Centro de Investigación Científica de Yucatán, A.C.

Posgrado en Ciencias Biológicas

“MINERÍA DE DATOS METAGENÓMICOS PARA IDENTIFICAR NUEVAS PROTEASAS S8A EN COMUNIDADES BACTERIANAS DEL ACUÍFERO DE YUCATÁN”

Tesis que presenta:

Perla Analuz Contreras de la Rosa

En opción al título de

MAESTRA EN CIENCIAS

(Ciencias Biológicas: Opción Biotecnología)

Asesoras:
Dra. Elsa Góngora Castillo
Dra. Aileen O’Connor Sánchez

Mérida, Yucatán, México

2020

CENTRO DE INVESTIGACIÓN CIENTÍFICA DE YUCATÁN, A. C.
POSGRADO EN CIENCIAS BIOLÓGICAS

RECONOCIMIENTO

Por medio de la presente hago constar que el trabajo de tesis de Perla Analuz Contreras de la Rosa, titulado “Minería de datos metagenómicos para identificar nuevas proteasas S8A en comunidades bacterianas del acuífero de Yucatán”, fue realizado en los laboratorios de Metagenómica y Biotecnología computacional de la Unidad de Biotecnología, del Centro de Investigación Científica de Yucatán, A.C., en la línea de Biotecnología de microorganismos bajo la dirección de las Dras. Elsa Góngora Castillo y Aileen O’Connor Sánchez, dentro de la opción de Biotecnología, perteneciente al Programa de Posgrado en Ciencias Biológicas de este Centro.

Atentamente:

Dra. Cecilia Hernández Zepeda
Directora de Docencia

Mérida, Yucatán, México, a 2 de Diciembre de 2020

DECLARACIÓN DE PROPIEDAD

Declaro que la información contenida en la sección de Materiales y Métodos Experimentales, los Resultados y Discusión de este documento proviene de las actividades de experimentación realizadas durante el período que se me asignó para desarrollar mi trabajo de tesis, en las Unidades y Laboratorios del Centro de Investigación Científica de Yucatán, A.C., y que a razón de lo anterior y en contraprestación de los servicios educativos o de apoyo que me fueron brindados, dicha información, en términos de la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, le pertenece patrimonialmente a dicho Centro de Investigación. Por otra parte, en virtud de lo ya manifestado, reconozco que de igual manera los productos intelectuales o desarrollos tecnológicos que deriven o pudieran derivar de lo correspondiente a dicha información, le pertenecen patrimonialmente al Centro de Investigación Científica de Yucatán, A.C., y en el mismo tenor, reconozco que si derivaren de este trabajo productos intelectuales o desarrollos tecnológicos, en lo especial, estos se regirán en todo caso por lo dispuesto por la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, en el tenor de lo expuesto en la presente Declaración.

Perla Analuz Contreras de la Rosa

AGRADECIMIENTOS

Al Centro de Investigación Científica de Yucatán y a la Unidad de Tecnologías de Información y Comunicación, por el asesoramiento y las instalaciones prestadas para el uso del equipo de súper cómputo “Hobón”, esencial para el desarrollo de la parte computacional de este trabajo.

A la Unidad de Biotecnología por las instalaciones prestadas para el desarrollo de la parte experimental de este trabajo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca No. 931932.

A mis asesoras, Dra. Elsa Góngora Castillo y Dra. Aileen O’Connor Sánchez, gracias por la paciencia, los ánimos y por todos los conocimientos propios de su área que han compartido conmigo, crecí con su asesoría.

A mi comité tutoral y revisores de tesis: Dr. Jorge Humberto Ramírez Prado, gracias por los múltiples tips que me proporcionó para el procesamiento de los datos de este trabajo, así como sus retroalimentaciones. Dra. Alejandra Prieto Davó por las observaciones realizadas a este trabajo, las cuales fueron de mucha ayuda para no perder de vista el enfoque biológico. Dr. Ramón Pacheco por la revisión de este trabajo.

Un agradecimiento especial al Dr. César de los santos Briones, por toda su ayuda técnica proporcionada en el laboratorio, para la extracción del ADN metagenómico. A la M. en C. Adriana Quiroz, por sus valiosos consejos, revisiones, retroalimentaciones y apoyo en todo el desarrollo de este trabajo.

A mí, por el gusto de aprender nuevas cosas y porque a pesar de las situaciones externas e internas he llegado a esta meta.

A mis compañeros de laboratorio: M. en C.C. Cynthia Soto, IBQ. Juan Martínez e IBQ. Daniel Macías, por su amistad y las múltiples retroalimentaciones que han aportado a este trabajo.

A mis compañeros de la Unidad de Biotecnología: M. en C. Eddy Turrent, Biol. Francisco Cárdenas, Biol. Diana Jacinto e IBQ. Jair Díaz, por brindarme su ayuda y amistad.

Y por último pero no menos importante, a mis padres Perla y Luciano por siempre apoyarme en todo para conseguir mis metas. IMEC Ferdy gracias por estar en mi vida.

Este trabajo se re direccionó y finalizó en medio de la pandemia causada por el virus SARS-CoV-2, la cual nos recuerda el alcance y la importancia que las Ciencias Biológicas y la Biotecnología poseen.

Productos académicos

2019. Presentación de póster: Mining genomic data to identify novel proteases S8 in plants. En co-autoría con los Doctores: Jorge Humberto Ramírez-Prado, Aileen O'Connor-Sánchez, Elsa Góngora-Castillo. En el XVIII National Congress of Biochemistry and Plant Molecular Biology, XI Symposium México/USA & 1st ASPB México Section Meeting held, en Mérida Yucatán México, October 28-31

2019. Asistencia: 4th International Symposium on Functional Genomics and Systems Biology Mérida, Yucatán November 20-22.

2020. Publicación de artículo, coautoría: Góngora-Castillo, E., L.A. López-Ochoa., M.M. Apolinar-Hernández., A.M. Caamal-Pech, **P.A. Contreras-de la Rosa**., A. Quiroz-Moreno, J.H. Ramírez-Prado y A. O'Connor-Sánchez (2020). Data mining of metagenomes to find novel enzymes: a non-computationally intensive method. 3 Biotech, 10. 1-8. doi:10.1007/s13205-019-2044-6

.

ÍNDICE

LISTADO DE FIGURAS

LISTADO DE CUADROS

ABREVIATURAS

ARN	Ácido ribonucleico
E.C	Enzyme Commission numbers
ATP	Adenosín trifosfato
pH	Potencial de hidrógeno
ARNr	Ácido ribonucleico ribosomal
PKS	Polyketide synthase
WMS	Secuenciación Total de Metagenoma (Whole metagenome sequencing)
BWA	Alineador de Burrows- Wheeler (Burrows-Wheeler Aligner)
GC	Guanina-Citocina
BLAST	Basic Local Alignment Search Tool
BLASTP	Basic Local Alignment Search Tool, protein
HMM	Modelos Ocultos de Markov (Hidden Markov Model)
HTML	HyperText Markup Language
ADN	Ácido desoxirribonucleico
S8A	Familia serina 8A
PS8A	Proteasa S8A
pb	Pares de base
DHS	Ácido aspártico, histidina, serina

Phylip Nombre del formato originalmente usado por el paquete bioinformático **“**PHYLogeny Inference Package”

PhyML Nombre de una aplicación de inferencia filogenética basada en Máxima Verosimilitud (Phylogenetic Maximum Likelihood method)

BIONJ Nombre de una variante del método de inferencia filogenética Neighbor Joining

RESUMEN

Las proteasas (EC.3.4.21) son enzimas hidrolíticas que catalizan la ruptura de los enlaces peptídicos de las proteínas utilizando una molécula de agua. Las proteasas S8A se caracterizan porque para ello utilizan una triada catalítica compuesta por tres aminoácidos en el orden: Ácido aspártico (D), Histidina (H), y Serina (S). Poseen múltiples aplicaciones en diversos campos industriales, especialmente las de origen bacteriano, por lo que hay un gran interés en el descubrimiento de nuevas proteínas que puedan tener nuevas propiedades para uso biotecnológico. En este estudio, se realizó una búsqueda *in silico* de secuencias del *cor*e de proteasas S8A a través del análisis de metagenomas, que provinieron de muestras de agua pertenecientes a tres puntos geográficos del acuífero de Yucatán. Para lo cual se utilizaron herramientas bioinformáticas, y se emplearon dos estrategias de minería de datos: expresiones regulares y modelos ocultos de Markov. La primera se define como un conjunto de reglas que especifican elementos que se desean encontrar en un archivo de texto, por lo que estuvieron diseñadas de acuerdo con los motivos conservados de las proteasas S8A y con los aminoácidos variables ubicados entre ellos. La segunda es un modelo probabilístico que otorga puntuaciones a las secuencias que se desea identificar, cuando se comparan con una familia de proteínas. Se obtuvieron 135 secuencias cuando se usaron las expresiones regulares y 245 cundo se usaron los modelos ocultos de Markov. En el análisis de anotación funcional se determinó que más del 50% de las PS8A obtenidas se encontraron en un solo punto geográfico, el cual tuvo el mayor número de contigs y el mayor número de phyla. Para la inferencia filogenética de las secuencias se utilizaron: 245 obtenidas de la minería de datos y 13 de proteasas conocidas. Se seleccionaron como putativas 16 secuencias proteasas S8A, pues se ubicaron en dos clados independientes a los que agrupan las 242 restantes. Se seleccionaron 16 secuencias *core* putativas novedosas de acuerdo al análisis filogenético y al organismo al que pertenecen, en estas se observaron variaciones de los aminoácidos ubicados justo al lado de cada motivo conservado, los cuales difieren del resto de las secuencias obtenidas y presentaron una identidad entre 37% y 91% con respecto a la PS8A más cercana. Las secuencias putativas recuperadas en este trabajo, resaltan el alcance que las tecnologías de secuenciación masiva, la bioinformática y la metagenómica pueden tener.

ABSTRACT

Proteases (EC.3.4.21) are hydrolytic enzymes that catalyze the breaking of peptide bonds in proteins using a water molecule. S8A proteases are characterized by using a catalytic triad composed of three amino acids in the order: aspartic acid (D), histidine (H) and serine (S). They have multiple applications in various industrial fields, especially those of bacterial origin, so there is great interest in the discovery of new proteins that may have new properties for biotechnological use. In this study, a silico search for sequences of the nucleus of S8A proteases was carried out through the analysis of metagenomes, which came from water samples belonging to three geographical points of the Yucatán aquifer. For which it will be used in bioinformatics tools, and two data mining strategies were used: regular expressions and hidden Markov models. The first is defined as a set of rules that specify the elements to be found in a text file, designed from the conserved motifs of the S8A proteases and with the variable amino acids located between them. The second is a probabilistic model that scores the sequences to be identified in comparison to a family of proteins. 135 sequences were obtained when the regular expressions were used and 245 when the hidden Markov models were used. In the functional annotation analysis, it was determined that more than 50% of the PS8A obtained was found in a single geographical point, which had the highest number of contigs and the highest number of phyla. For the phylogenetic inference from the sequences, the following were used: 245 obtained from data mining and 13 from known proteases. 16 S8A protease sequences were selected, since they will be located in two independent clades to those that encompass the remaining 242 PS8A. For the 16 supposed nucleus sequences were selected according to the phylogenetic analysis and the organism to which they belong, in these variations of the amino acids located right next to each conserved motif were observed, which differ from the rest of the sequences obtained and found an identity between 37% and 91% compared to the closest PS8A. The putative sequences recovered in this work highlight the scope that massive sequencing technologies, bioinformatics, and metagenomics have.

INTRODUCCIÓN

El avance de las tecnologías de secuenciación de ADN ha permitido que las ciencias como la genómica, transcriptómica y metagenómica, entre otras, realicen grandes aportaciones científicas, tales como la caracterización de genes, que dirigen la producción de proteínas, antibióticos y/o la caracterización de ambientes microbianos desconocidos (Vailati *et al.,* 2017).

Este trabajo se centró en el uso de la metagenómica, la cual estudia los conjuntos de genomas bacterianos presentes en un sitio de interés, de organismos cultivables y no cultivables, puesto que solo el 1% de las bacterias que son obtenidas de una muestra ambiental se pueden cultivar. La metagenómica se apoya de las tecnologías de secuenciación de nueva generación del ADN y de herramientas bioinformáticas (Walshaw *et al.,* 2011).

Entre las amplias aplicaciones que el estudio de los metagenomas genera, se encuentran: identificación de patógenos del ser humano (Zhao y Shen, 2011), la identificación de nuevos virus (Montoya *et al.,* 2011), así como la apertura a la exploración de nichos como desiertos (Neveu *et al.,* 2011), aguas subterráneas (Apolinar-Hernández, 2016), esponjas marinas (O'Connor-Sanchéz *et al*., 2014) e (Meirelles *et al.,* 2016), por citar aglunas, con el fin de caracterizar la comunidad microbiana y/o sus genes.

El mercado global de enzimas industriales en el año 2015 representó $4.4 mil millones de dólares y el 35% de este mercado es representado por las Proteasas S8 o subtilisinas. Estas se emplean en la elaboración de biodertegentes debido a que son altamente termoestables y actúan en ambientes alcalinos.

Este trabajo se enfoca en la identificación de genes novedosos de Proteasas S8A, mediante el uso de la minería de datos metagenómicos (Shalwan y Sharma, 2019). Las estrategias bioinformáticas para ello incluyen el diseño y uso de expresiones regulares y Modelos Ocultos de Márkov.

Las expresiones regulares buscan patrones en un archivo de texto a diferencia de los modelos ocultos de Márkov que se encargan de comparar secuencias en este caso metagenómicas con un modelo estadístico que describe un familia de secuencias (Proteasas S8).

CAPÍTULO I
ANTECEDENTES

1. Enzimas

Las enzimas son macromoléculas de naturaleza proteica y peso molecular elevado, que catalizan reacciones biológicas tanto anabólicas como catabólicas. Todas ellas presentan dos propiedades fundamentales: (i) aumentan la velocidad de las reacciones sin ser consumidas o alteradas permanentemente por la reacción y (ii) aumentan la velocidad de reacción sin alterar el equilibrio químico entre los reactivos y los productos (Cooper, 2000).

Según la definición de Lehninger (2013), todas las enzimas son proteínas, exceptuando a aquellas moléculas de ARN catalítico. La actividad catalítica que desempeñan depende de su secuencia y de la conformación nativa que presenten, algunas además requieren un componente químico adicional, llamado cofactor, así como otros iones inorgánicos como Fe^{2+}, Mg^{2+} o Zn^{2+}, o de complejos orgánicos denominados coenzimas, los cuales son derivados de vitaminas y de nutrientes orgánicos.

Muchas enzimas son nombradas con el sufijo “asa” antecedido de alguna palabra relacionada con su actividad catalítica, como ejemplo la Ureasa, la cual cataliza la hidrólisis de urea. Lo anterior provocó que algunas enzimas tuvieran el mismo nombre, debido a que se han incrementado las enzimas descubiertas, por lo que para homogeneizar el criterio para nombrarlas, se creó un sistema de clasificación. Este las divide dentro de seis clases, con subclases, de acuerdo con el tipo de reacción que catalizan (Cuadro 1.1); por lo que a cada enzima se le asigna un nombre sistemático que identifica la reacción catalizada y un número clasificatorio de cuatro dígitos (número E.C.). Por ejemplo, si el nombre sistemático de una enzima es ATP: Glucosa fosfotransferasa indicando que cataliza la transferencia de un grupo fosforilo desde el ATP a la glucosa, entonces el número de clasificación de esta enzima es 2.7.1.1. El primer número (el 2), corresponde al nombre de la clase (transferasa); el segundo (7), a la subclase fosfotransferasa; el tercero (1), a las fosfotransferasas con un grupo hidroxilo como aceptor y el cuarto digito (1), D- glucosa como aceptor del grupo fosforilo (Lehniger, 2013).

Cuadro 0.1 Clasificación internacional de las enzimas

Número	Clase	Tipo de reacción que catalizan
1	Oxidorreductasas	Transferencia de electrones
2	Transferasas	Transferencia de grupos
3	Hidrolasas	Transfiere grupos funcionales al agua
4	Liasas	Escinde, elimina o agrega grupos, formando doble enlace
5	Isomerasas	Transfiere grupos para la formación de formas isoméricas
6	Ligasas	Une por reacciones de condensación, liberando ATP o un cofactor similar.

Las enzimas están conformadas por dominios funcionales los cuales son regiones de la cadena polipeptídica que pueden plegarse de manera estable e independiente y pueden ser (i) funcionales si es una unidad que lleva a cabo una función bioquímica determinada y (ii) estructurales si se refiere a un componente estable de la estructura (Tyson y Novák, 2010), así como por los denominados motivos proteicos, los cuales son elementos conservados en la secuencia de aminoácidos o nucleótidos, asociados a una función por ejemplo la catálisis. Suelen formar parte del sitio activo de las enzimas.

Las enzimas, además de ayudar al metabolismo a realizar sus múltiples funciones controlando eficazmente la velocidad de las reacciones químicas dentro de las células, tienen características de interés industrial pues se considera que en los sectores industriales en donde se involucra al menos una reacción química, es posible integrar una enzima, por lo tanto son útiles para transformar diferentes moléculas en productos específicos. Se utilizan por ejemplo, en las industrias cervecera, láctea y peletera desde finales del siglo XIX y principios del XX, cuando fueron establecidos sus mecanismos de acción. En el año 2015, las enzimas industriales representaron una cantidad en el mercado global de aproximadamente $ 4.4 mil millones de dólares. Dentro de estas, las proteasas tienen un 60% de participación debido a su amplio uso como bioaditivos en la lavandería,

farmacéutica, cuero, industria alimentaria y agricultura, es por ello que se trabaja en la búsqueda de nuevas enzimas que desempeñen más satisfactoriamente las demandas industriales (Shalwan y Sharma, 2019).

1.1 Proteasas S8

Las proteasas, también llamadas peptidasas, son un grupo de enzimas hidrolasas (EC-3) presentes en la naturaleza, constituyentes esenciales de procariotas, protozoarios, hongos, plantas y animales, catalizan la ruptura de los enlaces peptídicos de las proteínas, utilizando una molécula de agua. Se encuentran divididas en dos grandes grupos: (i) endopeptidasas, las cuales actúan dentro de la cadena polipeptídica y dependiendo del aminoácido principal que se ubica en su sitio activo, se dividen en cuatro grupos: serina proteasas (EC.3.4.21), acido aspártico proteasas (EC.3.4.23), cisteía proteasas (EC.3.4.22) y metaloproteasas (EC.3.4.24); y (ii) exopeptidasas, que actúan en el enlace peptídico más cercano al amino o carboxilo terminal de los extremos de la cadena (Hartley, 1960).

Las proteasas poseen múltiples aplicaciones en diversos campos industriales, especialmente las de origen bacteriano, puesto que las bacterias elaboran enzimas intracelulares las cuales son importantes en procesos metabólicos, como la esporulación, recambio de proteínas y maduración de hormonas y las extracelulares, que permiten hidrolizar proteínas permitiendo que la célula absorba y utilice productos hidrolizados. Debido a lo anterior y al alto rendimiento que poseen, se trabaja en el descubrimiento de nuevas cepas productoras de enzimas novedosas que puedan tener uso biotecnológico (Furhan y Sharma, 2014).

