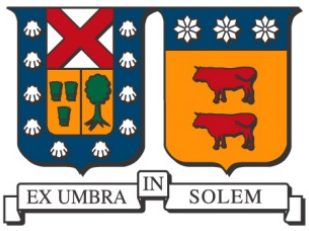
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



“DESARROLLO DE UN PROTOCOLO DE ANÁLISIS PARA VERIFICAR PATRONES DE COEXISTENCIA Y CO-ABUNDANCIA ENTRE ESPECIES VIRALES Y BACTERIANAS EN MUESTRAS METAGENOMICAS DE AMBIENTES ACUÁTICOS”

RODRIGO ALEJANDRO ESCAR SALINAS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Carlos Buil Aranda
Profesor Correferente: Roberto Orellana Román

Agosto - 2020

To the reader of the future,
thou whose God mandated blursed endeavors landed thou upon this cursed work.
Ars artis gratia, multa paucis

AGRADECIMIENTOS

A mis profesores de tesis por la oportunidad, el apoyo y la paciencia.

Al Departamento de Informática por el apoyo en gestión.

Al Hare-Team por la paciencia, el espacio y la calma.

Al equipo SCM I+D por la espera y el desinterés.

A mi familia y amigos por la presencia y la ausencia.

Nietzsche ist mein Gott und Gott ist tot, ich tötete ihn

Resumen

Los virus son los organismos mas abundantes y ubicuos en el planeta. Virus que infectan bacterias son llamados bacteriófagos y son fundamentales en el desarrollo de los ecosistemas que sustentan la vida. El conocimiento sobre los bacteriófagos esta limitado por la dificultad del cultivo de estos en condiciones de laboratorio. Se presenta un protocolo estructurado para hacer uso de diversas herramientas basadas en metagenomica para facilitar el estudio de bacteriófagos en el medio ambiente, consistentes en determinación de presencia de organismos en metagenomas a través de la asignación de unidades taxonómicas operativas (OTU) y la confirmación de relaciones de coexistencias usando variaciones del coeficiente de Jaccard. Se valida un desempeño minimamente suficiente de las herramientas listando la ausencia de información sobre bacteriófagos y sesgo de las bases de datos de referencia como principal causa en los problemas de entrenamiento y validación de las herramientas. Se presenta un análisis de metagenomas de ambientes agua dulce usando el protocolo presentado, limitado a grupos taxonómicos de interés.

Palabras Clave— Bioinformatica; Metagenoma; Redes; Ecologia; Fago

Abstract

Viruses are the most abundant and ubiquitous organisms on the planet. Viruses that infect bacteria are called bacteriophages and are essential in the development of ecosystems that support life. Knowledge about bacteriophages is limited by the difficulty of cultivating them under laboratory conditions. A structured protocol is presented to make use of various tools based on metagenomics to facilitate the study of bacteriophages in the environment, consisting of determining the presence of organisms in metagenomes through the assignment of operative taxonomic units (OTU) and the confirmation of coexistence relationships using variations of the Jaccard coefficient. A minimally sufficient performance of the tools is validated by noting the absence of information on bacteriophages and bias from the reference databases as the main cause of problems in the training and validation of the tools. An analysis on metagenomes of freshwater environments is presented using the presented protocol, limited to taxonomic groups of interest.

Keywords— Bioinformatics; Metagenome; Networks; Ecology; Phage

Glosario

ADN Ácido Desoxirribonucleico

ARN Ácido Ribonucleico

BLAST Basic Local Alignment Search Tool

BRIM Bipartite, Recursively Induced Modules

CRISPR clustered regularly interspaced short palindromic repeats

DNA Deoxyribonucleic Acid

dsDNA Double-Stranded DNA

FTP File Transfer Protocol

ICTV International Committee on Taxonomy of Viruses

INSDC International Nucleotide Sequence Database Collaboration

JGI Joint Genome Institute

MG-RAST MetaGenomics Rapid Annotation using Subsystem Technology

NCBI National Center for Biotechnology Information

NGS Next Generation Sequencing

NIH National Institutes of Health

OTU Operational Taxonomic Unit

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Los microorganismos son las formas más abundantes, diversas y ubicuas de la Tierra. A pesar de que son considerados por la opinión pública primariamente como agentes de enfermedad, los microorganismos tienen una enorme relevancia para el funcionamiento del planeta. Diariamente, diversos procesos microbianos impulsan los procesos biogeoquímicos de relevancia que sostienen la vida y directamente influencian nuestro clima. En una escala menor, las comunidades microbianas que habitan nuestro tracto digestivo han evolucionado conjuntamente con el humano durante miles de años para formar un equilibrio mutuamente beneficioso, del cual dependemos para mantenernos sanos. El estudio de los microorganismos ha enfrentado diversos desafíos a través del tiempo, tales como el descubrimiento de nuevas especies y la comprensión de cómo estos interactúan con su entorno. Desde la implementación de técnicas de cultivo, la microbiología ha enfocado el estudio de los microorganismos en el ambiente usando el mismo enfoque tradicional. Esto es a través del uso de medios de cultivo, en donde puedan crecer aquellos microorganismos que residen en el ambiente, y así permitir el estudio de estos a través de técnicas de laboratorio. Sin embargo, la primera evidencia de que una gran proporción de los microorganismos de un ambiente no son capaces de crecer en condiciones de laboratorio provino de la microscopía. Diferentes estudios evidenciaron que el número de células microbianas que se observaron microscópicamente en un ambiente superó ampliamente el número de colonias que crecen en una placa de Petri con muestras obtenidas desde el mismo ambiente (Achtman & Wagner, 2008), generando lo que se conoce como la ”Gran Anomalía del Conteo de Placas”. Esta anomalía no es constante y dependiendo del medioambiente podría alcanzar varios órdenes de magnitud, limitando así severamente la profundidad de la información recolectada en este tipo de estudios (Staley & Konopka, 1985).

Desde su descubrimiento, la biología molecular aplicada al estudio de la microbiología permitió eludir esta anomalía proporcionando una nueva visión de la ecología microbiana, consolidando una plataforma de estudio de comunidades altamente complejas. Adicionalmente, los recientes avances en las tecnologías de secuenciación genética masiva y su masificación han producido una revolución en la manera en que se estudian la composición y estructura de las comunidades microbianas. El impacto de este desarrollo tecnológico se ha visto reflejado en la expansión significativa del campo de investigación conocido como ecogenómica ambiental, el que incorpora las exploraciones de comunidades microbianas en el ambiente, el descubrimiento de nuevas especies de bacterias, arqueas y virus, y la determinación de su función en el ecosistema (Eloe-Fadrosh y col., 2016; Yarza y col., 2014).

Durante las últimas dos décadas, la ecogenómica ambiental se ha volcado a la exploración de virus en el medioambiente y de cómo estos juegan un rol indispensable en la regulación ecológica de los ecosistemas. Los virus son agentes infecciosos que solo se replican dentro de células vivas de otro organismo. Los virus pueden infectar todas las formas conocidas de vida, desde animales, plantas y otros microorganismos como bacterias y arqueas, incluso otros virus (Koonin, Senkevich & Dolja, 2006). Mientras no están infectando a una célula hospedera, los virus existen como partículas independientes llamadas *"viriones"*. Las formas de estas partículas virales van desde simples hélices e icosaedros, hasta formas más complejas, como filamentosas, según su especie o más específicamente su ecotipo. La mayoría de los virus poseen viriones tan pequeños que no pueden ser observados con un microscopio óptico, alcanzando tamaños tan pequeños como una centésima del tamaño de una célula bacteriana promedio.

Existen opiniones divididas sobre si los virus conforman una forma de vida o solo son estructuras orgánicas que interactúan con organismos vivos (Moreira & López-García, 2009). Poseen similitud con otros organismos ya que poseen genes, evolucionan por selección natural y se reproducen creando múltiples copias de sí mismos. A pesar de que tienen genes, carecen de estructura celular, la cual comúnmente se considera la unidad básica de la vida. Los virus no tienen su propio metabolismo y requieren de una célula hospedera para crear nuevos productos, por consiguiente no se pueden reproducir fuera de un hospedero (Wimmer, Mueller, Tumpey & Taubenberger, 2009).

Los virus corresponden, por lejos, a las formas de vida más abundantes en el planeta. Por ejemplo, en ambientes acuáticos, la abundancia de virus libres, excede a la de células procariotas por varios órdenes de magnitud (Bergh, BØrsheim, Bratbak & Heldal, 1989; Martha R.J. Clokie, Millard, Letarov & Heaphy, 2011; Robert A Edwards & Rohwer, 2005). Debido a su naturaleza como parásitos obligados, los virus que infectan bacterias (*bacteriófagos*) juegan un papel clave en la modulación de las comunidades microbianas en el ecosistema (G. F. Hatfull, 2008; Sharon y col., 2011).

El impacto ecológico global de los bacteriófagos, o sea virus que infectan bacterias y arqueas, se evidencia notablemente en la dinámica de los ciclos de nutrientes primarios en el océano. Por ejemplo, el ciclo del carbono es mediado por las cianobacterias, que corresponde a un grupo de microorganismos fotosintéticos que transforman el Carbono proveniente del Dióxido de Carbono (CO_2) en el aire a Carbono orgánico, constitutivo de la materia orgánica que es base de la cadena alimentaria del océano. Estos microorganismos son muy antiguos ya que se han registrado en fósiles de 2.7 billones de años y en la actualidad son muy abundantes (tanto como 10^5 bacterias por ml de agua), representando a la mitad de la biomasa fotosintética en algunas áreas (Hetherington & Raven, 2005). Diversos estudios han demostrado en ambientes marinos que cada día entre el $10 - 66$ % de los microorganismos que viven en estos ambientes son desintegrados por lisis celular producida por la acción de los bacteriófagos (Breitbart, Thompson, Suttle & Sullivan, 2007; Fuhrman & Noble, 1995; C. A. Suttle, 2007). Particularmente, la actividad lítica de cianófagos, o sea virus que infectan a cianobacterias, tienen un impacto directo en la cantidad de CO_2 fijado a nivel global.

Adicionalmente, evidencia obtenida desde estudios de la relación cianófagos-cianobacterias en ambientes marinos ha desafiado algunos conceptos básicos de la dinámica ecológica microbiana. Un claro ejemplo de esto corresponde a la capacidad sistemática de los cianófagos de retener en sus genomas una serie de genes con capacidad metabólica, conocidos como genes auxiliares, los cuales son obtenidos desde sus hospederos. Análisis genéticos preliminares revelaron la presencia de estos genes que no habían sido descrito anteriormente en genotipos virales y que cuya presencia hasta ese momento era inexplicable, ya que su función no poseía ninguna consecuencia aparente en el ciclo de vida del virus (Mann, Cook, Millard, Bailey & Clokie, 2003). Sin embargo, investigaciones posteriores han permitido identificar que estos genes auxiliares pertenecen a varias familias de genes que codifican proteínas relacionadas con fotosíntesis, ciclo del Nitrógeno y la reprogramación metabólica de los sistemas bacterianos para nitrógeno y azufre (Roux y col., 2016). Aún más importante, corresponde al hecho de que el acarreo de estos genes auxiliares les confiere una ventaja competitiva a aquellos cianófagos que los poseen, debido a que estos promueven procesos metabólicos en el hospedero para que este aumente su tasa de supervivencia, es decir, su adaptación al medio y ayude a crear un ambiente más propicio para la replicación viral (Lindell, Jaffe, Johnson, Church & Chisholm, 2005).

Desde hace muchos años, los bacteriófagos se consideran la materia oscura del mundo biológico debido a su gran abundancia, ubicuidad, población dinámica, diversidad genética y arquitectura genómica del tipo mosaico que los convierte en indescifrable (Graham F. Hatfull, 2015; Pedulla y col., 2003). El término bacteriófago comprende fagos templados, no templados y profagos defectuosos. Los fagos templados son bacteriófagos genéticamente capaces de exhibir ciclos tanto lisogénicos como productivos. El estado lisogénico se caracteriza por la incorporación del genoma viral en el cromosoma del huésped para convertirse en profagos que no resultan directamente en la producción y liberación de viriones (Bobay, Rocha & Touchon, 2012; Hobbs & Abedon, 2016). En cambio, una fracción de los profagos incorporan su cromosoma al hospedero en donde persisten hasta la activación por la ocurrencia de algún estrés, cambios químicos o físicos en el ambiente o por inducción espontánea. Posteriormente, los bacteriófagos se replican y lisan las células del huésped liberando nueva progenie viral (Bobay, Touchon & Rocha, 2014). Los fagos no templados son incapaces de tener ciclos lisogénicos. Esto incluye aquellos fagos que causan infecciones productivas en las que los viriones se liberan en periodos cortos de tiempo generando lisis de las células hospederas, y aquellos fagos que producen infecciones en las que los viriones se liberan durante largos intervalos de tiempo sin disrupción sustancial del hospedero, proceso conocido como infección crónica (Hobbs & Abedon, 2016). Además de esos, aquellos fagos que permanecen incompletos o que han perdido su capacidad de infectar, como los profagos defectuosos, los restos de profagos, los virus satélite y genes de fagos aislados, también se pueden encontrar en los genomas microbianos como resultado de la eliminación gradual o la racionalización genómica (*streamlining*), en que muchos de los genes que comprenden los profagos pueden eliminarse lentamente (Canchaya, Proux, Fournous, Bruttin & Brüssow, 2003).

Es evidente la relevancia de los bacteriófagos ambientales dada su relación con microor-

ganismos con alto impacto ecológico. Asimismo, una de las principales limitaciones de este campo corresponde a una gran carencia existente en el entendimiento de los fagos debido a dificultades técnicas asociadas con el aislamiento de estos y sus hospederos.

En el Capítulo 1 se exploran los diversos problemas que motivan esta memoria y se definen los objetivos del proyecto. En el Capítulo 2 se introducen y explican conceptos generales utilizados en el resto del trabajo. En el Capítulo 3 se detalla la solución propuesta y como está es un aporte substancial al campo de investigación bioinformática. En el Capítulo 4 se hace una breve validación de la primera parte de la solución que puede ser evaluada cuantitativa y cualitativamente, además de mostrar una aplicación experimental de la solución usando datos brutos reales junto a sus resultados. En el Capítulo 5 se entregan las conclusiones finales del trabajo, las ventajas y desventajas de la solución, junto a comentarios generales sobre esta.

Capítulo 1

DEFINICIÓN DEL PROBLEMA

Los virus son las entidades orgánicas más abundantes y diversas en el planeta. Existen aproximadamente 10 veces más virus que el número combinado total de organismos celulares, y la gran mayoría de esos virus son bacteriófagos (fagos), virus que infectan bacterias (Aziz, Dwivedi, Akhter, Breitbart & Edwards, 2015). Los bacteriófagos ocupan un rol crítico para la ecología y son clave en importantes descubrimientos en biología molecular. A pesar de su importancia el número actual de genomas de bacteriófagos completamente secuenciados es escasa en comparación con sus contrapartes los organismos celulares, y la información respecto a su abundancia y distribución en varios ecosistemas continúa siendo limitada. Como ejemplo del desconocimiento sobre la abundancia y distribución de bacteriófagos es que la prevalencia de dos fagos con presencia casi universal en los océanos (Zhao y col., 2013) y las heces humanas (Dutilh y col., 2014) eran parte de "la materia oscura" biológica (desconocimiento total) hasta hace muy recientemente. Al no poder reproducirse de forma libre, los bacteriófagos son parásitos estrictos u obligados, es decir, sólo pueden replicarse cuando estos ingresan al interior de una célula procariota huésped. Dentro de esta, el virus utiliza los recursos de la célula hospedera para duplicarse exponencialmente llevando a la destrucción (lisis) de ella y a la liberación de su progenie (nuevos bacteriófagos). Esta característica biológica determina que, para poder aislar un cultivo viral, es necesario poseer un cultivo bacteriano inicial que sea hospedero de este. Considerando que menos del 1 % de las bacterias que residen en el medioambiente han sido cultivadas en laboratorio, este requerimiento es sin duda la mayor limitación técnica que impide la mayor dispersión de estudios enfocados en bacteriófagos (Schloss & Handelsman, 2005).

Las estrategias tradicionales de recuperación tienden a subestimar la diversidad de bacteriófagos, mayoritariamente por qué los métodos basados en co-cultivo pierden la mayoría de los fagos. Los métodos tradicionales dependen de la detección de claros producidos por medio del co-cultivo en placas de fagos extraídos desde muestras ambientales junto a sus hospederos procariotas cultivables. Una vez el fago había generado lisis a su hospedero, este podía ser aislado (Lederberg & Lederberg, 1953). Sin embargo, cultivar fagos es experimentalmente demandante, dado que los fagos suelen requerir condiciones muy apropiadas y específicas para crecer, tales como suplementos químicos, temperatura y

medios de crecimiento específicos (Martha RJ Clokie, Kropinski & Lavigne, 2009). Además, fagos presentes en la muestra no necesariamente son capaces de infectar a alguna bacteria cultivable, perdiéndose en el proceso. O con fagos infectantes que no desintegran al hospedero inmediatamente, el hospedero tiene la posibilidad de desarrollar resistencia, derivando en placas nubosas o hasta ausencia total de señales físicas de infección (Hanna, Matthews, Dinsdale, Hasty & Edwards, 2012).

A pesar de que métodos modernos, basados en filtración, concentración y enriquecimiento de viriones por centrifugación ha mejorado la efectividad de la recuperación de partículas virales, una inmensa mayoría de fagos ambientales continúa siendo incultivable al día de hoy (Wommack, Williamson, Helton, Bench & Winget, 2009) y aun solo somos capaces de detectar aquellos virus que son parásitos de una bacteria cultivable.

Sumado a esto, la evaluación y clasificación de la diversidad de especies virales es mucho más compleja de lo que es con bacterias y otras células procariotas. Mientras las bacterias y arqueas pueden ser clasificadas por medio de marcadores genéticos como el gen 16S rRNA (presente en todas las bacterias), no existe gen compartido entre todas las especies virales, ni mucho menos específicamente en bacteriófagos, por lo que el estudio viral usando marcadores genéticos se limita a grupos virales específicos (Sullivan, 2015).

Dadas estas limitaciones, el estudio de la ecología viral actualmente es sumamente dependiente de métodos de análisis independientes de técnicas de cultivo, gracias al desarrollo de las tecnologías de secuenciación de siguiente generación (NGS), muchas estrategias basadas en la secuenciación y metagenómica han sido desarrolladas. Metagenómica que presenta nuevas problemáticas dada la dificultad de extracción de DNA utilizable para el proceso de secuenciación y evaluación de la diversidad taxonómica.

Existen protocolos y procedimientos bien establecidos para la investigación microbial. Mientras existen muchas metodologías para el procesamiento de muestras metaviromicas, todas siguen la misma estructura: Secuenciación, Ensamblado y Homología con bases de datos.

Los pasos de secuenciación y ensamblaje dependen de aspectos prácticos de la preparación de la muestra y análisis computacionales iniciales. Como ya se mencionaba, el proceso de aislación de un fago es experimentalmente demandante y muchos sesgos pueden introducirse accidentalmente, el muestreo de virus ambientales frecuentemente entrega muy poco material en comparación a librerías estándares de secuenciación. El filtrado, un aspecto importante de la preparación de muestras acuáticas tiene el potencial de excluir gran cantidad de virus dsDNA (Angly y col., 2006; Mohiuddin & Schellhorn, 2015; Steward y col., 2013). Adicionalmente, diversos tratamientos con diversos compuestos favorecen a ciertos tipos de virus tolerantes a estos compuestos, o aquellos que son abundantes en la muestra (López-Bueno, Rastrojo, Peiró, Arenas & Alcamí, 2015) causando una sobre representación de estos en algunos estudios. El almacenamiento que se les da a las muestras también puede excluir algunos virus ambientales, que decaen a diferentes velocidades con el tiempo y la temperatura (Angly y col., 2006).

Después de la secuenciación y antes de comenzar los análisis, las secuencias crudas deben ser inspeccionadas para para remover artefactos. Si bien las plataformas de secuenciación modernas tienen bajos índices de error, algunos sesgos pueden ocurrir. Diversas herramientas de código abierto se han desarrollado para facilitar el control de calidad (p. ej. khmer (Crusoe y col., 2015)), pero no están exentas de error.

Otro problema importante es la clasificación de una secuencia proveniente de un fago. Ya sea intentando identificar la taxonomía presente o la funcionalidad putativa de la región codificante, todo depende de la disponibilidad de una secuencia representativa en la base de datos usada como referencia. Los metagenomas virales son dominados por secuencias desconocidas, entre un 63 y 93 % de las *reads* carecen de anotaciones taxonómicas o funcionales (Hurwitz & Sullivan, 2013). A Julio del 2020, la base de datos de nucleótidos del NCBI (nuccore)[1] reporta aproximadamente 12 mil secuencias virales, < 1 % en comparación a las más de 20 millones de secuencias bacterianas. Es por esto, que los análisis metaviromicos dependen de las secuencias disponibles en bases de datos y del limitado número de especies caracterizadas. La pequeña fracción correspondiente a fagos que ha sido secuenciada también tiene un sesgo inherente. Fagos que infectan hospederos comúnmente encontrados (p. ej. *Pseudomonas* y *Enterobacteriaceae* como la Salmonella y la Escherichia Coli) están significativamente sobre-representados. Este desbalance en las bases de datos presenta un serio desafío al momento de asignarle importancia a secuencias virales nuevas, previamente no clasificadas (Angly y col., 2006; López-Bueno y col., 2015).

Incluso con secuencias ya conocidas, se debe guiar el estudio bajo gran cuidado. Por ejemplo, una inspección viral realizada en el Lago Michigan mostró que mientras una larga fracción de las *reads* recuperadas indicaban a un único gen del fago *Planktothrix phage PaV-LD*, otros dos genes específicos a este fago fueron encontrados con muy baja frecuencia (Bruder y col., 2016). Esta evidencia enfatiza el hecho de que la identificación por homología de un gen viral específico no indica la presencia de dicha especie viral en la muestra, sino solo la presencia de dicho gen específico (S. C. Watkins & Putonti, 2017).

Es más, muchas anotaciones en bases de datos actualmente son incorrectas. Encontrarse con secuencias incompletas o mal caracterizadas en bases de datos públicas es bastante común. Por ejemplo, una búsqueda por homología (blastx) de DNA de partículas virales purificadas contra la base de datos GenBank (Benson y col., 2012) produjo resultados con asignaciones no virales (Rosario, Nilsson, Lim, Ruan & Breitbart, 2009). Asimismo, un significativo grupo de fagos se integran al genoma hospedero en un estado conocido como *lisogenia*, formando "profago". Los profagos suelen pasar desapercibidos durante exámenes metaviromicos, estando muchos de ellos mal clasificados en bases de datos como genes bacteriales (Casjens, 2003; G. F. Hatfull & Hendrix, 2011).

Hasta ahora una gran escala de estudios en ecología viral se han concentrado en los océanos, la metagenómica de fagos se inició en el ambiente marino (Breitbart y col., 2002),

[1] Estadisticas tomadas de RefSeq https://www.ncbi.nlm.nih.gov/refseq/statistics/ (Pruitt, Tatusova & Maglott, 2005)

esencialmente donde los bacteriófagos marinos juegan roles importantes en la modulación de comunidades que generan consecuencias globales en el carbono oceánico (Bruder y col., 2016; C. Suttle, 2005; Curtis A Suttle, 2005). A pesar de su relevancia, por mucho tiempo estudios relacionados a la microbiota asociada a medios acuáticos fueron enfocados principalmente a ambientes marinos, siendo la información ecológica y funcional de los ambientes de agua dulce relegados a un segundo plano. Sin embargo, gracias a la aplicación de herramientas moleculares a un mayor número de ambientes, la ecología microbiana asociada a biomas de aguas dulce ha emergido como un campo de estudio de mucho interés. Es así, como un vasto cuerpo de evidencia indica que las bacterias que habitan ambientes de agua dulce, tales como lagos, lagunas, estuarios y ríos juegan un rol importante en la regulación de los ciclos biogeoquímicos. Esto se debe a que las bacterias corresponden a los principales degradadores y mineralizadores de compuestos orgánicos, constituyéndose así como un factor fundamental de la producción de biomasa y acoplamiento trófico a los depredadores eucariotas, que, al alimentar la red alimentaria, tiene un profundo impacto en los flujos de energía y la calidad del agua (Cole, Findlay & Pace, 1988). Estos estudios moleculares han permitido aumentar sustantivamente la información y conocimiento respecto a la ecología, la ecofisiología y la distribución de bacterias y arqueas en estos ambientes. En general, los lagos de agua dulce están íntimamente relacionados con los biomas terrestres y marinos a través del ciclo hidrológico y otros modos de dispersión. Es así que el estudio liderado por (Newton, Jones, Eiler, McMahon & Bertilsson, 2011a) se estableció que especies pertenecientes a los filos Actinobacterias, Bacteroidetes, Alphaproteobacteria y Betaproteobacterias parecen corresponder a especies nativas, y no de tránsito, asociadas a la epilimnion (zona superficial de mayor temperatura) (Allgaier & Grossart, 2006; Newton y col., 2011a). Otro estudio que incorporó análisis microbianos de 13 biomas de agua dulce (lagos y ríos) indicó que las divisiones microbianas más abundante en estos ambientes Proteobacterias (subdivisiones alfa y beta), el grupo Cytophaga-Flavobacterium-Bacteroides, Actinobacterias y Verrucomicrobia (Zwart, Crump, Agterveld, Hagen & Han, 2002).

La metagenómica viral le ha cambiado radicalmente la cara al descubrimiento viral, permitiendo la identificación de secuencias virales sin necesidad de aislar los virus. El siguiente paso es comprender las interacciones dentro de los sistemas fago-hospedero, siendo estas las que guían los ciclos biogeoquímicos más importantes del planeta. Las pregunta fundamental es: ¿A que hospedero infecta este virus? Cuantificar quien infecta a quién es esencial para comprender cómo las infecciones a nivel celular, escalan para influenciar la función de los ecosistemas en ambientes complejos (Robert A. Edwards, McNair, Faust, Raes & Dutilh, 2015; Weitz y col., 2013). La diversidad de la virosfera global y los volúmenes de información generados por proyectos de secuenciación metagenómica demandan herramientas computacionales (*in-silico*) para la predicción de relaciones fago-hospedero.

