cleaving enzyme phosphonoacetaldehyde hydrolase derived from gene sequence analysis and mutagenesis. *Biochemistry*, *37*(26), 9305–9315. https://doi.org/10.1021/bi972677d
3. Barajas, H. R., Martínez-Sánchez, S., Romero, M. F., Álvarez, C. H., Servín-González, L., Peimbert, M., Cruz-Ortega, R., García-Oliva, F., & Alcaraz, L. D. (2020). Testing the Two-Step Model of Plant Root Microbiome Acquisition Under Multiple Plant Species and Soil Sources. *Frontiers in Microbiology*, *11*, 17. https://doi.org/10.3389/fmicb.2020.542742
4. Buchanan, B., Gruissem, W., & Jones, R. (2015). *Biochemistry and molecular biology of plants* (2nd ed.). John Wiley & sons.
5. Bundalovic-Torma, C., Whitfield, G. B., Marmont, L. S., Howell, P. L., & Parkinson, J. (2019). A systematic pipeline for classifying bacterial operons reveals the evolutionary landscape of biofilm machineries. *PLOS Computational Biology*, 1–32. https://doi.org/10.1101/769745
6. Carles, L., Gardon, H., Joseph, L., Sanchís, J., Farré, M., & Artigas, J. (2019). Meta-analysis of glyphosate contamination in surface waters and dissipation by biofilms. *Environment International*, *124*, 284–293. https://doi.org/10.1016/j.envint.2018.12.064
7. Christensen, D. G., Meyer, J. G., Baumgartner, J. T., D'Souza, A. K., Nelson, W. C., Payne, S. H., Kuhn, M. L., Schilling, B., & Wolfe, A. J. (2018). Identification of Novel Protein Lysine Acetyltransferases in *Escherichia coli*. *MBio*, *9*(5), 1–23. https://doi.org/10.1128/mBio.01905-18
8. Christoffels, A., & Van Heusden, P. (2019). Genome Annotation: Perspective From Bacterial Genomes. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 152–156. https://doi.org/10.1016/B978-0-12-809633-8.20092-7
9. Ciccarelli, F. D., Doerks, T., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, *311*(5765), 1283–1287. http://www.ncbi.nlm.nih.gov/pubmed/16513982
10. Clark, L. L., Ingall, E. D., & Benner, R. (1998). Marine Phosphorus is Selectively Remineralized. *Nature*, *393*, 426.
11. Demarchi, F., Bertoli, C., Copetti, T., Tanida, I., Brancolini, C., Eskelinen, E. L., & Schneider, C. (2006). Calpain is required for macroautophagy in mammalian cells. *Journal of Cell Biology*, *175*(4), 595–605. https://doi.org/10.1083/jcb.200601024
12. Drzyzga, D., & Lipok, J. (2018). Glyphosate dose modulates the uptake of

inorganic phosphate by freshwater cyanobacteria. *Journal of Applied Phycology*, 30(1), 299–309. https://doi.org/10.1007/s10811-017-1231-2