Las Proteasas S8A de acuerdo con la base de datos MEROPS-The peptidase database (https://www.ebi.ac.uk/merops/) (Rawligns *et al*., 2017), la cual clasifica a las proteasas en clanes y familias según la secuencias y similitudes estructurales; pertenecen a la familia S8 y son nombradas de igual manera serin proteasas o subtilisinas. Estas se encuentran involucradas en la nutrición de bacterias, arqueas y hongos, y poseen aplicaciones industriales constituyendo más de un tercio de la participación en el mercado de enzimas. Generalmente se utilizan en la fabricación de biodetergentes, ya que son termoestables por ser activas en temperaturas que van de 10°C a 70°C y en un intervalo de pH neutro a alcalino (6 a 11), propio de los mismos, también se emplean en la industria farmacéutica y en el tratamiento de desechos industriales (Gupta y Lorenz, 2002).

Los miembros de la familia S8 poseen una triada catalítica en su secuencia de aminoácidos caracterizada por los aminoácidos Ácido aspártico (D), Histinida (H), y Serina (S), la cual está muy conservada y es específica para este tipo de proteínas, siendo variables los aminoácidos que separan cada miembro de la triada.

La estructura tridimensional que presentan las proteasas S8 consiste en tres capas con una lámina β de siete cadenas intercalada entre dos capas de hélices. La mayoría son endopeptidasas y mayormente son inhibidas por los inhibidores generales de la serina peptidasa, como Diisopropilfluorofosfato (DFP) y Phenylmethylsulfonyl fluoride (PMSF).

Debido a su alto uso en el mercado e industria se han implementado estudios metagenómicos como alternativas útiles y viables para el descubrimiento de enzimas de diversos nichos ecológicos, y en particular para microbios no cultivables. Encontrarlas representaría poder contar con enzimas novedosas (Shalwan y Sharma, 2019) con el potencial de ser más adecuadas para algunos procesos industriales.

1.2 Metagenómica y el acuífero de Yucatán

En la década de los 90's, se dio a conocer que el 99% de las bacterias presentes en una muestra obtenida de cualquier ecosistema no se pueden cultivar bajo condiciones de laboratorio (Sharma *et al.,* 2008), debido posiblemente a que no se conocen las condiciones fisicoquímicas, nutricionales o fisiológicas bajo las cuales se desarrollan, o bien, los ambientes en donde habitan presentan características extremas que dificultan igualarlas dentro de un medio de cultivo (Hernández-León *et al.,* 2010), por lo que las metodologías de estudios genómicos mediante el cultivo implican sesgos hacia información genómica exclusivamente de organismos cultivables.

Sin embargo, mediante la metagenómica se puede acceder a la información genómica de una muestra ambiental, incluyendo a las bacterias no cultivables.

Actualmente, los estudios metagenómicos son cada vez más comunes, como ejemplo se pueden mencionar aquellos que se centran en la caracterización de ambientes como la cavidad oral (Serrano y Cardona, 2015) el intestino humano (Gill *et al.,* 2006) el rumen de las vacas (Stewart *et al.,* 2018) el mar de los sargazos (Venter *et al.,* 2004), el microbioma de pacientes con fibrosis quística (Lim *et al.,* 2013), entre otros, con el fin de poder

desarrollar nuevas estrategias de diagnósticos y tratamientos médicos y generar productos industriales novedosos.

En este trabajo se analizaron metagenomas provenientes del acuífero de la Península de Yucatán, debido a que se ha reportado que la conformación de los sistemas acuáticos subterráneos está dada principalmente por la presencia de microorganismos (Perry *et al.,* 2009). El estudio de los metagenomas presentes en esta zona es prometedor ya que se trata de sistemas acuáticos únicos en el mundo, muy poco explorados y ricos en diversidad de microorganismos, aunado al hecho de que se han encontrado proteínas nuevas a partir del estudio de estos como son los trabajos de (Apolinar-Hernández *et al*., 2016; Marfil-Santana *et al.,* 2016) y reportándose en el primero Proteasas S8A novedosas, de metagenomas presentes en el agua de zonas del acuífero de Yucatán, mientras que en el segundo la diversidad de los genes PKS Tipo I útiles en la medicina

1.3 Tecnologías de secuenciación masiva

Las tecnologías de secuenciación de ADN existen desde principios de la década de 1970 y han evolucionado a paso acelerado otorgando en la actualidad un amplio abanico de opciones (Besser *et al.,* 2018).

Existen dos estrategias principales de secuenciación para el estudio de los metagenomas, la primera se basa en la secuenciación de amplicones de un gen objetivo en el metagenoma mediante el uso de cebadores específicos como el 16s ARNr utilizado para las especies procariotas, y el espaciador transcrito interno (ITS) para hongos y otros eucariontes, otorgando la ventaja de dirigir la obtención de solamente las secuencias de esas regiones de los genomas presentes en el metagenoma.

Es importante señalar que esta estrategia cuenta con la desventaja de que en ocasiones se obtienen secuencias quiméricas (formadas por dos segmentos de secuencias de diferentes organismos) y el sesgo de información hacia un gen blanco, puesto que todas las secuencias representarán la misma región genómica (Sedlar *et al.,* 2017).

La segunda estrategia es la secuenciación de metagenomas en escopeta (Shotgun o WMS por sus siglas en inglés); esta consiste en la generación y secuenciación de millones de fragmentos genómicos que se derivan de los genomas de todos los microorganismos presentes en la muestra estudiada, no va dirigida a un gen específico. Esta estrategia

permite obtener millones de secuencias en donde cada secuencia representa una parte aleatoria, de un genoma que se encuentra presente en la muestra (Vincent y Charette, 2015).

Por lo que cualquier estrategia seleccionada según la pregunta de investigación a responder, dará como resultado datos masivos que representan fragmentos de secuencias de longitud variable. Dependiendo de la tecnología de secuenciación que se utilice. Dichos fragmentos deben ser ensamblados para recrear segmentos más grandes de los genomas de la muestra. Los programas de ensamblado de metagenomas actualmente están desarrollados exclusivamente para el manejo de dichos datos ya que en contarste del ensamblador de un genoma, el algoritmo para el ensamblaje de metagenomas debe ser capaz de poder ensamblar de cientos a millones de genomas de una misma muestra (Vollmers *et al.,* 2017).

Existen dos grupos diferentes de estrategias para la agrupación de datos derivados de WMS. El primero: dependiente de la taxonomía, se basa en la comparación de secuencias con las bases de datos de referencia, de tres maneras (1) a nivel de referencia usando algoritmos de alineación como BLAST, Bowtie y/o BWA; (2) a nivel de modelo de un origen filogenético conocido, usando el Modelo oculto de Márkov (HMM) y bases de datos específicas como Pfam; y (3) en un nivel de composición de secuencia utilizando el contenido de GC y patrones de oligonucleótidos.

Este grupo de estrategias tiene algunas desventajas, como que el uso de los algoritmos de comparación requiere mucho tiempo, y que las bases de datos de referencia que contienen secuencias genómicas no están completas, por lo que un gran número de secuencias posiblemente arrojen resultados no asignados (Sedlar *et al.,* 2017).

El segundo, independiente de la taxonomía, basado en la extracción de parámetros específicos para un taxón determinado a partir de los contigs ensamblados, los parámetros obtenidos se comparan y agrupan directamente mediante el uso de algoritmos, sin la necesidad de una base de datos de referencia.

1.4 MG-RAST: servidor de análisis metagenómicos

El servidor RAST de metagenómica (MG-RAST) (https://www.mg-rast.org/) (Wilke *et al*., 2017), desde el año 2008, es un repositorio y analizador de datos metagenómicos y metadatos. Está desarrollado para colaboraciones multidisciplinarias orientadas hacia objetivos comunes, por ello en dicho servidor los usuarios pueden hacer públicos datos de secuencias en bruto de metagenomas, para realizar un análisis y compartirlos con otros usuarios. MG-RAST, acepta las secuencias metagenómicas en archivo tipo FASTA así como secuencias ensambladas y no ensambladas. Realiza un análisis automatizado proporcionando conocimientos cuantitativos de las poblaciones microbianas. Con base en las secuencias analizadas (Meyer *et al.,* 2008): calcula la abundancia relativa de los diferentes niveles de las categorías taxonómicas, así como los niveles de clasificación funcional y filogenia, con referencia en las secuencias anotadas mediante la realización automática de un BLAST con diversas bases de datos. Como ejemplos de dichas bases de datos se podrían mencionar la *Gene Ontology*, la cual es la fuente de información más grande del mundo sobre las funciones de los genes, KEGG, (*Kyoto Encyclopedia of Genes and Genomes*) la cual es una colección de bases de datos de genomas y rutas enzimáticas y la NCBI, (*National Center for Biotechnology Information*) la cual es una biblioteca de información genómica, entre otras (Keegan *et al*., 2007). De igual forma MG-RAST ofrece otras funciones como servicios de limpieza de secuencias y dereplication utilizando un k-mer (un número determinado de nucleótidos) para identificar rápidamente todas las secuencias idénticas de prefijo de 20 caracteres, logrando la disminución del tamaño del archivo y de errores en el ensamblado.

De igual manera MG-RAST provee metadatos de los metagenomas, los cuales son información contextual del significado y propiedades de los datos, en este caso la información adicional de las secuencias metagenómicas, como lugar de colecta, plataforma de secuenciación, entre otros. Por ello el uso de metadatos es una herramienta muy útil en el área de la metagenómica debido a que al generar inventarios de genes microbianos, se pueden utilizar para buscar y acceder a registros que cumplan ciertos criterios (Keegan *et al.,* 2007).

MG-RAST actualmente ha analizado más de 60 tera-bases de datos de más de 150,000 conjuntos de datos, de los cuales 23,000 se encuentran disponibles al público. Debido a la

gran cantidad de información que maneja está sometido a revisiones e innovaciones bioinformáticas frecuentes, para seguir el ritmo de crecimiento en el número y tamaño de los datos, que va de la mano con el desarrollo de la metagenómica y las ciencias relacionadas como la metatranscriptómcia (Keegan *et al.,* 2007).

1.5 Bioinformática

Los estudios genéticos han experimentado una evolución constante en diversos grados donde la bioinformática posee un papel cada vez más importante, debido a que cada día se requiere analizar una mayor cantidad de datos y las raíces de la bioinformática son la matematización de la biología (Searls, 2010).

Dependiendo del autor, la bioinformática presenta varias definiciones: “Rama de la ciencia relacionada con el flujo de información en sistemas biológicos, especialmente el uso de métodos computacionales en genética y genómica" (Vincent y Charette, 2015) “Aplicación de herramientas computacionales para organizar, analizar, comprender, visualizar y almacenar información asociada con macromoléculas biológicas” (Diniz y Canduri, 2017).

Las definiciones anteriores presentan ciertas constantes como la relación de los sistemas computacionales con los biológicos, lo cual no está separado de la realidad ya que actualmente se utilizan análisis bioinformáticos para responder diversas preguntas biológicas.

Cabe destacar que la bioinformática para la resolución de una pregunta de investigación, no trabaja sola, requiere de diversas herramientas que cumplan los requisitos que se necesiten en el caso específico. Un ejemplo de una herramienta que suele ser esencial para el estudio de las secuencias génicas posteriormente a la secuenciación son los programas encargados del pre-procesamiento y ensamblado de las secuencias para su posterior análisis.

1.6 Pre-procesamiento de las secuencias metagenómicas

1.6.1 Análisis de calidad y dereplicación

El pre-procesamiento consiste en analizar la calidad de las secuencias provenientes de la secuenciación masiva, con la finalidad de verificar si cuentan con problemas o sesgos en los datos que puedan afectar el análisis biológico posterior.

Para ello se han desarrollado herramientas como FastQC (Andrews, 2010), que proporciona un informe de calidad de las secuencias en formato de texto y HTML.

Los módulos que el programa utiliza para realizar la evaluación de calidad de las secuencias se describen a continuación:

- **Estadísticas básicas:** nombre del archivo, tipo del archivo, valores de calidad, número total de secuencias procesadas, longitud de las secuencias (más corta y más larga) y el porcentaje de GC.
- **Calidad de secuencias por base:** muestra el rango de valores de calidad de cada base de las secuencias. Para cada posición se dibuja un diagrama de tipo caja de bigote, donde el eje Y representa la calidad phred (puntaje de calidad Phred), la cual es una medida de calidad originalmente desarrollada para la identificación de una nucleobase a partir de señales de fluorescencia generadas por un secuenciador de ADN automatizado. El puntaje Phred expresa la probabilidad de error en una secuencia como una fracción del tipo 1 error en N nucleótidos. Esta medida es logarítmica, por lo que tener una calidad de phred de 10 significa que la probabilidad de error es de 1 base errónea en 10, para Phred 20 la probabilidad es 1 en 100, y 30 será 1 en 1000.
- El color verde de fondo del gráfico divide el eje Y en buena calidad (Phred >30), naranja calidad razonable (20<Phred>30) y en rojo baja calidad (Phred<20).
- **Niveles de calidad por secuencia:** permite ver si un subconjunto de las secuencias tienen valores de calidad universalmente bajos.
- **Contenido de secuencia por base**: traza la proporción de cada posición de base.
- **Contenido de GC por base:** traza el contenido de GC de cada posición base .
- **Contenido de GC por secuencia:** mide el contenido de GC en toda la longitud de cada secuencia y lo compara con una distribución normal modelada del contenido de GC.
- **Contenido de N por base:** Si un secuenciador no puede “llamar” identificar una base con suficiente confianza, lo sustituirá con una N esté módulo traza el porcentaje de llamadas de base en cada posición para la cual fue N.
- **Distribución de longitud de secuencia:** Algunos secuenciadores de alto rendimiento generan fragmentos de secuencia de longitud uniforme, pero otros pueden contener lecturas de longitudes muy diferentes, este módulo genera un gráfico que muestra la distribución de tamaños de fragmentos.

- **Secuencias duplicadas**: cuenta el grado de duplicación para cada secuencia y creando un gráfico que muestra el número relativo de secuencias con diferentes grados de duplicación.
- **Secuencias sobrerrepresentadas**: si hay secuencias sobrerrepresentadas puede indicar contaminación, este módulo enumera todas las secuencias que representan más del 0.1% del total.
- **Kmers sobrerrepresentados:** detecta un aumento en cualquier duplicado de secuencias exacto.

El programa trabaja con diversos formatos de archivos, sin embargo el más utilizado es el formato FASTQ, conformado por cuatro líneas por secuencia:

La línea 1 inicia con el carácter '@' seguido por un identificador de secuencia

La línea 2 contiene la secuencia

Línea 3 inicia con el caracter '+' separa la secuencia de los valores de calidad

Línea 4 codifica los valores de calidad de la secuencia (línea 2), y contiene el mismo número de símbolos que hay de letras en la secuencia.

En el caso de los datos metagenómicos y metatranscriptómicos existe un paso adicional a los anteriores y es el de la dereplication, la cual se basa en la eliminación de lecturas duplicadas artificialmente (ADR), las cuales se generan durante la reacción de PCR en la secuenciación, estas duplicaciones pueden ser exactas o presentarse en los extremos 5' o 3', por lo cual se opta por eliminarlas (Schmieder *et al.,* 2011; Gómez-Álvarez *et al*., 2009) y como se mencionó previamente algunos servidores como MG-RAST lo llevan a cabo. En general, se busca que los programas de ensamblaje trabajen en un mínimo de tiempo y utilizando poco poder computacional. Aplicando los pasos anteriormente mencionados es posible, por un lado mejorar la calidad del ensamblado de las secuencias al eliminar secuencias de baja calidad, y por otro reducir el tiempo de procesamiento al eliminar secuencias redundantes.

1.6.2 Procesamiento de las secuencias metagenómicas

1.6.2.1 Ensamblado y calidad del ensamblado.

Se le denomina ensamblaje de secuencias al proceso de unir pequeños fragmentos de ADN (lecturas), que presumiblemente en el genoma del cual provienen estaban contiguos, para la formación de secuencias más largas (contigs), inclusive hasta la construcción del genoma completo del organismo de estudio.

Una de las principales dificultades en el ensamblaje de secuencias genómicas es desarrollar un algoritmo capaz de detectar secuencias de ADN repetitivas sin causar errores de ensamblaje (Scheibye-Alsing *et al.,* 2009).

Existen dos categorías para el ensamblaje de genomas: (i) ensamblaje por comparación, en el que se utiliza un genoma como referencia y (ii) el ensamblaje *de novo*, utilizando la información obtenida en la secuenciación, sin conocimiento previo de la organización del mismo. Según Aguilar-Bultet y Flaquet (2015) los programas ensambladores se agrupan en tres paradigmas principales de acuerdo con su método de ensamblaje:

•**Greedy:** el ensamblador conecta las lecturas que mejor se sobreponen de manera iterativa, siempre y cuando no contradigan el ensamblaje ya construido, sin embargo el método no es ampliamente empleado ya que no resuelve eficientemente las regiones largas repetidas en los genomas.

•**Consenso de diseño superpuesto (*Overlapping consensus*):** identifica todos los pares de lectura que se sobreponen bien, organizando esta información en un grafo, en el cual hay un nodo por cada uno de ellos y un conector por cada sobrepuesto entre los mismos, por lo que se definen caminos, que corresponden con los segmentos del genoma que están siendo ensamblados, reconstruyéndose el genoma mediante la búsqueda de un único camino que atraviese todos los nodos solo una vez.

•**Grafos De Bruijn (*De Brujin graph*):** similar al método anterior, trabajan bajo el uso de nodos y conectores, donde los primeros representan k-mers y los segundos indican qué k-mers adyacentes se sobreponen por k-1 letras, por lo que la longitud del k-mer correlaciona con la longitud sobrepuesta que el ensamblador es capaz de detectar. No se modelan

directamente las lecturas, sino que están implícitamente representadas por los conectores en el grafo De Bruijn, y se incorporan métodos de corrección de errores para mejorar la calidad del ensamblaje.

En este trabajo se emplea el paradigma de los gráficos De Bruijn para el ensamble de las secuencias metagenómicas.

Por otro lado, en cuanto a la verificación de la calidad del ensamblado, existen indicadores que permiten evaluarlo cuantitativamente y cualitativamente, siendo los principales el N50 y el % de mapeo.

Para el primero, se ordenan de mayor a menor tamaño los contigs construidos en el ensamblado, y el N50 corresponderá al número de pares de bases que conforme el contig que cubra la mitad del ensamblado; sin embargo en el caso de los metagenomas el N50 dependerá del tamaño inicial con el que se ensamble.

Para obtener el porcentaje de mapeo se realiza un alineamiento de las lecturas contenidas en el archivo original (sin ensamblar) con el archivo ensamblado, lo que permite analizar la consistencia del ensamblado para conocer cuántas de las lecturas originales están contenidas en los contigs (Aguilar-Bultet y Falquet, 2015)

1.6.2.2 Traducción de las secuencias ensambladas

En la minería de datos para la búsqueda de proteínas es necesario que posteriormente al ensamblado se traduzcan los contigs a aminoácidos ya que de esta manera se verifica si contienen proteínas completas y/o parciales.