Como ha sido discutido en este capítulo, el impacto ecológico de bacteriófagos en otros ambientes acuáticos menos explorados y su coevolución con hospederos microbianos es aún un campo que requiere mayor atención, aún existe muy poca información respecto a cómo los fagos interactúan con comunidades microbianas en ambientes de agua dulce, la escasez de secuencias genómicas de fagos caracterizadas limita la capacidad de clasificar

la mayoría de secuencias metavirales. Los fagos son extremadamente plásticos y son capaces de transferir genes entre organismos (Canchaya, Fournous, Chibani-Chennoufi, Dillmann & Brüssow, 2003), por lo que observar únicamente los resultados de homología a un único gen puede indicar equivocadamente la presencia de una especie. Carentes de marcadores genéticos, asegurar la presencia de una especie en un metagenoma viral es un desafío aún pendiente (Bruder y col., 2016).

Es más, la exploración de sistemas fago-hospedero en comunidades virales complejas, presentes naturalmente, se limita completamente por la calidad y cantidad de información disponible. Diversificar, balancear y mantener la información disponible en bases de datos es un desafío multidisciplinario necesario de abordar desde las herramientas biológicas y de tecnologías de la información, dadas las últimas tendencias en ciencias de datos, estas metodologías *in-silico* puede adaptarse al estudio metaviromico. Cavar más profundo en los patrones de distribución de fagos en ecosistemas de agua dulce y contrastándolos con aquellos marinos mejorará nuestro entendimiento de las estrategias biológicas y ecológicas a través de diversos ambientes y proveerá conocimiento en el porqué algunos fagos tienden a estar equitativamente distribuidos globalmente cuando otros son endémicos a ambientes específicos (Thurber, 2009).

1.1. Objetivos

1.1.1. Objetivo general:

Desarrollar un protocolo o framework que permita, a través de la aplicación de métricas basadas en el análisis de metagenomas obtenidos desde ambientes acuáticos continentales, observar patrones de coexistencia de genes pertenecientes a especies virales y procariotas (bacteria y arqueas) en estos ambientes.

1.1.2. Objetivos específicos:

- Describir y desarrollar un procedimiento jerarquizado basado en herramientas bioinformáticas que permita obtener conocimiento acerca de posibles interacciones entre fago-hospedero en ambientes acuáticos continentales.
- Desarrollar y aplicar una heurística para determinar la presencia de genes pertenecientes a especies virales y procariotas en muestras de metagenomas obtenidas desde ambientes acuáticos continentales.
- Extraer conocimiento sobre la co-existencia e interacciones entre especies virales y procariotas en estos ambientes a través de la visualización de grupos de interés.

Entendiendo las dificultades mencionadas, el propósito de esta memoria es desarrollarse como una colección de herramientas y repetibles y abiertos, sin ostentar lograr un

descubrimiento relevante, sean de utilidad para la normalización de procedimientos y procesos de investigación al establecer un flujo centralizado de análisis y datos, que ante la disponibilidad de herramientas de mejor rendimiento y calidad, puedan ser fácilmente integradas, reemplazando aquellas originales presentadas en esta memoria de forma simple y compatible.

Es decir, la interoperabilidad y compatibilidad de las herramientas presentadas en el Capítulo 3 es un factor a considerar con igual o mayor relevancia que la calidad de los resultados de la aplicación de estas mismas.

A su vez, como ejemplificación del protocolo presentado, se explora un caso de interés real en el Capítulo 4.

Capítulo 2

MARCO CONCEPTUAL

2.1. Conceptos Generales

2.1.1. Bacteria

Las bacterias son microorganismos unicelulares que carecen de núcleo (procariotas). Las bacterias son una de las primeras formas de vida en aparecer en la Tierra y están presentes en la mayoría de los biomas del planeta, incluyendo sedimentos, agua dulce, hasta el interior de géiseres y volcanes y muy profundamente en la corteza terrestre.

Las bacterias no solo viven de forma libre en el ambiente, sino también pueden residir de forma simbiótica con organismos superiores como plantas y animales. En humanos y animales, una gran cantidad de comunidades bacterianas existen en el tracto intestinal y muchas de estas son fundamentales en el procesamiento y absorción de nutrientes. Adicionalmente, existen algunos grupos microbianos que al encontrarse en mayor proporción pueden tener efectos perjudiciales a la salud y causar enfermedades. Las bacterias también son importantes para aplicaciones industriales y medioambientales por medio de la biotecnología, tal como el tratamiento de aguas servidas y la contención de derrames de petróleo, la producción de alimentos como yogurt y queso y la recuperación de minerales en procesos mineros (for General Microbiology, 2010).

Siendo uno de los tipos de organismos más antiguos y abundantes de la Tierra, las bacterias son vitales para el funcionamiento del planeta. Un ejemplo claro de los efectos a nivel global ocurre en los sedimentos marinos en donde el equilibrio entre la generación de la materia orgánica y la mineralización de ésta en los sedimentos tiene implicaciones importantes para la concentración de O2 y CO2 de la biosfera y, a escalas de tiempo geológicas, un profundo impacto en las condiciones climáticas y la vida tal como la conocemos hoy en día (Hamilton, Bryant & Macalady, 2016). Estos mismos efectos se han registrado en una serie de otros ciclos biogeoquímicos, tales como fotosíntesis llevada a cabo por cianobacterias, reducción de metales llevada a cabo por bacterias metaloreductoras, entre otros.

Las bacterias poseen una gran variedad de formas y tamaños. Las células bacterianas,

al carecer de núcleo, son mucho más pequeñas que sus contrapartes eucariotas (cerca de una décima parte del tamaño promedio), en el rango de 0.5 a 5 micrómetros de largo, salvo excepciones que pueden llegar a ser extremadamente grandes o pequeñas (Williams, 2011).

2.1.2. Arquea

Al igual que las bacterias, las arqueas son microorganismos del tipo procariota. A diferencia de las bacterias, las arqueas están más relacionadas a los microorganismos del tipo eucariotas que a las bacterias. La gran mayoría de arqueas no han sido aisladas en condiciones de laboratorio y muchas de estas han sido detectadas por medio de análisis moleculares.

Las arqueas son microorganismos unicelulares morfológicamente muy similares a las bacterias y poseen un potencial metabólico y condiciones fisiológicas relativamente similares a estas, siendo muy común la co-ocurrencia de ambos dominios en el ambiente.

El hallazgo de las primeras arqueas ocurrió en ambientes extremófilos, los que son biomas comunes para estas especies, que incluyen ambientes extremos tales como géiseres o salares. Actualmente, su alta prevalencia se ha expandido también a los océanos, en donde participan de los ciclos de carbono y nitrógeno y son un importante miembro de la microbiota humana tanto en el tracto digestivo y la piel (Bang & Schmitz, 2015).

2.1.3. Virus

Los virus son, por lejos, las entidades biológicas más abundantes en el planeta. Los virus están constituidos por partículas llamadas viriones. El virión está compuesto por ADN o ARN encapsulado en una envoltura de varias subunidades estructurales oligoméricas hechas de proteínas llamada cápsides. Esta puede poseer distintas morfologías, aunque la mayoría de los virus poseen cápsides con estructura helicoidal o icosaédrica. La cápside posee el mecanismo de penetración y transferencia desde genoma viral hacia la célula hospedera.

Una de las clasificaciones más comunes utilizadas en virus corresponde a la clasificación de Baltimore, la considera las características del sistema genético que permite la replicación de su genoma (Baltimore, 1971). En particular. la clasificación de Baltimore incluye los siguientes parámetros relevantes: (i) tipo de ácido nucleico (ADN o ARN), (ii) tipo de hebra (hebra simple o doble), (iii) sentido de la hebra y (iv) el método de replicación.

Aquellos virus que son capaces de infectar bacterias son los más abundantes en el planeta y son llamados bacteriófagos. El ciclo de vida de los bacteriófagos funciona como un buen modelo para comprender los aspectos biológicos que determinan la ecología de estos grupos en el ambiente, muchos de los cuales poseen aspectos similares a los observados en hospederos animales.

Durante una infección, un fago se adhiere a una bacteria o arquea e inserta su material

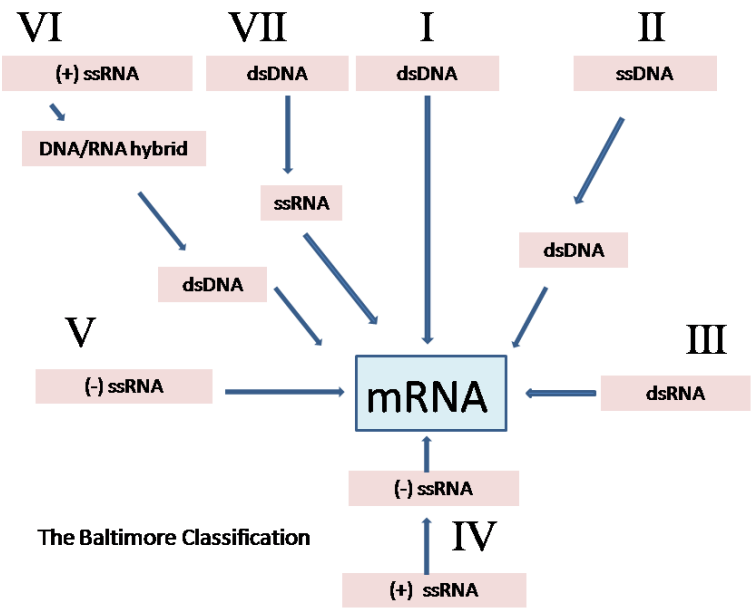


Figura 21: La clasificación de virus de acuerdo con Baltimore (Baltimore, 1971), la que ubica a los virus en uno de los siete grupos según la combinación de los siguientes parámetros relevantes: (i) ácido nucleico (ADN o ARN), (ii) tipo de hebra (hebra simple o doble), (iii) sentido de la hebra y (iv) el método de replicación.

genético en la célula. Después de eso, el fago proseguirá uno de los dos ciclos de vida, lítico (virulento) o lisogénico (templado). Los fagos líticos inmediatamente toman control de la maquinaria de la célula hospedera para replicar el ADN viral, fabricar los componentes del fago y ensamblarlos en viriones. Posterior a esto, la célula hospedera procede a destruirse debido a la acción de enzimas conocidas como endolisinas, que son capaces de degradar las complejas estructuras de la pared celular de sus huéspedes bacterianos, liberando así a los viriones (progenie viral) hacia el ambiente para asi poder encontrar nuevos hospederos para infectar.

Luego destruyen o lisan la célula, liberando nuevas partículas de fago. Los fagos lisogénicos incorporan su ácido nucleico en el cromosoma de la célula huésped y se replican con él como una unidad sin destruir la célula. Bajo ciertas condiciones, los fagos lisogénicos pueden ser inducidos a seguir un ciclo lítico.

Los fagos templados son bacteriófagos que tienen dos estados durante su ciclo de vida. El estado lítico consiste en la replicación y posterior lisis de las células huésped procariotas que liberan nueva progenie viral. Por el contrario, el estado lisogénico se caracteriza por la incorporación del genoma viral en el cromosoma del huésped para convertirse en profágicos (fagos integrados)

Los bacteriófagos pueden seguir dos tipos de formas de desarrollo, fase lítica o fase lisogénica. Los fagos virulentos poseen solo fase lítica. Esta fase se caracteriza por que los

virus inyectan su material genético en la célula hospedera y comienzan a producir proteínas virales y copias del genoma viral usando los aparatos biosintéticos del hospedero que usualmente termina provocando su muerte por medio de la lisis celular, y una vez completado el ciclo, expulsando la progenie viral al exterior. En contraste, los fagos temperados poseen tanto fase lítica como también fase lisogénica, Esta última se caracteriza por que virus integra su material genético al cromosoma del hospedero, y es replicado junto con el, hasta el momento en que son inducidos a crear su progenie, la cual es expulsada al exterior por medio de lisis celular.

2.1.3.1. Fase Lítica

Una de las posibles rutas de reproducción que puede tomar un fago es la fase lítica. Durante esta fase, el fago toma control de la maquinaria metabólica de la célula hospedera, se reproduce en nuevos fagos y termina destruyendo la célula. Hay cinco etapas en el ciclo lítico del bacteriofago: fijación, penetración, biosíntesis, maduración y lisis. Durante la etapa de fijación, el fago interactúa con receptores especiales en la superficie de las bacterias. En general, los fagos son específicos con las bacterias que infectan debido principalmente a la especificidad de estos receptores. La segunda etapa, la penetración consiste en la contracción de la funda de la cola del fago, que actúa como una aguja hipodérmica que inyecta el genoma viral a través de la membrana celular del hospedero.Durante este proceso la cabeza del fago permanece fuera de la bacteria. La tercera etapa es la síntesis de nuevos componentes virales. Después de ingresar al hospedero, el virus secuestra los mecanismos de la célula para replicar, transcribir y traducir los componentes y genes virales necesarios para ensamblar nuevos virus. Durante la cuarta etapa, la Maduración, nuevos viriones son creados. En la etapa final, los virus maduros rompen la pared celular en un proceso llamado lisis y son liberados al ambiente.

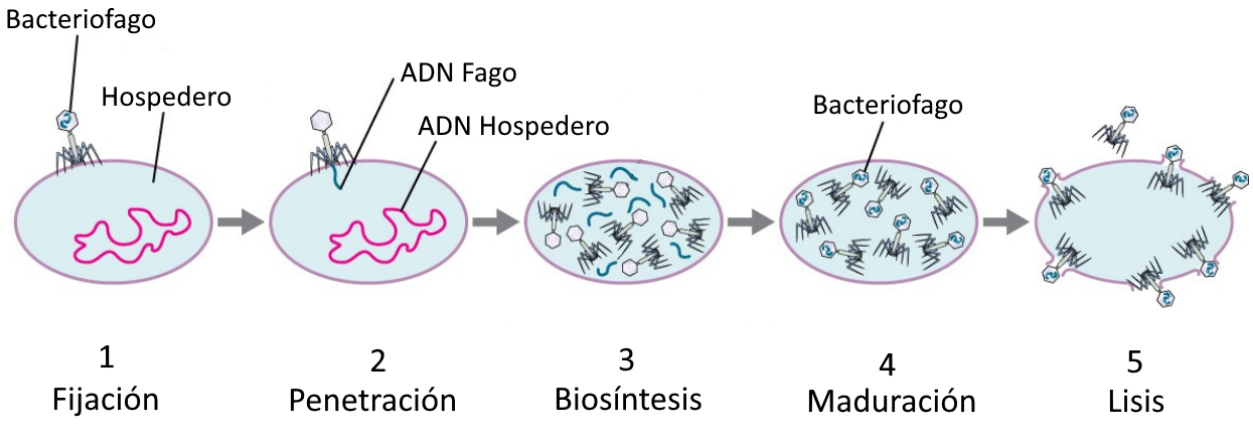


Figura 22: Ciclo Lítico («The Viral Life Cycle», s.f.)

2.1.3.2. Ciclo Lisogénico

En el ciclo lisogénico, las etapas de fijación y penetración son idénticas a las del ciclo lítico, donde el fago inserta su genoma en la célula hospedera, pero a diferencia de éste, el

genoma del fago se integra al cromosoma del hospedero, formando un profago, en vez de producir la lisis celular. Una célula hospedera con un profago es denominada lisogeno y el proceso de infección por un fago temperado se denomina lisogenia. A medida que la bacteria replica su genoma, también replica el ADN del fago y lo traspasa a la siguiente generación. La presencia del profago puede alterar el fenotipo del hospedero al disturbar genes que existían en el genoma, o incorporar nuevos genes y/o nuevas posibles funciones al hospedero. El profago en la bacteria persiste hasta que producto de una señal ambiental, por ejemplo un cambio abrupto en el ambiente o estrés, se produce la inducción, etapa que resulta en la expulsión del genoma viral desde el cromosoma hospedero. Luego de la inducción, el fago temperado procede a finalizar el ciclo lítico, produciéndose la lisis de la célula hospedera.

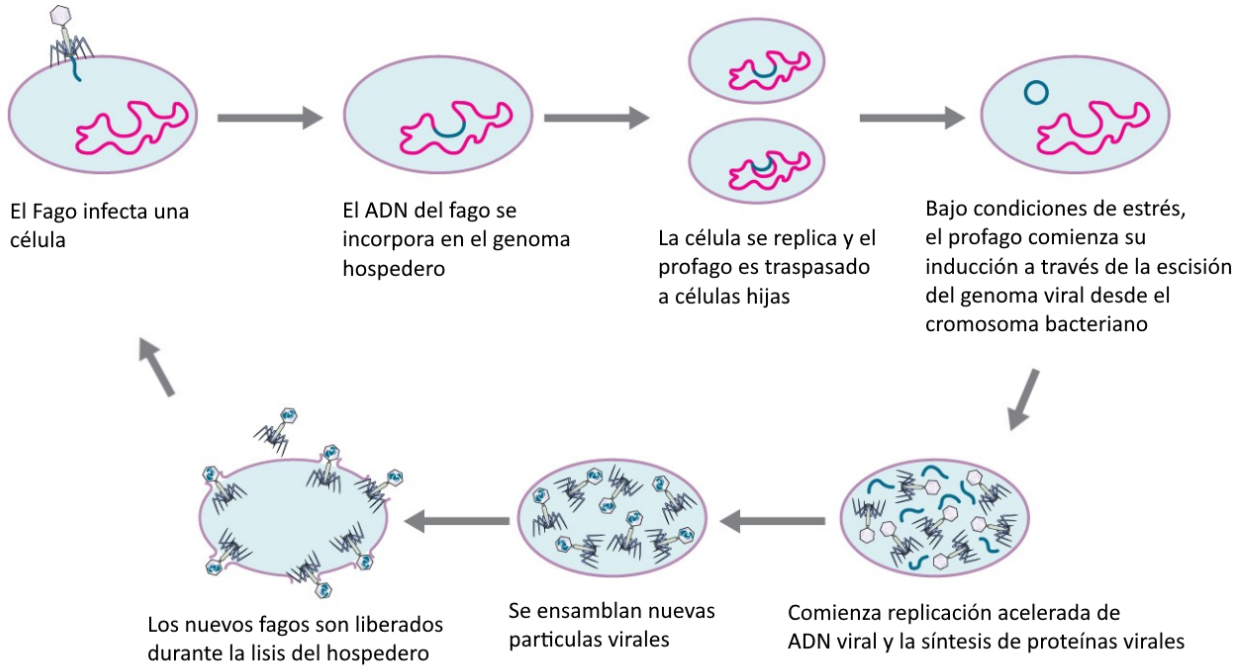


Figura 23: Ciclo Lisogénico («The Viral Life Cycle», s.f.)

2.1.4. ADN

El Ácido Desoxirribonucleico (ADN) es una macromolécula compuesta de dos hebras que se enrollan encima de la otra para formar una doble hélice. Esta molécula almacena y transmite instrucciones para el desarrollo, funcionamiento, crecimiento y reproducción de todos los organismos vivos y muchos virus.

La información contenida en el ADN se encuentra codificada en forma de combinaciones de cuatro bases químicas: Adenina (A), Guanina (G), Citosina (C) y Timina (T). La secuencia en que estas bases son dispuestas determina la información disponible para construir y mantener un organismo, de manera similar a como las letras de un alfabeto aparecen en orden para formar palabras.

Debido a su estructura química, estas cuatro bases son conocidas como ”bases nitrogenadas” dado que cada macromolécula está compuesta de moléculas de Nitrógeno (N), las que se combinan con moléculas conocidas como ”desoxirribosas” (tipo de azúcar) y ”fosfatos” para conformar lo que es técnicamente un nucleótido. Es importante destacar que los azúcares se unen químicamente con los fosfatos de otro nucleótido para formar cadenas que conforman cada una de las hebras del ADN. Para mayor simplicidad, existe la convención de que las hebras de ADN son anotadas usando la letra de la base nitrogenada del nucleótido desde una dirección “5’ a 3’ ”, siendo el extremo 5’ el que frecuentemente contiene un grupo fosfato unido al carbono 5’ del anillo de ribosa, y el extremo 3’, el que típicamente no se modifica.

Las bases nitrogenadas de cada nucleótido se concatenan químicamente unas con otras de acuerdo a las ”reglas de emparejamiento” (A con T y C con G), formando el ADN de doble hebra (dsDNA o Double-Stranded DNA), también conocido como ”modelo de doble hélice”. Ambas hebras de la doble hélice son complementarias (dadas las reglas de emparejamiento) y comparten la misma información biológica . Esto es importante debido a la capacidad de este modelo de replicarse al momento en que ambas hebras son separadas, cada hebra volverá a emparejarse con el complemento correcto generando copias exactas del modelo original.

2.1.5. ARN

El ARN es otra macromolécula conformada por ácidos nucleicos que al igual que el ADN, posee la capacidad de contener y transmitir información biológica. Sin embargo, el ARN posee varias características que lo distinguen del ADN. Las moléculas de ARN corresponden a cadenas de una sola hebra que no se enrollan para formar una doble hélice. Sin embargo, producto de la naturaleza fisicoquímica de estas moléculas poseen la tendencia de que los nucleótidos se emparejen con nucleótidos complementarios. Una analogía que representa muy bien este fenómeno, corresponde a la representación de la molécula de ARN como un trozo de scotch que está libre en una oficina. Debido a su naturaleza, es muy probable que luego de un rato ese scotch termine pegado a algo más, no dejando ninguna parte del pegamento expuesto al aire. Otro aspecto que diferencia al ARN del ADN, es que la base complementaria a la adenina en el ARN es el uracil, a diferencia de la timina en el ADN. Una última característica diferente corresponde a que el ARN contiene ribosa, mientras que el ADN contiene desoxirribosa (Shukla, 2014).

2.1.6. Proteína

Las proteínas son macromoléculas que realizan la mayoría del trabajo estructural, funcional y regulador de las células. Están compuestas de cientos de pequeñas unidades conocidas como ”aminoácidos”, que son concatenados en largas cadenas. Existen 20 tipos diferentes de aminoácidos cuya combinatoria construye cualquier tipo de proteína, definiendo su estructura física, localización y función biológica.

Las proteínas son ensambladas con aminoácidos usando la información codificada en los

genes. Cada proteína tiene su propia secuencia única de aminoácidos que es especificada por la secuencia de nucleótidos en el gen. El código genético es la suma de información contenida en unidades de tres nucleótidos llamados codones, los que configuran un aminoácido. Dado que el ADN contiene cuatro nucleótidos en total, el total posible de codones es 64, con algunos aminoácidos siendo redundantes.

En general, las proteínas son unidades altamente abundantes en las células microbianas, conformando cerca de 50 % del peso seco de estas (O'Connor, Adams & Fairman, 2010)

Se estima que una bacteria promedio, por ejemplo Escherichia coli y Staphylococcus aureus, contiene alrededor de dos millones de proteínas por célula . Se estima que bacterias más pequeñas contienen menos moléculas, desde cincuenta mil hasta un millón.

2.2. Disciplinas que estudian a los microorganismos en el ambiente

2.2.1. Ecogenomica Ambiental

El nuevo campo de la ecogenómica busca comprender los fundamentos de la adaptación y la variación fenotípica mediante el uso de técnicas genómicas, enfocándose en las bases de interacción entre las especies, e identificando aquellos genes afectados por la evolución (Ouborg & Vriezen, 2007).

Esta disciplina utiliza un conjunto de técnicas y metodologías, experimentales y computacionales, enfocadas primordialmente en el estudio y modelamiento de las comunidades microbianas y su impacto en el funcionamiento del ecosistema. Esta disciplina ha tenido un alto impacto en distintas áreas de estudio, incluyendo desde el estudio de procesos biogeoquimicos a nivel global, hasta el estudio de procesos de biodegradación de contaminantes, los cuales pueden ser utilizados como una solución biotecnológica para pasivos ambientales (sitios contaminados). Debido a la imposibilidad de cultivo de una gran diversidad de microorganismos, las técnicas independientes de cultivos han tenido una importante contribución al área de ecogenomica ambiental. Entre estas técnicas, el desarrollo e implementación de métodos basados en secuenciación de moléculas como ADN y ARN, extraídas directamente desde muestras ambientales (metagenomica y metatranscriptomica) han sido claves para una interpretación mas completa y robusta del funcionamiento ecológico.

De manera análoga, métodos de espectrometría en masa para la identificación de pequeñas moléculas han avanzado en la identificación de proteínas (metaproteomica) y metabolitos (metabolómica) desde muestras ambientales complejas. Estas y otras técnicas combinadas e integradas en protocolos ecológicos apoyados en herramientas y capacidades computacionales y estadísticas, han sustentado robustamente el surgimiento del campo de ecogenómica ambiental.

2.2.2. Genómica

La genómica es el campo de la biología que se enfoca en el estudio de la estructura, función y evolución del material genético de los organismos. Un genoma es el conjunto de ADN completo de un organismo único y por extensión, con genómica nos referimos al estudio del ADN de los organismos de forma individual.

El rápido declive en el costo de la secuenciación genética, hizo posible generar secuencias genómicas para una gran variedad de organismos. El primer genoma microbiano secuenciado fue publicado en 1995 (Fleischmann y col., 1995). Desde entonces el total de genomas microbianos secuenciados ha ido incrementando de forma exponencial. Organismos patogénicos causantes de enfermedades han recibido gran atención, pero muchas arqueas y bacterias que han sido secuenciadas incluyen organismos beneficiosos como diversas especies de *Prochlorococcus* y *Synechococcus*, los mayores productores de oxígeno en los océanos.

La genómica busca reunir toda la información que necesita sobre un organismo determinado, enfocándose en el análisis en todos los genes de un genoma específico. En contraste, estos nuevos recursos también promueven comparaciones de genomas completos de especies distintas, lo que sienta las bases de un nuevo campo de estudio llamado genómica comparativa.