13. Dumora, C., Lacoste, A., & Cassaigne, A. (1983). Purification and Properties of 2-Aminoethy1phosphonate : Pyruvate Aminotransferase from *Pseudomonas aeruginosa*. *European Journal of Biochemistry*, *133*(1983), 119–125.
14. Eddy, S. (1998). *HMMER user's guide: biological sequence analysis using prole hidden Márkov models* (p. 229). citeseer.ist.psu.edu/eddy98hmmer.html
15. Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
16. Edwards, W. M., & Owens, L. B. (1991). Large storm effects on total soil erosion. *Journal of Soil and Water Conservation*, *46*(1), 75–78. https://www.jswconline.org/content/46/1/75
17. Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136. https://doi.org/10.1093/nar/gkr1178 https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=2&lvl=3&keep=1&srchmode=1&unlock
18. Galili, T., O'Callaghan, A., Sidi, J., & Sievert, C. (2018). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, *34*(9), 1600–1602. https://doi.org/10.1093/BIOINFORMATICS/BTX657
19. Gama, S. R., Vogt, M., Kalina, T., Hupp, K., Hammerschmidt, F., Pallitsch, K., & Zechel, D. L. (2019). An Oxidative Pathway for Microbial Utilization of Methylphosphonic Acid as a Phosphate Source [Research-article]. *ACS Chemical Biology*, *14*(4), 735–741. https://doi.org/10.1021/acschembio.9b00024
20. Gebhard, S., Tran, S. L., & Cook, G. M. (2006). The Phn system of *Mycobacterium smegmatis*: A second high-affinity ABC-transporter for phosphate. *Microbiology*, 152(11), 3453–3465. https://doi.org/10.1099/mic.0.29201-0
21. Ghodge, S. V., Cummings, J. A., Williams, H. J., & Raushel, F. M. (2013). Discovery of a cyclic phosphodiesterase that catalyzes the sequential hydrolysis of both ester bonds to phosphorus. *Journal of the American Chemical Society*, *135*(44), 16360–16363. https://doi.org/10.1021/ja409376k
22. Gilliam, J. W., Logan, T. J., & Broadbent, F. E. (2015). Fertilizer Use in Relation to the Environment. *Fertilizer Technology and Use*, 561–588. https://doi.org/10.2136/1985.FERTILIZERTECHNOLOGY.C16
23. González, O. F. (2016). *Distribución filogenética de las proteínas de la biosíntesis de carotenoides en bacterias.* Universidad Nacional Autónoma de México.
24. Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., … Pruitt, K. D. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, *46*(D1), D851–D860. https://doi.org/10.1093/nar/gkx1068

25. Hayes, J. E., Richardson, A. E., & Simpson, R. J. (2000). Components of organic phosphorus in soil extracts that are hydrolysed by phytase and acid phosphatase. *Biology and Fertility of Soils 2000 32:4*, *32*(4), 279–286. https://doi.org/10.1007/S003740000249
26. Horsman, G. P., & Zechel, D. L. (2017). Phosphonate Biochemistry. *Chemical Reviews*, *117*(8), 5704–5783. https://doi.org/10.1021/acs.chemrev.6b00536
27. Hove-Jensen, B., Zechel, D. L., & Jochimsen, B. (2014). Utilization of Glyphosate as Phosphate Source: Biochemistry and Genetics of Bacterial Carbon-Phosphorus Lyase. *Microbiology and Molecular Biology Reviews*, *78*(1), 176–197. https://doi.org/10.1128/mmbr.00040-13
28. Hove-Jensen, Bjarne, McSorley, F. R., & Zechel, D. L. (2011a). Physiological role of phnP-specified phosphoribosyl cyclic phosphodiesterase in catabolism of organophosphonic acids by the carbon-phosphorus lyase pathway. *Journal of the American Chemical Society*, *133*(10), 3617–3624. https://doi.org/10.1021/ja1102713
29. Hove-Jensen, Bjarne, McSorley, F. R., & Zechel, D. L. (2011b). Physiological role of phnP-specified phosphoribosyl cyclic phosphodiesterase in catabolism of organophosphonic acids by the carbon-phosphorus lyase pathway. *Journal of the American Chemical Society*, *133*(10), 3617–3624. https://doi.org/10.1021/ja1102713
30. Hove-Jensen, Bjarne, McSorley, F. R., & Zechel, D. L. (2012). Catabolism and Detoxification of 1-Aminoalkylphosphonic Acids: N-Acetylation by the phnO Gene Product. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0046416
31. Huang, J., Su, Z., & Xu, Y. (2005). The evolution of microbial phosphonate degradative pathways. *Journal of Molecular Evolution*, *61*(5), 682–690. https://doi.org/10.1007/s00239-004-0349-4
32. Kononova, S. V., & Nesmeyanova, M. A. (2002). Phosphonates and their degradation by microorganisms. *Biochemistry (Moscow)*, *67*(2), 184–195. https://doi.org/10.1023/a:1014409929875
33. Kulakova, A. N., Wisdom, G. B., Kulakov, L. A., & Quinn, J. P. (2003). The purification and characterization of phosphonopyruvate hydrolase, a novel carbon-phosphorus bond cleavage enzyme from *Variovorax* sp. Pal2. *Journal of Biological Chemistry*, *278*(26), 23426–23431. https://doi.org/10.1074/jbc.M301871200
34. Makino, K., Kim, S. K., Shinagawa, H., Amemura, M., & Nakata, A. (1991). Molecular analysis of the cryptic and functional phn Operons for phosphonate use in *Escherichia coli* K-12. *Journal of Bacteriology*, *173*(8), 2665–2672. https://doi.org/10.1128/jb.173.8.2665-2672.1991
35. Manav, M. C., Sofos, N., Hove-Jensen, B., & Brodersen, D. E. (2018). The Abc of Phosphonate Breakdown: A Mechanism for Bacterial Survival. *In BioEssays* (Vol. 40, Issue 11). https://doi.org/10.1002/bies.201800091
36. Martinez-Guerrero, C. E., Ciria, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., & Merino, E. (2008). GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic acids research*, 36(suppl_2), W176-W180. http://biocomputo.ibt.unam.mx:8080/GeConT/