La traducción se debe de realizar en los seis marcos de lectura, tres de sentido y tres de anti sentido, pues se pueden codificar proteínas diferentes, pues según el marco de lectura con el que se lea, los aminoácidos que las conforman pueden variar (Austin,2020)

De acuerdo con la pregunta de investigación es necesario traducir las secuencias según el código genético del grupo específico de los organismos que se estén estudiando debido a que entre grupos biológicos o la procedencia del ADN existen variaciones en la codificación de las proteínas. Tal es el caso del código genético bacterial que a diferencia del estándar posee variaciones en los codones de inicio (Nakamoto, 2009), por lo tanto al usar un código genético “inapropiado” puede propiciar disminuir la calidad de la posterior minería de datos

1.6.3 Minería de datos

La minería de datos es el proceso de descubrir estructuras y patrones de interés en grandes conjuntos de datos. Se apoya de otras disciplinas como la estadística, el aprendizaje automático y el reconocimiento de patrones, las cuales poseen sus propios enfoques. Actualmente a la minería de datos en el campo científico se le ha dado el término de "minería de grandes datos" (Big data minning) pues con las tecnologías de secuenciación masiva actuales es posible la generación de información biológica en gran medida, la cual es posible de analizar con ayuda de herramientas computacionales (Hand y Adams, 2014; Kanchi y Krishna, 2013).

En este trabajo se utilizaron expresiones regulares y los modelos ocultos de Markov para llevar a cabo la minería de los datos metagenómicos, con la finalidad de detectar secuencias con motivos proteasas S8A.

1.6.3.1 Expresiones regulares

Linux es un sistema operativo gratuito que deriva de UNIX, es de tiempo compartido, lo que quiere decir que es (i) interactivo multiusuario, permitiendo que múltiples usuarios utilicen simultáneamente el computador y (ii) multiproceso, lo que indica que puede ejecutar diferentes procesos a la vez, por lo que incrementa la utilización y el rendimiento de los recursos del sistema, reflejándose en el número de procesos que finalizan por unidad de tiempo (Sarwar *et al.,* 2003).

Entre las utilerías de Linux se encuentran las expresiones regulares, las cuales según señalan Goyvaerts y Levithan (2012), son un conjunto de reglas que especifican elementos que se desean encontrar, reemplazar y/o reorganizar en un cuerpo de texto más grande y existen muchas herramientas de Linux como los comandos grep, egrep entre otras que admiten expresiones regulares.

Un comando es un mensaje enviado por parte del usuario hacia el ordenador para provocar una respuesta, la cual dependerá de lo que se le solicite.

Por lo tanto, estas utilerías y herramientas de Linux son útiles en la bioinformática para la localización de secuencias biológicas de interés expresadas como textos simples; el diseño

de expresiones regulares para la búsqueda de motivos de proteínas de interés se basan en las secuencias de aminoácidos que se encuentren más conservados dentro de cada motivo, pudiendo ser consultadas en bases de datos, tales como la CDD (Conserved Domain Database) de la NCBI (National Center for Biotechnology Information) (Caamal-Pech *et al.,* 2018)

Las herramientas anteriores, brindan al usuario de manera eficiente, el manejo de información masiva para encontrar secuencias que posean genes que codifiquen nuevas proteínas con probable potencial biotecnológico, dentro de archivos con información masiva como los generados por la metagenómica.

1.6.3.2 Modelos Ocultos de Markov

El análisis computacional es cada vez más importante para inferir las funciones y estructuras de las proteínas ya que la velocidad de la secuenciación del ADN supera por mucho la velocidad a la que la función biológica de las secuencias se puede dilucidar experimentalmente.

La selección natural ha favorecido que algunos residuos de las secuencias de enzimas estén muy conservados evolutivamente y que otros sean muy variables, algunas posiciones están más conservadas que otras y algunas regiones parecen tolerar inserciones y eliminaciones más que otras regiones. Lo anterior es posible de dilucidar cuando se observan los patrones de conservación en las múltiples alineaciones de una familia de secuencias. En bioinformática una alineación de secuencias es la comparación de dos o más cadenas de ADN, ARN o estructuras primarias proteicas para visualizar las zonas que posean similitud y las que no, con el fin de dilucidar las relaciones funcionales y/o flogenéticas entre las secuencias alineadas. Por lo que es posible obtener información específica de una posición haciendo múltiples alineaciones de secuencias homólogas o sea similares y que se presumen con un mismo origen evolutivo, obtenidas en bases de datos que almacenan secuencias genómicas (Eddy, 1996).

Por lo cual se introdujeron los métodos de perfil para construir modelos a partir de alineaciones, que utilizan los Modelos Ocultos de Marcov (HMM por sus siglas en inglés). Los modelos Ocultos de Markov son una clase de modelos de probabilística generalmente aplicables a series de tiempo o secuencias lineales, los cuales se introdujeron en biología computacional a fines de la década de 1980.

Basándose en el principio de los HMM, se han creado perfiles HMM, los cuales contienen estados para coincidencia, inserción o eliminación, que son usados para modelar una familia de secuencias. Cada estado en el modelo tiene distribuciones de probabilidad y cada transición tiene una probabilidad. Entonces, si se tiene un aminoácido comúnmente representado en una posición particular en el alineamiento múltiple de secuencias, este obtiene un puntaje más alto y de igual manera se otorgan puntajes a los residuos insertados o eliminados; por lo tanto el puntaje resultante es la probabilidad de que la secuencia esté relacionada con el modelo dado y la probabilidad es usada para encontrar un valor (e-value) para la coincidencia (Eddy, 1996).

La ventaja de usar perfiles HMM para la búsqueda de motivos en bases de datos es que se compara una secuencia con un modelo estadístico que describe una familia o patrón de secuencias, al contrario de la comparación de aminoácidos individuales de dos secuencias. Basándose en esta información, es más fácil saber si una secuencia particular está relacionada con una familia o no.

Existen diferentes *softwares,* como por ejemplo el HMMER que utilizan este tipo de modelado y bibliotecas de HMM las cuales requieren un gran número de múltiples alineaciones de dominios de proteínas comunes. Actualmente hay dos grandes colecciones de perfiles de HMM anotados: la base de datos Pfam (Sonnhammer *et al*., 1998, 1997) y la base de datos de perfiles PROSITE (Bairoch *et al*., 1997). Cabe señalar que tanto el *software* de búsqueda como las bases de datos de perfiles están mejorando y cambiando rápidamente (Eddy, 1998).

HIPÓTESIS

La minería de datos, utilizando diferentes estrategias bioinformáticas, permitirá identificar secuencias de nuevas proteasas S8A en datos metagenómicos que provienen del acuífero de Yucatán.

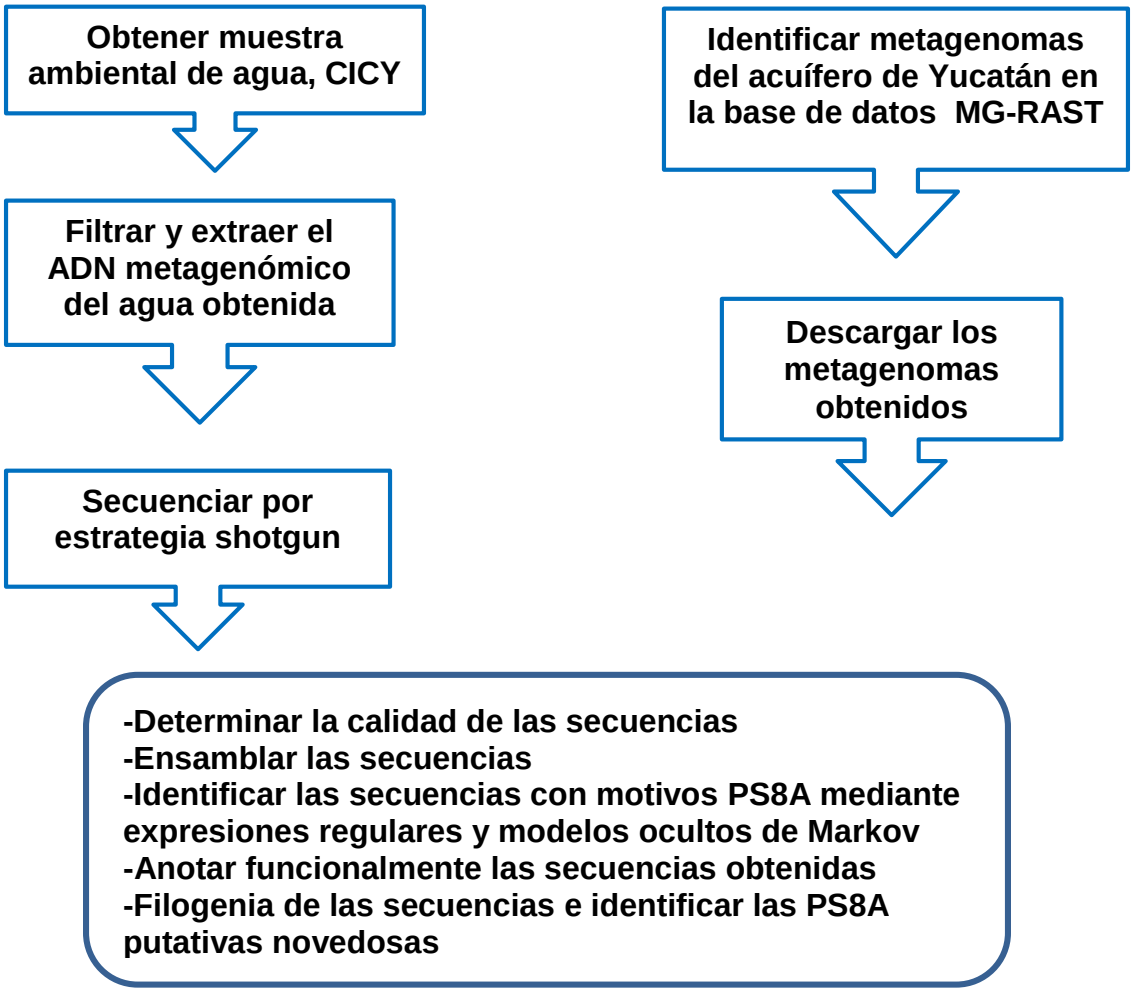
OBJETIVO GENERAL

Desarrollar una estrategia bioinformática que permita minar datos para identificar secuencias nuevas de proteasas S8A, en metagenomas del acuífero de Yucatán.

Objetivos específicos

1. Desarrollar una estrategia de análisis bioinformáticos que permita pre-procesar datos metagenómicos.
2. Obtener secuencias que contengan los tres motivos característicos de las proteasas S8A a través de dos estrategias de minería de datos.
3. Identificar nuevas secuencias de proteasas S8A putativas a través de un análisis filogenético.

ESTRATEGIA EXPERIMENTAL



CAPÍTULO II

MATERIALES Y MÉTODOS

El trabajo bioinformático se llevó a cabo utilizando el equipo de súper cómputo, “Hobón”, del Centro de Investigación Científica de Yucatán, A.C (CICY). “Hobón” utiliza el sistema operativo (SO) LINUX/Red-Hat (https://www.redhat.com/es) y tiene una arquitectura de 18 nodos de 128 GB de RAM y 2 nodos de 512 GB de RAM.

2.1 Descarga y metadatos de las secuencias metagenómicas utilizadas

Los metagenomas almacenados en el servidor MG-RAST (versión 4.0.3, Wiket *et al.,* 2017), bajo el proyecto “Yucatán groundwater”, dispuestos al público por Anni Moore, de la Universidad del Norte de Illinois, fueron descargados utilizando la siguiente línea de comandos:

```
$ curl  http://api.metagenomics.anl.gov/1/download/mgmxxxx.x?file=050.1 > salida.fastq
```

Siete metagenomas pertenecen al cenote Xcolac, el cual está localizado en la carretera Izamal-Cenotillo, del estado de Yucatán (20º 54’ 35” Norte y 88º 51’57” Oeste), y seis al cenote Calica localizado en la carretera Tulum-Playa del Carmen Km. 282 a un costado del Parque Xcaret, del estado de Quintana Roo (20º35'8.4"Norte 87º10'28" Oeste), dentro de las inmediaciones de la empresa Calica, dedicada a la producción de agregados pétreos y que tiene una cantera en esa zona (Figura 2.1).

Las muestras de donde se extrajeron los trece metagenomas fueron obtenidas de distintos estratos de la columna de agua de los puntos geográficos antes descritos. Como control de identificación, cada muestra se nombró con tres letras, siendo la primera X para Xcolac y C para Calica; seguida de una letra que identifica el estrato del que fue tomado, el cual fue obtenido de los metadatos de cada metagenoma: m (mesotrófico), a (anóxico), s (salobre) o d (sedimento) y por último el número 1 o 2, según el número de muestras que se tiene. Todos fueron secuenciados por estrategia de escopeta y con la plataforma Illumina, CASAVA 1.8.

Los metadatos de los metagenomas descargados tales como las condiciones ambientales de los puntos geográficos y la composición de las comunidades bacterianas presentes en dichos puntos, se encuentran documentados en el trabajo de Moore *et al.,* (2020). Se utilizó la información pertinente de dichos metadatos conforme se fue requiriendo para analizar los datos sobre las secuencias PS8A que se obtuvieran en este estudio.

2.2 Obtención de la muestra ambiental para la extracción de ADN metagenómico

La muestra de DNA se obtuvo, de un pozo localizado dentro de las instalaciones del Centro de Investigación Científica de Yucatán, A.C. (CICY), en las coordenadas 21º 01' 44" Norte 89º 38' 19" Oeste (Figura 2.1), a través del uso de una bomba instalada por el CICY y localizada dentro del pozo. Se utilizó un bidón de 20 L de capacidad para la colecta de la muestra previamente desinfectando con un lavado de etanol al 70%. El bidón utilizado para la colecta de la muestra se enjuagó dos veces con agua de la misma muestra.

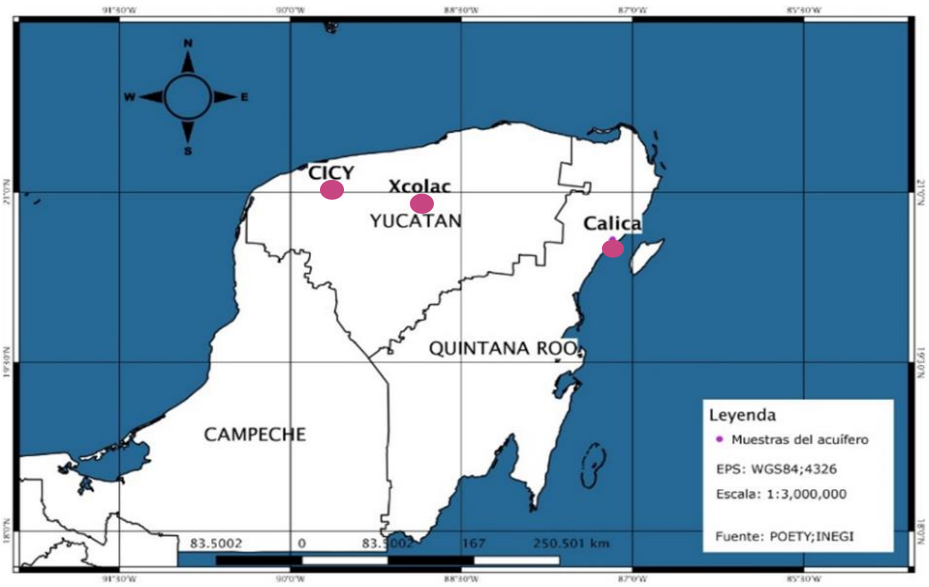


Figura 0.1 En círculos morados, se encuentra señalada la localización aproximada de los puntos geográficos de donde procedieron las secuencias metagenómicas utilizadas en este proyecto**.**

2.3 Filtrado del agua y extracción del ADN metagenómico

El agua obtenida se pasó serialmente primero por un filtro marca Millipore de 20 µm de diámetro de poro para eliminar pequeños trozos de vegetación y tierra, y posteriormente, para retener la masa procariota se utilizó un filtro marca Sterivex GV^{MR} de 0.22 µm de diámetro de poro. Se saturaron cinco filtros de 0.22 µm.

Para almacenar los filtros, se les retiró el agua contenida y con una pipeta se les añadió etanol al 70% hasta que se llenaran por completo. Se almacenaron en el congelador, hasta el día de la extracción de ADN.

Para la extracción cada filtro fue roto manualmente con ayuda de unas pinzas y se utilizó el kit "ZymoBIOMICS™ DNA Miniprep, No. de catálogo D4304". Los pasos de la extracción se presentan a continuación:

I. Con tijeras estériles, se cortó la membrana del filtro en pedazos de pequeño tamaño (aprox. 0.1 a 0.5 mm), para colocarlos dentro de un tubo de lisis, posteriormente se adicionaron 750 µl de solución de lisis para mezclar en el vortéx por 5 minutos. Se centrifugó a 10,000 g por un minuto, y se transfirieron 400 µl de sobrenadante al filtro III F en un tubo de colección, y se centrifugó a 8,000 x g por 1 minuto.
II. Se descartó el filtro III F y se añadieron 1,200 µl de Buffer de unión al filtrado del tubo del paso anterior, mezclándose bien.
III. Se transfirieron 800 µl de la mezcla anterior a la columna IIC en un tubo de colección y se centrifugó a 10,000 g por 1 minuto.
IV. Se descartó el sobrenadante y se repitió el paso anterior, para añadir 400 µl de Buffer de lavado 1 de ADN a la columna IIC en un tubo de colección y se centrifugó a 10,000 g por 1 minuto.
V. Se descartó el sobrenadante y se añadieron 700 µl de Buffer de lavado 2 a la columna IIC en un tubo de colección, se centrifugó a 10,000 g por 1 minuto y de nuevo se descartó el sobrenadante. Al finalizar, se agregaron 200 µl de Buffer de lavado 2 a la columna IIC en un tubo de colección para centrifugar a 10,000 g por un minuto.
VI. Se transfirió la columna IIC a un tubo para microcentrifuga de 1.5 ml y se añadieron 50 µl de agua libre de ADNasa y ARNasa, se incubó por un minuto y se centrifugó a 10,000 g por un minuto para eluir el ADN.

VII. Se colocó el filtro HRC en un nuevo tubo de colección, se agregaron 600 µl de la solución preparada del kit, se centrifugó a 8,000 g por 3 minutos.

VIII. Para finalizar, se transfirió el ADN eludido, a un micro tubo nuevo de 1.5 ml para centrifugar exactamente a 16,000 g por 3 minutos.

2.4 Validación de la integridad, concentración y pureza del ADN metagenómico

Para validar y visualizar la integridad del ADN metagenómico, se realizó una electroforesis en gel de agarosa al 1%, y se utilizó la escalera O'gene ruler 1 kb plus de Thermo Scientific™ como marcador de peso molecular y 2 µl por cada muestra (5 muestras en total).

Para la determinación de la concentración, pureza y calidad del ADN metagenómico se realizó un análisis con NanoDrop 1000 Spectrophotometer. Utilizando la razón de absorbancia de 260/280 nm para determinar que el ADN no estuviera contaminado con proteínas, la absorbancia de 260 corresponde a ácidos nucleicos y la de 280 a proteínas y/o fenoles.

De igual manera se utilizó la razón 260/230 nm para determinar la presencia y/o ausencia de contaminantes como carbohidratos, la absorbancia de 260 a ácidos nucleicos y 230 a carbohidratos y/o sales, por lo que un resultado obtenido de 1.8 a 2.2 indicará que el ADN se encuentra libre de este tipo de contaminantes.