2.2.2.1. Aislamiento

La técnica tradicional para el estudio de microorganismos consiste en la aislación del organismo de interés en un medio de cultivo bajo condiciones de laboratorio. El cultivo microbiano ha sido y sigue siendo el principal método de estudio de la fisiología de los microorganismos y su capacidad de adaptarse a nuevos desafíos medioambientales.

2.2.3. Metagenómica

El término ”metagenómica” no es nuevo, pero no fue hasta muy recientemente cuando se define como ”la aplicación de técnicas genéticas modernas sin necesidad de la aislación y cultivo de especies individuales ni amplificación de un gen específico” (Chen & Pachter, 2005). Desde entonces, el estudio microbiológico entró de lleno en la era de la ciencia de datos, donde la materia prima del análisis es un gran volumen de datos brutos en forma de secuencias (cadenas de caracteres).

Con el advenimiento de las tecnologías modernas de secuenciación de siguiente generación (Next Generation Sequencing (NGS)) y el desarrollo de las tecnologías informáticas, se ha ampliado la visión y alcance de la investigación microbiológica en medio ambientes complejos. Como se explicaba previamente, el proceso para comprender la diversidad de una comunidad microbiológica comenzaba con el intento de aislar y cultivar la mayor cantidad de organismos posibles desde una muestra, para luego identificar taxonómicamente cada individuo para entonces estudiar sus características metabólicas. Posteriormente, y de la mano de la aparición de técnicas de secuenciación, este proceso ha surgido como una

aproximación capaz de incorporar distintas capas de información (genómica, fisiológica, metabólica y ecológica) integrandolos y conformando así análisis más robustos.

2.2.3.1. Flujo de un proyecto metagenómico

Toma de muestras y extracción de ADN El procesamiento de las muestras es el primer y más crucial paso en cualquier proyecto metagenómico. El ADN extraído debería ser representativo de todas las células presentes en la muestra y disponer de cantidades suficientes de ácidos nucleicos de alta calidad para enviar a secuenciar.

Secuenciación Las tecnologías de secuenciación de siguiente generación (NGS) ofrecen bajas tasas de error a un relativo bajo costo, haciéndolas la elección más popular para secuenciación metagenómica *shotgun*.

Ensamblado Si la investigación necesita recuperar el genoma completo de un organismo incultivable, o al menos obtener regiones codificantes de genes (CDS) con el afán de obtener una caracterización más profunda que una descripción funcional de la comunidad, se debe realizar el ensamblado de los fragmentos obtenidos desde la secuenciación para obtener secuencias más largas llamadas contigs. Existen dos estrategias utilizadas por los softwares de ensamblado: Basados en referencia, que aprovechan el conocimiento de un genoma conocido para unir los fragmentos secuenciados, y los ensamblados *De novo*, que son capaces de unir fragmentos sin necesidad de referencia utilizando metodologías estocásticas. Esta última es la preferida en metagenómica dada la inexistencia de referencias al momento de estudiar comunidades de organismos desconocidos, incultivables y heterogéneos.

Binning Se refiere al proceso de agrupar secuencias de ADN en grupos que podrían representar el genoma de un individuo particular o genomas de organismos cercanamente relacionados.

Anotación Taxonómica y Funcional Este es el proceso de etiquetar las secuencias como genes o elementos genómicos, y asociarlas a un organismo o grupo de organismos particular (unidad taxonómica operacional, OTU). Diversos algoritmos y herramientas basados en diferentes estrategias se encuentran disponibles, pero todos representan un desafío computacional, estadístico y práctico, principalmente por que la unica forma de anotar metagenomas recae en la comparación por homología frente a datos conocidos. Muchas bases de datos se encuentran disponibles para dar contexto funcional y taxonómico a datasets metagenómicos. Sin embargo ninguna cubre todas las posibilidades, además de encontrarse implícitamente sesgadas organismos y elementos genéticos de alta presencia y fácil estudio como bacterias y arqueas que poseen marcadores genéticos comunes además de ser cultivables en condiciones de laboratorio.

Análisis estadístico Uno de los objetivos definitivos de la metagenómica es unir la información funcional y filogenética con la química, física y otros parámetros biológicos

que caracterizan un ambiente. A pesar de que medir todos estos parámetros es costoso en tiempo y dinero, permite realizar análisis correlativos y retrospectivos de data metagenómica que no inicialmente no eran parte de los objetivos del proyecto pero que pueden ser de interés para otras preguntas investigativas. El valor de tal metadata no debe ser subestimado y se ha vuelto parte importante de la deposición de data metagenómica en bases de datos (Markowitz y col., 2008; Sun y col., 2011).

2.2.3.2. Secuenciación total Shotgun

La secuenciación shotgun es un método utilizado para secuenciar hebras aleatorias de ADN. se hace la analogía con el patrón obtenido al disparar una escopeta, de expansión veloz, cuasi-aleatoria. En la secuenciación shotgun, el ADN es fragmentado de forma aleatoria en numerosos segmentos pequeños que son secuenciados usando el método Sanger para obtener *reads*. Se obtienen múltiples reads solapantes repitiendo el proceso por varias veces. Posteriormente estos fragmentos solapados son ensamblados en secuencias más largas usando software ensamblador (Anderson, 1981; Staden, 1979). Esta es una de las tecnologías precursoras que han permitido la secuenciación de genomas completos, superando los inconvenientes de la secuenciación Sanger a un relativo bajo costo, aprovechando el posterior procesamiento computacional.

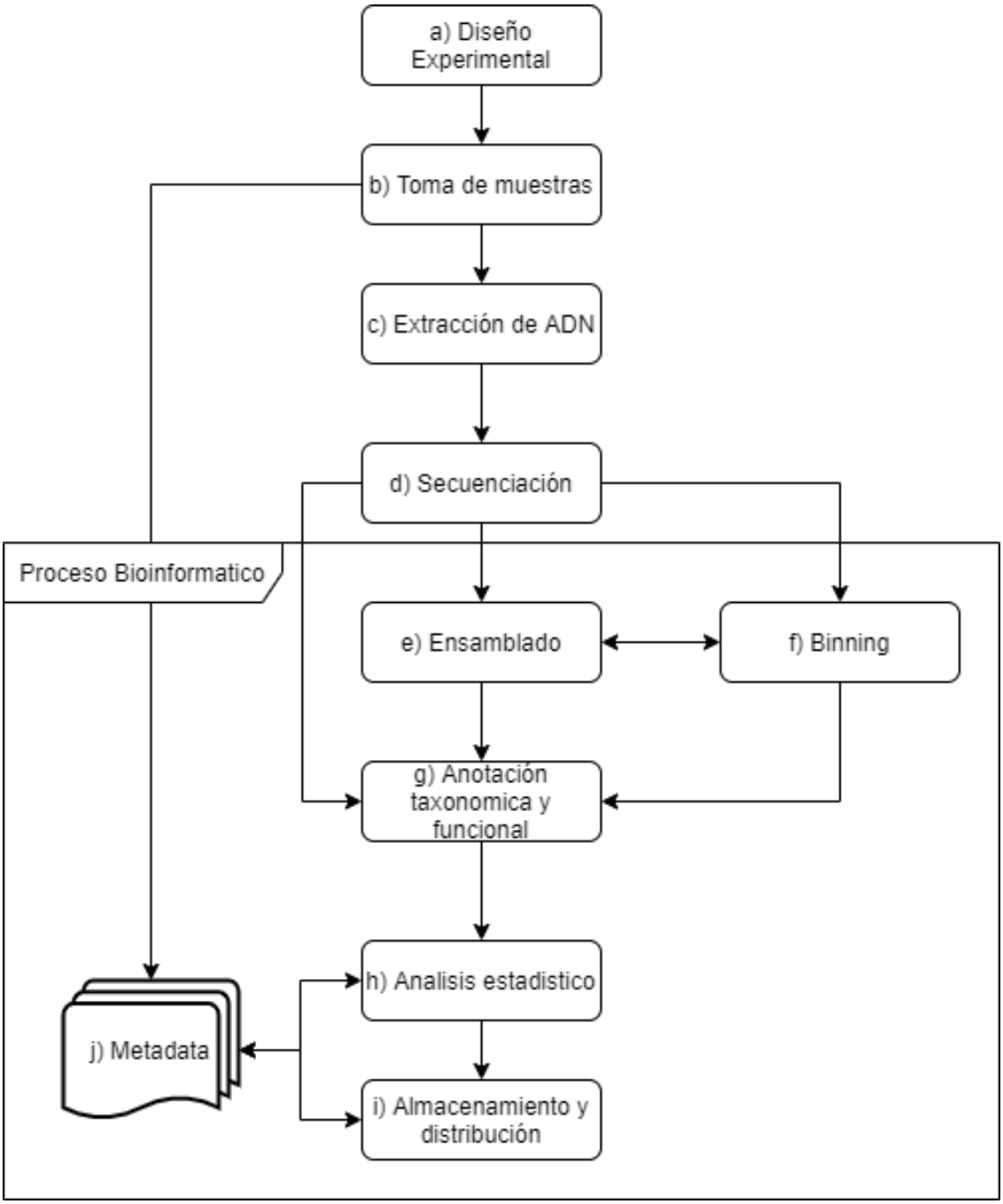


Figura 24: Pasos comunes de un proyecto metagenómico, extraído de (Thomas, Gilbert & Meyer, 2012). Los pasos a), b), c) y d) corresponden a procesos experimentales de extracción y secuenciación del material genético a estudiar. Una vez obtenidas las secuencias se procede a la etapa de análisis usando herramientas computacionales, pasos e), f), g), h) e i).

2.3. Herramientas Computacionales

"Ciencia de Datos" es un término que ha aparecido en diversos contextos en diversas disciplinas. Sólo recientemente se ha establecido como una agrupación de técnicas, procedimientos, algoritmos y sistemas para extraer conocimiento y estadísticas de diversas fuentes de data estructurada y no estructurada. Comúnmente se suele entender la ciencia de datos como una disciplina en sí misma: una "ciencia enfocada en datos" entendida como un "cuarto paradigma" de ciencia que usa el análisis computacional de grandes volúmenes de datos como método científico (Bell, Hey & Szalay, 2009; Hey, Tansley & Tolle, 2009).

Si bien esencialmente la ciencia de datos puede aplicarse en cualquier volumen de datos dado a que principalmente corresponde a una colección de técnicas, su relación con los "macrodatos" o Big Data es intrínseca pues es esta la principal fuente de informacion.

Su relación con otras disciplinas, particularmente la genómica y metagenómica, es evidente dado el exponencial crecimiento que estos datasets están teniendo.

2.3.1. Preprocesamiento de datos

Frecuentemente (y siendo pragmáticos, siempre), los datos a analizar se presentan de forma desorganizada lo que dificulta el trabajo. Es fundamental tratar con las imperfecciones en la data antes de enviarlas a la rutina de procesamiento. El preprocesamiento y estructuración de la información es clave para no solo lograr modelos estadísticos eficientes, sino que también análisis fidedignos, esto consiste en el $80\,\%$ del trabajo de un analista de datos. Algunos ejemplos de imperfecciones incluyen datos faltantes, formatos inconsistentes, asi como tambien la falta de datos curados y fidedignos. La Accesibilidad de los lenguajes de programación modernos ponen a disposición varias herramientas que apoyan la transformación de datos en masa con un nivel de control bastante fino (Hengtee Lim, 2020).

2.3.1.1. Proyecto Jupyter

El proyecto Jupyter desarrolla la herramienta Jupyter Notebook y sus herramientas asociadas. Jupyter Notebook es un sistema que facilita la distribución y comunicación de proyectos en computación científica y ciencia de datos permitiendo presentar de forma simple e interactiva segmentos de código con explicaciones en texto y gráficos. El proyecto Jupyter es totalmente libre y se ha vuelto el estándar de facto como herramienta de desarrollo y distribución de proyectos científicos.

2.3.1.2. Pandas

Pandas es la herramienta principal para la manipulación de datos sobre el lenguaje de programación Python. Pandas toma el concepto de "data frame" desarrollado por R y lo adapta a las necesidades y singularidades de Python (pandas development team, 2020). Es

otro proyecto libre masivamente usado para el análisis de datos por sus funcionalidades para manipulación de datos brutos.

2.3.1.3. BioPython

El proyecto BioPython surge en respuesta al creciente interés de los investigadores en la biología computacional. Principalmente es una colección de herramientas y estructuras de datos para representar secuencias biológicas y anotaciones biológicas, así como también leer y escribir desde varios formatos y acceder rápidamente a diversas bases de datos en línea. Es integrable con otras herramientas biológicas por medio de módulos lo que vuelven a este proyecto útil al momento de reducir la duplicación de código en biología computacional.

2.3.2. Alineamiento genético

Basic Local Alignment Search Tool (BLAST) (Altschul, Gish, Miller, Myers & Lipman, 1990) es el algoritmo y herramienta primordial en el estudio de secuencias genéticas. Permite comparar una secuencia de información biológica como ADN, ARN y proteínas (llamada ”query”) con una biblioteca o base de datos de secuencias, identificando las secuencias en la biblioteca que más se asemejan. Este análisis determina la similitud de ambas secuencias por sobre cierto umbral, lo que determina que sean ”secuencias homólogas”.

BLAST específicamente es un algoritmo heurístico que realiza la tarea de ”alineamiento local de secuencias”, que consiste en determinar regiones similares entre dos cadenas de secuencias de ácido nucleico o proteínas. Antes del desarrollo de algoritmos veloces como BLAST, FASTA (Lipman & Pearson, 1985) o DIAMOND, la búsqueda en bases de datos se realizaba con el procedimiento de alineamiento completo usando el algoritmo de programación dinámica Smith-Waterman (Smith & Waterman, 1981), que es computacionalmente costoso en comparación a los modelos heurísticos y estocásticos ofrecidos por estos desarrollos más recientes.

En la gran mayoría de los casos, BLAST y similares son sumamente rápidos en tiempo y uso de recursos que cualquier implementación de Smith-Waterman, pero no pueden garantizar la optimalidad de los alineamientos. Pero esta falta de precisión es negligible dada la practicidad de búsquedas veloces en las bases de datos genómicas actualmente disponibles.

En la actualidad, una serie de implementaciones de BLAST se encuentran disponibles y son ampliamente usadas por la comunidad de investigadores. Actualmente es común el uso de DIAMOND como alternativa veloz y precisa para alineamientos ADN-proteína en el contexto de secuencias cortas (150bp-250bp), principalmente asociadas a proyectos de metagenómica (Buchfink, Xie & Huson, 2015).

Entre los principales usos de BLAST (de la búsqueda de secuencias homólogas) se incluyen la identificación de especies, por ejemplo cuando se trabaja con secuencias de una especie desconocida y se desea identificar a qué organismo corresponde. Identificar dominios en

secuencias proteicas, que corresponden a secuencias de aminoácidos que cumplen con ciertas características estructurales conocidas.

2.3.3. Identificación de Profagos

Un profago corresponde a un bacteriofago que se ha incorporado en el genoma de su hospedero en fase lisogénica. La presencia del profago puede alterar el fenotipo del hospedero al disturbar genes que existían en el genoma, o incorporar nuevos genes y/o nuevas posibles funciones al hospedero. La detección de un profago en el genoma bacteriano puede dilucidar varios detalles sobre la relación ecológica entre el fago y un hospedero, como la contribución del fago a la supervivencia del hospedero y la regulación de su expresión genética . Existe evidencia de que después de integrarse en el genoma bacteriano, los profagos sufren de importantes procesos consistentes en la activación y desactivación de mutaciones puntuales, reordenamientos genómicos, invasión por parte de elementos móviles de ADN y borrados masivos de ADN (Canchaya, Proux y col., 2003).

La gran mayoría de las secuencias genómicas depositadas en bases de datos públicas como la de NCBI contienen secuencias de profagos, las cuales pueden ser detectadas por medio de algoritmos heurísticos y estocásticos.

2.3.3.1. PhiSpy

PhiSpy es un algoritmo y herramienta bioinformática implementada en Python y C++ tomando en consideración, siete características distintivas de los profagos, por ejemplo: largo proteico, direccionalidad de la hebra transcriptora, sesgos de AT y GC, la abundancia de secuencias únicas de fagos, puntos de inserción y similitud con otras proteínas de fagos (Akhter, Aziz & Edwards, 2012). Las primeras cinco características son capaces de identificar profagos sin necesidad de usar referencias y similitud con genes de fagos conocidos. El código fuente de PhiSpy está disponible como código abierto y se debe utilizar de forma local.

2.3.3.2. Phaster

Phaster es otra herramienta bioinformática que combina diversas otras herramientas para identificar y anotar profagos en secuencias de ADN. Phaster se encuentra únicamente disponible a través de una plataforma web. Phaster considera los profagos como clústeres de genes similares a fagos dentro de un genoma bacteriano, usando el algoritmo DBSCAN (Arndt y col., 2016; Zhou, Liang, Lynch, Dennis & Wishart, 2011).

2.3.4. Identificación de grupos Taxonómicos

El concepto de “Unidad taxonómica operacional” (OTU) hace referencia a clústeres de organismos, agrupados por similitud de ADN respecto a genes marcadores taxonómicos específicos. En otras palabras, las OTUs corresponden a sinónimos de especies a diferentes niveles taxonómicos, en ausencia de sistemas de clasificación biológica como la disponible para organismos macroscópicos (Blaxter y col., 2005).

En metagenómica es importante la asignación de las secuencias presentes a OTUs y rangos taxonómicos permitiendo la identificación y caracterización de grupos de organismos (taxones) presentes en la muestra.

La base de datos de taxonomía de NCBI corresponde al repositorio estandar de clasificacion y nomenclatura del INSDC, conteniendo los nombres de los organismos y sus linajes taxonómicos asociados a cada una de las secuencias nucleicas y proteicas también disponibles en las bases de datos de nucleótidos y proteínas del INSDC (Federhen, 2012).

La asociación secuencia-taxón presente en NCBI convierten a esta base de datos en el punto de partida para la identificación de secuencias en metagenomas. Sin embargo una simple asignación por búsqueda BLAST podría no ser indicativo de la real presencia de un organismo en el metagenoma, ya que solo estaría indicando la presencia del específico gen al que se está asignando, el cual podría o no estar en copresencia de los genes del organismo completo (Siobhan C. Watkins y col., 2016). Diversos métodos y algoritmos heurísticos ofrecen mayor confiabilidad en la asignación de clasificacion taxonomica.

2.3.4.1. Phylosift

Phylosift es una herramienta bioinformática que permite el análisis filogenético de muestras metagenómicas y la comparación de la estructura de la comunidad entre múltiples muestras relacionadas. El método se basa en modelos filogenéticos estadísticos. Adicionalmente el método propone un conjunto de treinta y siete marcadores genéticos que utiliza como referencia para la estimación de OTUs presentes (Darling y col., 2014). Este método no ha recibido actualizaciones desde 2014 y se limita únicamente a organismos bacterianos no incluyendo organismos virales.

2.3.4.2. GTDB-Tk

El “Genome Taxonomy Database Toolkit” es otra herramienta que provee asignaciones taxonómicas para genomas bacterianos y de arqueas basado en la base de datos de Taxonomía Genomica (GTDB) que incorpora la última información genómica disponible en las bases de datos de NCBI (Parks y col., 2018). GTDB-Tk clasifica los genomas según una combinación de factores como su similitud con genomas de referencia bajo diversas métricas (Chaumeil, Mussig, Hugenholtz & Parks, 2019).

Un punto clave en común entre todas las herramientas disponibles, es la incapacidad de entregar asignaciones taxonómicas a metagenomas virales.

2.3.5. Visualización y Comunicación de la Información

Con los grandes volúmenes de datos producidos por experimentos genómicos y metagenómicos, se vuelve aún más compleja la condensación de estos datos de una forma que haga sentido. A través de gráficos, se pueden ingerir rápidamente estos grandes volúmenes de información y derivar conclusiones a partir de ellos, acelerando el proceso de análisis.

Si bien el mercado ofrece diversos productos enfocados directamente en la visualización de Big Data, por ejemplo Qlik, Tableau y PowerBI, estos están completamente enfocados en visualizaciones relacionadas a inteligencia de negocios y son totalmente carentes de poder para presentar visualizaciones de interés biológico o son limitantes al momento de adaptar la visualización para responder preguntas específicas de interés. Se vuelve necesario recurrir a herramientas de más bajo nivel para generar visualizaciones útiles.

2.3.5.1. Matplotlib

Matplotlib es la herramienta que sirve de base a todos los sistemas de generación de gráficos utilizados en Python a pesar de que también puede ser usado directamente. Es sumamente poderosa permitiendo un control fino sobre la generación de gráficos altamente personalizables y para grandes volúmenes de datos (Hunter, 2007).

2.3.5.2. Graphviz

En el contexto de visualización de grafos, Graphviz es una colección de poderosos programas y algoritmos para la disposición espacial de nodos de un grafo en el plano bidimensional.

Todos los programas están escritos en C++ y existen como instancias independientes, sin embargo interfaces para lenguajes como Python existen (Gansner & North, 2000). Graphviz es la herramienta que presenta mejor desempeño con un alto grado de personalización y control de entre la variedad de opciones disponibles, especialmente para grandes volúmenes de datos, en contraste, la compleja curva de aprendizaje es su principal desventaja.

2.4. Teoría de Redes

La teoría de redes es una ciencia interdisciplinaria que ofrece un enfoque para estudiar las interacciones entre diferentes entidades de sistemas complejos. Considerar las propiedades estructurales de las redes de interacción puede guiar el entendimiento de cómo llegaron a ser, las reglas que las gobiernan, como se espera que cambien durante el tiempo y como se espera que respondan a perturbaciones (Beckett, 2015).

En particular las redes bipartitas describen interacciones entre dos conjuntos diferentes de nodos donde las conexiones solo ocurren entre grupos y no dentro de los grupos. Esto es particularmente útil ya que en el estudio de comunidades fago-hospedero poseemos dos grupos evidentes que interactúan entre si.

A pesar de su utilidad, el entendimiento teórico de estas redes no esta tan desarrollado como el de aquellas que solo involucran un único tipo de nodo.

2.4.1. Patrones de estructura en redes bipartitas

Como cualquier otro tipo de red que ocurre de forma natural, en el caso de las redes ecológicas, se sospecha que estas no están formadas al azar, sino que poseen organización interna. En la Figura 25 se ejemplifican diferentes tipos de interacciones entre ambos grupos de la red.

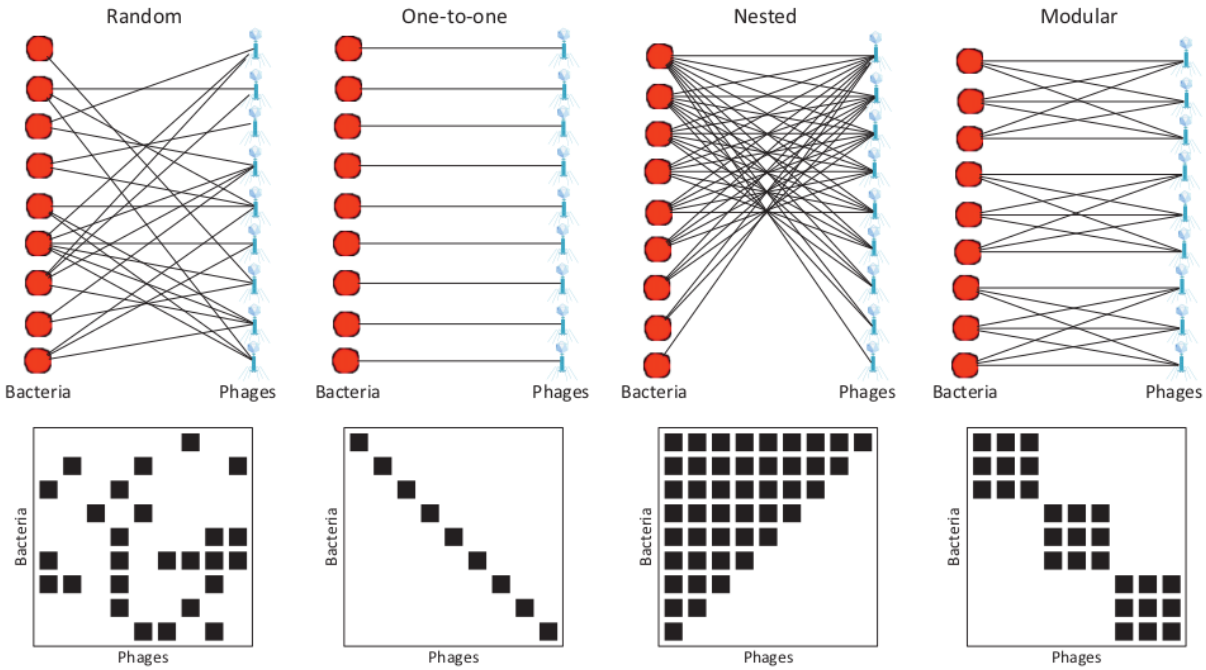


Figura 25: Ejemplos de estructura interna en redes bipartitas de interacciones (Weitz y col., 2013)

Los patrones de mayor importancia son los que se relacionan con la *nestedness* (Anidamiento) y la *modularity* (Modularidad). En el anidamiento los nodos de la red presentan estrategias de interacción del tipo generalista, en la modularidad la tendencia de los nodos de la red es a desarrollar grupos de miembros altamente conectados, llamados módulos.

2.4.1.1. Anidamiento

El anidamiento de la red bipartita es una propiedad que describe la tendencia de un nodo especialista de un tipo a interactuar con un nodo generalista del otro tipo. Esta definición es amplia y el concepto y difícil de definir operacionalmente, puede ser satisfecho de diferentes maneras, como diferentes tipos de anidamiento son posibles (Beckett, 2015).