37. Martínez-Sánchez, D. S. (2017). *Análisis del microbioma de rizósfera visto a través de diferentes suelos y hospederos*. Universidad Nacional Autónoma de México.
38. McMullan, G., & Quinn, J. P. (1994). In vitro characterization of a phosphate starvation-independent carbon-phosphorus bond cleavage activity in *Pseudomonas fluorescens* 23F. *Journal of Bacteriology*, *176*(2), 320–324. https://doi.org/10.1128/jb.176.2.320-324.1994
39. McSorley, F. R., Wyatt, P. B., Martinez, A., Delong, E. F., Hove-Jensen, B., & Zechel, D. L. (2012). PhnY and PhnZ comprise a new oxidative pathway for enzymatic cleavage of a carbon-phosphorus bond. *Journal of the American Chemical Society*, *134*(20), 8364–8367. https://doi.org/10.1021/ja302072f
40. Menzel, P., Ng, K. L., & Krogh, A. (2016a). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 9. https://doi.org/10.1038/ncomms11257 https://github.com/bioinformatics-centre/kaiju#:~:text=Kaiju%20is%20a%20program%20for,from%20microbial%20and%20viral%20genomes.
41. Menzel, P., Ng, K. L., & Krogh, A. (2016b). S:Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 1–6. https://doi.org/10.1038/ncomms11257
42. Metcalf, W. W., & Wanner, B. L. (1991). Involvement of the *Escherichia coli* phn (psiD) gene cluster in assimilation of phosphorus in the form of phosphonates, phosphite, P(i) esters, and P(i). *Journal of Bacteriology*, *173*(2), 587–600. https://doi.org/10.1128/jb.173.2.587-600.1991
43. Mohapatra, M., Yadav, R., Rajput, V., Dharne, M. S., & Rastogi, G. (2021). Metagenomic analysis reveals genetic insights on biogeochemical cycling, xenobiotic degradation, and stress resistance in mudflat microbiome. *Journal of Environmental Management*, *292*, 112738. https://doi.org/10.1016/j.jenvman.2021.112738
44. Morales, M. E., Allegrini, M., Basualdo, J., Villamil, M. B., & Zabaloy, M. C. (2020). Primer design to assess bacterial degradation of glyphosate and other phosphonates. *Journal of Microbiological Methods*, 169, 105814. https://doi.org/10.1016/j.mimet.2019.105814
45. Nowack, B. (2003). Environmental chemistry of phosphonates. *Water Research*, 37(11), 2533–2546. https://doi.org/10.1016/S0043-1354(03)00079-4
46. O'leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., Mcveigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*, 733–745. https://doi.org/10.1093/nar/gkv1189
47. Quinn, J. P., Kulakova, A. N., Cooley, N. A., & McGrath, J. W. (2007). New ways to break an old bond: The bacterial carbon-phosphorus hydrolases and their role in biogeochemical phosphorus cycling. *Environmental Microbiology*, *9*(10), 2392–2400. https://doi.org/10.1111/j.1462-2920.2007.01397.x

48. Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., et al. (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306: 1344–1350.
49. Repeta, D.J., Ferrón, S., Sosa, O.A., Johnson, C.G., Repeta, L.D., Acker, M., et al. (2016) Marine methane paradox explained by bacterial degradation of dissolved organic matter. *Nature Geoscience* 9: 884–887. doi: 10.1038/ngeo2837
50. Sobecky, P.A., & Hazen, T.H. (2009) Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Molecular Biology* 532: 435–453.
51. Stosiek, N., Talma, M., & Klimek-Ochab, M. (2019). Carbon-Phosphorus Lyase—the State of the Art. *Applied Biochemistry and Biotechnology*, 1–28. https://doi.org/10.1007/s12010-019-03161-4
52. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., ... & von Mering, C. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1), D605-D612. https://string-db.org/cgi/network.pl?taskId=VkWMOCKZSjV6
53. Taboada, B., Estrada, K., Ciria, R., & Merino, E. (2018). Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*, 34(23), 4118-4120. https://biocomputo.ibt.unam.mx/operon_mapper/
54. Tapia-Torres, Y., & García-Oliva, F. (2013). La disponibilidad del fósforo es producto de la actividad bacteriana en el suelo en ecosistemas oligotróficos: una revisión crítica. *TERRA Latinoamericana*, *31*, 231–242.
55. Tran, S. L., Rao, M., Simmers, C., Gebhard, S., Olsson, K., & Cook, G. M. (2005). Mutants of *Mycobacterium smegmatis* unable to grow at acidic pH in the presence of the protonophore carbonyl cyanide m-chlorophenylhydrazone. *Microbiology*, *151*(3), 665–672. https://doi.org/10.1099/mic.0.27624-0
56. Treangen TJ & Rocha EP. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 2011;7(1):e1001284.
57. Ulrich, E. C., Kamat, S. S., Hove-Jensen, B., & Zechel, D. L. (2018). Methylphosphonic Acid Biosynthesis and Catabolism in Pelagic Archaea and Bacteria. In *Methods in Enzymology* (Vol. 605). https://doi.org/10.1016/bs.mie.2018.01.039
58. Van Staalduinen, L. M., McSorley, F. R., Schiessl, K., Séguin, J., Wyatt, P. B., Hammerschmidt, F., Zechel, D. L., & Jia, Z. (2014). Crystal structure of PhnZ in complex with substrate reveals a di-iron oxygenase mechanism for catabolism of organophosphonates. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(14), 5171–5176. https://doi.org/10.1073/PNAS.1320039111
59. Villarreal-Chiu, J. F., Quinn, J. P., & McGrath, J. W. (2012). The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. *Frontiers in Microbiology*, *3*(JAN), 19. https://doi.org/10.3389/fmicb.2012.00019
60. Wanner, B. L., & Boline, J. A. (1990). Mapping and molecular cloning of the

phn (psiD) locus for phosphonate utilization in *Escherichia coli. Journal of Bacteriology*, *172*(3), 1186–1196. https://doi.org/10.1128/JB.172.3.1186-1196.1990
61. Wang, Q., Dore, J. E., & McDermott, T. R. (2017). Methylphosphonate metabolism by Pseudomonas sp. populations contributes to the methane oversaturation paradox in an oxic freshwater lake. *Environmental Microbiology*, 19(6), 2366–2378. https://doi.org/10.1111/1462-2920.13747
62. Wilkins, Marc (2009). Proteomics data mining. *Expert Review of Proteomics. England.* 6 (6): 599–603. doi:10.1586/epr.09.81.
63. Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Márkov models. *BMC Bioinformatics*, *15*(1), 1–9. https://doi.org/10.1186/1471-2105-15-7 http://skylign.org/
64. White, A. K., & Metcalf, W. W. (2007). Microbial Metabolism of Reduced Phosphorus Compounds. *Annual Review of Microbiology*, *61*(1), 379–400. https://doi.org/10.1146/annurev.micro.61.080706.093357
65. Wörsdörfer, B., Lingaraju, M., Yennawar, N. H., Boal, A. K., Krebs, C., Bollinger, J. M., & Pandelia, M. E. (2013). Organophosphonate-degrading PhnZ reveals an emerging family of HD domain mixed-valent diiron oxygenases. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(47), 18874–18879. https://doi.org/10.1073/pnas.1315927110
66. Wu, X., Peng, J., Liu, P., Bei, Q., Rensing, C., Li, Y., Yuan, H., Liesack, W., Zhang, F., & Cui, Z. (2021). Metagenomic insights into nitrogen and phosphorus cycling at the soil aggregate scale driven by organic material amendments. *Science of the Total Environment*, *785*, 147329. https://doi.org/10.1016/j.scitotenv.2021.147329

11. Anexos

Los anexos en forma de tablas y figuras complementarias citadas en esta tesis se encuentran disponibles en: https://github.com/LChora/LChora