2.5 Secuenciación del ADN metagenómico de la muestra de agua del pozo del CICY

La muestra de ADN metagenómico extraído se envió a secuenciar a la compañía de secuenciación BGI (Instituto genómico de Pekín), donde se utilizó la estrategia de secuenciación de escopeta, la plataforma de Ilumina HiSeq X Ten System en paired end y secuencias de 150 pb.

2.6 Evaluación de la calidad de las secuencias

Para la evaluación de la calidad de las secuencias se utilizó el programa FastQC (versión 0.11.5; Andrews, 2010), el cual proporciona verificaciones de control de calidad de los datos

de secuencia provenientes de los procesos de secuenciación, mediante la generación de resumen de gráficos y tablas para una evaluación rápida. El reporte se genera en formato HTML, que son posibles de visualizar con cualquier navegador de internet. Se tomaron como secuencias de buena calidad aquellas ubicadas en la zona verde que representan un rango de valores que va de 30 a 40 de puntaje Phred, del reporte del módulo de “calidad de secuencias por base” y por lo tanto que se encuentren dentro de un phred score mayor a 30 es decir una base errónea de cada 1000.

Para ejecutar el programa se utilizó la siguiente línea de comandos:

```
$ fastqc  -f <indica que el formato del archivo es fastq> “dirección del archivo que contiene las lecturas” -nombre del archivo final
```

Los archivos en los que se observaron distribuciones anormales fueron modificados utilizando los comandos head (para visualizar la parte superior del archivo), grep (para encontrar patrones dentro del archivo) tail (para visualizar la parte inferior del archivo) y sed (para editar el archivo), la metodología resultante se ubica en el apartado de resultados.

2.7 Ensamblado de las secuencias metagenómicas

Para el ensamblado de las secuencias, se utilizó el programa Megahit versión 1.1.4 (Li *et al.,* 2015), utilizando un tamaño mínimo de contig de 500 pb, pues se llevó a cabo una revisión de las Proteasas S8A depositadas en la base de datos de la NCBI, y todas cuentan con un tamaño que va de 600 a 1,600 nucleótidos.

La línea de comandos utilizada se muestra a continuación:

```
$ megahit -1 <lecturas en sentido>, -2 <lecturas antisentido> - r <lecturas sin par> -o <nombre de archivo de salida>
                                        $ megahit -1 -2 –r –o
```

Los parámetros anteriores se utilizaron de acuerdo con el tipo de lecturas metagenómicas que contenía cada archivo.

Para medir la calidad de cada metagenoma ensamblado se alinearon las secuencias sin ensamblar a las secuencias ensambladas, para obtener el porcentaje de mapeo, utilizando el programa BBMAP versión 35.34 (DOE Joint Genome Institute), mediante el

uso de la siguiente línea de comandos:

```
$ bbmap/bbwrap ref= <dirección del archivo con las secuencias ensambladas> in= <dirección del archivo con las lecturas sin ensamblar> out= <nombre del archivo de salida> kfilter= <longitud mínima de coincidencias exactas consecutivas para una alineación> subfilter= <Limita el número de desajustes para una alineación> maxindel= <limita el tamaño de la inserción o deleción>
$ bbmap/bbwrap ref= in= out= kfilter= subfilter= maxindel=
```

2.8 Traducción de las secuencias

Para traducir las secuencias metagenómicas de nucleótidos a aminoácidos se utilizó la herramienta transeq de EMBOSS Toolkit (v6.6.0.0), la cual se aplicó en los seis marcos de lectura (+1, +2, +3, -1, -2, -3) y se empleó el código genético bacteriano, debido a que se está analizando material genético proveniente de procariotas.

La siguiente línea de comando fue la utilizada:

```
$ transeq –sequence <archivo con las secuencias en nucleótidos> –outseq <nombre del archivo de secuencias en aminoácidos> –frame <número de marcos de lectura> –table 11 <código genético de las bacterias> –nomethionine <no convierte los primeros codones en metionina>
$ transeq –sequence –outseq –frame –table –nomethionine
```

Una vez ensambladas y traducidas de nucleótidos a aminoácidos las secuencias, se procedió a la minería de datos metagenómicos con la finalidad de identificar secuencias de proteasas S8 (PS8).

2.9 Minería de datos utilizando expresiones regulares

2.9.1 Obtención de las expresiones regulares.

Las expresiones regulares utilizadas en este estudio fueron tomadas del trabajo de Góngora-Castillo *et al,* (2020). Dichas expresiones fueron diseñadas de acuerdo con visualización de múltiples alineaciones (Figura 2.2) de las secuencias de PS8 más diversas y depositadas en la base de datos de NCBI (https: //www.ncbi.nlm.nih.gov/) y de secuencias PS8 completamente curadas de la base de datos MEROS.

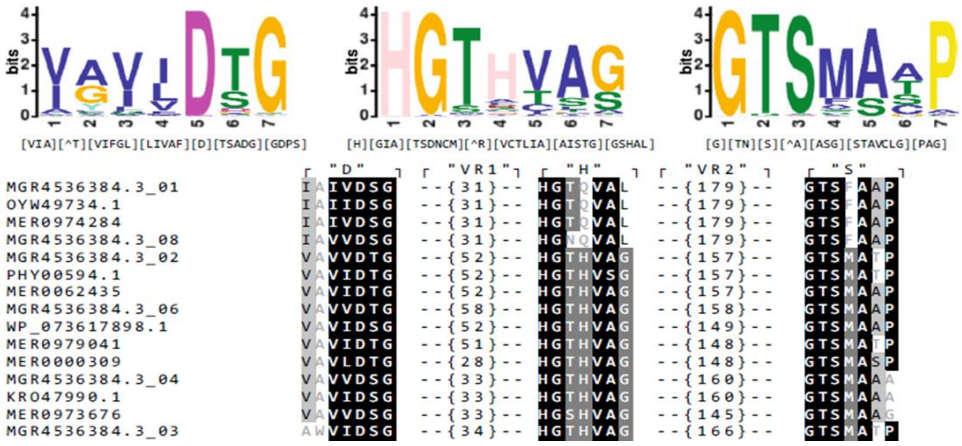


Figura 0.2 Construcción de las Expresiones Regulares**.** En la parte superior se aprecian los logos de secuencia que muestran los motivos conservados "D", "H" y "S" y la variabilidad de los aminoácidos entre los motivos (generados por el programa MEME suite en línea a partir de la colección de proteasas S8A de MEROPS). Debajo de cada logo hay una expresión regular diseñada a partir de lo observado en los logos y de las alineaciones múltiples (parte inferior) de las secuencias pertenecientes a PS8A. Tomado de Góngora *et al*., (2020).

Las expresiones fueron realizadas para cada uno de los tres motivos conservados, y se presentan a continuación:

Motivo D: [VIA][T][VIFGL][LIVAF][**D**][TSADG][GDPS][A-Z]

Motivo H: [**H**][GIA][TSDNCM][^R][VCTLIA][AISTG][GSHAL][A-Z]

Motivo S: [G][TN][**S**][^A][ASG][STAVCLG][PAG]

En donde los [ ] representan una posición dentro de la secuencia y las letras en el interior de los [ ] representan los posibles aminoácidos que se encuentran en esa posición en específico. Por lo tanto a través del implemento de las expresiones regulares se busca obtener solamente las secuencias que cumplan estrictamente con ellas.

2.9.2 Validación de las expresiones regulares

Previamente al uso de las expresiones regulares en este trabajo, se verificó su eficacia, para lo cual se descargó de la base de datos MEROPS, la librería S08a.lib que contiene secuencias PS8A, con la finalidad de analizar cuánto porcentaje de secuencias PS8A de esa librería, era recuperado por la expresión regular.

Para ello se utilizó la línea de comando descrita del apartado 2.9.3 al 2.9.5. El número resultante de las secuencias PS8A obtenidas se utilizó en una regla de tres, para calcular el porcentaje de PS8 que la expresión podía recuperar.

2.9.3 Modificación del formato de los archivos metagenómicos

Se utilizó la siguiente línea de comandos, para modificar el formato de las secuencias y organizarlas de manera linear, con la finalidad de que cada secuencia esté contenida en una sola línea, de esta manera la expresión regular se aplica a cada una.

```
$ sed e 's/\(^>.*$\)/#\1#/' metagenome_aa_file.faa | tr d "\r" | tr d "\n" | sed  e  's/$/#/' | tr "#" "\n" | sed e '/^$/d' > lineal_metagenome_aa_file.faa
```

2.9.4 Identificación de los contigs que cumplen con la expresión regular

Las expresiones regulares utilizadas para identificar las secuencias que contienen los tres motivos PS8 conservados fueron tomadas de Góngora-Castillo *et al., (*2020) y son las siguientes:

```
$ grep –B 1 <imprime en este caso una línea ubicada antes de la que contiene el motivo> --no-group-separator <arrojará la búsqueda sin separadores intermediarios> <expresión regular de motivos S8> lineal.fasta > file_noB.fasta
    $ grep –B 1 --no-group-separator '[VIA][^T][VIFGL][LIVAF][D][TSADG][GDPS]
                [A-Z]*[H][GIA][TSDNCM][^R][VCTLIA][AISTG][GSHAL]
    [A-Z]*[G][TN][S][^A][ASG][STAVCLG][PAG]' archivo.fasta > archivo_salida.fasta
```

En donde el carácter ^ seguido de un aminoácido indica que en esa posición puede estar cualquier aminoácido excepto el que se indica, mientras que la combinación [A-Z] * indica que hay variación de aminoácidos y que podría ser cualquiera (o incluso ninguno) repetido tantas veces sea necesario.

2.9.5 Obtención de las triadas catalíticas (región *core*) de PS8A a partir de los contigs que cumplen con la expresión regular

Por último, se extrajo el segmento que contiene los tres motivos conservados (región *core* PS8A) de los contigs que cumplieron con la expresión regular utilizada, con la finalidad de eliminar las regiones hipervariables ubicadas fuera de la triada catalítica de cada secuencia, dicha región *core* es la que fue utilizada para el análisis filogenético.

La línea de comandos empleada fue la siguiente:

```
$ sed –e <especifica el uso de varios comandos> 's/[A-Z\*]*\ <sustituye cualquier aminoácido nombrado de la A-Z en esta posición> <expresión regular de la triada de las S8> [A-Z\*]*/\1/' <sustituye cualquier aminoácido nombrado de la A-Z en esta posición>  <archivo del que se quiere ejecutar la acción> <archivo de salida>

$ sed –e 's/[A-Z\*]*\ ([VIA][^T][VIFGL][LIVAF][D][TSADG][GDPS]
[A-Z]*[H][GIA][TSDNCM][^R][VCTLIA][AISTG][GSHAL]
[A-Z]*[G][TN][S][^A][ASG][STAVCLG][PAG]\) [A-Z\*]*/\1/' archivo.fasta > archivo_salida
```

2.10 Minería de datos utilizando los Modelos Ocultos de Markov

Los modelos Ocultos de Markov son una clase de modelos de probabilística aplicables a secuencias lineales. A partir de un archivo de secuencias de aminoácidos se otorgan valores probabilísticos de coincidencias, inserciones o eliminaciones (estados) para cada residuo que conforme cada secuencia. Entonces, si se tiene un aminoácido comúnmente representado en una posición particular en el alineamiento múltiple de secuencias, este obtiene un puntaje más alto y de igual manera se otorgan puntajes a los residuos insertados o eliminados; por lo tanto el puntaje resultante es la probabilidad de que la secuencia esté relacionada con el modelo dado y la probabilidad es usada para encontrar un valor (e-value) para la coincidencia (Eddy, 1996).

Para el análisis con los modelos ocultos de Markov se utilizó el software HMMER, versión 3.3, (Finn *et al.,* 2011) el cual fue desarrollado para analizar cualquier conjunto de datos a través del uso de los modelos ocultos de Markov, Otorgando al final de cada análisis un reporte donde se puede apreciar detalladamente todos los estados que contiene cada secuencia.

2.10.1 Preparación de las secuencias metagenómicas para la implementación de los modelos Ocultos de Markov.

Para utilizar el programa HMMER es necesario contar con una base de datos respaldada para alinear las secuencias de interés con dicha base de datos.

La base de datos de Pfam (https://pfam.xfam.org/) 32.0 (El-Gebali *et al.*, 2018) es una gran colección de familias de proteínas, cada una representada por múltiples alineamientos de secuencia y modelos ocultos de Markov.

Se realizó la descarga desde el servidor de la base de datos, las secuencias correspondientes a la familia de las proteasas S8A nombradas con el identificador: Peptidase_S8 (PF00082).

2.10.2 Identificación de las secuencias con motivos PS8A implementando los modelos ocultos de Markov

Se utilizó la librería Hmmsearch de HMMER, la cual realiza un alineamiento de una sola familia de proteínas contra un conjunto de secuencias tomando como base el principio de los modelos ocultos, asignando puntajes a cada residuo de acuerdo a su estado.

El alineamiento se llevó a cabo utilizando las secuencias *core* PS8A de cada metagenoma ensamblado y la base de datos de las PS8A (PF00082) descargada, de manera que en el resultado final se obtendrán las secuencias ubicadas dentro del archivo de los metagenomas que cumplan con los estados que corresponden solamente a la familia de las PS8A.

La línea de comando utilizada fue la siguiente:

```
$ hmmsearch –o <nombre del archivo de salida> <base de datos PS8A> <archivo que contiene las secuencias a identificar>
```

2.10.3 Extracción de las secuencias identificadas con motivos PS8A mediante los modelos ocultos de Markov

HMMER reporta las secuencias que contengan al menos un motivo de la triada catalítica, pero para este trabajo se tomaron en cuenta solamente aquellas que contenían los tres motivos.

Por lo tanto las que cumplieron con dicha característica se aislaron en un nuevo archivo a través de la siguiente línea de comandos:

```
$ grep –w –A1 <“identificador de la secuencia”> <archivo de donde se quieren extraer las secuencias > <nombre de archivo de salida>
        $ grep –w –A1 “identificador de la secuencia” archivo.fasta > archivo_salida
```

Posteriormente, para unir en un solo archivo las secuencias extraídas se utilizó el comando cat para concatenar los archivos resultantes del paso anterior.

Las instrucciones para el uso de ese comando son las siguientes:

```
$ cat  <une archivos de texto en uno solo> <nombre del archivo a unir> <nombre del archivo a unir> <nombre de archivo de salida>
                $ cat  nombre_archivo nombre_archivo > archivo_salida
```

2.10.4 Modificación de la expresión regular para obtener la triada catalítica de las secuencias identificadas con los modelos ocultos de Markov

Se realizó la modificación de la expresión regular utilizada en el apartado 2.9.3, mediante la adición de aminoácidos en las zonas variables y adjuntas a los motivos conservados, de acuerdo con lo observado en el alineamiento de esas secuencias, con la finalidad de obtener solamente la región *core* de las secuencias identificadas como PS8A mediante los modelos ocultos de Markov ya que a diferencia de las expresiones regulares, HMMER identifica dentro del archivo de secuencias aquellas que sean PS8A, sin extraerlas en un archivo nuevo.

La expresión regular utilizada fue la siguiente:

```
‘[VILFMPA][^T][VIFDGNMHATLQS][LIVACPMNFGQWHYST][D][TSFEVKNADQG][GQADIPS][A-Z]*[H][GIA][TSDAVNILCM]
[A-Z][VCTLSQMIA][AISTGMCVLNY][GSNCQHATL][A-Z]*[GAV][TSAMN][S]
[A-Z][MASFG][STAVCLG][PNAG]*’
```

2.10.5 Confirmación de la identidad de todas las secuencias *core* PS8A obtenidas

Mediante el uso de más de un perfil de proteínas, se corroboró que las secuencias *core* PS8A obtenidas fueran identificadas de nuevo como PS8A. Se entenderá por perfil a un conjunto de secuencias alineadas de una familia específica de proteínas utilizadas para construir un perfil HMM.

Desde el servidor, se realizó la descarga de las 17,929 familias de proteínas contenidas en la base de datos de Pfam versión 32.0, para realizar un análisis con la librería Hmmscan, la cual ejecuta un alineamiento mediante el uso de toda la base de datos con el archivo que contiene las secuencias *core* identificadas como PS8A.

Según el manual del *software* es necesario cambiar el formato de la base de datos descargada a binario, por lo tanto se utilizó la herramienta hmmconvert mediante la siguiente línea de comando:

```
$ hmmerconvert < nombre del archivo a convertir > -o < nombre del archivo final >
$ hmmerconvert archivo_fasta -o
```

Una vez convertido se procedió al alineamiento antes mencionado, utilizando la siguiente línea de comando:

```
$ hmmscan –o <nombre del archivo final> <base de datos pfam en binario> <archivo que contiene las secuencias a identificar>
```

2.11 Anotación funcional de las PS8A

A las secuencias que no fueron identificadas mediante hmmscan como PS8A se les realizó un blastp desde el servidor de la NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) para verificar de manera adicional su identidad.

2.12 Anotación funcional de las secuencias *core* identificadas y verificadas como PS8A

Para el análisis funcional de las secuencias *core* PS8A se utilizó BLASTp (Madden., 2013). Las secuencias obtenidas se compararon con secuencias públicas que se encuentran en la base de datos de MEROPS the Peptidase Database (https://www.ebi.ac.uk/merops/index.shtml). Se utilizó la librería s08a.lib, que corresponde a la familia de las subtilisinas. Utilizándose la siguiente línea de comandos para dar formato “blast” a la librería descargada de MEROPS.

Se utilizó la opción “blastp” la cual es propia para el análisis de proteínas, por lo que se alinearon las triadas PS8A obtenidas con las proteasas de la base de datos MEROPS, utilizando la siguiente línea de comandos:

$ makeblastdb –in <dirección de la librería descargada> –input_type <tipo de archivo en este caso fasta> –dbtype <tipo de base de datos, en este caso proteínas> -out <nombre de la base de datos de salida>

$ makeblastdb –in –input_type –dbtype -out

Posteriormente, se realizó la anotación funcional de las triadas PS8A mediante el uso del programa BLASTp con las siguientes instrucciones:

$ blastp –db <dirección de la base de datos .phr, pin y .psq> –query <dirección del archivo que contiene los contigs pertenecientes a la triada de las proteasas S8> -outfmt '6 <formato tabular> qseqid <ID> stitle <nombre de la proteasa identificada> qcovs <covertura> nident <número de coincidencias idénticas> pident <porcentaje de coincidencias idénticas> qstart <inicio de la secuencia que coincide con el alineamiento> sstart <sitio donde inicia el alineamiento con la secuencia de la base de datos> send <sitio donde finaliza> evalue <valor evalue> bitscore <puntaje de bits>'– max_target_seqs <número de alineamientos reportados>

$ blastp –db –query –outfmt qseqid stitle qcovs nident pident qstart sstart send evalue bitscore max_target_seqs > archivo_salida

2.12.1 Normalización de los datos metagenómicos

Debido a que para cada punto geográfico (exceptuando al punto CICY), se obtuvieron varios metagenomas. Para la normalización de los datos, fue necesario unir las secuencias, por lo que se crearon tres archivos con secuencias metagenómicas correspondientes a cada punto.