Una forma común de implementar la medida de anidamiento son los métodos caracterizados como ”Temperatura” de la red. Los métodos basados en temperatura apuntan a re-ordenar la matriz de biadjacencia de la red de forma que la mayoría de asociaciones calcen con la curva llamada *”isoclina de anidamiento perfecto”*, que es una línea dibujada entre esquinas opuestas de la matriz y curvada según la conectancia de la matriz. Entonces la

temperatura se calcula contando el número de *"sorpresas"*, es decir, el número de ausencias arriba de la curva y el número de presencias bajo la curva. Implementaciones populares son los métodos `BITMANTEST` (Rodríguez-Gironés & Santamaría, 2006), `ANINHADO` (Guimaraes Jr & Guimaraes, 2006) y `NTC` (Oksanen y col., 2015).

Otra forma bastante común de medir anidamiento es con los métodos de *solapamiento*. Estos métodos se basan en revisar filas y columnas de la matriz de biadjacencia por similaridad de solapamiento. Se realizan comparaciones entre todo par de filas y columnas para establecer el número de elementos solapados y cuantificar la medida de anidamiento. El algoritmo más utilizado es `NODF` (Almeida-Neto, Guimaraes, Guimaraes Jr, Loyola & Ulrich, 2008).

2.4.1.2. Modularidad

La modularidad fue descrita originalmente como una propiedad de redes unipartitas (Newman & Girvan, 2004), sin embargo aplica de igual manera a redes bipartitas. La modularidad asume que hay nodos dentro de una comunidad que tienen más probabilidad de interactuar entre si que con los nodos del resto de la red, de esta forma la modularidad busca identificar comunidades compuestas de nodos densamente clusterizados (Leger, Vacher & Daudin, 2014).

El método `BRIM` *Bipartite, Recursively Induced Modules* (Barber, 2007; Guimerà, Sales-Pardo & Amaral, 2007) intenta descomponer la matriz de biadjacencia en cierto número de módulos, re-ordenando columnas y filas. El número óptimo de módulos es aquel en que las interacciones ocurren más frecuentemente dentro de un módulo mientras ocurren muy infrecuentemente fuera de un módulo.

Capítulo 3

PROPUESTA DE SOLUCIÓN

3.1. Racionalización

Como ha sido expuesto por otros estudios, la exploración de la relación fago-hospedero basado en estudios del ADN ambiental de comunidades complejas es limitada por la cantidad y calidad de información disponible (Casjens, 2003; G. F. Hatfull & Hendrix, 2011; Rosario y col., 2009). Una de las principales limitaciones está relacionada con la capacidad en obtener suficiente cantidad, calidad y cobertura del ADN total secuenciado, material que constituye la base de todos los análisis posteriores realizados en metagenómica. Esta deficiencia es particularmente notoria en la fracción constituida por virus, que dado su tamaño, diversidad, variabilidad genética y naturaleza polifilética, no poseen una amplia proporción de información biológica y genómica descrita e incluida en las actuales bases de datos (Bruder y col., 2016; Casjens, 2003; G. F. Hatfull & Hendrix, 2011). Una mayor incertidumbre se produce al describir la microbiota viral en ambientes acuáticos, en donde la concentración de la información contenida en bases de datos y evidencia científica recae en procariontes y hongos que poseen un rol directo ya sea en los ciclos biogeoquímicos, o como posibles amenazas de transmisión de enfermedades.

La presente memoria tiene como propósito desarrollar una colección de herramientas y procedimientos capaces de interrogar la presencia de genes pertenecientes a genotipos virales y microbianos (procariotes) en muestras metagenómicas realizadas en el ambiente, y de esta manera, obtener información relevante acerca de la relación bacteriófago-hospedero (virus-procarionte). Asimismo, esta colección incorpora de manera modular distintas herramientas y procedimientos de detección de genes, asignación de taxonomía y normalización que permite establecer un flujo centralizado de análisis de datos, sin embargo permite que ante la disponibilidad de herramientas de mejor rendimiento y calidad, puedan ser fácilmente integradas, reemplazando de forma simple y compatible aquellas originales presentadas en esta memoria. El protocolo realizado en esta memoria corresponde a un protocolo de tipo modular que integra una serie de etapas, cada una de estas consistente de herramientas reemplazables con tal de que se alimenten y entreguen datos en formatos estandarizados y compatible con las demás herramientas disponibles en la disciplina.

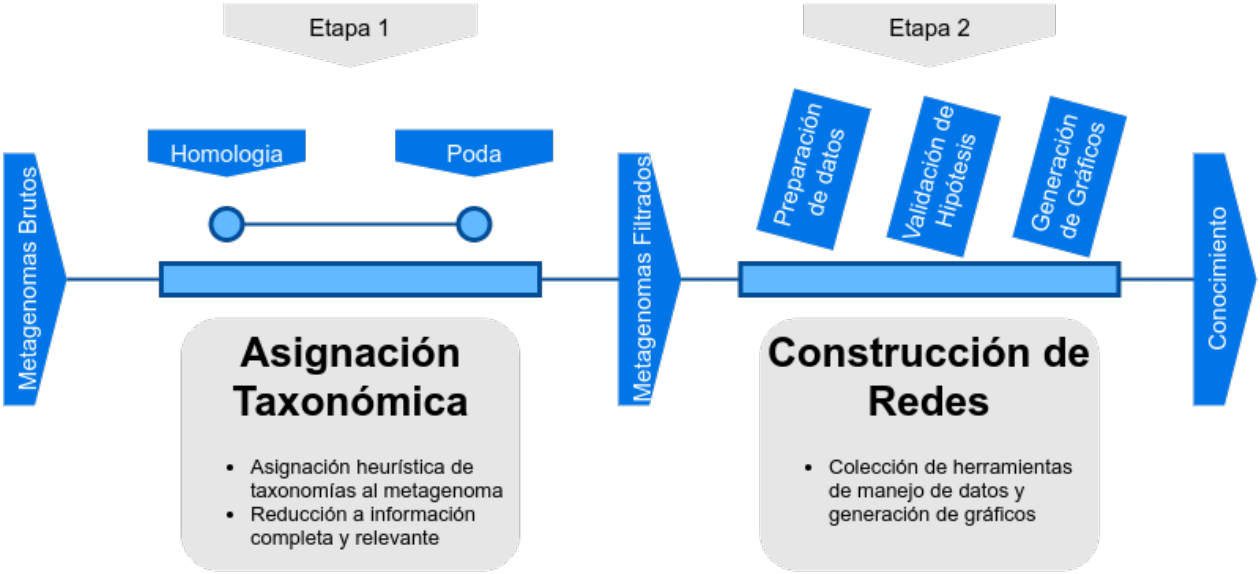


Figura 31: Flujo general del protocolo.

El protocolo propuesto en esta memoria consta de dos etapas, las que pueden ser realizadas de forma secuencial o paralela según sea posible. La primera de estas etapas corresponde al proceso de asignación de unidades taxonómicas operacionales (OTU) a las secuencias de los metagenomas originales que se pretende analizar. Esta etapa se compone de pasos secuenciales y utiliza una combinación de herramientas públicas en conjunto con herramientas desarrolladas para este propósito. La segunda etapa es una colección de herramientas para el tratamiento de datos y generación de gráficos de los cuales se puede extraer información acerca de la coexistencia de genes pertenecientes a virus-procariontes en diversos ambientes. Dada la naturaleza de los datos y el tipo de preguntas que se intentan responder, las herramientas utilizadas poseen características de aprendizaje máquina y aprendizaje no supervisado.

3.2. Etapa 1: Asignación Taxonómica

La primera pregunta científica al estudiar un microbioma es, generalmente, ¿Cuál es la estructura de la comunidad microbiana?. Esto es, en otras palabras, la identificación de los microorganismos presentes (riqueza) y determinación de su distribución (abundancia relativa). Existen diversas herramientas y protocolos en el estado del arte para abordar esta pregunta, basados en datos generados por el enfoque metagenómico.

Debido a que la herramienta propuesta en esta tesis se basa en la identificar patrones de distribución de genes de virus y procartiontes entre diversos ambientes, la primera tarea del protocolo consiste en recolectar metagenomas obtenidos desde los ambientes que se quiere obtener información. Los metagenomas corresponden a archivos que poseen secuencias trozadas de ADN de toda (o la mayor parte) de la comunidad microbiana obtenida desde un ambiente en particular. Posterior a la secuenciación del material biológico,

Figura 32: Resumen de la etapa de Asignación Taxonómica

la información se guarda en archivos en formato "FASTA", en donde se almacena un alto número (en el rango de miles de millones) de secuencias de nucleótidos. Por la naturaleza de nuestra herramienta, se busca que los metagenomas seleccionados deben poseer características que lo hagan lo más comparable posible (tipo de ambiente obtenido, tipo de obtención y tratamiento de las muestras biológicas, tipo de secuenciación, esfuerzo de muestreo, etc).

Posteriormente, siendo el propósito de esta memoria dar una aplicación práctica a este protocolo, posterior a la recolección se comienza progresivamente a eliminar aquellas secuencias que no se alineen con los grupos taxonómicos de interés, es decir, se filtraran secuencias asignadas a especies eucariotas (protozoos, hongos, plantas y animales) las que no contribuyen a establecer relaciones entre virus procariontes. Adicionalmente, los datos resultantes de los distintos pasos de esta herramienta se le aplicarán diversas transformaciones para reducirla a información útil en una estructura de datos estandarizada que puede ser utilizada en cada una de las etapas.

3.2.1. Paso 1: Búsqueda de Homólogos para las secuencias del Metagenoma

Como se mencionó previamente, el enfoque metagenómico genera como principal producto posterior al proceso de secuenciación archivos en formato "FASTA"(ver Figura 34). Cada uno de estos archivos constituye un "metagenoma" obtenido desde un ambiente y tiempo específico, una "toma de muestras". El proceso de búsqueda de homólogos se concentra

en el procesamiento individual e independiente de cada metagenoma y puede ejecutarse de forma paralela. Este proceso busca comparar las secuencias genéticas obtenidas desde los metagenomas a procesar en contra de una base de datos de secuencias previamente recopiladas, lo que nos permite asignar una función específica a cada una de estas. Este proceso se denomina anotación de las secuencias. El resultado es un archivo de texto plano llamado "tabla m8", que corresponde a la estructura de tablas reportadas por la herramienta "BLAST", este archivo es procesado por programas desarrollados *ad-hoc* para este protocolo, resultando en una versión reducida de la misma tabla m8 para cada metagenoma.

De las diversas herramientas de alineamiento de secuencias genéticas disponibles en el estado del arte, usaremos "Diamond BLAST" (Buchfink y col., 2015), específicamente la sub-herramienta "blastp" la cual ha sido diseñada para realizar alineamientos entre secuencias de ADN y bases de datos de secuencias de proteínas previamente descritas. A su vez, como base de datos usaremos la última versión disponible "REFSEQ" (Pruitt y col., 2005), la que para ser utilizada en Diamond debe ser primero descargada de forma local y compilada en un proceso explicado a continuación.

3.2.1.1. Reclutamiento y descarga de Bases de datos de proteínas REFSEQ

RefSeq corresponde a una base de datos de referencia que incluye secuencias de ADN, ARN y proteínas no redundantes, bastante extensiva y con una alta calidad de anotación (Pruitt y col., 2005). Esta base de datos se encuentra disponible públicamente en el servidor FTP de NCBI (ftp://ftp.ncbi.nlm.nih.gov) y contiene diferentes particiones y formatos. La herramienta propuesta en esta tesis utiliza la versión completa para proteínas (disponible en ftp://ftp.ncbi.nlm.nih.gov/refseq/release/complete/). Para iniciar la descarga se ejecuta el siguiente script en una terminal *BASH*:

```
wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/complete/*.protein.faa.gz
```

Los archivos descargados consisten en particiones de la base de datos en formato `FASTA` comprimido en `gzip`, adicionalmente se deben descargar:

`accession2taxid` Tabla de asignaciones indices de proteínas NCBI e identificadores de especies. Disponible en ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz

`nodes.dmp` Archivo disponible dentro del comprimido ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip.

`names.dmp` Disponible en el mismo comprimido ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip.

Estos archivos pueden ser enviados directamente a `Diamond` para generar una base de datos local:

Figura 33: Resumen del proceso de Búsqueda por Homología. Este proceso es independiente para cada archivo de entrada.

```
diamond makedb --in *.protein.faa.gz --db refseq --taxonmap
↪  prot.accession2taxid.gz --taxonnodes nodes.dmp --taxonnames names.dmp
```

Aqui, `Diamond` es invocado con las siguientes opciones para compilar la base de datos:

`--in` Se indica que se usarán `*.protein.faa.gz` como fuentes para la base de datos

`--db` La base de datos compilada será guardada como `refseq.dmnd`

`--taxonmap` Se indica la tabla de asignaciones de proteínas `accession2taxid`

`--taxonnodes` El archivo `nodes.dmp`

`--taxonnames` El archivo `names.dmp`

Se debe tener en consideración que tanto DIAMOND, como todas las variaciones de BLAST, son algoritmos que poseen un alto consumo en memoria, por lo que se recomienda ejecutarlo en un ambiente cluster con al menos $1{,}3$ veces el tamaño de la base de datos en memoria virtual (Para el caso de REFSEQ, $145GB$ de memoria virtual son recomendables).

Una vez compilada la base de datos, se procede a utilizar `Diamond BLASTp` para realizar el análisis de homologías que permitirá asignar una función y posterior anotación de cada una de las secuencias incluidas en cada metagenoma analizado. Durante este proceso, la mayoría de opciones se pueden dejar por defecto u optimizar a la respectiva plataforma donde se esté ejecutando la herramienta. Sin embargo, es muy importante estandarizar las columnas de salida de la `tabla m8`. De esta manera, se ejecuta `blastp` con las opciones mínimas requeridas para nuestro protocolo:

```
diamond blastp --db refseq --query metagenome.fasta \
    --out metagenome.m8 \
    --outfmt 6 qseqid sseqid evalue bitscore staxids
```

Donde:

`--db` La base de datos que recopilamos en previamente.

`--query` Este es el metagenoma al que se le buscarán homólogos.

`--out` Archivo donde se guardará la `tabla m8` resultante.

`--outfmt` Indica que se usará una tabla para reportar los resultados (formato 6) con las columnas indicadas en el orden indicado.

El resultado de esta ejecución es una `tabla m8` con cinco columnas (`qseqid`, `sseqid`, `evalue`, `bitscore` y `staxids`). Solo estas cinco columnas son necesarias para los siguientes pasos, por lo que se recomienda configurar `blast` para restringir la configuración de la tabla m8 a esta disposición mínima, reduciendo así el consumo de memoria. En caso de ser necesario o resultar de interés se puede ejecutar `blast` produciendo más columnas, siempre que antes de moverse a la siguiente parte, se procese el archivo para extraer únicamente las cinco columnas aquí mencionadas. Se sugiere ejecutar un ordenamiento de las filas de la tabla obtenida, con tal de que las secuencias del metagenoma sean contiguas y en orden descendente de puntaje de asignación de homología, con tal de facilitar el post-procesamiento en las etapas siguientes. En particular, este paso busca ordenar primero por la columna `qseqid`(la primera) y luego por la tercera (`e-value`) y cuarta columna (`bitscore`).

```
sort -k1,1 -k3,3g -k4,4nr metagenome.m8 \
    -o sorted.metagenome.m8
```

Donde:

`-k1,1` Realiza el ordenamiento por la primera columna

`-k3,3g` Indica hacer un segundo ordenamiento por la tercera columna en orden numérico general, esto lee los valores de expectativa y los ordena en orden ascendente, asegurando los calces de mayor expectativa (menor `e-value` liderando cada grupo de asignaciones).

`-k4,4nr` Indica hacer un tercer ordenamiento por la cuarta columna en orden numérico descendente, así obtenemos los calces con mayor puntaje liderando cada grupo de asignaciones.

`metagenome.m8` Tabla de origen

`-o sorted.metagenome.m8` Archivo donde se escribirá el ordenamiento. Este puede reemplazar al original una vez completado.

Consideraciones computacionales: Al ejecutar `sort` es importante tener en cuenta que el programa necesita memoria suficiente para escribir archivos temporales de al menos el tamaño del archivo original.

3.2.1.2. Sobre el formato FASTA

`FASTA` es abreviación de `Fast-All`, originalmente hacía referencia a una *suite* de programas para hacer búsqueda de homologías (similitud) entre secuencias, de la misma forma en que BLAST y similares, funcionan (Pearson & Lipman, 1988). Sin embargo, el uso actual del término hace referencia al formato de archivo de texto creado para este *software*, actualmente ubicuo en bioinformática.

`FASTA` es un formato de texto plano. La primera línea comienza con un símbolo $>$ (mayor que), seguido de un nombre que identifique a la secuencia (usualmente un código de índice en una base de datos), más un comentario o descripción de la secuencia. La segunda línea contiene la secuencia codificada como cadenas de letras del alfabeto según una tabla de conversión estandarizada según se trate de secuencias de ADN, ARN o proteínas. Los saltos de línea en la secuencia son ignorados, se acostumbra a usarlos para facilitar la visualización de la secuencia en procesadores de texto.

Una cantidad arbitraria de archivos pueden concatenarse en un único archivo común de forma simple para facilitar su manejo. También es común presentar y usar archivos `FASTA` comprimidos en formato `gzip` debido a que permite concatenar los archivos sin la necesidad de descomprimirlos previamente.

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK

>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
```

Figura 34: Ejemplo de un archivo FASTA

3.2.1.3. Sobre la familia BLAST

BLAST es la herramienta universal usada en bioinformática para realizar búsquedas de homologías (similitud) entre secuencias usando el algoritmo de Smith-Waterman (Smith & Waterman, 1981) junto a diversas heurísticas y aproximaciones. Implementaciones más modernas están recientemente disponibles para ser ejecutadas de forma local utilizando bases de datos locales o en servicios web con bases de datos públicas, así como también permiten reportar sus resultados en diversos formatos que son interoperables con otros servicios y programas. En general, todas las implementaciones de BLAST ofrecen una o más de los programas específicos de la herramienta, cuya única diferencia es la forma en que aceptan los archivos de entrada. Estas son:

`blastn` Busca en bases de datos de nucleótidos (ADN), homólogos para secuencias de nucleótidos (ADN).

`blastx` Busca en bases de datos de aminoácidos (proteínas), homólogos para secuencias de nucleótidos (ADN).

`blastp` Busca en bases de datos de aminoácidos (proteínas), homólogos para secuencias de aminoácidos (proteínas).

`tblastn` Busca en bases de datos de nucleótidos (ADN), homólogos para secuencias de aminoácidos (proteínas).

3.2.1.4. Sobre las Tablas m8

Corresponde al formato estándar en que se reportan los resultados de BLAST. Es un archivo de texto plano ordenado en columnas separadas por tabuladores (`.tsv`). Las columnas no presentan cabeceras, y por defecto solo se reporta un limitado subconjunto de columnas

posibles, cada implementación de BLAST ofrece la posibilidad de seleccionar las columnas reportadas. Al ser un formato de texto plano estructurado, la integración con cualquier tipo de procesador de texto es trivial.

La columna `E-Value` es una de las columnas que usualmente se reporta y es una métrica que reporta el número de calces de calidad similar al resto, que pueden haber sido hallados por azar. Es decir, es una prueba de hipómemoria respecto a validez del calce, un `E-Value` de 10 indica que hasta 10 asignaciones podrían ser halladas solo usando azar en una base de datos del mismo tamaño. Se puede usar esta métrica como un filtro inicial para cortar la cantidad de calces reportados. Mientras menor sea el `E-Value`, mejor es el calce.

`bitscore` es otra columna usualmente reportada y es un valor que normaliza y resume al `E-Value` tomando en cuenta el tamaño de la base de datos que se está revisando, dejando de depender de esta. De esta forma se obtiene un valor normalizado que puede ser usado para comparar resultados entre distintas bases de datos (Fassler & Cooper, 2011).

3.2.1.5. Sobre REFSEQ y otras bases de datos

La base de datos de Secuencias de Referencia (`REFSEQ` [Pruitt y col., 2005]) es una colección, anotada y curada de acceso público de secuencias de nucleótidos (ADN,ARN) y aminoácidos (proteínas). Esta base de datos es mantenida por el NCBI (como parte del National Institutes of Health (NIH) de Estados Unidos) y tiene como objetivo proveer referencias no redundantes, curadas y de alta calidad de organismos de interés, desde virus, bacterias hasta eucariotas.

La ventaja de `REFSEQ` respecto a las alternativas (como `GENBANK` o `UniProt`) es que la información incorporada a RefSEQ corresponde a la revisión manual de la información presentada.

3.2.2. Paso 2: Selección de asignaciones taxonómicas

El proceso de búsqueda de homologías (similitud) entre secuencias genera la tabla m8 que contiene una lista de secuencias homólogas candidatas para cada secuencia presente en el metagenoma. Estas secuencias homólogas poseen una asignación taxonómica entregada por la base de datos de referencia, sin embargo, no todos los homólogos posee una asignación taxonómica, y si la poseen no necesariamente se encuentra en el mismo nivel del ranking taxonómico. En primera instancia se ejecutará el programa `select-first-n.py` (ver Apéndice A) que tomará las primeras N secuencias con mejores `E-Value` y `bitscore`. Este programa automáticamente eliminará secuencias sin asignación taxonómica y añadirá nombres a las columnas para facilitar su revisión manual.

```
python3 select-first-n.py metagenome.m8 out.metagenome.m8 3
```

Donde:

metagenome.m8 Es la tabla m8 a procesar.

out.metagenome.m8 Es donde se guardará la tabla m8 procesada.

3 Indica el número calces a conservar. El programa conservará los calces de menor E-Value y mayor bitscore.

3.2.2.1. Consideraciones computacionales

Es importante destacar que BLAST y su familia de programas similares, disponen de herramientas para reportar una asignación taxonómica a cada secuencias analizadas. Por ejemplo, DIAMOND permite reportar una tabla solo con las columnas qseqid, E-Vaue y staxids asignando una afiliación taxonómica a cada una de estas. Si bien esta tarea se puede realizar con DIAMOND, cuando tratamos con virus la simple asignación por este método tiene problemas de fiabilidad. Ya mencionado en el Capítulo 1, los virus son entidades que poseen alta variabilidad genética, por tanto carecen de marcadores genéticos u otros elementos deterministas que permitan garantizar la asociación proteina/gen/(micro)organismo, relación que es mucho más fácil y posible establecer con claridad con secuencias de bacterias, arqueas y eucariotas. En Canchaya, Fournous y col., 2003 se menciona que encontrar un homólogo para una secuencia no garantiza la presencia de ese (micro)organismo en el ambiente, sino más bien la presencia del gen individual que fue identificado, idea que es reforzada por otros autores (Siobhan C. Watkins y col., 2016).

No existe hasta el momento de escribir esta memoria un procedimiento de detección de virus a través de herramientas metagenómicas que garantice con una confiabilidad similar que aquellas optimizadas para organismos más complejos. Por esta razón, la posibilidad de dejar que Diamond sea la herramienta encargada en realizar asignaciones taxonómicas de virus directamente es poco confiable.

La alternativa de nuestra propuesta a este problema corresponde a la generación de un reporte que incorpore un conjunto arbitrario de homologías en vez de un único. El número de secuencias (o de asignaciones taxonómicas) dependerá de cada situación en particular y no se puede establecer *a priori*, sin embargo no se recomienda una cantidad muy grande, dado que se está permitiendo la introducción de "ruido", definido como la presencia de genotipos virales que no están presentes en el ambiente, al interior de los análisis. El programa select-first-n.py (Apéndice A) realiza este filtrado además de remover aquellas homologías sin información taxonómica disponible.

3.2.3. Estandarización de niveles Taxonómicos y Selección de Dominios Taxonómicos de interés

El enfoque de esta memoria se basa en el estudio de bacteriófagos y sus respectivos hospederos procariontes (bacteria y arqueas). Debido a esto, la información metagenómica extraída directamente del ambiente, que contiene secuencias de ADN perteneciente a otros organismos que son totalmente irrelevantes en el estudio microbiano, deben ser extraída.

El proceso explicado a continuación busca remover esta información irrelevante para el procesamiento de los metagenomas analizados.

Figura 35: Resumen del Proceso de poda de datos

Posterior a la asignación taxonómica realizada por BLAST, se asocia a cada una de las secuencias un índice identificador (en forma número identificador [`taxid`]), el que especifica la identidad taxonómica de la secuencias homóloga más cercana a la analizada. Sin embargo, es común que este identificador no indique el nivel de especie, sino que hace referencia a un ranking taxonómico más amplio, tal como un género o familia, o en diversas circunstancias se hallan identificadores apuntando a niveles más específicos como sub-especies o cepas del organismo en particular. Para lo que nos concierne en esta memoria nos basta con información confiable y segura a nivel de especies, es por esto que el propósito del programa `filter-domains.py` (ver Apéndice B) es normalizar aquellos identificadores hiper-específicos (sub-especies y cepas) generalizándolos a los identificadores de nivel de especie correspondientes. Aquellos identificadores hiper-generalizados (superiores a especie) son descartados debido a la imposibilidad de asegurar una identificación concreta y debido al ruido que aportan a los datos.

Adicionalmente, se realiza una comparación entre todas aquellas secuencias que son homólogas a virus con aquellas secuencias genómicas que pertenecen a virus y que están anotados en la base de datos del ICTV. Dentro de este paso, se realiza un filtrado por aquellos virus que estén clasificados en los grupos I y II del sistema de Baltimore (DNA virus Figura 21), que poseen ADN tanto de hebra sencilla como doble como molécula que almacena información, los que corresponde a los más prevalentes dentro del grupo de virus capaces de infectar tanto bacterias, como arqueas.

Sobre la base de datos ICTV: El International Committee on Taxonomy of Viruses (ICTV) es un organismo que se encarga de desarrollar, refinar y mantener una base de datos estandarizada para taxonomía viral. Esto incluye la clasificación de especies virales como también niveles más generales de clasificación (Lefkowitz y col., 2017). Cada año el ICTV pública un archivo en formato `Microsoft Excel` resumiendo el listando todas las especies reconocidas por el organismo junto a su linaje taxonómico. Este archivo se encuentra disponible en https://talk.ictvonline.org/files/master-species-lists/m/msl y es actualizado año a año. Antes de comenzar el procesamiento, se debe procesar el archivo `excel` para extraer la tabla correspondiente al listado de especies, en un formato de texto plano (separado por tabuladores `.tsv`) y eliminar cualquier especie que se desea eliminar del metagenoma. Esta preparación puede ser realizada en el mismo editor de hojas de cálculo. Una vez preparado, se puede ejecutar Apéndice B con:

```
python3 filter_domains.py ICTVRank.tsv metagenome.m8 filt.metagenome.m8
```

Donde:

`ICTVRank.tsv` Es la base de datos del ICTV preparada.

`metagenome.m8` Metagenoma proveniente del paso anterior.

`out.metagenome.m8` Archivo de salida.

3.2.3.1. Umbral de repetición de asignaciones

Dado que la presencia de un único gen perteneciente al genoma de un organismo, es decir, la asignación de una secuencia a una especie no garantiza la real presencia de esa especie en el metagenoma (Canchaya, Fournous y col., 2003; Siobhan C. Watkins y col., 2016), muchas de las asignaciones reportadas corresponden a ruido y no representan la presencia de virus en un ambiente. Si bien es posible realizar un filtrado preciso revisando si efectivamente las secuencias homólogas asignadas al metagenoma se corresponden con genes conocidos reportados por genomas modelo de las especies correspondientes, en el caso de muchos organismos, especialmente virales, no existe un genoma de referencia con el cual se pueda verificar la confianza de la asignación. Esta carencia de información nos limita al momento de realizar las asignaciones y eliminar aquellas que no se condicen con la realidad. Como mitigación a esta limitación, proponemos una heurística simple relacionada a la frecuencia

de repetición en la que se presentan las asignaciones. Se cortan aquellas secuencias homólogas, que para una misma asignación taxonómica, fueron alcanzadas menos de un número arbitrario de veces. El razonamiento detrás de este caso está en eliminar aquellas asignaciones que se hayan obtenido por mero azar, desconociendo las propiedades del genoma de la especie en cuestión el umbral se establece de forma arbitraria.

Ejecutamos Apéndice C con:

```
python3 remove_low_freq.py metagenome.m8 out.metagenome.m8 3 100
```

Donde:

`metagenome.m8` Metagenoma proveniente del paso anterior.

`out.metagenome.m8` Archivo de salida.

`3` Umbral de corte para especies virales, por defecto: 3

`3` Umbral de corte para especies bacterianas y arqueas, por defecto: 100

Debido al conocido sesgo por parte del las bases de datos para reportar especies bacterianas por sobre todas las demas, el umbral de corte es independiente para organismos hospederos y virales, pudiendo usar umbrales diferentes para filtrar ambos grupos, compensando de cierta forma el desbalance.

3.3. Etapa 2: Construcción de redes de interacción ecológicas

Una vez identificada la presencia de virus en el ambiente, se realizará una sistematización de los datos en matrices que incorporen la presencia de genotipos virales y procariontes con el objetivo de establecer relaciones de coexistencia en el ambiente. Este proceso consta de una etapa de preparación de datos, volcándolos en estructuras de datos que permitan su fácil operación, seguido de una colección tareas para la generación de gráficos y visualizaciones útiles para nuestro objetivo, tales como redes de interacción, así como métricas sobre atributos estructurales de estas .

En esta etapa unificaremos el sistema completo de metagenomas en un manojo de estructuras de datos que nos permitirán extraer conocimiento desde los datos de una forma más eficiente.

3.3.0.1. Codificación

La Figura 37 representa la estructura de datos propuesta en esta memoria. Esta estructura corresponde a una matriz que codifica la presencia de una especie en cada uno de los

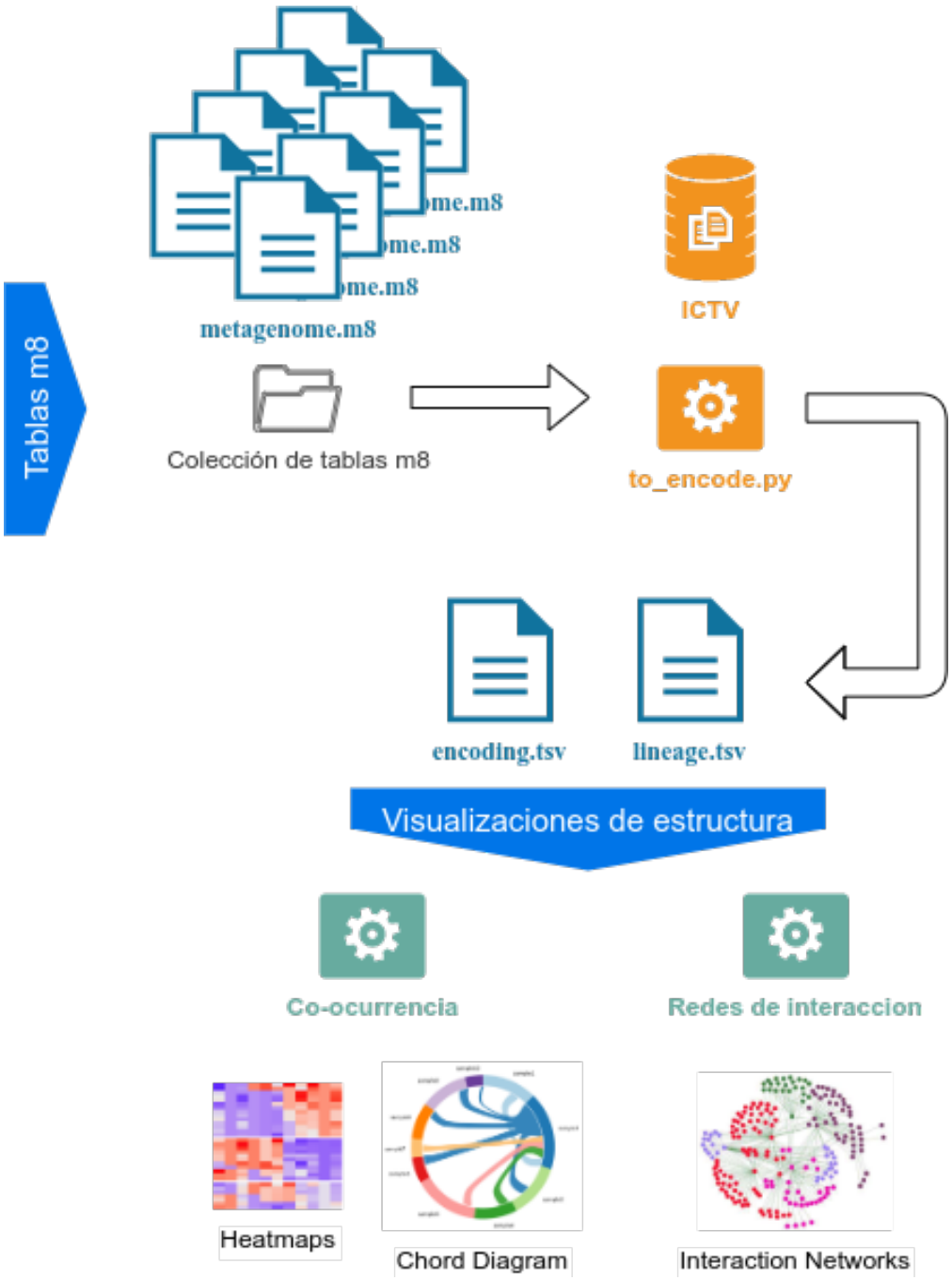


Figura 36: Etapa 2: Preparación de datos y colección de programas

ambientes en codificación *one-hot* (Wikipedia contributors, 2020). Aquí, cada fila representa un ambiente o metagenoma, y cada columna uno de los *taxid* de cada especie presente en el sistema completo de metagenomas.

Como es de interés estudiar la relación entre virus y hospederos (bacterias y arqueas), el programa `to_encoding.py` (ver to_encoding.py) realiza dos tareas. Primero, el programa

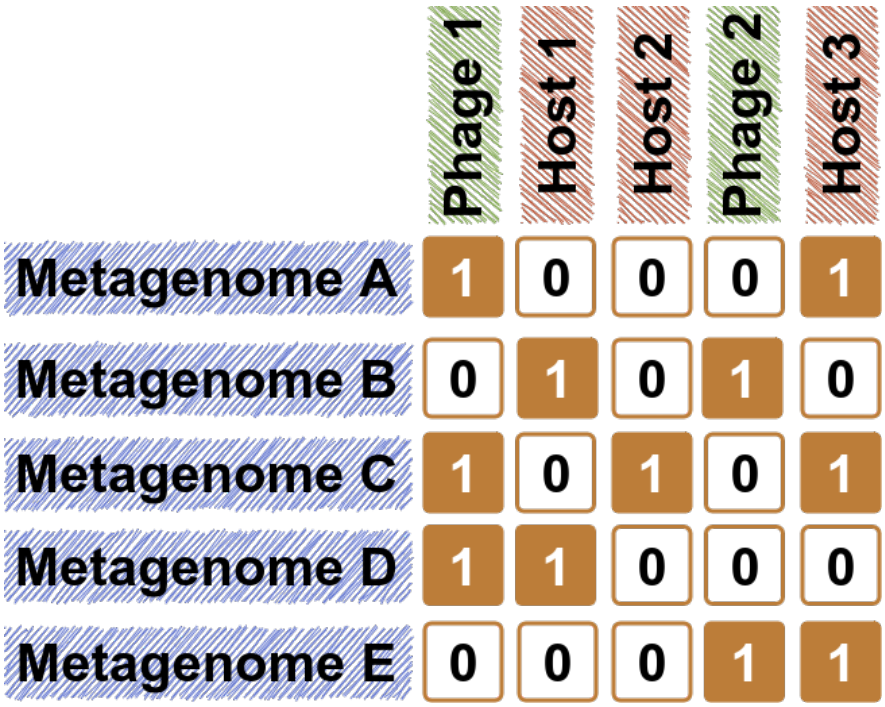


Figura 37: Estructura de datos: Codificación de existencia de especies en ambientes

comprime el sistema completo de metagenomas en la estructura mencionada en la Figura 37. En el segundo paso, el programa genera una tabla con los linajes taxonómicos de cada especie presente en el set completo. Esto genera dos archivos de texto plano: la codificación *one-hot* de las especies presentes en los metagenomas y la tabla de información taxonómica de todas las especies presentes en formato `.tsv`.

Teniendo los metagenomas de interés, ya procesados por la Etapa 1: Asignación Taxonómica, en la misma carpeta, se ejecuta:

```
python3 to_encoding.py METAGENOMES_DIR encoding.tsv lineage.tsv
```

Donde:

`METAGENOMES_DIR` Directorio donde se encuentran las tablas `m8` procesadas.

`encoding.tsv` Archivo donde se guardará la codificación `one-hot` del sistema.

`lineage.tsv` Archivo donde se guardará la información del linaje taxonómico de las especies del sistema.

3.3.1. Co-ocurrencia

La co-ocurrencia puede entenderse simplemente como la confirmación de que dos especies se encuentran presentes simultáneamente en el sistema, es decir, ocurren simultáneamente. Esto puede verse tanto como un valor *booleano* como medirse con algún coeficiente que agregue la información correspondiente a los diversos ambientes del sistema. En la Figura 38 podemos apreciar un ejemplo de visualización de la Co-ocurrencia booleana entre órdenes de un sistema simple.

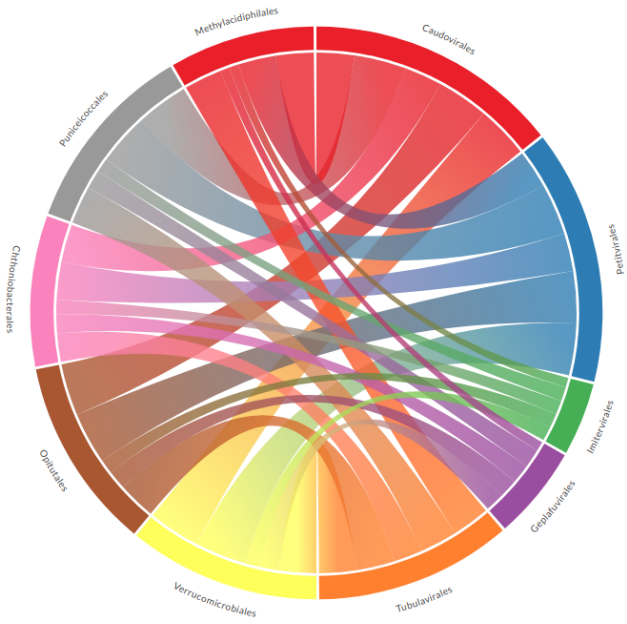


Figura 38: Ejemplo: Diagrama de cuerdas - Co-ocurrencia de Ordenes en un sistema simple

3.3.2. Métricas de Co-ocurrencia

3.3.2.1. Similitud de Jaccard y Tanimoto

Se establece la similitud de Tanimoto como una métrica que cuantifica la similitud entre conjuntos binarios, como entre la presencia de una especie en los diversos ambientes del sistema con las presencias de otra. La similitud de Tanimoto (Rogers & Tanimoto, 1960) es usualmente descrita como equivalente a la similitud de Jaccard para conjuntos booleanos. El coeficiente de similitud de Jaccard, también conocido como "intersección sobre unión" se usa para cuantificar la similitud y diversidad de conjuntos de muestras. La similitud de Tanimoto si bien nunca fue definida formalmente con ese nombre, se entiende como el cociente de *bits* comunes entre dos mapas de bits. Matemáticamente:

El Indice de Jaccard:

$$\mathcal{J}_s(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$$

La Similitud de Tanimoto:

$$\mathcal{T}_s\left(X,Y\right)=\frac{\sum_i\left(X_i\bigwedge Y_i\right)}{\sum_i\left(X_i\bigvee Y_i\right)}$$

La Similitud de Jaccard con Peso:

$$\mathcal{J}_w\left(X,Y\right)=\frac{\sum_i \text{mín}\, X_i,Y_y}{\sum_i \text{máx}\, X_i,Y_y}$$

Dada la codificación `one-hot`(Figura 37) es posible fácilmente aplicar la fórmula de similitud de Tanimoto para cuantificar la similitud de la presencia ambiental entre dos especies en los ambientes analizados. En nuestro caso, esto es cuantificar coexistencia entre bacteriófagos y sus hospederos (arqueas y bacterias), y representar estas relaciones de forma gráfica como el ejemplo de la Figura 39. Sin embargo, esta solo es útil al momento de realizar comparaciones directas entre especies, a un nivel muy granular, lo cual, dado el número de especies diferentes presentes en el sistema no solo es computacionalmente costoso, sino que aporta información que difícilmente puede ser discernible por su amplia complejidad. Es por esto, que nuestro enfoque recomienda que posterior a este paso, se indague en uno o más subconjuntos de especies en particular restringiéndolo a grupos taxonómicos a niveles más altos en la clasificación.

Con las métricas aquí presentadas se puede dar una cuantificación general sobre las relaciones fago-hospedero de los ambientes analizados.

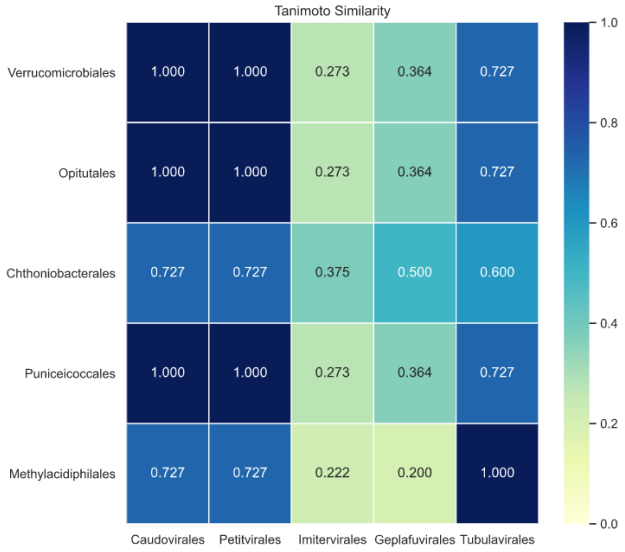


Figura 39: Ejemplo: Mapa de calor, Métrica de co-ocurrencia de Ordenes en un sistema simple

3.3.3. Estructura de redes ecológicas

3.3.3.1. Redes de Interacción Bipartitas

El análisis de relaciones fago-hospedero propuesto por esta tesis se completa a través de la representación gráfica de las relaciones binarias (virus-hospedero) respecto a los ambientes analizados. Es por esto que, a través de este paso, se genera una red de interacción de subconjuntos de las relaciones bipartitas principalmente por las limitaciones computacionales y de visualización dado el volumen de los datos. Esta red se representa computacionalmente en forma de una matriz de biadyacencia. Adicionalmente, algunas hipótesis sobre la estructura de redes ecológicas basadas en la relación virus-hospedero pueden ser testeadas, tales como la identificación de patrones de "anidamiento" de red (Subsubsección 2.4.1.1) y su "modularidad"(Subsubsección 2.4.1.2).

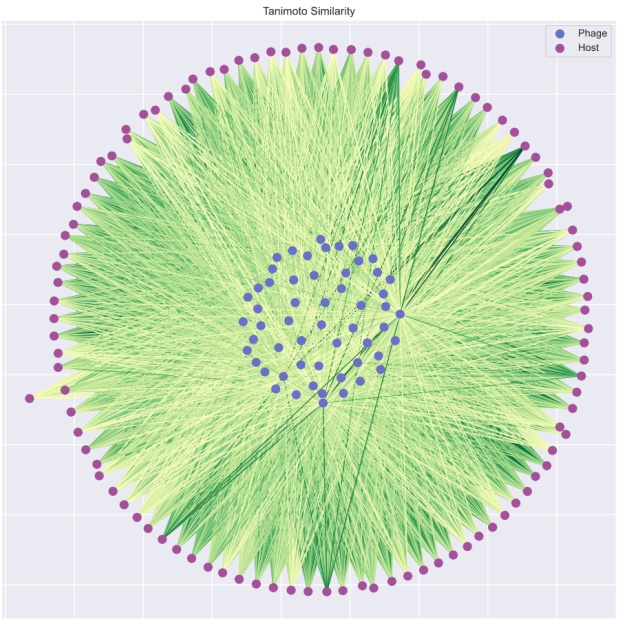


Figura 310: Ejemplo: Una red de interacción bipartita Fago-Hospedero

3.3.3.2. Anidamiento

Explicado en la Subsubsección 2.4.1.1, el anidamiento o *nestedness* es una propiedad que se utiliza para cuantificar el grado de interacción entre organismos especialistas (aquellos que solo interactúan con un único tipo de organismo) con organismos generalistas en (aquellos que no se restringen a un único organismo para interactuar) la red bipartita.

Para efectos de esta memoria se ha desarrollado una implementación propia de los algoritmos `NODF` (Almeida-Neto y col., 2008) y `WNODF` (Almeida-Neto y col., 2008) (ver Implementación NODF en numpy y Implementación WNODF en numpy).

La elección de los algoritmos `NODF` y `WNODF` es puramente arbitraria y debido a su facilidad de implementación, esto no quiere decir que sea mejor o peor que otros métodos, sino

solo simple de preparar. Dada la estructura modular de este protocolo, el algoritmo usado aquí puede ser intercambiado por cualquier otro disponible. Cabe destacar que en general las métricas de anidamiento se calculan sobre redes de interacción binarias, es decir, para nuestro caso podemos utilizar aquella versión bidimensional de la red que resulta del cálculo de la similitud de Jaccard o Tanimoto, convertidas a binarias, seleccionando la ausencia o presencia de una relación bajo un umbral arbitrario de la métrica elegida. NODF es un algoritmo que calcula la metrica de anidamiento a partir de esta matriz binarizada, adicionalmente WNODF ofrece la posibilidad de calcular una métrica de anidamiento directamente sobre la matriz de biadyacencia original.

Previamente a calcular el anidamiento, se debe preparar la matriz de biadyacencia, moviendo filas y columnas de forma de maximizar la presencia de celdas no nulas a la izquierda superior de la matriz. De esta forma se estandariza el cálculo, que depende del orden en que se presentan las filas y columnas en la matriz.

Sea `M` la matriz de biadyacencia:

```
import numpy as np

M = M[np.argsort(np.sum(M, axis=1))[::-1]]\
  [:,np.argsort(np.sum(M, axis=0))[::-1]]
```

Luego, sea `M` la matriz de biadyacencia preparada y `t` el umbral de decisión de presencia u ausencia de interacción arbitrario:

```
NODF(M > t)
WNODF(M)
```

3.3.3.3. Modularidad

Para el cálculo de modularidad aprovechamos la implementación ofrecida por los paquetes `python-louvain` y `NetworkX` (Hagberg, Schult & Swart, 2008) para el lenguaje de programación `Python`. Estos programas utilizan el algoritmo de Louvain para detectar particiones dentro de una red que maximicen su modularidad.

Sea `M` la matriz de biadyacencia:

```
import community as community_louvain
import networkx as nx
import scipy.sparse as sp
from networkx.algorithms import bipartite