Para evaluar la proporción de las triadas PS8A en los tres puntos geográficos estudiados se utilizó la siguiente fórmula

$$\frac{No.\ total\ de\ pb\ de\ las\ triadas\ \ PS8}{No.\ total\ de\ pb\ de\ los\ contigs}\ X\ 100{,}000$$

La cual se aplicó para cada punto geográfico, mediante la sumatoria del número total de pares de bases contenidas en cada secuencia *core* PS8A de cada punto geográfico, dicho resultado se dividió entre el número total de pares de bases contenidas en los contigs ensamblados de cada punto geográfico y el resultado se multiplicó por 100,000 para el manejo de números enteros.

Por otro lado para la normalización de los datos metagenómicos para evaluar la distribución de las secuencias por cada punto geográfico y phylum fue necesario la anotación funcional de todos los metagenomas ensamblados (14), mediante el uso del servidor RAST de metagenómica (MG-RAST) (https://www.mg-rast.org/) por lo que para el análisis se elaboraron tres archivos que contenían: (i) 7 metagenomas pertenecientes al punto Xcolac, (ii) 6 metagenomas de Calica y (iii) 1 metagenoma de CICY, para posteriormente ser dispuestos en dicho servidor.

Posteriormente a la obtención de los resultados se aplicó la siguiente fórmula:

$$\frac{No.\,de\ contigs\ de\ x\ phylum}{No.\,de\ contigs\ de\ x\ phylum\ que\ contiene\ triada\ PS8}$$

Lo anterior se aplicó para cada phylum que hubiera contado con una o más secuencias *core* PS8A por cada archivo.

Posteriormente se sumaron las proporciones resultantes de cada phylum en cada archivo y mediante una regla de tres se calculó el porcentaje de cada phylum y dicho resultado fue el graficado.

2.12.2 Selección y alineamiento de las secuencias *core* PS8A para el análisis filogenético.

Para analizar la diversidad de las secuencias *core* PS8A obtenidas y conocer su relación con PS8 conocidas, se llevó a cabo una inferencia filogenética.

Para lo cual se utilizaron las 245 secuencias *core* PS8A obtenidas y 13 proteasas conocidas como control, las cuales se describen en el Cuadro 2.1

Cuadro 0.1 Proteasas conocidas y utilizadas en el análisis filogenético

Base de datos	Identificador	Organismo de procedencia	Proteína
MEROPS	AT1G01900	*Arabidopsis thaliana*	PS8A
	LOC_Os07g48650	*Oriza sativa*	PS8A
	MER0000364	*Saccharomyces cerevisiae*	PS8B
	MER0000375	*Homo sapiens*	S8B
	MER0002053	*Gallus gallus*	S8B
	MER0000885	*Homo sapiens*	Pepsina (Ácido aspártico proteasa)
	MER0011652	*Deinococcus radiodurans*	M8 (Metaloproteasa)
	MER0000309	Carlsberg	PS8A
	MER0972600	Euryarchaeota	PS8A
	MER0123791	Actinobacteria	PS8A
	MER0077555	Firmicutes	PS8A
	MER0143948	Protobacteria	PS8A
	MER0984325	Cyanobacteria	PS8A

Siguiendo la metodología de Góngora *et al*., (2020) para la inferencia filogenética, se utilizaron secuencias proteasas S8B, la proteasa Carlsberg y cinco de sus homólogos. Los homólogos se obtuvieron a través de un alineamiento de secuencias. Se utilizó como *query* (secuencia a comparar) la Carlsberg, y la librería S08a.lib, mediante BLASTp y se seleccionaron a las primeras cinco PS8 que fueran de diferente phylum.

La proteasa Carlsberg y sus homólogos fueron utilizados como referentes ya que esta es la primera proteasas S8A descubierta y se encuentra bien caracterizada (Shalwan y Sharma., 2019).

Y para agregar mayor variabilidad al análisis fueron utilizadas secuencias de proteasas S8A provenientes de (i) plantas (Beers *et al*., 2004); (ii) secuencias S8B, debido a que pertenecen a la misma familia que las S8A, sin embargo se ubican en un grupo aparte porque son exclusivas de hongos y a pesar de que sus motivos conservados son iguales a las S8A, presentan variaciones particulares en los aminoácidos ubicados entre los motivos, los cuales difieren de las S8A; y por último (iii) 2 proteasas pertenecientes a otras familias, la familia ácido aspártico proteasa y metaloproteasa. La primera comparte el motivo conservado (D) de las S8A, mientras que las metaloproteasas son completamente diferentes ya que tienen tres motivos conservados pero estos pueden variar entre cinco aminoácidos.

Posteriormente, se creó un archivo con las secuencias *core* de las 245 PS8A identificadas en este trabajo y las 13 PS8 conocidas, para la elaboración de un alineamiento múltiple de esas secuencias, con la finalidad de alinear los tres motivos conservados (DHS) de las 258 triadas.

Para lo cual se utilizó el programa MAFFT (https://mafft.cbrc.jp/aligment/server/) (Katoh y Standley 2013) en su versión en línea. Utilizando los parámetros por *default* con la opción “E-INS-i” como método iteractivo, el cual es recomendado si se poseen secuencias con motivos conservados, regiones variables y vacíos entre cada motivo.

El archivo se descargó en formato PHYLIP versión 4.0, mediante el convertidor Readseq (https://mafft.cbrc.jp/alignment/server/cgi-bin/readseq.txt) (Gilbert, 2010).

Las secuencias que el programa no pudo alinear, se alinearon manualmente, para lo cual se descargó y utilizó el *software* Jalview ver. 2.11. (https://www.jalview.org/) (Waterhouse *et al*., 2009) recomendado por MAFFT.

2.12.3 Análisis filogenético.

Para realizar la construcción del árbol filogenético se utilizó la metodología descrita en Góngora-Castillo *et al*.,(2020) para lo cual se utilizó el servidor en línea del programa PhyML (http//www.atgc.montpellier.fr/phyml/) (Guindon *et al.* 2010). Para predecir el mejor modelo de sustitución para la construcción de la filogenia se utilizó la herramienta *Smart Model,* incluida en PhyML, eligiendo el criterio de información Akaike. Para la inferencia filogenética se utilizaron los siguientes criterios: (i) algoritmo BIONJ el cual a partir del alineamiento de las secuencias, toma en cuenta sus características biológicas es decir las distancias evolutivas de estas para poder hacer la inferencia de sus relaciones. (ii) reordenamiento de árboles SPR por sus siglas en inglés: *Subtree Pruning and Regrafting*, el cual a partir de un árbol principal, se corta un subárbol, creándose un nuevo nodo y por último la opción aLRT SH-like para el soporte de las ramas.

CAPITULO III

RESULTADOS

3.1 Extracción y calidad del ADN metagenómico.

A partir de la extracción de ADN metagenómico de la muestra del punto CICY, se obtuvo aprox. 50 µl de ADN por cada uno de los 5 filtros. Se procedió a visualizar su integridad en una electroforesis en gel de agarosa al 1%, utilizándose 2 µl por cada muestra (Figura 3.1 y 3.2) y de acuerdo con los cálculos utilizando la escalera de peso molecular, se obtuvo una concentración de ADN de aproximadamente 40 ng x µl, lo cual implica que en cada muestra se contó con 240 µl equivalentes a 10 µg.

M 1 2 3 4

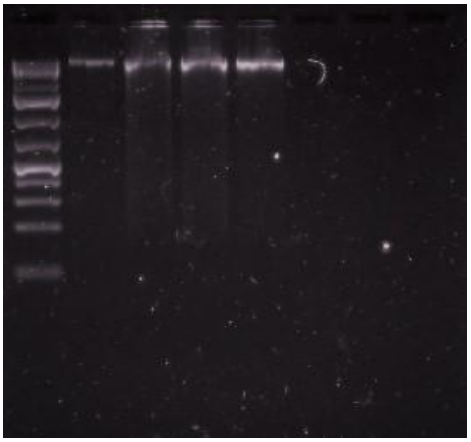


Figura 0.1 Visualización del ADN metagenómico. De izquierda a derecha: M= marcador, 5 número de filtro

M 5

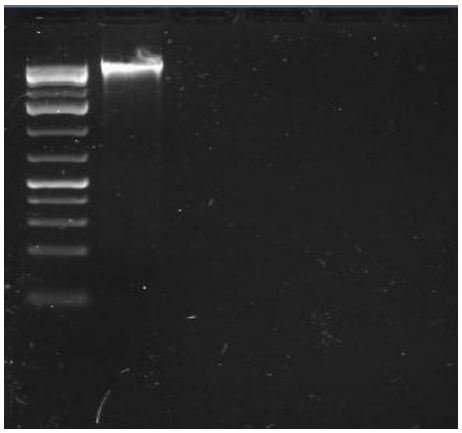


Figura 0.2 Visualización del ADN metagenómico. De izquierda a derecha: M= marcador, 1-4 número de filtro

Los resultados del NanoDrop 1000 Spectrophotometer arrojaron que el ADN total extraído contenía 112 ng/UL, debido a que el ADN metagenómico resultante por cada muestra se juntó en uno solo. La relación 260/280 fue de 1.89, indicando que no hay presencia de contaminación por proteínas, mientras que la relación 260/230 fue de 1.093 indicando la

presencia de carbohidratos en la muestra de DNA. Siendo de 16 millones el número total de lecturas obtenidas.

3.2 Metagenomas descargados

Se descargaron 13 metagenomas de la base de datos de MG-RAST en formato fastq y de la extracción del ADN metagenómico de la muestra CICY se obtuvo un metagenoma adicional. O sea que en este trabajo se analizaron 14 metagenomas en total.

Los metagenomas descargados presentaron un valor de calidad de Phred arriba de 30 es decir 1 base errónea por cada 1000, mientras que el metagenoma CICY obtuvo 40, es decir 1 base errónea por cada 10,000 (Cuadro 3.1).

Cuadro 0.1 Información de los metagenomas obtenidos

Punto geográfico	ID MGRAST	Número de lecturas (millones)	Tamaño de las lecturas (pb)	Puntaje de calidad Phred
Xcolac	mgm4536389.3	4	102-194	30
Xcolac	mgm4536390.3	9	101-194	30
Xcolac	mgm4536384.3	11	101-194	30
Xcolac	mgm4536385.3	18	101-194	30
Xcolac	mgm4536387.3	21	101-194	30
Xcolac	mgm4536388.3	25	101-194	30
Xcolac	mgm4536386.3	27	101-194	30
Calica	mgm4536373.3	2	102-194	30
Calica	mgm4536377.3	2	102-194	30
Calica	mgm4536378.3	3	102-194	30
Calica	mgm4536375.3	8	101-194	30
Calica	mgm4536376.3	13	101-194	30
Calica	mgm4536374.3	43	101-194	30
CICY		16	150	40

3.3 Calidad de las secuencias

De acuerdo con el reporte del programa FastQC para visualizar la calidad de las secuencias metagenómicas, el eje Y representa la calidad Phred, las secuencias se encuentran representadas con las cajas de bigote de color amarillo. El eje X correspondiente a las pares de bases y se encontraron distribuidas dentro de la sección verde, que representa una calidad Phred entre 30 y 40, por lo tanto todos los metagenomas contenían secuencias de buena calidad.

En el reporte de FastQC se observó que los datos de los 13 metagenomas obtenidos de la base de datos de MG-RAST, presentaban una curva bimodal de distribución, en lugar de una de distribución normal, (Figura 3.3)

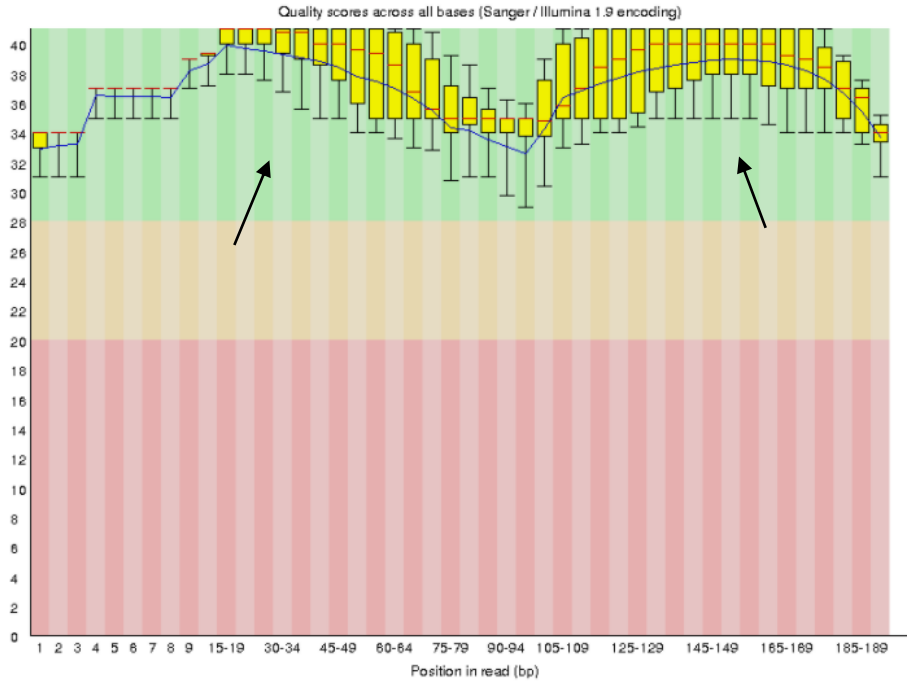


Figura 0.1 Reporte del módulo de calidad de secuencias por base del programa FastQC perteneciente al metagenoma Ca1. Las flechas señalan las crestas de la curva bimodal que se obtuvo.

Por lo que observando lo anterior se procedió a analizar cada archivo fastq (los 13 descargados de MG.Rast y el del CICY), encontrando que de los 14 archivos, nueve contenían tres conjuntos de secuencias: (i) “singles” o únicas, (ii) lectura 1 o secuencia directa y (iii) lectura 2 o secuencia indirecta; tres archivos contenían secuencias singles o

únicas, mientras que el archivo de la muestra CICY contó con secuencias correspondientes a la lectura 1 y 2. Para separar el conjunto de secuencias mezcladas de los 9 primeros archivos se desarrolló la siguiente metodología:

Para la determinación de cada conjunto de secuencias se identificaron los caracteres que eran únicos para cada uno.

Las secuencias singles poseían los caracteres: 1:N:0

Las de la lectura 1: .1

Las de la lectura 2: .2

Y se procedió a crear intervalos con los números de línea correspondientes a cada conjunto y dividir en archivos de la siguiente manera:

1 .Enumerar las líneas
$ nl <enumera líneas> file.fastq > file.fastqn

2. Identificar el número de línea con el último 1:N:0
$ grep <busca un modelo que satisfaga la expresión regular> '1:N:0' file
+3 será la última línea perteneciente a los singles
+4 será la primera del r1

3. Identificar el número de línea con el último .1
$ grep '\.1$' file
+3 será la última perteneciente a los r1
+4 será la primera del r 2
4. Identificar la última línea perteneciente al r2
$ tail <visualiza las últimas diez líneas del archivo> file

Por último se dividió cada conjunto de secuencias:
5. Obtener read 1
a) Elimino líneas de r2
$ sed '$,$d' <elimina el intervalo de líneas que se desee> file.fastq > filesinr2.fastq
b) Elimino líneas de singles
$ sed '$, $d' filesinr2.fastq > filer1.fastq

6. Obtener read 2
$ tail –n <visualiza el número de líneas que se indique, del final del archivo> $ file.fastq > filer2.fastq
3. Obtener singles
$ head -$ <visualiza el número de líneas que se indique, del principio del archivo> file.fastq > filesing.fasta

Posteriormente a la separación de cada conjunto, se realizó un análisis FastQC para observar de nuevo la calidad, los resultados se ilustran en las siguientes figuras:

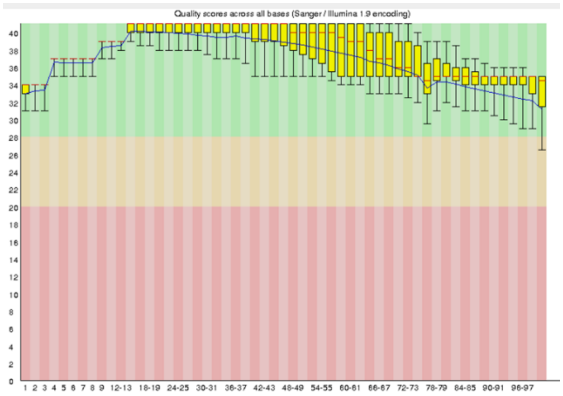


Figura 0.3 Reporte del módulo de calidad de secuencias por base del programa FastQC de la lectura 1 del metagenoma Ca1.

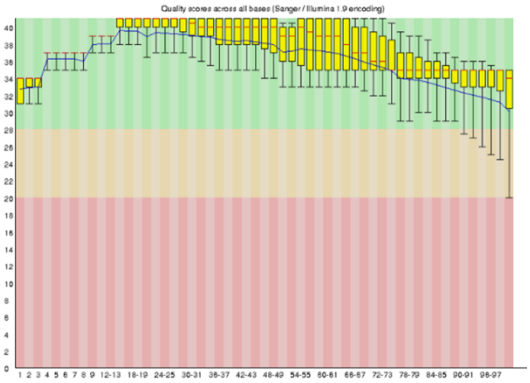


Figura 0.2 Reporte del módulo de calidad de secuencias por base del programa FastQC de la lectura 2 del metagenoma Ca1

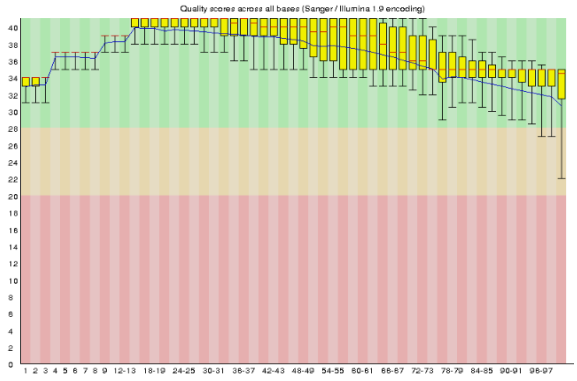


Figura 0.4 Reporte del módulo de calidad de secuencias por base del programa FastQC de las lecturas 1 y 2

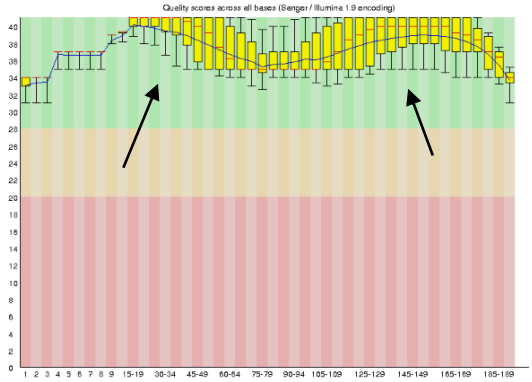


Figura 0.5 Reporte del módulo de calidad de secuencias por base del programa FastQC de las lecturas únicas del metagenoma Ca1.

En los reportes pertenecientes a las lecturas 1 y 2 separadas (Figura 3.4 y 3.5) y en el de las lecturas 1 y 2 juntas (Figura 3.6) se observó una distribución normal. Pero en el reporte de las secuencias únicas (Figura 3.7) se observó la distribución bimodal (señalando con flechas negras sus dos crestas). O sea que dicha distribución se encuentra presente solamente en el conjunto de datos de secuencias únicas.

Cabe señalar que el archivo perteneciente a la muestra Xa2, que contenía secuencias pareadas y solas, las pareadas no se encontraban completas, por lo que se optó por manejarlo como si todas las secuencias fueran solas.