G = bipartite.from_biadjacency_matrix(sp.csr_matrix(M))
```

```
partition = community_louvain.best_partition(G)
community_louvain.modularity(partition, G)
```

Con las herramientas presentadas en este protocolo, se resuelven dos problemas: el primero es la complejidad de la asignación de unidades taxonómicas a secuencias desconocidas, siendo este un problema particularmente importante en el estudio de organismos virales encontrados en el libres en medio-ambiente, lo cual resolvemos de cierta manera con la primera etapa de nuestra propuesta. En segundo lugar, entregamos herramientas para potenciar el estudio de relaciones fago-hospedero en el ambiente las cuales son de gran relevancia en ecología global.

Capítulo 4

VALIDACIÓN DE LA SOLUCIÓN

4.1. Racionalización

Dada la estructura modular del protocolo antes presentada, se presenta a continuación la verificación y validación del funcionamiento del protocolo, la cual fue desarrollada de forma independiente para cada módulo. En primer lugar, se evaluó la precisión de la asignación taxonómica propuesta usando un conjunto de metagenomas construidos aleatoriamente. En segundo lugar, se verificó la precisión en las predicciones de interacción fago-hospedero obtenidas a partir del protocolo, la cual fue basada en predicciones realizadas sobre la misma lista de metagenomas aleatorios usando herramientas de aprendizaje-máquina. Posteriormente, se estudió un caso práctico la aplicación de este protocolo en metagenomas reales recolectados desde ambientes de agua dulce obtenidos desde bases de datos públicas.

4.2. Etapa 1: Proceso de Asignación taxonómica

4.2.1. Preparación del caso

Se recopiló un set de datos de once metagenomas sintéticos desarrollados por el software MetaPhlAn (Segata y col., 2012). Este programa y sus datos de prueba se encuentran disponibles de forma pública [1] y contienen la información suficiente para determinar la certeza con la que nuestra herramienta identifica OTUs de manera adecuada. El dataset de metagenomas sintéticos ha sido generado tomando secuencias aleatorias de genomas de especies bacterianas, virales y eucariotas previamente descritas, siguiendo una distribución de probabilidad uniforme, los que fueron dispuestos en forma de archivos `FASTA` de forma equitativamente distribuida en abundancia respecto al total de especies que se usaron para estos efectos. De esta manera, el dataset sintético ha sido generado mezclando datos reales de forma equitativa y equiprobable. La distribución de organismos presentes se resume en la Figura 41.

[1] https://huttenhower.sph.harvard.edu/metaphlan

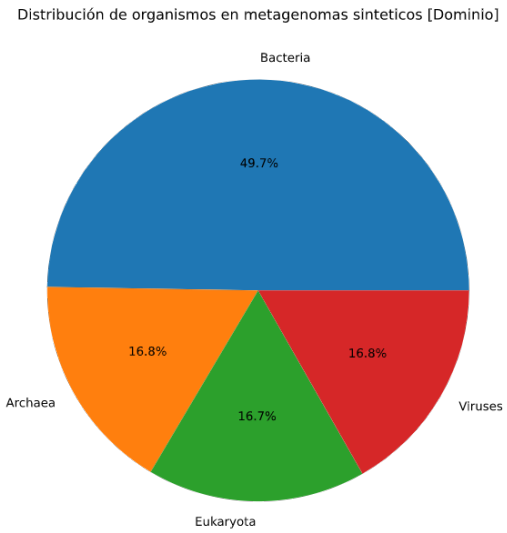


Figura 41: Se aprecia como distribución de organismos en los metagenomas sintéticos es dominada fuertemente por organismos bacteriales. El resto del los dominios principales se ha logrado construir con una distribución similar.

4.2.2. Reporte inicial de la ejecución del proceso de asignación

Un reporte inicial de la asignación taxonómica se obtuvo producto de la ejecución de los programas dispuestos para la etapa 1 utilizando parámetros por defecto, excepto para el paso de filtrado por según umbral repetición de asignaciones. Al momento de verificar la repetición de asignaciones se realizó una optimización de los parámetros del programa por medio de una búsqueda en grilla (*grid-search*), culminando en diferentes versiones de la codificación *one-hot* del sistema, resultados que se resumen a continuación.

El análisis de la asignación taxonómica obtenida con esos parámetros muestra que el número total de especies identificadas (no el número de asignaciones correctas) disminuye respecto al aumento del valor umbral de corte específico utilizado (Figura 42). Esta tendencia observada en virus y procariontes (bacterias y arqueas) corresponde a la esperada previamente ya que indica que en la medida que se es más exigente en la cantidad de genes requeridos para determinar la presencia de una especie, menor es la cantidad de especies encontradas en ese ambiente. Estas poseen una tendencia similar a la encontrada en la curva de abundancia de rango (o diagrama de Whittaker), la que es una forma gráfica utilizado por los ecólogos para mostrar la abundancia relativa de especies, un componente de la biodiversidad [2]. Respecto a eucariontes, tal como se espera, el programa no identificó ninguna especie, desechando a todas las secuencias identificadas como tal (Figura 42).

Adicionalmente, se observó que hay un alto nivel de sobre-estimación respecto a la asignación de especies procariontes, particularmente para especies clasificadas como bacterias, en todos los umbrales de corte. Esta sobre-estimación puede explicarse debido a al sesgo en las bases de datos como REFSEQ previamente explicada en el Capítulo 1, y se sugiere

[2]https://en.wikipedia.org/wiki/Rank_abundance_curve

que en optimizaciones de esta propuesta este sesgo pueda reducirse usando la estrategia heurística de filtrar por un número mínimo de repeticiones de asignaciones. La curva posee un comportamiento asintótico hacia el número efectivo de especies presentes en el ambiente (conjunto de metagenomas) analizado, lo que permite inferir que en algún momento de su proyección, nuestra herramienta se acercará a pronosticar certeramente el número real de especies presentes en el ambiente (Figura 42).

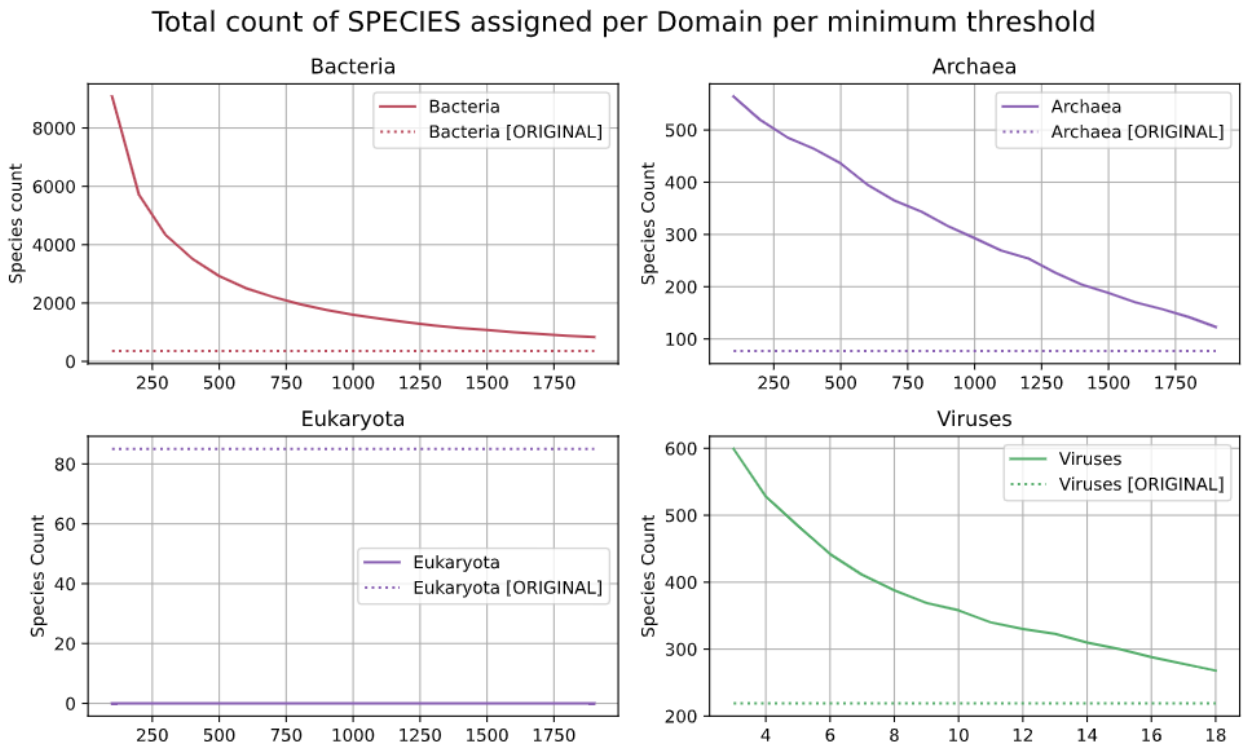


Figura 42: Variación del número total de especies identificadas (eje y) versus del valor umbral de corte específico utilizado (eje x). La línea punteada indica el número real de especies presentes en el total de las muestras (metagenoma sintético) y la línea continua indica el número de especies presentes identificadas por el protocolo.

4.2.3. Validación de la asignación taxonómica del sistema

Para cuantificar la precisión de la secuencia de programas presentados para la asignación taxonómica, se procede a calcular las métricas de contingencia entre los resultados obtenidos versus la información original presente en los metagenomas de prueba. Nos vemos enfrentados entonces a un problema de clasificación binaria, donde las clases a predecir son ”Especie Presente” o ’Especie No Presente”. Para validar esto verificamos que una especie que fue calificada presente por nuestro método efectivamente aparece como presente en el ambiente ya que disponemos de esa información (metagenomas sintéticos) . Sin embargo, al verificar que una especie no está presente en el ambiente, tenemos que considerar que aquellas especies que efectivamente no están presentes son todo el universo de especies anotadas en bases de datos (millones de especies). Esto genera un

sesgo natural y significativo produciendo un problema de clasificación desbalanceado, que prioriza la asignación de especies no presentes (mayor cantidad) por sobre las especies presentes (menor cantidad). Un ejemplo de este desbalance se aprecia en que en nuestro listado de metagenomas sintéticos, las 'Especies Presentes" representan menos del $1\,\%$ (precisamente $0{,}14\,\%$) de la base de datos de taxonomías de NCBI.

Posterior al procedimiento de búsqueda en grilla, se realizó la visualización de curvas ROC (Receiver Operating Characteristic) en la Figura 43 para diferentes umbrales de corte por repetición de asignaciones, a distintos niveles taxonómicos. Los resultados de este procedimiento indicaron que el desempeño del asignador de OTUs crece muy rápidamente hasta cierto nivel máximo (AUC $>$ a $80\,\%$) permitiendo una tasa reducida de falsos positivos (Figura 43, izquierda). Este comportamiento es reflejo de la restricción de los umbrales observada previamente (Figura 42) . Es decir que nuestra herramienta alcanza un desempeño máximo para asignar OTUs después de ajustar los umbrales para entregar un volumen cercano a la realidad de asignaciones de los microorganismos. Esto se condice con lo observado en el gráfico Precision-Recall dónde precisión varía respecto a una razón de falsos positivos invariantes sugiriendo que el desempeño depende fuertemente de la cantidad de información que se está utilizando (Figura 43, izquierda).

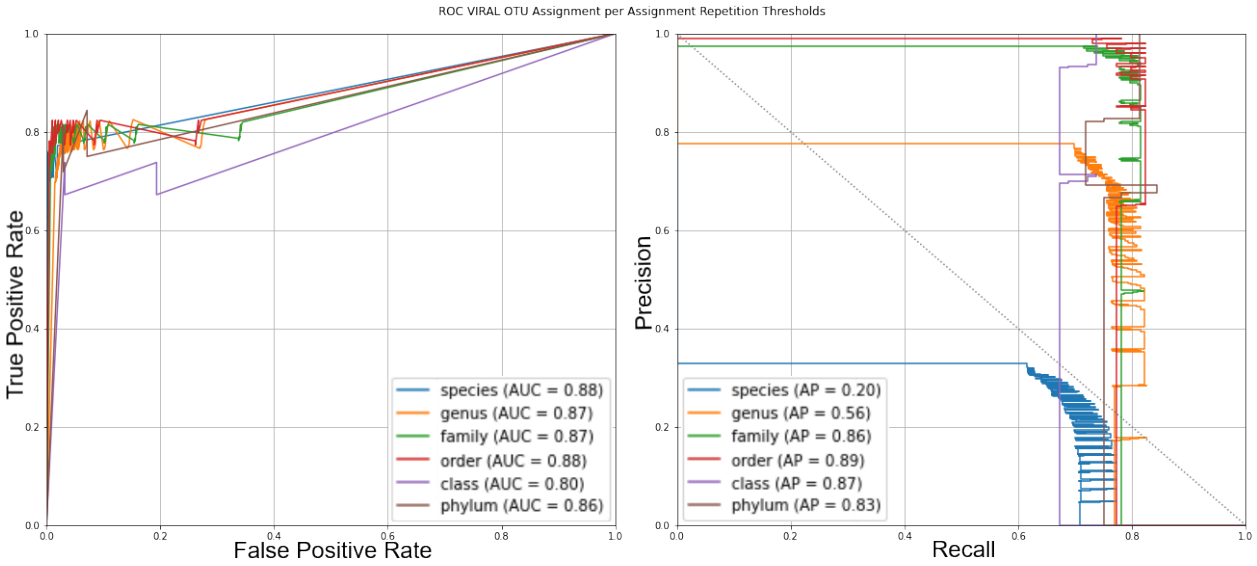


Figura 43: La curva ROC (izquierda) nos muestra como efectivamente hay un buen grado de correctitud en las asignaciones de existencia y taxonomía de especies sin embargo la curva Precision-Recall (izquierda) nos muestra que realmente esto correspondería a un caso con datos desbalanceados donde la mayor cantidad de las asignaciones están ocurriendo hacia elementos que se encuentran en el grupo negativo.

El adecuado desempeño de herramienta durante la selección de las especies presentes en los metagenomas sintéticos puede estar influenciado por el hecho de que los elementos presentes en el ambiente es un subconjunto muy acotado del universo (cantidad total de especies presentes en la base de datos), por lo que los buenos resultados pueden responder principalmente a las asignaciones de especies que están ausentes. Para evitar

este posible sesgo, se llevó a cabo una validación de la metodología por sí sola, haciendo *Leave-one-out cross-validation* en los metagenomas, dejando de depender de los valores de las asignaciones reales y validando que la técnica en sí sea capaz de entregar resultados consistentes. Los resultados de este procedimiento validan la consistencia del método, confirmando que hay una mejoría sustantiva respecto al azar (Figura 44, izquierda). Se llevó a cabo este mismo procedimiento de validación para microorganismos hospederos (Figura 45) y virus (Figura 46). En la primera de estas validaciones, se observó que la asignación de hospederos (bacterias) posee una alta consistencia Figura 45, a pesar de que son las bacterias las que aportan la mayor cantidad de volumen a las bases de datos de secuencias. La segunda validación referida a la asignación de virus, la efectividad en la asignación de especies se comportó de manera más variable que en bacterias. Esta situación responde a la "volatilidad" con la que se encuentran virus en el ambiente, entendida como la pobre representatividad de su biodiversidad en las bases de datos, generando así mayores dificultades en la asignación taxonómica que las encontradas en bacterias (Figura 46).

Los resultados de validación observados en las curvas ROC indican que el método de selección de taxonomías dependen directamente de los valores de los umbrales de filtrado por repetición de asignaciones. Esta heurística ha demostrado ser clave al momento de reducir la sobre-estimación inicial realizada por la búsqueda de homologías, lo cual se observó previamente (Figura nn). Estos resultados demuestran que el protocolo es capaz de determinar correctamente la presencia o ausencia de microorganismos bacterianos, arqueas y virus de manera confiable. Posteriormente, la segunda validación permite inferir que el protocolo permite la asignación taxonómica de las especies presentes en el ambiente, presentando una baja tasa de error. Los atributos de efectividad de este protocolo validados en metagenomas conteniendo información conocida permite presentar la herramienta diseñada durante la presente tesis como una plataforma adecuada para la determinación de presencia e identificación de especies microbianas en metagenomas realizados a partir de muestras medioambientales, las que presentan una mayor complejidad. (Casjens, 2003; G. F. Hatfull & Hendrix, 2011).

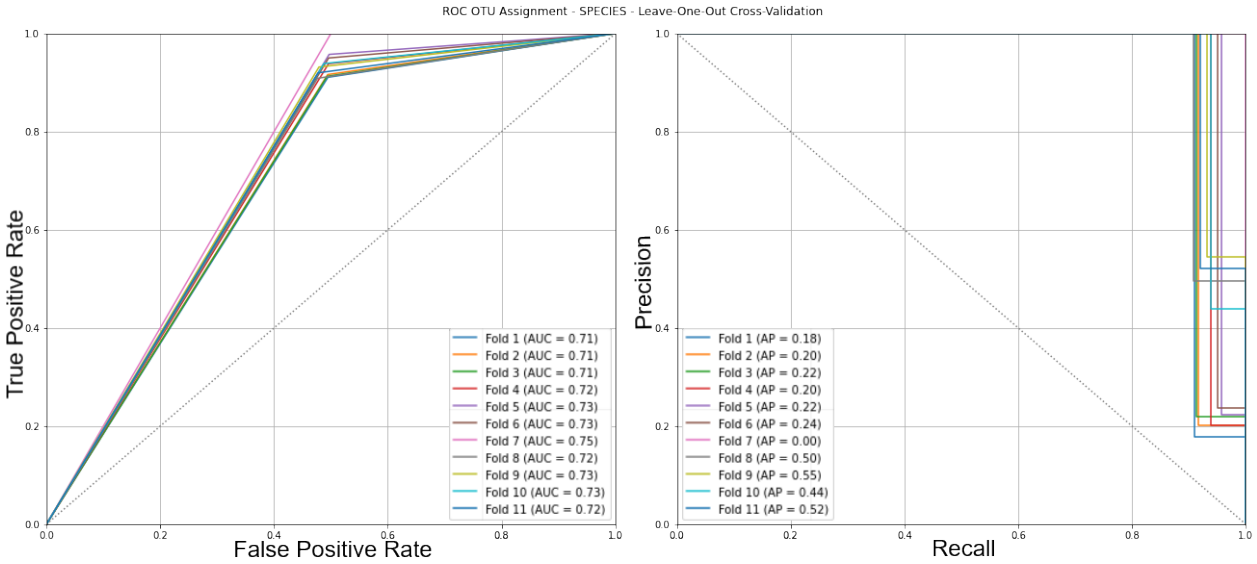


Figura 44: Se observa tanto en la curva ROC (izquierda) como en la curva Precision-Recall (derecha) una mejoría importante respecto a la linea base de validación (el azar), sin embargo en la curva Precision-Recall se sugiere fuertemente una influencia del volumen relativo de una clase sobre otra (asignaciones de no presencia) dandonos una baja precision con alta sensitividad.

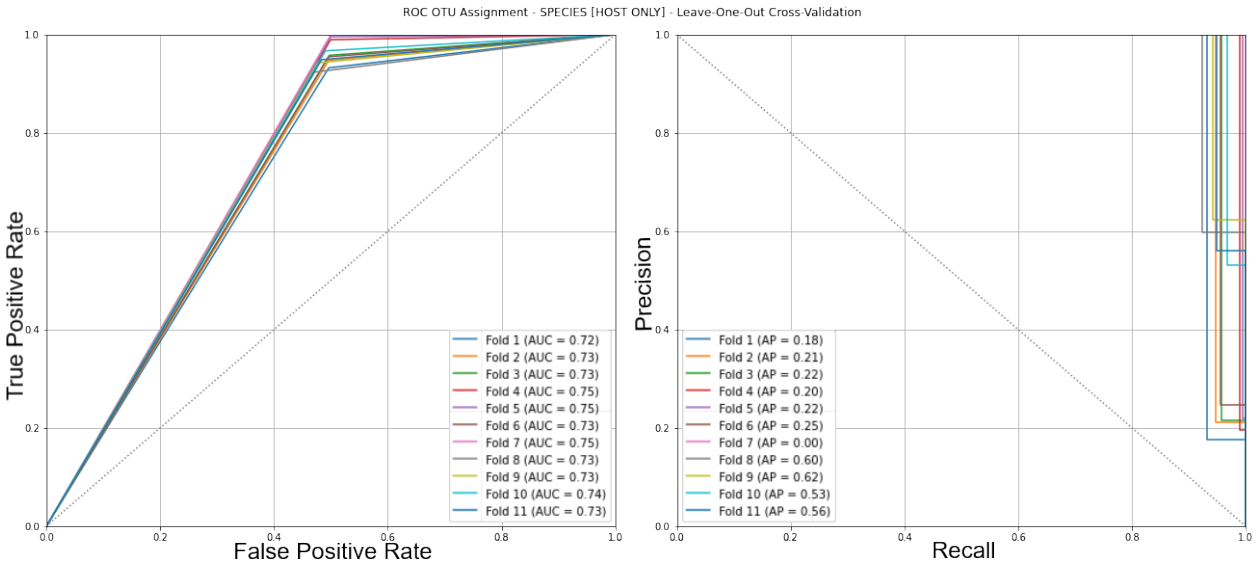


Figura 45: Se observa que la asignación de hospederos es muy consistente usando esta metodología, replicandos ambas curvas ROC (izquierda) y Precision-Recall (derecha) a las observadas en las asignaciones con la totalidad de las especies (Figura 44). Cabe recordar que son las bacterias las que aportan la mayor cantidad de volumen a las bases de datos de secuencias, lo cual se ve reflejado con la facilidad en que podemos asignar la presencia de una.

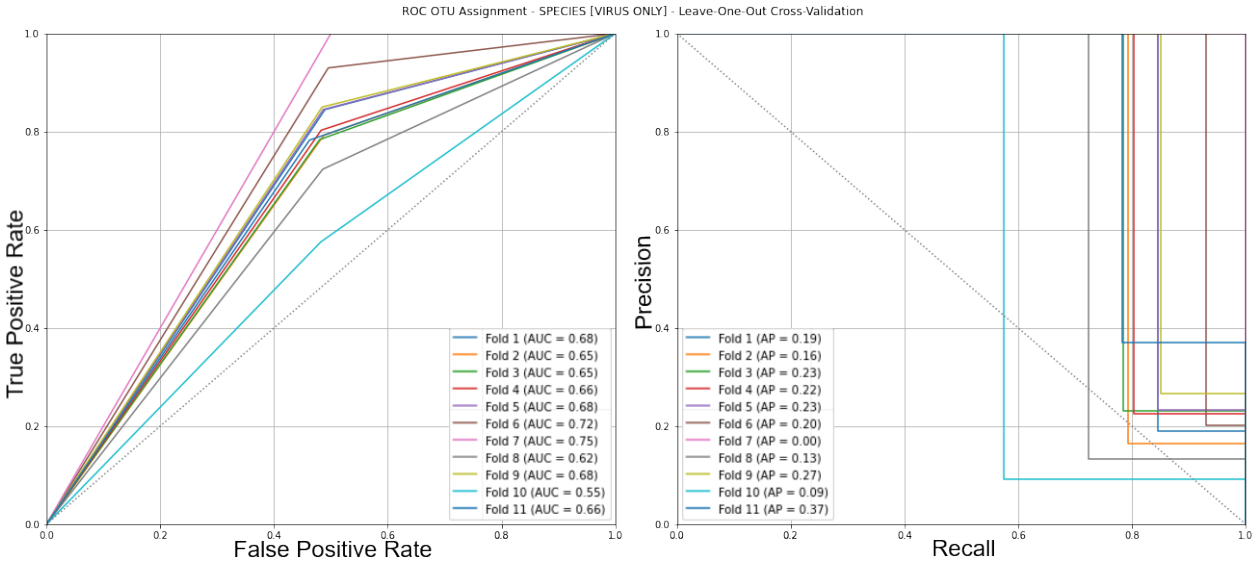


Figura 46: Observamos que la asignación de organismos virales sufre de más variabilidad, esto se condice con la volatilidad con la que se encuentran en el ambiente y son captados por las bases de datos en las cuales se buscan homologías, estando presente el mismo organismo en pocos metagenomas a la vez obteniendo esta variación mayor entre los distintos grupos de prueba. Sin embargo, la media sigue mejorando sobre el caso aleatorio en la curva ROC (izquierda), por lo que se puede verificar la consistencia de este método.

4.3. Etapa 2: Construcción de redes de interacción

4.3.1. Confirmación de interacciones a partir de las métricas de coexistencia

Para construir la red de interacciones fago-hospedero se utilizó la estrategia de calcular ambas similitudes de Tanimoto y Jaccard con peso para todas las posibles relaciones de coexistencia en el sistema, y se tomó como interacciones confirmadas aquellas que superan cierto umbral arbitrario para la métrica de similitud calculada. Al igual que en el caso de la asignación de OTUs, en esta situación disponemos de un problema de clasificación binaria en el que se intenta clasificar si una relación de interacción, es decir, un arco en el grafo denso generado por todas las interacciones posibles en el sistema, existe o no. El número máximo de posibles interacciones fago-hospedero corresponde a la coexistencia completa entre todos los virus registrados en la base de datos de taxonomías con todos las bacterias y arqueas, el cual configura un grafo de alta densidad formado por $\texttt{Nº Virus} \times \texttt{Nº Hospederos}$ arcos. Sin embargo, es evidente que no todas estas relaciones ocurren naturalmente o poseen alguna relevancia ecológica, siendo la realidad un subconjunto de este máximo.

Ya que no se dispone de una forma de validación de aquellas interacciones que no ocurren en el ambiente, como tampoco no es recomendable utilizar el universo completo como base de comparación, resulta imposible validar la efectividad de la herramienta en cuanto

exactitud. De esta manera, solo podemos validar la técnica propuesta en términos de su consistencia para confirmar interacciones de coexistencia.

Para evaluar la efectividad del protocolo propuesto en esta tesis se realizó una validación del tipo "*Hold-Out*", en la que se separaron los metagenomas de prueba en dos conjuntos: entrenamiento y validación, los que contenían el $70\,\%$ y $30\,\%$ de los metagenomas, respectivamente. Las métricas de Tanimoto y Jaccard se calcularon para ambos grupos de forma independiente. Posteriormente, se comparó los valores obtenidos de similitud de las redes obtenidas en ambos grupos, usando solo los organismos presentes en el grupo de prueba. Es decir, se ha reducido artificialmente el grupo de entrenamiento para ser contenido dentro del grupo de validación, para de esta forma evaluarlos a través de análisis de negativos-correctos y curvas ROC-AUC.

Los resultados de esta validación indican que la herramienta no posee una alta eficiencia en la asignación de OTUs a distintos niveles. Se observa como, especialmente para niveles taxonómicos bajos como *especies* y *géneros*, los predictores a diferentes niveles taxonomicos no mejoran significativamente respecto a la linea base de comparacion, a pesar de esto, niveles mas altos como *ordenes* y *clases* muestran una mejoría aceptable. Las curvas *Precision-Recall* que acompañan al análisis ROC-AUC responden a un comportamiento típicamente asociado a conjuntos de datos desbalanceados hacia una clase en particular, en nuestro caso, este grupo corresponde a los datos asignados como "No coexistencia". Las métricas utilizadas para determinar la confirmación de una interacción de co-existencia (Tanimoto y Jaccard) no toman en consideración este desbalance en los datos, por lo que aunque esta técnica puede considerarse un aceptable primer enfoque en un análisis metagenómico, especialmente observando grupos taxonómicos superiores, la aplicación correcta de este protocolo en muestras ambientales requiere del desarrollo de una metodología más efectiva para la confirmación de interacciones entre virus y hospederos.

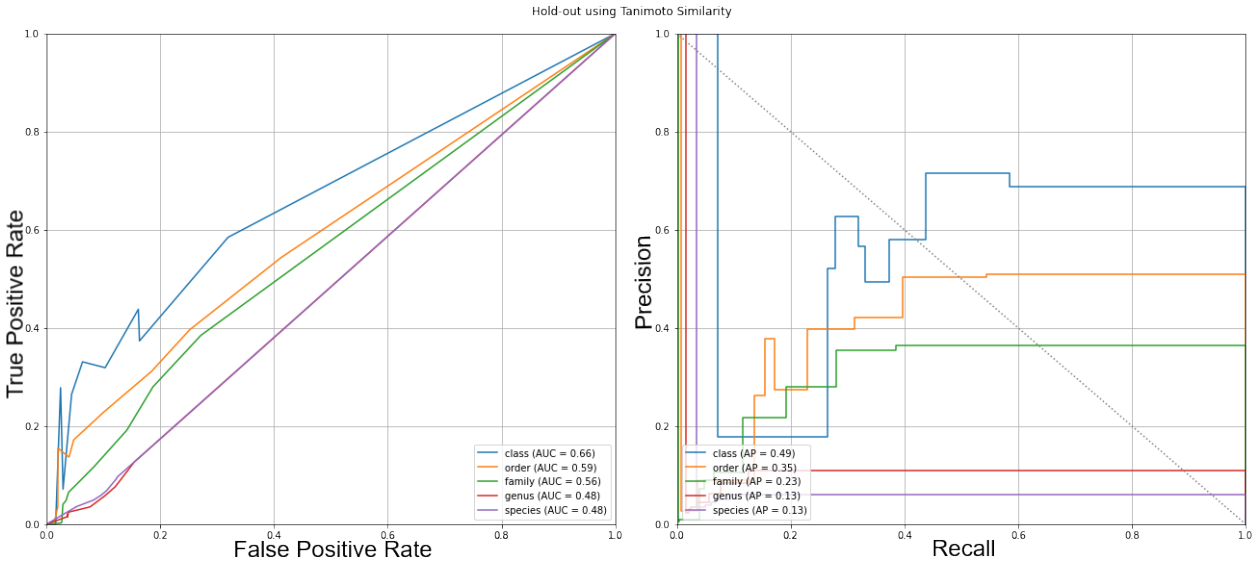


Figura 47: Curva ROC (izquierda) y Precision-Recall (derecha) para la asignación de taxonomías a diferentes rangos a partir de la similitud de Tanimoto

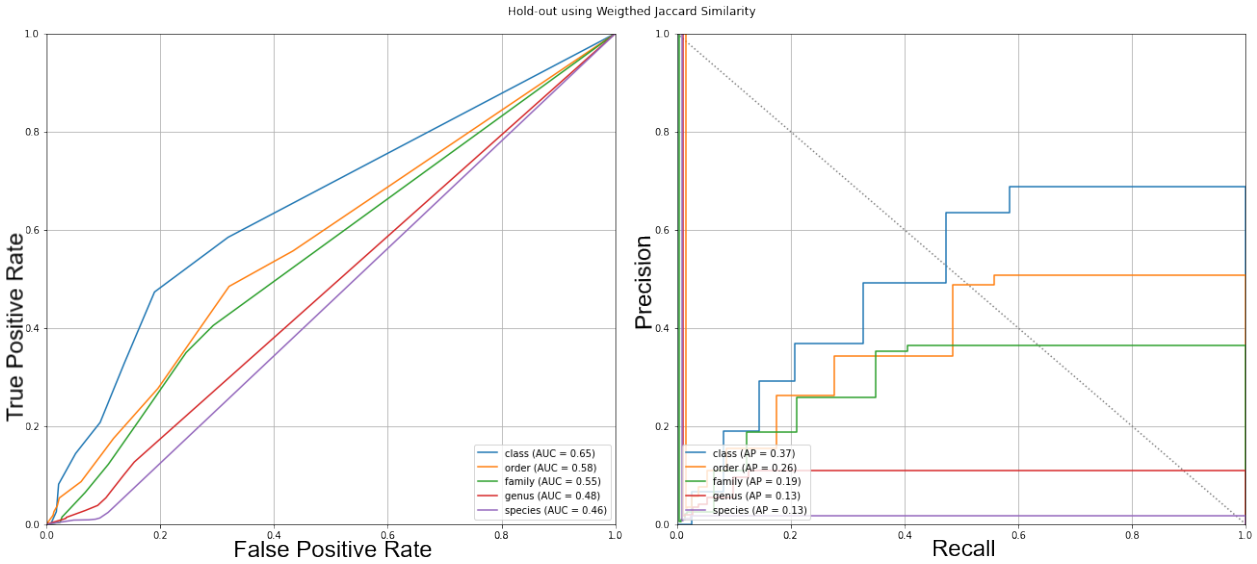


Figura 48: Curva ROC (izquierda) y Precision-Recall (derecha) para la asignación de taxonomías a diferentes rangos a partir de la similitud de Jaccard con peso

4.4. Caso de aplicación

4.4.1. Introducción

Como es de interés explorar las interacciones entre especies en sistemas fago-hospedero en ambientes de agua dulce, se han obtenido desde las bases de datos JGI y MG-RAST un total de ciento treinta y nueve (139) metagenomas sin procesamiento, sumando un

total de 185 GigaBytes de información. Estos metagenomas fueron elegidos como una muestra representativa de ambientes de agua dulce en diversas partes del mundo. Meta-información sobre estos metagenomas, como locación física y conteo de secuencias, se encuentra disponible en el Apéndice G.

Posterior a la recopilación de la información, se aplicó la primera etapa del protocolo propuesto para realizar la asignación taxonómica y selección de especies virales, bacterianas y de arqueas junto a la transformación y preparación de los datos para su visualización.

De los 139 metagenomas analizados se lograron identificar un total de 25006 especies, de las cuales 23670 corresponden a especies bacterianas, 618 a arqueas y 1336 a especies virales, registrándose el mismo problema de sesgo hacia los especímenes bacterianos observado durante la validación con metagenomas aleatorios. Con el objeto de ejemplificar la aplicación de nuestra propuesta con datos obtenidos desde estudios realizados, acotamos el análisis enfocándonos en un grupo taxonómico de interés específico detectado dentro de los metagenomas. Para esto, se optó por enfocarnos en el grupo de bacterias *phylum Bacteroidetes*, las que constituyen una proporción alta de la microbiota residente en el cuerpo humano y corresponden a especies mayormente nativas con alta prevalencia en estos ambientes de agua dulce, tales como lagos, lagunas, estuarios y ríos, jugando un rol importante en la regulación de los ciclos biogeoquímicos de estos ambientes. En consecuencia, se descartó al final de la primera etapa a todas aquellas especies de hospederos que no pertenecen al *phylum Bacteroidetes*. Adicionalmente, para enfocarnos en aquellos genotipos virales que más afectan a las Bacteroidetes, se descartó a todas aquellos virus que no pertenecen a los grupos de familias **Podoviridae**, **Myoviridae**, **Ackermannviridae**, **Microviridae** y **Siphoviridae**, siendo estos los grupos que poseen una mayor representatividad de información en la literatura como aquellas que contienen organismos con mayor importancia medioambiental.

4.4.2. Grupo de Interés Ecológico: Phylum Bacteroidetes

El *phylum* "**Bacteroidetes**" está compuesto de bacterias que están ampliamente distribuidas en el medioambiente, desde suelos, sedimentos y aguas oceánicas, así como también en asociado al sistema digestivo y la piel de los animales (Aislabie y col., 2006; Bäckhed, Ley, Sonnenburg, Peterson & Gordon, 2005).