Posteriormente al procesamiento anterior se continuó con el ensamblado de las secuencias.

3.4 Ensamblado

El resultado de los ensamblados de los metagenomas arrojó diferente número de contigs que van desde 5,000 hasta 77,000 para los metagenomas Cs1 y CICY, respectivamente (Figura 3.8). Como se puede observar no existe una correlación entre el número de lecturas y la cantidad de contigs ensamblados, ya que de los metagenoma con menor número de lecturas, Cs1, Cm1, Cm2 y Xa1, se obtuvieron 5,938, 10,798, 28,235 y 21,301 contigs, respectivamente. De los metagenomas con mayor número de lecturas, Xa2, Xs2 y Cs2, que tienen 25,414,201, 27,398,280 y 43,506,356 reads, se obtuvo un total 60,348, 53,674 y 59,632 contigs, respectivamente.

El contig N50 de los metagenomas ensamblados varía de 1.1 hasta 19.4 kb. Tampoco se observa una correlación evidente entre el número de contigs y el contig N50, ya que el metagenoma Xd1 que muestra el mayor contig N50 se encuentra entre los que tiene menor número de reads y contigs, y el metagenoma del CICY que tiene mayor número de contigs ensamblados tiene un contig N50 de 1.5 kb.

En el porcentaje de mapeo se observó un rango que va desde el 22 hasta el 74% de lecturas mapeadas a los metagenomas ensamblados. El metagenoma Xs2 mostró el mayor porcentaje de mapeo, este metagenoma tiene 27,398,280 lecturas que fueron ensambladas en 57,674 contigs y tiene un contig N50 de 7.5 kb. El metagenoma Xs1 fue el siguiente que tuvo un porcentaje de mapeo alto con 60% de las lecturas mapeadas al ensamblado, este metagenoma contiene un total de 21,392,723 reads que se ensamblaron en 40,848 contigs con un contig N50 de 1.9kb. El tercer metagenoma con un porcentaje alto de mapeo (60%) fue Cs2 y que contiene 43,506,356 reads ensamblados en 59,632 contigs con un contig N50 de 3.2 kb. Observándose que en aquellos metagenomas con mayor profundidad de secuenciación se ensambló más del 50% de las lecturas representados en los metagenomas secuenciados.

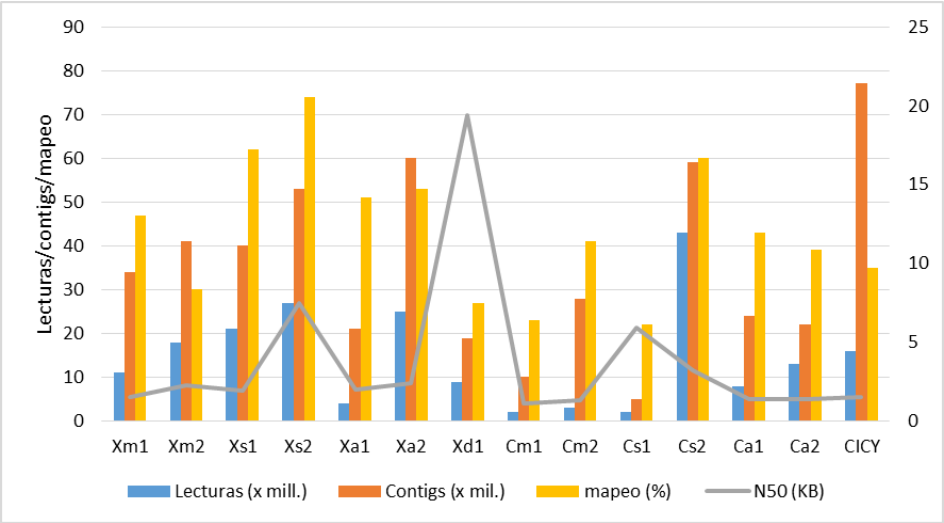


Figura 0.6 Valores cuantitativos de la calidad del ensamblado de los 14 metagenomas. **Nota:** Para simplificar los valores de la gráfica se utilizaron valores cerrados de número de lecturas, contigs, contig N50 y porcentaje de mapeo.

3.5 Validación de la eficacia de la expresión regular

Las expresiones regulares diseñadas con el propósito de identificar secuencias de proteasas S8A fueron validadas utilizando la librería S08a.lib descargada del servidor de MEROPS, The Peptidase Database (https://www.ebi.ac.uk/merops/). Esta librería contiene 20,806 secuencias PS8A y utilizando el comando de linux “grep” se ejecutaron las expresiones regulares para buscar todas las secuencias de triadas catalíticas de proteasas S8A contenidas en la librería. Del total de secuencias contenidas en la librería de MEROPS se recuperaron 12,625 triadas PS8, las cuales representan el 69% de las PS8 contenidas en dicha librería. Por lo que hubo un 31% de secuencias que no se recuperaron, sin embargo se decidió no modificar la expresión regular en este punto, para mantener el rigor de este método.

3.6 Secuencias con motivos de PS8A obtenidas a partir de la minería de datos de los metagenomas ensamblados

Las expresiones regulares se ejecutaron utilizando el comando “grep” de linux en todos los archivos FASTA de los 14 metagenomas ensamblados. De lo anterior se obtuvieron primeramente 135 contigs que contenían los motivos (DHS) y posteriormente se extrajo solamente la región de cada secuencia que representaba la secuencias *core* de dichos contigs (Cuadro 3.2).

A la par se ejecutó el programa HMMER que utiliza Modelos Ocultos de Markov (HMM) en los 14 archivos FASTA que contenían los metagenomas ensamblados. Los resultados de este análisis arrojaron primeramente un total de 251 secuencias de proteasas Posteriormente se utilizó la expresión regular modificada, para obtener exclusivamente todas las regiones catalíticas de cada una de las 251 secuencias identificadas, (Cuadro 3.2), pues HMMER solo arroja resultados de contigs completos.

A partir de este punto, los análisis que se realizaron, fueron a través del uso de las regiones catalíticas Identificadas de las PS8A (*cores* PS8A), no de las secuencias completas.

Se realizó una comparación de los identificadores de secuencias de los resultados obtenidos con las expresiones regulares y los obtenidos utilizando HMM y se observó que las 135 secuencias derivadas del primer análisis estaban contenidas en el grupo de las 251 secuencias identificadas a través de los modelos ocultos.

Se observó que el metagenoma Cs2 fue del que se obtuvo un mayor número de secuencias *core* de PS8A (39), seguido de Xa2 y Xs2 con 34 (Cuadro 3.2) y fueron los metagenomas que mayor número de lecturas presentaron 43, 25 y 27 millones de lecturas respectivamente (Figura 3.8). Mientras que Cs1 y Cm1 fueron los que presentaron un menor número de secuencias *core* de PS8A (3) para cada metagenoma y de igual manera fueron los que contaron con un número menor de lecturas, 2 millones. Sin embargo fue la única relación que se observó pues del metagenoma Ca1 que contó con 15 secuencias *core* de PS8A tuvo 8 millones de lecturas mientras que del metagenoma Xd1 que contó con 9 millones de lecturas se obtuvieron 11 secuencias *core* de PS8A.

Cuadro 0.2 Número de secuencias *core* identificadas como Proteasas S8A utilizando expresiones regulares y modelos ocultos de Markov.

Metagenoma	Identificadas con ER	Identificadas con HMMER
Ca1	7	15
Ca2	7	12
Cm1	1	3
Cm2	6	7
Cs1	2	3
Cs2	12	39
Xa1	5	16
Xa2	15	34
Xm1	13	13
Xm2	13	20
Xs1	13	18
Xs2	12	34
Xd1	6	11
CICY	23	26
TOTAL:	**135**	**251**

3.7 Confirmación de la identidad de las 251 secuencias identificadas en la minería de datos como PS8A.

A través del análisis de anotación funcional con los programas HMMER y BLASTp, se identificaron seis secuencias en el conjunto de datos obtenidos con los modelos ocultos de Markov que no correspondían a PS8A (Cuadro 3.3).

Estas seis secuencias fueron descartadas y los análisis subsecuentes se realizaron con un total de 245 secuencias.

Cuadro 0.3 Secuencias identificadas como no proteasas S8A obtenidas a partir de la minería de datos con los modelos ocultos de Markov.

ID triada	Hmmsearch	Hmmscan	BLASTp	% de identidad
k87_20623_4	No PS8A	Amidasa	Amidasa	75
k87_8660_6		Deacetylasa	Deacetylasa	69
k31313_5		Policétido cicalasa/dehidrasa	Proteína hipotética	100
k87_15212_4		Sin resultado	Proteína hipotética	98
k87_19472_5		Sin resultado	Proteína hipotética	98
k87_44215_5		Sin resultado	Proteína hipotética	67

3.8 Análisis de anotación funcional

Se realizó un análisis de anotación funcional para otorgarle identidad taxonómica a las 245 triadas PS8A, utilizando el programa BLASTp y la librería S08a.lib de MEROPS.

Los resultados mostraron que el 49% de las 245 secuencias *core* de PS8A identificadas taxonómicamente se encuentran en el punto geográfico Xcolac, seguido del punto Calica con 32%, mientras que el punto CICY presentó 19%, siendo el que obtuvo menor porcentaje (Figura 3.9). Estos resultados correlacionan con la cantidad total de contigs identificados en cada punto, pues Xcolac fue el punto geográfico que tuvo un mayor número de contigs formados (268 mil), seguido de Calica (167 mil) y CICY con (77 mil).

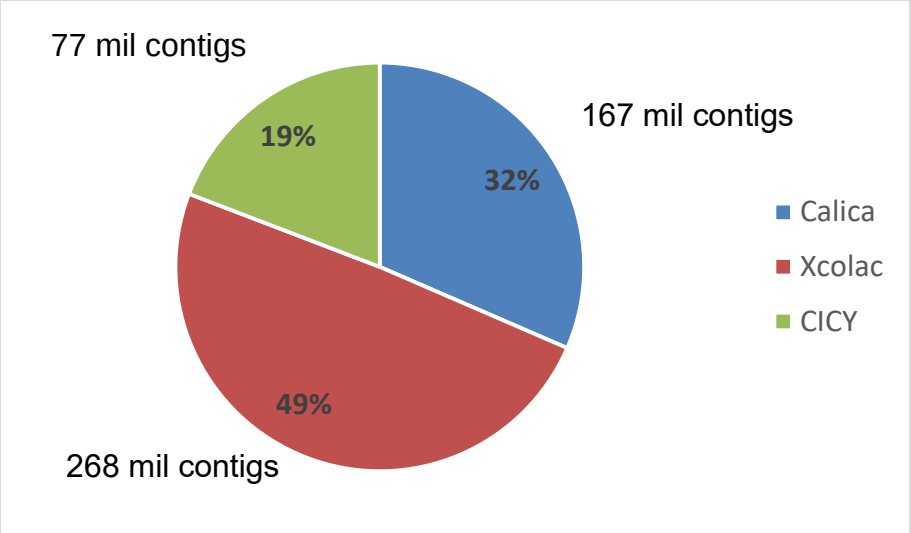


Figura 0.7 Distribución de las 245 secuencias *core* de PS8A identificadas en cada punto geográfico (en porcentaje) y número total de contigs obtenidos en cada uno. Datos normalizados.

El análisis de anotación funcional realizado con Blastp a nivel de phylum (Figura 3.10) reveló que el punto Xcolac fue el que obtuvo un mayor número de phyla (13), seguido del punto CICY (12), mientras que el punto Calica fue el que tuvo un menor número (10).

Lo anterior no mostró una relación entre el número total de contigs obtenidos y el número de phyla, pues Calica a pesar de ser el punto geográfico que se posicionó en segundo lugar respecto al número de contigs totales (167,000) y de PS8A (32%) fue el punto que menor número de phyla de PS8A obtuvo.

En cuanto a la distribución de los phyla en cada punto geográfico, hubo phyla que exclusivamente se encontraron en ciertos puntos, como el phylum Euryarchaeota el cual solo estuvo presente en el punto CICY y Xcolac. El phylum Cyanobacteria se encontró en menor proporción en el punto CICY y mayor en Xcolac. Actinobacteria, Bacteriodetes, Chlorobi, Chloroflexi, Firmicutes, Proteobacteria y no asignados, estuvieron presentes en los tres puntos geográficos (Figura 3.10).

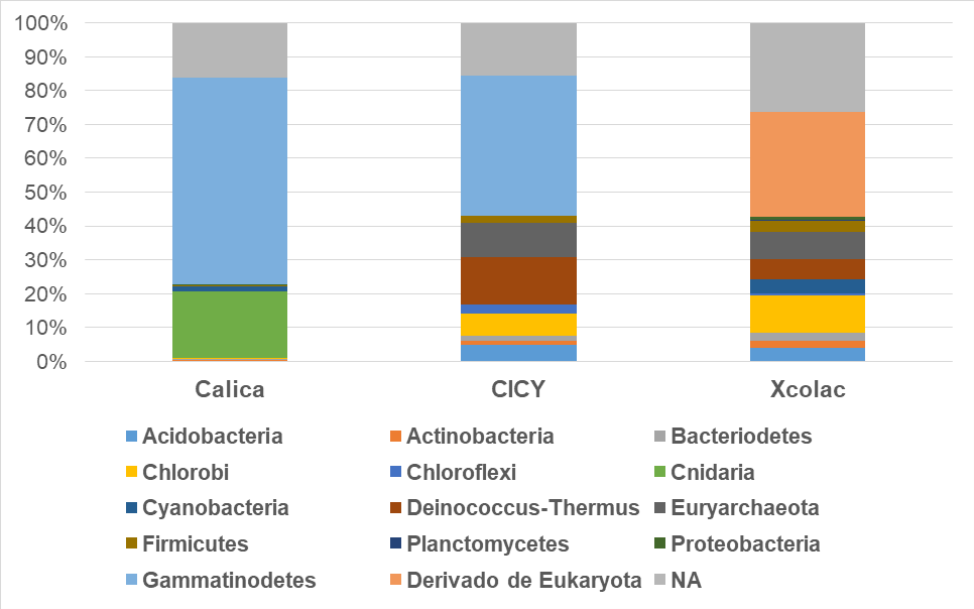


Figura 0.8 Distribución de las 245 secuencias *core* de PS8A por punto geográfico y phylum, los datos se encuentran normalizados de acuerdo con el número de contigs por phylum entre el número de contigs del phylum que contiene PS8A.

Dado que las expresiones regulares se diseñaron con base en las bases de datos existentes (Góngora *et al*.,2020) y en estas probablemente exista un sesgo que favorece la presencia de S8 provenientes de microorganismos cultivables, se seleccionaron los 89 miembros más diversos de la familia de las PS8A depositados en la base de datos de la NCBI, para conocer si con la minería de datos utilizada en este trabajo, se obtuvieron secuencias pertenecientes a estos mismos miembros o estamos ampliando la gama taxonómica de microorganismos a los cuales pertenecen las PS8A. Para esto se compararon los 14 phyla que obtuvimos en la minería de datos, con los phyla más diversos de la NCBI, utilizando su abundancia relativa (Figura 3.11).

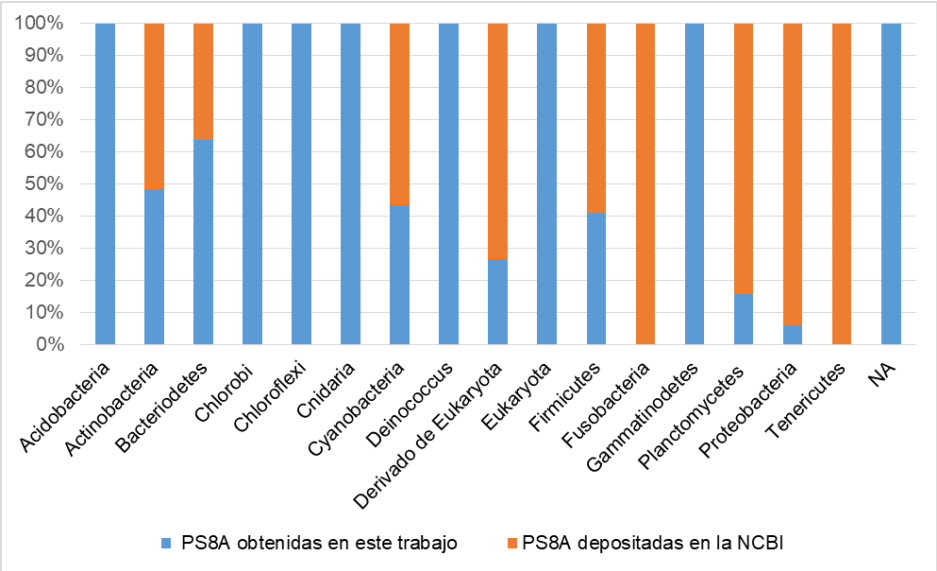


Figura 0.9 Abundancia relativa por phylum de los phyla PS8A más diversos presentes en la NCBI y los obtenidos en este estudio.

Al comparar los phyla, se observó que siete de los 14 phyla obtenidos en este estudio no están presentes dentro de los miembros más diversos de la NCBI (Figura 3.11).

Mientras que solo dos phyla: Fusobacteria y Tenericutes, no se obtuvieron en este trabajo y se encuentran dentro de los más diversos en la NCBI. Por lo que la gama taxonómica se amplió con las secuencias obtenidas en este trabajo.

3.9 Análisis filogenético

Con el fin de identificar las secuencias *core* de PS8A putativas, se realizó un análisis de inferencia filogenética, donde se alinearon 258 secuencias de PS8A utilizando solo la región *core* de la triada catalítica (245 PS8A obtenidas de la minería de datos y 13 de proteasas conocidas). Los resultados mostraron que 212 secuencias se alinearon de forma automática, y las 46 restantes fueron alineadas manualmente, posicionando los motivos conservados y las regiones variables con respecto al alineamiento realizado por el programa.

En el árbol filogenético resultante del alineamiento anterior (Anexo 5.3.1), se observó que las triadas PS8A se agruparon en tres grupos principales, el primero conformado por cuatro

PS8A, que cuenta con un valor de soporte de 99, el segundo conformado por 12 secuencias, tiene valor de soporte de 98 y el tercero da origen a las 242 restantes y presenta un valor de soporte de 55.

Las PS8A correspondientes a plantas se encontraron ubicadas en un mismo clado y compartieron ancestro con tres PS8A.

La secuencia Carlsberg y cuatro de sus secuencias homólogas se distribuyeron en un solo clado, compartiendo ancestro con una secuencia PS8A, mientras que la secuencia homóloga restante y perteneciente al phylum Cyanobacteria, se ubicó en un clado aparte, conformado por dos ramas, por lo que se encuentra estrechamente relacionada con una PS8A no asignada. Las tres secuencias pertenecientes a la familia S8B, se distribuyeron en un solo clado. Por último, la peptidasa y la metaloproteasa se ubicaron en clados separados y cada una estuvo estrechamente relacionada con una sola PS8A, sin embargo ambas secuencias mostraron ramas largas que indican valores altos en la escala, lo cual difiere de todas las secuencias utilizadas en este trabajo.

Se seleccionaron como putativas novedosas 16 secuencias proteasas S8A, pues se ubicaron en dos clados independientes a los que agrupan las 242 PS8A restantes., con valores de soporte de 99 para el primero y el segundo 98.

En las 16 secuencias putativas novedosas, se observaron variaciones de los aminoácidos ubicados justo al lado de cada motivo conservado, los cuales difieren del resto de las secuencias PS8A obtenidas, (Cuadro 3.4).

Las 16 presentaron una identidad entre 37% y 91% con respecto a la proteína PS8A más cercana. Dos pertenecieron a un grupo no asignado (NA), las cuales fueron idénticas entre sí en tamaño y aminoácidos que las conforman. Cuatro pertenecieron al phylum Proteobacteria (género *Thalassomonas*), las cuales tres fueron idénticas entre sí en tamaño y aminoácidos que las conforman. Las secuencias restantes estuvieron distribuidas en los phyla siguientes: Euryarchaeota (género *Methanocella*), Chlorobi (*Ignavibacterium*) y Bacteriodetes (género *Flavobacterium* y *Draconibacterium*).

De los cuales, *Thalassomonas* y *Methanocella* están reportados como géneros no cultivables, mientras que *Ignavibacterium*, *Flavobacterium* y *Draconibacterium* son cultivables.

Sin embargo, al analizar los clados donde se encuentran distribuidas las 242 PS8A restantes, se puede apreciar que probablemente se cuente con más PS8A putativas novedosas, pero se encuentran más cercanas a las PS8 conocidas. Por lo tanto, con mayor determinación se proponen como putativas novedosas solamente aquellas que se ubicaron en los clados descritos con anterioridad.

Cuadro 0.4 Información de las secuencias *core* de las PS8A putativas novedosas.

Identificador de la secuencia	Tamaño (aa)	Organismo respecto a la proteína más cercana	e-value	% de identidad respecto a la proteína más cercana	Motivos conservados y sus aminoácidos
C2406	191	*Thalassomonas actiniarum**	3.97e-71	53	**D**[VGSD]**H**[II]**S**
C607	188	*Thalassomonas actiniarum*	3.97e-71	54	**D**[VGSD]**H**[II]**S**
X7190	188	*Thalassomonas actiniarum*	3.97E-71	53	**D**[VGSD]**H**[II]**S**
X5049	188	*Thalassomonas actiniarum*	3.97E-71	53	**D**[VGSD]**H**[II]**S**
X3432	271	*Methanocella paludícola**	6.54e-27	37	**D**[GDEE]**H**[GTR]**S**
CY7895	207	*Ignavibacterium album* ▪	1.24e-68	48	**D**[TI]**H**[GTR]**S**
X1088	238	*Ignavibacterium album*	2.51e-80	55	**D**[GVS]**H**[GTY]**S**
X6108	254	*Ignavibacterium album*	1.37E-52	45	**D**[GFD]**H**[GTS]**S**
C1202	232	*Ignavibacterium album*	1.24e-68	48	**D**[SGFQ]**H**[GSI]**S**
X3771	229	*Ignavibacterium album*	3.36e-73	51	**D**[TGFN]**H**[GTA]**S**

X6843	232	*Flavobacterium beibuense* ▪	1.13e-120	73	**D**[NGFP]**H**[GTM]**S**
C7495	227	*NA*	8.31e-120	91	**D**[SGFQ]**H**[GTM]**S**
C1431	227	*NA*	8.31e-120	91	**D**[SGFQ]**H**[GTM]**S**
X2273	227	*Draconibacterium orientale*▪	1.15e 80	55	**D**[GGFS]**H**[GNS]**S**
X8626	220	*Draconibacterium orientale*	1.07e-85	58	**D**[DI]**H**[GTM]**S**
X5103	222	*Draconibacterium orientale*	7.76-e79	54	**D**[QI]**H**[GTS]**S**

Nota: * Organismos no cultivables, ▪ cultivables.

CAPITULO IV

DISCUSIÓN

De los 14 metagenomas que analizamos, observamos que las secuencias provenientes de los 13 metagenomas descargados de MG-RAST poseían intervalos de tamaño de lecturas de 101-194, los cuales se obtienen al utilizar para secuenciar la plataforma Illumina CASAVA 1.9. Mientras que en el caso del CICY, todas las lecturas fueron de 150 pb, que es el tamaño que se genera mediante la secuenciación con la plataforma Illumina HiSeqXTen System, por lo que dichas variaciones observadas en las características de cada conjunto de secuencias, fueron a causa del uso de diferentes versiones de secuenciación, lo cual implicó un procesamiento extra a nuestros datos, porque en el caso de los metagenomas se ha reportado que se tiene mejor rendimiento en el ensamblado si se trabaja con archivos separados, es decir archivos con secuencias pareadas y por otro lado secuencias solas. Por lo que es de importancia tomar en cuenta dichas diferencias en las plataformas de secuenciación para poder decidir si es necesario realizar algún paso adicional a las secuencias que se van a analizar.

Según Zhan, (2018) en un estudio metagenómico, el ensamblado de las secuencias es un proceso desafiante, pues se reconstruyen genomas de miles de microorganismos a partir de fragmentos genómicos, por lo que se han desarrollado diversos programas computacionales para su procesamiento, los cuales deben ser seleccionados de acuerdo con la pregunta biológica de investigación, (Van der Walt *et al.,* 2017).

En este estudio el ensamblador utilizado fue Megahit, tomando en cuenta los trabajos de (Gupta *et al*., 2019; Vollmers *et al.,* 2017), en los cuales pusieron a prueba diversos ensambladores de metagenomas. Megahit es uno de los mejores ensambladores actuales, porque no requiere gran poder computacional, es altamente sensible ante la diversidad de las secuencias, puesto que en el ensamblado se tiene alta cobertura taxonómica, y por lo tanto está reportado como útil para descubrir especies nuevas en hábitats complejos, lo cual se ajusta a este trabajo.

En este proyecto se determinó que el tamaño mínimo de contig para el ensamblado fuera de 500 pb para todos los metagenomas, ya que ninguna de las proteasas S8A conocidas es menor a 600 pb de acuerdo con la base de datos de las proteasas S8A de la NCBI.

Con los resultados del ensamblado se pudo apreciar que el número de lecturas por metagenoma estuvo en un intervalo de 2 (Cm1) a 43 (Cs2) millones. Estas diferencias en el número de lecturas obtenidas en la secuenciación son muy comunes, se cree que pueden deberse a la profundidad que el usuario solicite o en otros casos al manejo que tuvo la muestra al realizar la extracción del ADN, lo cual involucra más/menos micro litros o porque existió algún tipo de degradación.

En cuanto a los resultados del ensamblado, se pudo observar que los tres metagenomas que tuvieron mayor porcentaje de mapeo, Xs2 (74%), Xs1 (60%) y Cs2 (60%), tienen más de 20 millones de lecturas que fueron ensambladas en más de 40,000 contigs con un contig N50 promedio de 4.2 kb. De lo anterior podemos concluir que aquellos metagenomas con mayor profundidad de secuenciación (>20,000,000 lecturas) pueden ensamblar más del 50% de los genomas representados en la muestra secuenciada y esto se refleja en el porcentaje de lecturas alineadas al ensamblado.

Sin embargo, se observó de igual manera que cuando se contó con una mayor profundidad de secuenciación (>25, 000,000 de lecturas) se obtuvo un mayor número de secuencias PS8A (arriba de 30), lo cual es similar a lo observando en el trabajo de Gweon *et al*., (2019) en el cual analizaron como interfiere la profundidad de secuenciación en metagenomas con estrategia de escopeta, con respecto a la obtención de genes de resistencia a antimicrobianos (por sus siglas en inglés RAM). Observaron que para realizar una caracterización taxonómica requirieron de una profundidad de secuenciación relativamente baja (un millón de lecturas), pero para caracterizar la diversidad genética de los genes RAM fue insuficiente, pues al usar 200 millones de lecturas no se obtuvo la diversidad completa de dichos genes, debido a que contaban con un control del número de genes presentes.

Por lo que posiblemente sea necesario un mayor número de lecturas para poder ensamblar de manera más completa los genomas de los microorganismos y de esa manera tener el acceso a las regiones genómicas que contienen dichas proteínas de interés. Sin embargo, son necesarios análisis más profundos para poder concluir algo al respecto.

Las estructuras cristalinas de proteasas S8A, han revelado que utilizan una triada catalítica altamente conservada y que contiene los motivos de ácido aspártico (D), histidina (H) y serina (S) que poseen poca o ninguna similitud de secuencia con otras proteínas (Tripathi y Sowdhamini, 2009).

Lo anterior brinda la oportunidad de desarrollar y aplicar estrategias que permitan identificar nuevas secuencias proteasas S8A, a partir de la detección de dichos aminoácidos conservados (Neitzel *et al.,* 2010). En este estudio se utilizaron dos estrategias de minería de datos para su detección: las expresiones regulares y los modelos ocultos de Markov.

La evaluación realizada para probar la eficiencia de las expresiones regulares presentó una eficacia de 69% de efectividad, lo cual significó que de las 135 secuencias *core* PS8A putativas que obtuvimos de los metagenomas a partir del uso de las expresiones probablemente existió un 31% de secuencias *core* que no se lograron detectar.

Sin embargo el 30% que no se detectó fue debido a la rigidez que el método posee, pues el hecho de que hubiera secuencias que no se recuperaron fue a causa de que estas poseían aminoácidos distintos en las regiones variables a los motivos conservados, los cuales no se encontraban dentro de las expresiones regulares diseñadas.

Lo anterior contrasta con los modelos ocultos de Markov, que confiere mayor flexibilidad al ejecutar la búsqueda de secuencias de Proteasas S8A. En este estudio utilizando los Modelos ocultos de Markov obtuvimos 251 secuencias proteasas S8A putativas. Por lo que es un método más eficiente en cuanto a número de secuencias detectadas y se obtiene un reporte extenso de cada secuencia. A través de los resultados del análisis utilizando dicho método se obtuvo incluso secuencias que contiene un solo motivo, las cuales fueron descartadas ya que no hay reportes de proteasas S8A que tengan actividad catalítica con un menor número de motivos.

Sin embargo, debido a las características anteriormente descritas que dicha estrategia posee, tuvo que ser modificada la expresión regular de Góngora *et al*., (2020), a través de la incorporación de aminoácidos en las regiones variables, para poder extraer de las secuencias (obtenidas con los modelo ocultos) las triadas catalíticas correspondientes, pues en los resultados no se obtiene un archivo que contenga solamente las traídas PS8A, sino el contig completo.

De lo anterior, fue posible recuperar seis secuencias *core* que no correspondieron a PS8A, demostrando que al hacer menos rígida la expresión regular al agregar un mayor número de aminoácidos en las regiones hipervariables, es posible obtener secuencias no correspondientes a las de interés y que en este trabajo en cuanto a la detección de secuencias PS8A los modelos ocultos tuvieron un porcentaje de error de 2.4%.

En cuanto al análisis taxonómico de las secuencias PS8A putativas, en el punto Xcolac estuvieron presentes más del 45% de las secuencias PS8A obtenidas y fue el punto con que contó con más contigs formados, seguido de Calica y CICY.

De acuerdo con el trabajo de Moore *et al*., (2020) Xcolac se encuentra rodeado de bosques y campos agrícolas, es un sitios abierto que permite la entra de vegetación y flujo constante de materia orgánica, así como de luz solar. En el caso del punto geográfico Calica, está ubicado en un área con canteras, y parcialmente cubierto por lo que el flujo de materia orgánica y de luz solar es limitado. Mientras que el punto CICY no posee acceso alguno, pues el agua de su interior se obtiene a través de una bomba, por lo que hay mínima cantidad de luz solar y el flujo de materia orgánica es incierto.

Las características que posee cada punto geográfico, podrían estar determinando la distribución que las secuencias PS8A putativas tienen en cada uno, pues su presencia se encuentra relacionada con las comunidades bacterias que habitan cada punto, las cuales a su vez están siendo determinadas por las condiciones del medio.

Al respecto Muscarella *et al*., (2019) mencionan que las comunidades bacterianas se encuentran influenciadas por la heterogeneidad del conjunto de recursos disponibles del sitio donde se encuentran. La diversidad puede aumentar linealmente con la concentración de recursos, pues algunas bacterias acuáticas pueden especializarse en ciertos recursos y si hay cambios en estos hay una reestructuración de las comunidades bacterianas, y por ende puede disminuir o aumentar la diversidad. Lo cual se relaciona con los phyla a los que pertenecieron las secuencias PS8A putativas obtenidas en cada punto geográfico, pues el punto Xcolac es del que se obtuvo un mayor número de secuencias PS8A putativas pertenecientes al phylum Cyanobacteria, seguido del Calica y CICY, dicho phylum está reportado como prevalente en los sitios acuáticos con luz solar directa (Moore *et al*., 2020), lo que coincide con las condiciones de luz solar que cada punto geográfico presenta. Por lo

que se puede observar una relación entre la presencia de los microorganismos con los metadatos obtenidos de cada punto geográfico.

Otro ejemplo, es el phylum Euryarcheota, el único perteneciente a Archaea del que obtuvimos secuencias proteasas S8A putativas. Este está reportado como prevalente en el sedimento y posee miembros no cultivables. De acuerdo con Moore *et al*., (2020) la cantidad de miembros de ese phylum conocidos del acuífero es baja.

Las secuencias PS8A putativas pertenecientes al phylum Euryarchaeota fueron prevalentes en el sitio CICY, seguido de Xcolac y nulas en Calica.

Lo anterior, puede tener relación con la manera en la cual la muestra CICY fue obtenida, pues fue a través del uso de una bomba. La fuerza ejercida impulsa el agua hacia el exterior, por lo que se podría inferir que al aplicar dicha fuerza existe un levantamiento del sedimento depositado al fondo del pozo. Lo que propicia la obtención del material genómico de miembros de este phylum. Mientras que su presencia en el punto Xcolac podría estar dada porque un metagenoma de dicho punto perteneció al sedimento, mientras que del punto Calica no se utilizaron secuencias pertenecientes a ninguna muestra de sedimento, (Moore *et al*., 2020). Por último, podemos mencionar al phylum Proteobacteria, el cual comprende una de las divisiones más grandes y diversas de los procariotas y se encuentra distribuido en casi todos los ambientes (Gupta *et al*., 2000). En este estudio se obtuvieron secuencias de dicho phylum de los tres sitios, por lo que al ser un phylum grande y diverso, su presencia en diversos ambientes es esperada, a diferencia del phylum Euryarcheota que está conformado por miembros que se distribuyen específicamente en el sedimento.

De acuerdo con el análisis taxonómico desarrollado en este trabajo, se observó que obtuvimos secuencias PS8A putativas pertenecientes a phyla que no se encuentran entre los más diversos reportados en la NCBI, sin embargo no obtuvimos dos phyla (Tenericutes y Fusobacteria) que sí. No obstante están reportados como patógenos de tejidos y cavidades humanas (Booth *et al*., 2014; Berman., 2012), por lo que su ausencia en este trabajo es esperada. Por lo mencionado anteriormente el presente trabajo permitió ampliar la gama taxonómica de PS8A, al obtener secuencias pertenecientes a phyla poco diversos.

De acuerdo con los resultados obtenidos en el análisis filogenético, Krem y Di cera, (2010) mencionan que cuando se trabaja con secuencias de proteínas de gran similitud, estas tienden a agruparse en el mismo grupo taxonómico, así como aquellas que pertenecen al

mismo organismo o a organismos estrechamente relacionados. Lo anterior explica por qué en algunos clados hubo segregación de secuencias, de acuerdo a dos cosas: (i) tipo de proteína, como ejemplo se puede mencionar las S8B, que se agruparon en un solo clado, coincidiendo con lo obtenido en el trabajo de Góngora *et al*., (2020), pues las S8B que utilizaron de igual manera se agruparon en un clado. Lo mismo se observó con las secuencias pertenecientes a las PS8A de plantas, la Carlsberg y sus homólogos. (ii) relación filogenética, pues se encontraron clados que contenían secuencias provenientes de diferentes puntos geográficos, las cuales al consultar pertenecieron al mismo phylum.

En lo que concierne a las 16 secuencias PS8A propuestas como putativas novedosas, se distribuyeron en géneros reportados de recién descubrimiento con énfasis en *Methanocella paludícola,* pues son poco conocidas en el sedimento del acuífero de Yucatán *(*Moore *et al.,* 2020*),* ésta no es cultivable y es la secuencia que presentó el menor porcentaje de identidad, con 37%, por lo que la diferencia con las proteasas S8A existentes es alta.

Se observó que a pesar de contar con los tres motivos conservados (como el resto de las proteasas S8A encontradas), los aminoácidos contiguos a los motivos D y H difirieron de los que se encontraron en el resto de las secuencias, pues generalmente se encuentran de la siguiente manera: **D** [TGI] **H** [GTH] y solamente el motivo de la serina (S), no presentó variaciones en sus aminoácidos contiguos, pues en todas estuvieron presentes los aminoácidos G y T antes del motivo S**,** por lo que se infiere que dichas diferencias en los motivos D y H pudieron provocar que se ubicaran en clados distintos.

No obstante al comparar visualmente los aminoácidos que conformaban las secuencias *core* de todas las secuencias se observaron diferencias a nivel de género pero a nivel de especie parecen estar conservadas en ciertos casos. Como ejemplo se puede mencionar a *Paenibacillus polymyxa* del cual se recuperaron dos secuencias PS8A idénticas en tamaño y en aminoácidos. Una se obtuvo del punto Calica y la otra de Xcolac, lo cual descarta la idea de que haya existido algún error generalizado en el ensamblado o repetición de datos por algún motivo no identificado y aporta la idea de que posiblemente en algunas especies la región genómica perteneciente a la triada PS8A se encuentra conservada.

Otro ejemplo, son las secuencias obtenidas de *Thllassomonas actiniarum*, las cuales están propuestas como putativas novedosas. Tres fueron secuencias idénticas en aminoácidos y tamaño, la cuarta solo varió en un aminoácido, con respecto a las anteriores. La quinta

secuencia obtenida perteneció al mismo género pero a diferente especie: *T. viridians* y los aminoácidos que la conformaron variaron con respecto a *T. actiniarum*. Por lo que se continúa soportando la idea de que la región catalítica PS8A entre microorganismos de la misma especie está conservada.

Sin embargo, se observó que en algunos casos, las secuencias *core* variaban de tamaño dentro de la misma especie, como ejemplo se menciona a *Acinetobacter baumanii* de la cual se recuperaron cinco secuencias PS8A putativas, las cuales variaron en tamaño. Dos fueron de 194 aminoácidos, dos de 229 y una de 135.

Por último, se obtuvieron secuencias que tuvieron el mismo tamaño pero los aminoácidos variaron, como es el caso de *Micromospora carbonaceae* del que se recuperaron nueve triadas catalíticas, las cuales todas presentaron el tamaño de 196 aminoácidos, pero dichos aminoácidos variaron entre secuencias.

Por lo que se concluyó que no hay una regla general aplicable al tamaño específico o al patrón de aminoácidos observado en las secuencias *core*, sin embargo si se observó que ciertos microorganismos tienden a poseer dichas regiones altamente conservadas.

No obstante, la capacidad de predecir qué cambios trae consigo la sustitución de un aminoácido fuera de dichos motivos es uno de los desafíos más difíciles en bioquímica y biología celular. Sin embargo, Neitzel *et al* (2019) menciona que debido a que se conocen los motivos conservados de las PS8A es posible identificar nuevas, con solo examinar los aminoácidos de las regiones hipervariables que las conforman, por lo que los resultados de este estudio siembran las bases para que en futuros trabajos se analicen si dichos aminoácidos variables interfieren en la función y en el plegamiento de dichas proteínas a su vez analizar otros clados que contengan secuencias diferentes a las que proponemos como putativas novedosas.