En los ecosistemas oceánicos, los Bacteroidetes constituyen una fracción significativa del bacterioplancton especialmente en las zonas costeras, en donde representan entre el 10 % y el 30 % del total recuentos microscópicos (Alonso, Warnecke, Amann & Pernthaler, 2007). Se ha demostrado que los Bacteroidetes se han especializado en la biodegradación de compuestos de alto peso molecular y poseen una preferencia por el crecimiento asociado a partículas en suspensión, superficies u otras células de (micro)algas (Pedrós-Alió y col., 2013). Diversos estudios han demostrado que los Bacteroidetes exhiben una enorme diversidad fenotípica y metabólica. La mayoría de los aislados de Bacteroidetes descritos poseen metabolismo del tipo quimioorganótrofo, el cual es capaz de obtener energía a través de la oxidación de compuestos orgánicos, a través de la respiración y/o

fermentación (Newton, Jones, Eiler, McMahon & Bertilsson, 2011b). Sin embargo, algunos aislados marinos, tales como *Polaribacter sp. Cepa MED152* y *Dokdonia sp. Cepa MED134*, han demostrado capacidades fototróficas, comprobando una alta versatilidad metabólica y fenotípica del Phylum (Gómez-Consarnau y col., 2007).

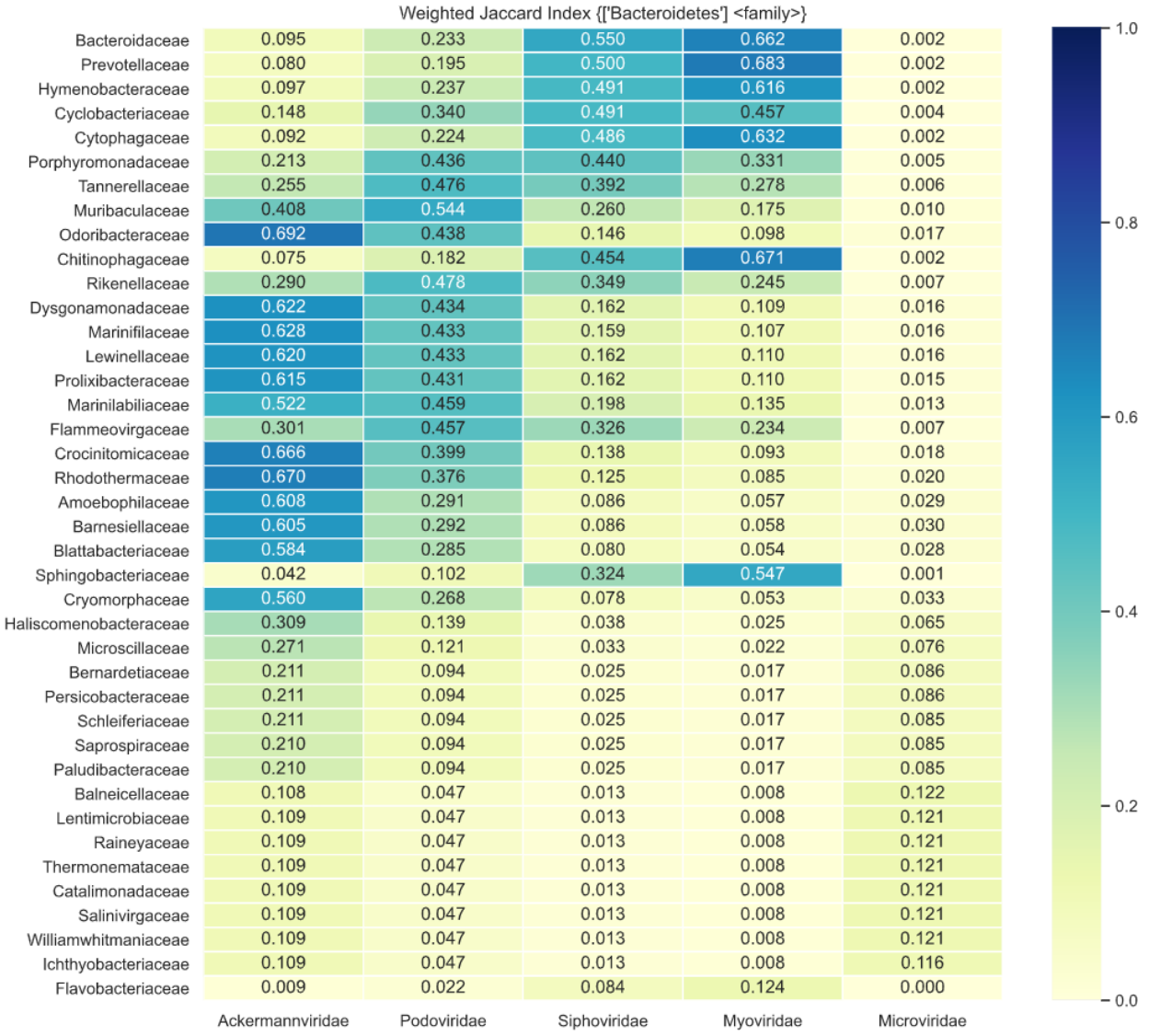


Figura 49: Se observan relaciones mayormente débiles entre todos las familias del grupo, especialmente para aquellos relacionados a Microviridae, sin embargo hay familias con una muy fuerte coexistencia como lo es el caso de Odoribacteraceae con Ackermannviridae, familias que presentan la coexistencia mas fuerte. El heatmap sugiere que la red podría tener un alto anidamiento.

4.4.2.1. Observaciones

Los resultados obtenidos por nuestros análisis indican que existen relaciones importantes entre ciertos linajes bacterianos y genotipos virales. Por ejemplo, una de las relaciones más

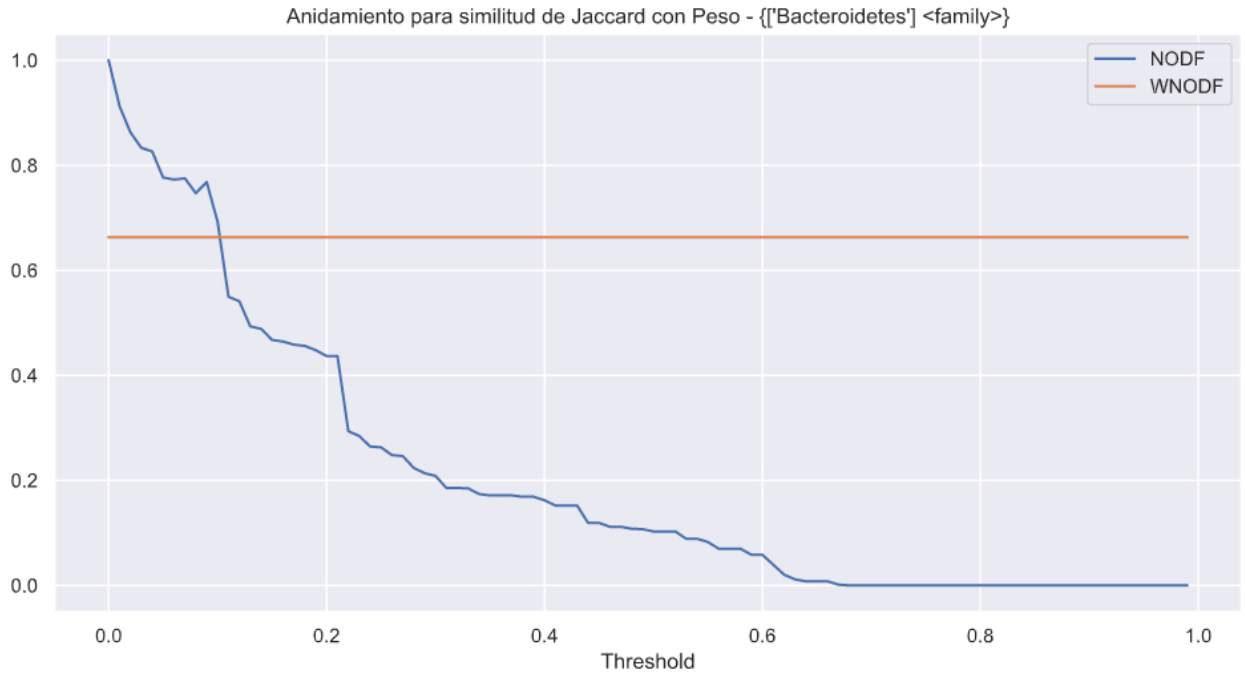


Figura 410: Ambos anidamientos con peso y sin peso se cruzan alrededor del umbral $0{,}1$, sugiriendo que en esta vecindad pueda encontrarse el umbral que significativamente decida si la relación de coexistencia es tal o es inexistente.

fuertes comprende al grupo de bacterias pertenecientes a la Familia *Prevotellaceae* y la de virus pertenecientes a la familia *Miovyridae*, relación que posee un alto valor de indice de Jaccard ($0{,}683$, Figura 49) lo que sugiere una importante coexistencia en el ambiente de genes pertenecientes a ambos grupos. Esta relación se observa también en otros ambientes en donde esta relación se establece de manera robusta. Un ejemplo de esto, se observa en la microbiota residente en tracto digestivo, en donde un alto número de los espaciadores incorporados en el sistema CRISPR presente en aquellas cepas pertenecientes a *Prevotellaceae* corresponden a virus de la familia *Miovyridae* (Fujimoto y col., 2020), como también esta misma relación ha sido demostrada a través de herramientas moleculares (Devoto y col., 2019). Una relación similar también se observó entre genes pertenecientes a bacterias de la Familia *Bacteroidaceae* y virus del mismo grupo, residentes del microbioma humano que incluso se transmite vericalmente desde madre a recién nacidos durante el proceso de parto (Maqsood y col., 2019). Otra relación en donde se registran altos valores del Indice de Jaccard corresponde a la registrada entre genes pertenecientes a bacterias de la Familia Odoribacteareceae y virus de la familia Ackermannviridae. Este tipo de relación debiesen ser exploradas por otros estudios en el futuro, ya que para nuestro conocimiento este tipo de relaciones no ha sido esclarecida en estudios previos (Zheng y col., 2019).

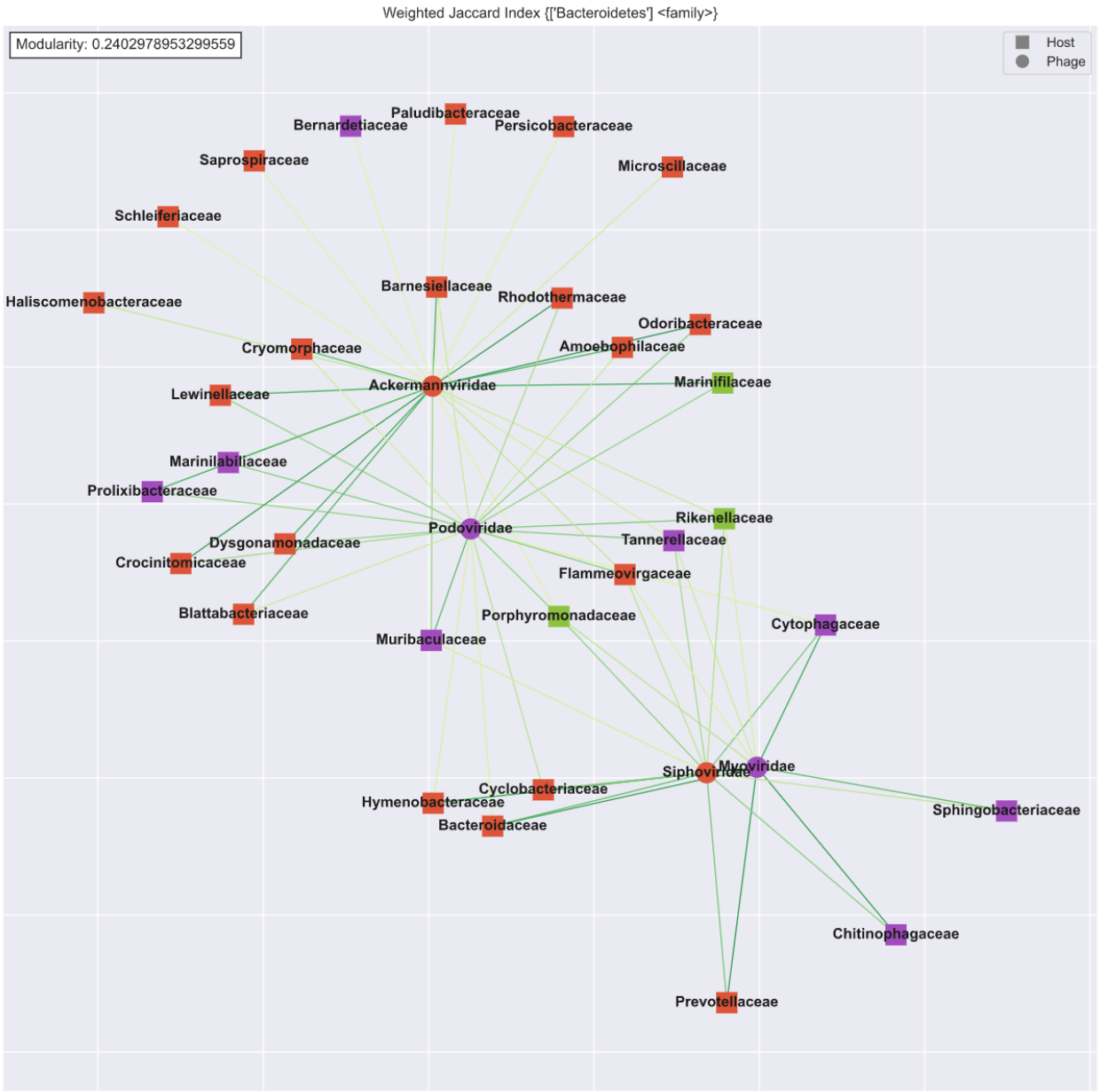


Figura 411: Se grafican solo aquellas relaciones que poseen un valor de similaridad de Jaccard con peso mayor a $0{,}2$. Podemos apreciar que se detectan tres clusteres o grupos de familias que interactúan fuertemente, sin embargo con un valor de $24\,\%$ de modularidad, se sugiere que estas interacciones modulares no son mas fuertes que las relaciones de anidamiento.

Capítulo 5

CONCLUSIONES

5.1. La colección de estampillas

Es común para las ciencias en general confiar en herramientas computacionales para realizar cálculos complejos, tales como aquellos que usan grandes números astronómicos, operaciones de punto flotante y complejas simulaciones de física de alta energía, o hasta encontrar soluciones analíticas a problemas de matemática simbólica. Todos estos problemas tienen en común el caso de trabajar con datos elegantemente estructurados y consistentes, es decir, recurrentemente nos encontramos con tablas, matrices y tuplas con números moviéndose en ejes comparables entre si, como el tiempo, temperaturas, posiciones en el plano, y a escalas comparables, como flotantes entre cero y uno u valores en escalas conocidas. Independientemente del problema al que nos estemos enfrentando, nos encontramos con alguna de las estructuras que puede ser fácilmente traducida a un problema computacional que puede ser resuelto usando las ultimas tendencias en computación de alto rendimiento, como operaciones en unidades gráficas y tensoriales.

En el caso de las ciencias de la vida tenemos suerte si logramos convertir parte de los datos en una estructura computacionalente amigable. Esto a simple vista pareciera no ser un problema importante, pero se vuelve uno cuando se toman en consideración los importantes volúmenes de datos no estructurados que los diversos experimentos biológicos están generando en la actualidad. Ya sean las cadenas de *strings* que son las secuencias de ADN, nodos de un árbol como las taxonomías o tratándose simplemente de aliases o diferencias en la nomenclatura de organismos, existen grandes diferencias entre diferentes bases de datos, e incluso mas frecuentemente de lo que parece, diferencias de formato dentro de una misma base de datos. De hecho, el volumen de información biológica no ha hecho más que incrementar exponencialmente desde los últimos diez años. Múltiples y ambiciosos intentos de estudiar sistemas complejos como la investigación del microbioma humano han ayudado a rellenar las bases de datos con miles de millones de estas secuencias que encontramos fácilmente en nuestro cuerpo. Estos estudios marcan los primeros pasos de la biología en la ciencia de datos.

El punto clave en el que el "*big data*" producido por las ciencias de la vida se diferencia del

generado por la física e ingeniería es el grado de estructuración de los datos y la cantidad de preparación que es necesaria realizar antes de extraer información de utilidad. Se puede decir que ”la biología es una colección de estampillas” en el sentido de que de la gran cantidad de datos que se producen, estos tienen muy poco en común y son incomparables.

El ejemplo mas fuerte y trascendental es el simple hecho de que los virus son organismos *polifiléticos*. En el árbol filogenetico, las características de los miembros son heredadas de sus antepasados directos, los virus, en cambio, no comparten ningún gen entre si o entre varios linajes virales, no tienen antepasados directos. Mientras toda la vida celular posee un único origen común, los virus poseen múltiples orígenes evolutivos. Un árbol es inútil para describir el traspaso de características genéticas a través de la evolución, el proceso evolutivo de los virus es mas parecido a un borroso grafo, sin raíces claras.

Este inconveniente ha sido el principal problema enfrentado durante el desarrollo de esta memoria, si bien es común tratándose de ciencia de datos que el 80 % del esfuerzo este concentrado en la transformación y preparación de datos, dedicando el otro 20 % restante para el verdadero análisis de datos, esto fue particularmente notorio y sentido en este proyecto. Efectivamente, puede considerarse que la ”Etapa 1: Asignación Taxonómica” por completo corresponde unicamente a transformar, validar y filtrar las tablas m8 resultadas de la ejecución de BLAST para obtener un volumen de información reducido y manejable, tratándose esta transformación (*data-wrangling*) anecdoticamente casi un 90 % del esfuerzo de trabajo, incluso durante la segunda etapa (Etapa 2: Construcción de redes de interacción ecológicas), el trabajo de preparación de datos realizado fue mayor al que se logra apreciar en este escrito.

5.2. El software según la gente que no es de software

La regla $80/20$ refiriéndose a la relación entre esfuerzo y resultados no es algo exótico en la ciencia de datos, de hecho, es un concepto bien interiorizado. Si bien esta pareciera ser una dificultad fuerte cuando se lidian con datos no estructurados (y lo es), el verdadero problema yace en el software escrito estudiantes de doctorado una decada atrás. Nadie se atreve a tocarlo por que nadie tiene el interés de aprender programación. Por si no fuera poco, el doctor que escribió el software original usualmente no tiene ninguna clase de indoctrinacion en buenas practicas de programación, tal que, laboratorios que son capaces de contratar a un informático con dedicación completa usualmente concluyen con el programador reescribiendo el software completo desde cero con el objetivo de que sea mantenible por otros. Asimismo, perdiendo la lógica del negocio y aquellos detalles que solo un investigador con los conocimientos apropiados tiene en consideración, los cuales están enterrados en el *código spaghetti* del software original.

Lo descrito anteriormente es totalmente no una exageración y un problema real al que esta memoria se vio enfrentado en varias instancias.

Las alternativas mencionadas en el Capítulo 2, Subsubsección 2.4.1.1 para el calculo de la propiedad de anidamiento, corresponden a un pequeño subconjunto de una gran cantidad

de algoritmos propuestos en el estado del arte con la misma finalidad, sin embargo solo unos pocos fueron posibles de encontrar empaquetados en software ”listo para usar”. Sin embargo, nos vimos en la necesidad de implementar nuestra propia version del único algoritmo que fue accesiblemente explicado en su respectiva publicación, debido a que el software ya disponible lo estaba para plataformas no compatibles con las nuestras por lo que requerían una enorme cantidad de trabajo en rediseñar las salidas y entradas de nuestros propios programas para tolerar la interoperabilidad con software que no ha sido actualizado en varios años, sin binarios disponibles para plataformas otras que MSWindows, que no cuenta con su código fuente disponible de manera abierta u esta integrado de manera *ad-hoc* a un paquete obscuro para el lenguaje R, el cual ha caído en deshuso para estos propósitos, dejando un volumen no despreciable de código olvidado, que esta siendo requerido por investigaciones modernas.

La base de datos del International Committee on Taxonomy of Viruses (ICTV), es un archivo MS Excel con tres hojas: Una hoja con varias tablas arbitrariamente posicionadas indicando información sobre el versionamiento. Otra hoja con una tabla con la definición de cada columna y finalmente, una hoja nombrada arbitrariamente con un numero, que contiene el listado de organismos virales. Si bien no es imposible ni complicado procesar un archivo MS Excel, es un abuso de recursos usar un formato basado en XML si lo único que se necesita guardar es una tabla correctamente formateada. Esta es la única forma de acceder al listado oficial de organismos virales presentado por esta organización, ya que no disponen de una API u otro medio por el cual obtenerlo. Independientemente de los problemas anteriores, es aun mas cuestionable el hecho de utilizar un formato privativo para presentar información de utilidad publica.

La búsqueda de profagos fue un proceso que originalmente estaba considerado en nuestro proyecto pero que finalmente no se terminó incluyendo. Una herramienta popular para este proceso es PHASTER[1], la cual se encuentra unicamente disponible como servicio web y dispone de una API publica a la cual se le pueden hacer consultas en masa. Sin embargo esta API es el ejemplo máximo de malas practicas. La API acepta archivos FASTA con el propósito de buscar y anotar secuencias de profagos en el, sin embargo solo acepta estos archivos por medio del protocolo `application/x-www-form-urlencoded`, el cual solo permite datos en el formato ”clave-valor”. Lo que sucede al momento de forzar un archivo cualquiera a pasar como si tuviera la estructura ”clave-valor” es que el contenido completo del archivo se codifica como la clave de un dato vacío. De alguna manera PHASTER solo procesa el string correspondiente a esta llave y lo interpreta como el archivo enviado, en vez de aceptar el archivo o las secuencias propiamente tales como se esperaría, usando el protocolo `multipart/form-data` o apropiadamente en forma de una request JSON.

Entre otros detalles menores es evidente que un problema presente y significativo es la ausencia de doctrina en buenas practicas de programación, que al del dia terminan lastimando tanto el uso de las mismas herramientas, como su mantención y mejora, lo cual afecta tanto a sucesores dentro del mismo laboratorio como usuarios generales del software.

[1] https://phaster.ca

5.3. Sobre el cumplimiento de los objetivos

Se presentan entonces objetivos a simple vista particularmente acotados, pero con diversas ramificaciones que solo se pueden ver una vez se adentra en su desarrollo.

En cuanto al objetivo general:

> ”Desarrollar un protocolo o framework que permita, a través de la aplicación de métricas basadas en el análisis de metagenomas obtenidos desde ambientes acuáticos continentales, observar patrones de coexistencia de genes pertenecientes a especies virales y procariotas (bacteria y arqueas) en estos ambientes.”

Con este objetivo hacemos referencia a la necesidad de establecer estructura en **proceso** investigativo, específicamente para metagenomas obtenidos de ambientes acuáticos continentales (o de agua dulce) de los que se espera observar especies virales y procariotas, y entonces estudiar sus patrones de coexistencia.

Se puede concluir que este objetivo se ha cumplido a cabalidad, en el Capítulo 3 se ha establecido la estructura y secuencia de herramientas y estructuras que tanto los datos como el investigador debería seguir para guiar su análisis, análisis de metagenomas, que si bien se espera de ellos presencia viral de relevancia ambiental, no se limitan solo a bacteriofagos en aguas continentales. Se presentan después en el mismo Capítulo 3 estrategias y herramientas para observar coexistencia de las especies.

La globalidad del objetivo se alcanzo a través del cumplimiento general de los objetivos específicos, siendo el primero de ellos:

> ”Describir y desarrollar un procedimiento jerarquizado basado en herramientas bioinformáticas que permita obtener conocimiento acerca de posibles interacciones entre fago-hospedero en ambientes acuáticos continentales.”

Esto es simplemente el grueso del Capítulo 3, en donde se explica el protocolo, que le da estructura a la investigación de interacciones fago-hospedero, usando herramientas bioinformáticas como BLAST, bases de datos como las de NCBI, tanto como herramientas y estructuras de datos desarrolladas *ad-hoc* para unir estas otras. El objetivo ha sido cumplido satisfactoriamente.

Respecto al segundo objetivo:

> ”Desarrollar y aplicar una heurística para determinar la presencia de genes pertenecientes a especies virales y procariotas en muestras de metagenomas obtenidas desde ambientes acuáticos continentales.”

Se hace referencia a la detectar y cuantificar el volumen de organismos, así como identificar sus especies y taxonomías. Concluimos que este objetivo se alcanzó tangencialmente, la única razón para hacer esta clase de aseveración son las dificultades encontradas en cuanto a disponibilidad de la información como la utilidad de esta: el problema del sesgo hacia bacterias y la ausencia de virus bacteriófagos. Problema de datos que nos dificultan la validación de nuestros modelos, imposibilitandonos en general de garantizar un buen desempeño, mejor del que se pueda probar anecdótica y experimentalmente, como se hizo en esta memoria.

Respecto al tercer y último objetivo:

> ”Extraer conocimiento sobre la coexistencia e interacciones entre especies virales y procariotas en estos ambientes a través de la visualización de grupos de interés.”

Sufrimos del mismo inconveniente que con la identificación de taxonomías. En este caso, tratando de identificar relaciones, nuestro problema y mayor dificultad yace en la total inexistencia de información probada sobre relaciones fago-hospedero, haciendo nuestro trabajo de validación puramente especulativo, tratando de hacer que un modelo estadístico prediga información desde donde no la hay.

Pese a los reparos en los logros alcanzados, se puede concluir en una aprobación aceptable del proyecto presentado, suficiente para esta memoria.

5.4. Trabajo Futuro

Apropiadamente hacerse cargo de datasets desbalanceados Para el caso de las conjuntos de datos desbalanceados, refiriéndonos a las tablas m8 resultantes de la asignación inicial realizada por BLAST, se pueden aplicar técnicas tradicionales para el manejo de clases desbalanceadas como *oversampling* y *undersampling*. Una vez conociendo u logrando estimar el real alcance del sesgo de la base de datos, se puede aplicar un sesgo artificial que contrarreste la subrepresentacion de una especie, genero, u familia, etc, en el modelo predictivo. Con la popularización del aprendizaje profundo como herramienta en la ciencia de datos, se debe considerar el desarrollo de una red generativa adversaria (GAN) que cumpla con el objetivo de realizar un remuestreo artificial o de los organismos encontrados en el metagenoma o de las secuencias presentes en el metagenoma, agregando un balance artificial al conjunto de datos.

Revalidar la metodologia con datos apropiados Uno de los problemas por los que se complejizo el proceso de validación es en si la varianza y poca representatividad de los metagenomas de prueba, lo cual en si se debe a que estos no estaban originalmente pensados para el desarrollo de una herramienta que trate el caso de organismos

virales como la de este proyecto. Se vuelve relevante la necesidad de generar metagenomas de prueba específicamente usando especies secuencias de organismos virales. Se sugiere el desarrollo de un generador de metagenomas aleatorios que utilice como fuente secuencias de organismos de grupos virales, y realice alteraciones aleatorias en los strings como forma de añadir variabilidad.

Mejorar Tecnicas de Asignacion taxonomica Siendo evidente que diferentes grupos de organismos presentan diferentes dificultades para realizar la detección y asignación de unidades taxonómicas (como se aprecia en la Figura 42), una estrategia diferenciada para cada grupo se vuelve apropiada. Protocolos para la asignación de especies bacterianas existen en la actualidad, se recomienda combinar diferentes metodologías junto a la propuesta en esta memoria para lograr la mayor precision posible.

Mejorar Tecnicas prediccion de interacciones de co-existencia Como se pudo apreciar durante el Capítulo 4, en particular durante la detección de relaciones fago-hospedero, usar las similaridades de Jaccard y Tanimoto como decididores de existencia de relación de coexistencia no es efectivo al nivel que se espera o necesita, es imperativo investigar y desarrollar una metodología diferente para tratar con los problemas encontrados en estas métricas y técnicas. Una de las posibles alternativas es un cambio en la representación de la red en forma de *graph-embedding* en vez de la actual matriz de biadyacencia. El *graph-embedding* es una reducción dimensional de la información presente en el grafo en forma vectorial, facilitando la aplicación de múltiples herramientas de aprendizaje maquina como análisis PCA entre muchas otras.