Por lo tanto, a través del uso de las expresiones regulares y los modelos ocultos de Markov se tuvo la capacidad de detectar secuencias putativas novedosas, sobresaltando el alcance que la aplicación de las tecnologías de secuenciación masiva, la bioinformática y la metagenómica poseen.

CAPITULO V

CONCLUSIONES Y PERSPECTIVAS

5.1 CONCLUSIONES

- Mediante la implementación de estrategias bioinformáticas para la minería de datos metagenómicos fue posible obtener secuencias que estuvieran los motivos característicos de proteasas S8A en metagenomas del acuífero de Yucatán.
- En este trabajo se describe una serie de pasos bioinformáticos para el análisis de secuencias metagenómicas que permitió la identificación de secuencias putativas de proteasas S8A. Las expresiones regulares y los modelos ocultos de Markov son herramientas que pueden ser aplicadas en cualquier tipo de organismo para buscar e identificar cualquier secuencia de interés.
- La variación en los aminoácidos contiguos a los motivos conservados fue la principal diferencia observada en las 16 secuencias PS8A putativas novedosas, con respecto al resto de las secuencias.
- Se observó una relación entre los metadatos de los metagenomas con la distribución de las secuencias PS8A obtenidas, pues estuvieron relacionadas con los phyla presentes en cada punto geográfico, los cuales a su vez son específicos para las condiciones particulares que cada punto geográfico posee.

5.2 PERSPECTIVAS

Es de suma importancia tomar en cuenta los metadatos de los metagenomas con los que se esté trabajando pues aportan gran información que puede ser utilizada para trabajos posteriores, como se observa en la figura 5.1, en la que se representa la distribución de secuencias *core* de PS8A por phylum, en cuatro estratos de la columna de agua. Lo que podría direccionar la búsqueda de dichas proteínas a estratos específicos o puntos geográficos.

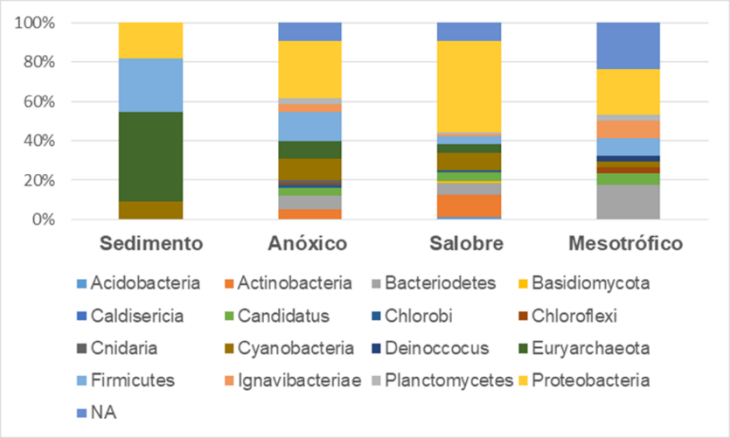


Figura 0.1. Distribución de 222 secuencias *core* de PS8A por estrato acuático

Con este estudio se siembran bases para futuros trabajos para analizar si las secuencias proteicas obtenidas presentan aplicaciones biotecnológicas, pues con la identificación de las secuencias putativas en la inferencia filogenética se abre camino a diversos análisis tales como: (i) búsqueda de genes PS8A completos, con la identificación de los marcos abiertos de lectura (por sus siglas en inglés: ORF'S), (ii) expresión de las proteínas para verificación de la catálisis, (iii) patentes y aplicaciones biotecnológicas industriales.

5.3 ANEXOS

5.3.1. Inferencia filogenética de las relaciones entre las proteasas S8A obtenidas por la minería de datos y proteasas conocidas.

Descripción de los identificadores utilizados en el árbol: X: pertenece al punto Xcolac, C: punto Calica, CY: al punto CICY. Los números siguientes a las letras anteriores corresponden al identificador del contig al que pertenecen.

La secuencia correspondiente a la proteasa Carlsberg fue nombrada como Carls. Las secuencias homólogas a esta, se nombraron de acuerdo al phylum al que pertenecen. EuryH: Euryarchaeota, FirmH: firmicutes, ProtH: proteobacteria, ActinoHC: actinobacteria, CyanH: Cyanobacteria.

La secuencia perteneciente a la pepsina**:** pepsin, a la familia de las metalo proteasas: metall. Por último las S8B, fueron nombradas como furiS8B y ScS8B.

El árbol está presentado en escala y los valores de soporte están representados en los números ubicados en cada clado.

5.4 BIBLIOGRAFÍA

Aguilar-Bultet, L., y L. Flaquet. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. Revista de Salud Animal, 37, 125–132.

Andrews., S. (2010). FastQC: a quality control tool for high throughput sequence data. [Online] (Actualizado 2020). Disponible en: http://www.bioinformatics.babraham.ac.uk/projects/ fastqc [Acceso 2019].

Apolinar-Hernández, M., J. Peña-Ramírez., E. Pérez-Rueda., B. Canto- Canché., C. De Los Santos-Briones., A. O'Connor-Sánchez. (2016). Identification and in silico characterization of two novel genes encoding peptidases S8 found by functional screening in a metagenomic library of Yucatán underground water. Gene, 593, 154-161.

Austin, C. (2020). National Genome Research Institute. Traslation. [Online] (Actualizado 2020). Disponible en: https://www.genome.gov/genetics-glossary/Translation [Acceso agosto 2020].

Beers, E., A. Jones., y A. Dickerman (2004). The S8 serine, C1A cysteine and A1 aspartic protease families in Arabidopsis. Phytochemistry, 65, 43–58.

Berman, J J. (2012). Taxonomic Guide to Infectious Diseases (Second Edition). pp. 374.

Besser, J., H.A. Carleton., P. Gerner-Smidt., R. L, Lindsey., y E. Trees. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical Microbiology and Infection, 24, 335-341. doi.org/10.1016/j.cmi.2017.10.013.

Booth, S J. (2014). Reference Module in Biomedical Sciences. pp. 356.

Caamal-Pech, Aldo., J. Ramírez-Prado., E. Góngora-Castillo y A. O'Connor-Sánchez. (2018). Análisis de la diversidad de proteasas S8 en un metagenoma acuático mediante herramientas bioinformáticas. Revista Del Centro de Graduados e Investigación, Instituto Tecnológico de Mérida, 33, 263–266.

Cooper G.M. (2000). The Cell: A Molecular Approach. Sinauer Associates. pp. 570.

Diniz, W y F. Canduri (2017). Bioinformatics: An overview and its applications. Genetics Molecular Research, 16. doi: 10.4238/gmr16019645

DOE: departament of energy, Joint Genome Institute. BBmap Guide. [Online]. (Actualizado 2020). Disponible en: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/ [Acceso 2020].

Eddy, S. R. (1996). Hidden Markov models. Current Opinion in Structural Biology, 6, 361-365. doi:10.1016/s0959-440x(96)80056-x .

Eddy. R. S. (1998). Profile Hidden Markov Models. Oxford University Press. Bioinformatics Review. 9, 755-763.

El-Gebali, J. M., A. Bateman., Eddy, R.S., A. Luciani., S. Potter., M. Qureshi., L. Richardson., G. Salazar., A. Smart., E. Sonnhammer., L. Hirsh., L. Paladin, D. Piovesan., S. Tosatto y R. Finn (2018). The Pfam protein families database in 2019, Nucleic Acids Research. 47, 427- 432. doi.org/10.1093/nar/gky995.

Finn, R. D., J. Clements, y S. R. Eddy (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research. 39, 29-37. doi.org/10.1093/nar/gkr367.

Furhan, J., y S. Sharma (2014). Microbial alkaline proteases: Findings and applications. International Journal of Inven (2011). Aplication of Metagenomics to Bioremediation. Great Britain: Craister Academic Press. 119-140.

Gilbert, D (2010). Sequence File Format Conversion with Command-Line Readseq. Current Protocols in Bioinformatics, 1,1-4. doi:10.1002/0471250953.bia01es00.

Gill, S., M. Pop., R. DeBoy., P. Eckburg., P. Turnbaugh., B, Samuel y J. Gordon (2006). Metagenomic analysis of the human distal gut microbiome. Science, 31, 1355-1359. doi:10.1126/science.1124234.

Gomez-Alvarez, V., T. Teal y T. Schmidt (2009). Systematic artifacts in metagenomes from complex microbial communities. ISME. 1314-1317. https://doi.org/10.1038/ismej.2009.72.

Góngora-Castillo, E., L.A. López-Ochoa., M.M. Apolinar-Hernández., A.M. Caamal-Pech, P.A. Contreras-de la Rosa., A. Quiroz-Moreno, J.H. Ramírez-Prado y A. O'Connor-Sánchez (2020). Data mining of metagenomes to find novel enzymes: a non-computationally intensive method. 3 Biotech, 10. 1-8. doi:10.1007/s13205-019-2044-6.

Goyvaerts, J y S. Levithan (2012). Regular Expressions Cookbook. O'REILLY. pp. 575.

Guindon, S., J.F., Dufayard., V. Lefort., M. Anisimova., W. Hordijk y O. Gascuel (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology, 59.

Gunawan, R., A. Rahmatulloh., I. Darmawan y F. Firdaus (2019). Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath. doi.10.2991/icoiese-18.2019.50.

Gupta, R. S. (2000). The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. Microbiology Reviews. 24, 367-402.

Gupta, R., Q. Beg y P. Lorenz (2002). Bacterial alkaline proteases: Molecular approaches and industrial applications. Applied Microbiology and Biotechnology. 59, 15-32. doi.org/10.1007/s00253-002-0975-y.

Gupta, S. K., S. Raza y T. Unno (2019). Comparison of de-novo assembly tools for plasmid metagenome analysis. Genes & Genomics. doi:10.1007/s13258-019-00839-1.

Gweon, H. S., L. P. Shaw., N. De Maio., M. Abuon y N. Stoesser (2019). The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. Environmental Microbiome, 14. doi:10.1186/s40793-019-0347-1.

Hand, D.J y N.M. Adams (2015). Data Mining. Statistics. 1-7. doi:10.1002/9781118445112.stat06466.pub2.

Hartley, B. S (1960). Proteolytic enzymes. Anual Review of Biochemistry, 29,45-72. doi:10.1146/annurev.bi.29.070160.000401.

Hernández-De Lira, H., M.P, Luévanos-Escareño., F. Hernández-Terán., J. Sáenz-Mata y N. Balagurusamy (2016). Metagenómica: Concepto y Aplicaciones en el Mundo Microbiano. Fronteras en microbiología aplicada. 154-170.

Hernández-León, R., I. Velázquez-Sepúlveda., M. Orozco., G. Santoyo (2010). Metagenómica de suelos: grandes desafíos y nuevas oportunidades biotecnológicas. Phyton. 79, 133-139.

Hulo, N (2006). The PROSITE database. Nucleic Acids Research. 34,227-230. doi:10.1093/nar/gkj063.

Katoh, K y D. M. Standley (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30, 772-780.

Keegan, K.P., E.M, Glass y F. Meyer (2007). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Microbial Environmental Genomics. 207-233. doi: 10.1007/978-1-4939-3369-3_13.

Krem, M. y Di Cera, E (2010). Molecular markers of serine protease evolution. The EMBO Journal, 20, 3036–3045. doi:10.1093/emboj/20.12.3036.

Kanchi, S y Krishna P (2013). Utility Based Pattern Matching Approach for Data Mining. International Journal of Computer Science and Information Technologies. 4, 917 – 921.

Li, D., C.M. Liu., R. Luo., K, Sadakane y T.W, Lam (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Lim, Y. W., R, Schmieder., M. Haynes., D. Willner., M. Furlan., M. Youle y F. Rohwer (2013). Metagenomics and metatranscriptomics: Windows on CF- associated viral and microbial communities. Journal of Cystic Fibrosis. 12, 154–164. doi:10.1016/j.jcf.2012.07.009.

Marfil-Santana, M.D., A. O'Connor-Sánchez., J.H. Ramírez-Prado., C. De Los Santos-Briones., K.L. López-Aguiar., R. Rojas-Herrera y A. Prieto-Davó (2016). A computationally simplistic poly-phasic approach to explore microbial communities from the Yucatan aquifer

as a potential sources of novel natural products. Journal of Microbiology, 54, 774-781. doi: 10.1007/s12275-016- 6092-x.

Meirelles, L. A., Q.S. McFrederick., A. Rodrigues., J.D. Mantovani., C. de Melo Rodovalho y U.G. Mueller (2016). Bacterial microbiomes from vertically transmitted fungal inocula of the leaf-cutting ant Atta texana. Environmental Microbiology Reports, 8, 630-640. doi: 10.1111/1758- 2229.12415.

Meyer, F., D. Paarmann y M. D' Souza (2008). The metagenomics RAST server a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9. https://doi.org/10.1186/1471-2105-9-386.

Montoya, V., Y.C, Chiu y P. Tang (2011). Metagenomics for the Identification of Novel Viruses. Great Britain: Craister Academic Press.145-160.

Moore, A., M, Lenczewski., R.M. Leal-Bautista y M, Duvall (2020). Groundwater microbial diversity and antibiotic resistance linked to human population density in Yucatan Peninsula, Mexico. Canadian Journal of Microbiology. doi:10.1139/cjm-2019-0173.

Muscarella, M.E., C.M. Boot y C.D. Broeckling. Resource heterogeneity structures aquatic bacterial communities (2019). ISME 13, 2183–2195. https://doi.org/10.1038/s41396-019-0427-7.

Nakamoto, T. (2009). Evolution and the universality of the mechanism of initiation of protein synthesis. Gene, 432, 1–6. doi:10.1016/j.gene.2008.11.001

National Center for Biotechnology Information NCBI. Peptidase_s8. [Online] (Actualizado en 2020]. Disponible en: https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?hslf=1&uid=173793&#seqhrch [Acceso en 2020].

Rawlings, N. D., A.J. Barrett., P.D. Thomas., X, Huang., A. Bateman y R.D. Finn (2017). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Research. 46, 624-632. doi:10.1093/nar/gkx1134.

Neitzel, J. J. (2010) Enzyme Catalysis: The Serine Proteases. Nature Education 3.

Nelson, L., A. Lehninger y M. Cox (2013). Lehniger principios de Bioquímica. Omega, pp 19.

Neveu, J., C. Regeard y M.S. Dubow (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Applied Microbiology and Biotechnology. 9, 635-644. doi: 10.1007/s00253-011-3256-9.

O'Connor-Sánchez A., A.J. Rivera-Domínguez., C. De los Santos-Briones., L. López-Aguiar., J. Peña-Ramírez y A. Prieto-Davó (2014). Acidobacteria appear to dominate the microbiome of two sympatric Caribbean Sponges and one Zoanthid. Biological Research, 47. doi: 10.1186/0717-6287-47-67.

Perry, E., A. Paytan., B. Pedersen y G. Velazquez-OlimaN (2009). Groundwater geochemistry of the Yucatan Peninsula, Mexico: Constraints on stratigraphy and hydrogeology. Journal of Hydrology, 367, 27–40. doi:10.1016/j.jhydrol.2008.12.026.

Sharma, P., H. Kumar., M. Verma., M. Kumari, M., y S. Malhotra (2008). From bacterial genomics to metagenomics: concept, tools and recent advances. Indian Journal of Microbiology. 48, 173–194. doi:10.1007/s12088-008.

ReadSeq: biosequence conversion tool. [Online] (Actualizado 12 mayo 2010]. Disponible en https://mafft.cbrc.jp/alignment/server/cgi-bin/readseq.txt [Acceso abril 2020].

Salwan, R y Sharma V. (2019). Trends in extracellular serine proteases of bacteria as detergent bioadditive: alternate and environmental friendly tool for detergent industry. Springer. doi: org/10.1007/s00203-019-01662-8.

Salwan, R., V. Sharma., M. Kasana., S.K. Yadav y A. Gulati (2018). Heterologous expression and structure-function relationship of low-temperature and alkaline active protease from Acinetobacter sp. IHB B 5011(MN12). International Journal of Biological Macromolecules, 107, 567–574.

Sarwar, S.M., R. Kpretsky y S.A. Sarwar (2003). El libro de LINUX. Pearson, pp.839.

Scheibye-Alsing, K., S. Hoffmann., A. Frankel., P. Jensen., P.F. Stadler., Y. Mang y J. Gorodkin (2009). Sequence assembly. Computational Biology and Chemistry, 33, 121–136. doi:10.1016/j.compbiolchem.2008.11.003.

Searls, D.B. (2010). The Roots of Bioinformatics. PLoS Computational Biology, 6. doi:10.1371/journal.pcbi.1000809.

Sedlar, K., K. Kupkova y I. Provaznik (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and Structural. Biotechnology Journal, 15, 48-55. https://doi.org/10.1016/j.csbj.2016.11.005.

Serrano, H. A., M. Sánchez y N. Cardona (2015). Conocimiento de la microbiota de la cavidad oral a través de la metagenómica. Revista CES Odontología, 28, 112-118. doi:10.21615/3681.

Tripathi, L. P y R. Sowdhamini (2006). BMC Genomics, 7, 193-200. doi:10.1186/1471-2164-7-200.

Tyson, J. J., y B. Novák (2010). Functional motifs in biochemical reaction networks. Annual review of physical chemistry, 61, 219–240. https://doi.org/10.1146/annurev.physchem.012809.103457.

Van der Walt, A., M. Van Goethem y J. Ramond (2017). Assembling metagenomes, one community at a time. BMC Genomics 18, 515-521. https://doi.org/10.1186/s12864-017-3918-9.

Venter, J.C., K. Remington., J.F. Heidelberg., A.L. Halper., D. Rusch, J.A. Eisen y H.O. Smith (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304, 66-74. doi: 10.1126/science.1093857.

Vincent, A y S.J. Charette (2015). Who qualifies to be a bioinformatician? Frontiers:Genetics. 6, 1–3. doi:10.3389/fgene.2015.00164.

Vollmers, J., S. Wiegand y A.K. Kaster (2017) Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS ONE. doi.org/10.1371/journal.pone.0169662.

Walshaw, J., G. Etherington y D. MacLean (2011). Next-Generation Sequencing Approaches to Metagenomics. Great Britain: Caister Academic Press. pp. 63-88.

Waterhouse, A. M., J.B. Procter., D. Martin., M. Clamp y G.J. Barton (2009). Jalview Version 2, a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25.

Wilke A., W. Gerlach., T. Harrison., T. Paczian., W. Trimble y F. Meyer (2017). MG-RAST: metagenomics analysis server. pp. 126.

Xu, J. (2011). Metagenomics and Ecosystems Biology: Conceptual Frameworks, Tools adn Methods. Great Britain: Craiser Academic Press. pp. 1-14.

Zhang Q. (2018) Metagenome Assembly and Contig Assignment. Methods in Molecular Biology. doi.org/10.1007/978-1-4939-8728-3_12.

Zhao, L y J. Shen, (2011). Functional Metagenomics and Systems Biology: Understanding the Human Organismal Complexity in Disease and Health. GreatBritain: Craister Academic Press. pp. 49-62.