Contenerización y despliegue La ausencia de portabilidad del software de uso científico fue uno de los problemas mencionados anteriormente. Al igual que en muchas otras áreas la academia puede beneficiarse con múltiples soluciones desarrolladas actualmente en la industria. *Docker* y la contenerización si bien fueron desarrollados para resolver el problema de la reproducibilidad del software, también mitigan el problema de su portabilidad. Empaquetando los programas en contenedores se simplifica su distribución, actualización, ejecución y orquestración, tanto en ambientes locales, en la nube y clusters.

Oportunidades de optimización y paralelización Siendo la mayoría de los programas aquí generados y ejecutados principalmente dominados por operaciones de escritura y lectura de archivos no presentan muchas oportunidades de paralelización, sin embargo, una reestructuración del flujo de los programas, reduciendo la dependencia en bibliotecas externas, da espacio a un control mas fino sobre los procesos de lectura y escritura de los archivos, como al uso de memoria de los programas. Se recomienda una revision completa al orden en el que los pasos del protocolo son ejecutados e integrados entre si.

REFERENCIAS BIBLIOGRÁFICAS

Achtman, M. & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, *6*(6), 431-440. doi:10.1038/nrmicro1872

Aislabie, J. M., Chhour, K.-L., Saul, D. J., Miyauchi, S., Ayton, J., Paetzold, R. F. & Balks, M. R. (2006). Dominant bacteria in soils of Marble Point and Wright Valley, Victoria Land, Antarctica. *Soil Biology and Biochemistry*, *38*(10), 3041-3056. Antarctic Victoria Land Soil Ecology. doi:https://doi.org/10.1016/j.soilbio.2006.02.018

Akhter, S., Aziz, R. K. & Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, *40*(16), e126-e126. doi:10.1093/nar/gks406. eprint: http://oup.prod.sis.lan/nar/article-pdf/40/16/e126/25343008/gks406.pdf

Allgaier, M. & Grossart, H.-P. (2006). Seasonal dynamics and phylogenetic diversity of free-living and particle-associated bacterial communities in four lakes in northeastern Germany. *Aquatic Microbial Ecology*, *45*(2), 115-128. doi:10.3354/ame045115

Almeida-Neto, M., Guimaraes, P., Guimaraes Jr, P. R., Loyola, R. D. & Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, *117*(8), 1227-1239.

Alonso, C., Warnecke, F., Amann, R. & Pernthaler, J. (2007). High local and global diversity of Flavobacteria in marine plankton. *Environmental microbiology*, *9*(5), 1253-1266.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol. 215*(3), 403-410. [DOI:10.1016/S0022-2836(05)80360-2] [PubMed:2231712].

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, *9*(13), 3015-3027. doi:10.1093/nar/9.13.3015. eprint: https://academic.oup.com/nar/article-pdf/9/13/3015/6166406/9-13-3015.pdf

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., … Rohwer, F. (2006). The Marine Viromes of Four Oceanic Regions. *PLOS Biology*, *4*(11), e368. doi:10.1371/journal.pbio.0040368

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*, *44*(W1), W16-W21. gkw387[PII]. doi:10.1093/nar/gkw387

Aziz, R. K., Dwivedi, B., Akhter, S., Breitbart, M. & Edwards, R. A. (2015). Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. *Frontiers in Microbiology*, *6*, 381. doi:10.3389/fmicb.2015.00381

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. (2005). Host-Bacterial Mutualism in the Human Intestine. *Science*, *307*(5717), 1915-1920. doi:10.1126/science.1104816. eprint: https://science.sciencemag.org/content/307/5717/1915.full.pdf

Baltimore, D. (1971). Expression of animal virus genomes. *Microbiology and Molecular Biology Reviews*, *35*(3), 235-241. eprint: https://mmbr.asm.org/content/35/3/235.full.pdf. Recuperado desde https://mmbr.asm.org/content/35/3/235

Bang, C. & Schmitz, R. A. (2015). Archaea associated with human surfaces: not to be underestimated. *FEMS Microbiology Reviews*, *39*(5), 631-648. doi:10.1093/femsre/fuv010. eprint: https://academic.oup.com/femsre/article-pdf/39/5/631/10740965/fuv010.pdf

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, *76*(6), 066102.

Beckett, S. (2015). *Nestedness and modularity in bipartite networks* (Tesis doctoral).

Bell, G., Hey, T. & Szalay, A. (2009). Beyond the Data Deluge. *Science*, *323*(5919), 1297-1298. doi:10.1126/science.1170411. eprint: https://science.sciencemag.org/content/323/5919/1297.full.pdf

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36-D42. doi:10.1093/nar/gks1195. eprint: http://oup.prod.sis.lan/nar/article-pdf/41/D1/D36/3680750/gks1195.pdf

Bergh, Ø., BØrsheim, K. Y., Bratbak, G. & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467-468. doi:10.1038/340467a0

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1462), 1935-1943. L3416270197T513H[PII]. doi:10.1098/rstb.2005.1725

Bobay, L.-M., Rocha, E. & Touchon, M. (2012). The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular biology and evolution*, *30*. doi:10.1093/molbev/mss279

Bobay, L.-M., Touchon, M. & Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences*, *111*(33), 12127-12132. doi:10.1073/pnas.1405336111. eprint: https://www.pnas.org/content/111/33/12127.full.pdf

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., … Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences*, *99*(22), 14250-14255. doi:10.1073/pnas.202488399. eprint: https://www.pnas.org/content/99/22/14250.full.pdf

Breitbart, M., Thompson, L., Suttle, C. & Sullivan, M. (2007). Exploring the Vast Diversity of Marine Viruses. *OCEANOGRAPHY*, *20*, 135-139. doi:10.5670/oceanog.2007.58

Bruder, K., Malki, K., Cooper, A., Sible, E., Shapiro, J. W., Watkins, S. C. & Putonti, C. (2016). Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. *Evolutionary bioinformatics online*, *12*(Suppl 1), 25-33. ebo-suppl.1-2016-025[PII]. doi:10.4137/EBO.S38549

Buchfink, B., Xie, C. & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59-60. doi:10.1038/nmeth.3176

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L. & Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, *6*(4), 417-424. doi:https://doi.org/10.1016/S1369-5274(03)00086-9

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. (2003). Prophage Genomics. *Microbiology and Molecular Biology Reviews*, *67*(2), 238-276. doi:10.1128/MMBR.67.2.238-276.2003. eprint: https://mmbr.asm.org/content/67/2/238.full.pdf

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol. 49*(2), 277-300. [DOI:10.1046/j.1365-2958.2003.03580.x] [PubMed:12886937].

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925-1927. doi:10.1093/bioinformatics/btz848. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/6/1925/32915144/btz848.pdf

Chen, K. & Pachter, L. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLOS Computational Biology*, *1*(2). doi:10.1371/journal.pcbi.0010024

Clokie, M. R. [Martha R.J.], Millard, A. D., Letarov, A. V. & Heaphy, S. (2011). Phages in nature. *Bacteriophage*, *1*(1), 31-45. PMID: 21687533. doi:10.4161/bact.1.1.14942. eprint: https://doi.org/10.4161/bact.1.1.14942

Clokie, M. R. [Martha RJ], Kropinski, A. M. & Lavigne, R. (2009). *Bacteriophages*. Springer.

Cole, J., Findlay, S. & Pace, M. (1988). Bacterial Production in Fresh and Saltwater Ecosystems – a Cross-System Overview. *Marine Ecology - Progress Series*, *43*, 1-10. doi:10.3354/meps043001

Crusoe, M., Alameldin, H., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., … Brown, C. (2015). The khmer software package: enabling efficient nucleotide sequence analysis [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Research*, *4*(900). doi:10.12688/f1000research.6924.1

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A. 4., Bik, H. M. & Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*, e243-e243. 243[PII]. doi:10.7717/peerj.243

Devoto, A. E., Santini, J. M., Olm, M. R., Anantharaman, K., Munk, P., Tung, J., … Banfield, J. F. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature Microbiology*, *4*(4), 693-700. doi:10.1038/s41564-018-0338-9

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., … Edwards, R. A. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*, *5*(1), 4498. doi:10.1038/ncomms5498

Edwards, R. A. [Robert A.], McNair, K., Faust, K., Raes, J. & Dutilh, B. E. (2015). Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, *40*(2), 258-272. doi:10.1093/femsre/fuv048. eprint: http://oup.prod.sis.lan/femsre/article-pdf/40/2/258/23905864/fuv048.pdf

Edwards, R. A. [Robert A] & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, *3*(6), 504. doi:10.1038/nrmicro1163

Eloe-Fadrosh, E. A., Paez-Espino, D., Jarett, J., Dunfield, P. F., Hedlund, B. P., Dekas, A. E., … Ivanova, N. N. (2016). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun*, *7*, 10476. [PubMed Central:PMC4737851] [DOI:10.1038/ncomms10476] [PubMed:24008419].

Fassler, J. & Cooper, P. (2011). BLAST Glossary. Recuperado el 27 de marzo de 2020, desde https://www.ncbi.nlm.nih.gov/books/NBK62051/

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic acids research*, *40*(Database issue), D136-D143. gkr1178[PII]. doi:10.1093/nar/gkr1178

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., … Merrick, J. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. *Science (New York, N.Y.) 269*, 496-512. doi:10.1126/science.7542800

for General Microbiology, S. (2010). Metal-mining bacteria are green chemists. Recuperado el 2 de septiembre de 2010, desde www.sciencedaily.com/releases/2010/09/100901191137.htm

Fuhrman, J. A. & Noble, R. T. (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnology and Oceanography*, *40*(7), 1236-1242.

Fujimoto, K., Kimura, Y., Shimohigoshi, M., Satoh, T., Sato, S., Tremmel, G., … Nakano, Y. y col. (2020). Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts. *Cell Host & Microbe*.

Gansner, E. R. & North, S. C. (2000). An open graph visualization system and its applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, *30*(11), 1203-1233.

Gómez-Consarnau, L., González, J. M., Coll-Lladó, M., Gourdon, P., Pascher, T., Neutze, R., … Pinhassi, J. (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature*, *445*(7124), 210-213. doi:10.1038/nature05381

Guimaraes Jr, P. R. & Guimaraes, P. (2006). Improving the analyses of nestedness for large sets of matrices. *Environmental Modelling & Software*, *21*(10), 1512-1513.

Guimerà, R., Sales-Pardo, M. & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E*, *76*(3), 036102.

Hagberg, A. A., Schult, D. A. & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. En G. Varoquaux, T. Vaught & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11-15). Pasadena, CA USA.

Hamilton, T. L., Bryant, D. A. & Macalady, J. L. (2016). The role of biology in planetary evolution: cyanobacterial primary production in low-oxygen Proterozoic oceans. *Environmental Microbiology*, *18*(2), 325-340. doi:10.1111/1462-2920.13118. eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.13118

Hanna, L. F., Matthews, T. D., Dinsdale, E. A., Hasty, D. & Edwards, R. A. (2012). Characterization of the ELPhiS prophage from Salmonella enterica serovar Enteritidis strain LK5. *Applied and environmental microbiology*, *78*(6), 1785-1793. AEM.07241-11[PII]. doi:10.1128/AEM.07241-11

Hatfull, G. F. [G. F.]. (2008). Bacteriophage genomics. *Curr. Opin. Microbiol.* *11*(5), 447-453. [PubMed Central:PMC2706577] [DOI:10.1016/j.mib.2008.09.004] [PubMed:10645443].

Hatfull, G. F. [G. F.] & Hendrix, R. W. (2011). Bacteriophages and their genomes. *Curr Opin Virol*, *1*(4), 298-303. [PubMed Central:PMC3199584] [DOI:10.1016/j.coviro.2011.06.009] [PubMed:17289101].

Hatfull, G. F. [Graham F.]. (2015). Dark Matter of the Biosphere: the Amazing World of Bacteriophage Diversity. *Journal of virology*, *89*(16), 8107-8110. JVI.01340-15[PII]. doi:10.1128/JVI.01340-15

Hengtee Lim. (2020). Interview with Fujitsu Cloud Technologies: 80 % of Data Science is Pre-processing. Recuperado el 15 de julio de 2020, desde https://lionbridge.ai/articles/interview-with-fujitsu-cloud-technologies-80-of-data-science-is-pre-processing/

Hetherington, A. & Raven, J. (2005). The biology of carbon dioxide. *Current biology : CB*, *15*, R406-10. doi:10.1016/j.cub.2005.05.042

Hey, A., Tansley, S. & Tolle, K. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research. Recuperado desde https://books.google.cl/books?id=oGs%5C_AQAAIAAJ

Hobbs, Z. & Abedon, S. (2016). Diversity of phage infection types and associated terminology: the problem with “Lytic or Lysogenic”. *FEMS Microbiology Letters*, *363*, fnw047. doi:10.1093/femsle/fnw047

Huerta-Cepas, J., Serra, F. & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635-1638. doi:10.1093/molbev/msw046. eprint: https://academic.oup.com/mbe/article-pdf/33/6/1635/7953632/msw046.pdf

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90-95. doi:10.1109/MCSE.2007.55

Hurwitz, B. L. & Sullivan, M. B. [M. B.]. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE*, *8*(2), e57355. [PubMed Central:PMC3585363] [DOI:10.1371/journal.pone.0057355] [PubMed:16794078].

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biology Direct*, *1*(1), 29. doi:10.1186/1745-6150-1-29

Lederberg, E. M. & Lederberg, J. (1953). Genetic Studies of Lysogenicity in Escherichia Coli. *Genetics*, *38*(1), 51-64. [PubMed Central:PMC1209586] [PubMed:17247418].

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G. & Smith, D. B. (2017). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(D1), D708-D717. doi:10.1093/nar/gkx932. eprint: https://academic.oup.com/nar/article-pdf/46/D1/D708/23162757/gkx932.pdf

Leger, J.-B., Vacher, C. & Daudin, J.-J. (2014). Detection of structurally homogeneous subsets in graphs. *Statistics and computing*, *24*(5), 675-692.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, *438*(7064), 86-89. doi:10.1038/nature04111

Lipman, D. & Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science*, *227*(4693), 1435-1441. doi:10.1126/science.2983426. eprint: https://science.sciencemag.org/content/227/4693/1435.full.pdf

López-Bueno, A., Rastrojo, A., Peiró, R., Arenas, M. & Alcamí, A. (2015). Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. *Molecular Ecology*, *24*(19), 4812-4825. doi:10.1111/mec.13321. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13321

Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature*, *424*(6950), 741-741. doi:10.1038/424741a

Maqsood, R., Rodgers, R., Rodriguez, C., Handley, S. A., Ndao, I. M., Tarr, P. I., … Holtz, L. R. (2019). Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome*, *7*(1), 156. doi:10.1186/s40168-019-0766-7

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., … Kyrpides, N. C. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic acids research*, *36*(Database issue), D534-D538. gkm869[PII]. doi:10.1093/nar/gkm869

Mohiuddin, M. & Schellhorn, H. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology*, *6*, 960. doi:10.3389/fmicb.2015.00960

Moreira, D. & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, *7*(4), 306-311.

Newman, M. E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. (2011a). A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews*, *75*(1), 14-49. doi:10.1128/MMBR.00028-10. eprint: https://mmbr.asm.org/content/75/1/14.full.pdf

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. (2011b). A guide to the natural history of freshwater lake bacteria. *Microbiology and molecular biology reviews : MMBR*, *75*(1), 14-49. 75/1/14[PII]. doi:10.1128/MMBR.00028-10

O'Connor, C. M., Adams, J. U. & Fairman, J. (2010). Essentials of cell biology. *Cambridge, MA: NPG Education*, *1*, 54.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'hara, R., … Wagner, H. (2015). vegan: Community Ecology Package. R package version 2.0-10. 2013. *There is no corresponding record for this reference*.

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol Publishing USA.

Ouborg, N. J. & Vriezen, W. H. (2007). An ecologist's guide to ecogenomics. *Journal of Ecology*, *95*(1), 8-16. doi:10.1111/j.1365-2745.2006.01197.x. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2745.2006.01197.x

pandas development team, T. (2020). pandas-dev/pandas: Pandas (Ver. latest). doi:10.5281/zenodo.3509134

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996-1004. doi:10.1038/nbt.4229

Pearson, W. R. [W R] & Lipman, D. J. [D J]. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*(8), 2444-2448. doi:10.1073/pnas.85.8.2444. eprint: https://www.pnas.org/content/85/8/2444.full.pdf

Pedrós-Alió, C., Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Fernàndez-Guerra, A., … González, J. M. (2013). Ecology of marine Bacteroidetes: a comparative genomics approach.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., … Hatfull, G. F. (2003). Origins of Highly Mosaic Mycobacteriophage Genomes. *Cell*, *113*(2), 171-182. doi:10.1016/S0092-8674(03)00233-2

Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *33*(Database issue), D501-D504. 33/suppl_1/D501[PII]. doi:10.1093/nar/gki025

Rodríguez-Gironés, M. A. & Santamaría, L. (2006). A new algorithm to calculate the nestedness temperature of presence–absence matrices. *Journal of Biogeography*, *33*(5), 924-935. doi:10.1111/j.1365-2699.2006.01444.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2699.2006.01444.x

Rogers, D. J. & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science*, *132*(3434), 1115-1118. doi:10.1126/science.132.3434.1115. eprint: https://science.sciencemag.org/content/132/3434/1115.full.pdf

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology*, *11*(11), 2806-2820. doi:10.1111/j.1462-2920.2009.01964.x. eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2009.01964.x

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., … Coordinators, T. O. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, *537*(7622), 689-693. doi:10.1038/nature19366

Schloss, P. D. & Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology*, *6*(8), 229-229. gb-2005-6-8-229[PII]. doi:10.1186/gb-2005-6-8-229

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811-814. doi:10.1038/nmeth.2066

Sharon, I., Battchikova, N., Aro, E. M., Giglione, C., Meinnel, T., Glaser, F., … Beja, O. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J*, *5*(7), 1178-1190. [PubMed Central:PMC3146289] [DOI:10.1038/ismej.2011.2] [PubMed:11413011].

Shukla, R. (2014). *Analysis Of Chromosome*. Agrotech Press. Recuperado desde https://books.google.cl/books?id=7-UKCgAAQBAJ

Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195-197. doi:https://doi.org/10.1016/0022-2836(81)90087-5

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601-2610. doi:10.1093/nar/6.7.2601. eprint: https://academic.oup.com/nar/article-pdf/6/7/2601/7063509/6-7-2601.pdf

Staley, J. T. & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual review of microbiology*, *39*(1), 321-346.

Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M. & Poisson, G. (2013). Are we missing half of the viruses in the ocean? *The ISME Journal*, *7*(3), 672-679. doi:10.1038/ismej.2012.121

Sullivan, M. B. [Matthew B.]. (2015). Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *Journal of Virology*, *89*(5), 2459-2461. doi:10.1128/JVI.03289-14. eprint: https://jvi.asm.org/content/89/5/2459.full.pdf

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., … Wooley, J. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research*, *39*(Database issue), D546-D551. gkq1102[PII]. doi:10.1093/nar/gkq1102

Suttle, C. A. [C. A.]. (2007). Marine viruses–major players in the global ecosystem. *Nat. Rev. Microbiol. 5*(10), 801-812.

Suttle, C. (2005). The viriosphere: the greatest biological diversity on Earth and driver of global processes. *Environmental Microbiology*, *7*(4), 481-482. doi:10.1111/j.1462-2920.2005.803_11.x. eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2005.803_11.x

Suttle, C. A. [Curtis A]. (2005). Viruses in the sea. *Nature*, *437*(7057), 356. doi:10.1038/nature04160

The Viral Life Cycle. (s.f.). Recuperado el 6 de marzo de 2010, desde https://courses.lumenlearning.com/microbiology/chapter/the-viral-life-cycle/

Thomas, T., Gilbert, J. & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. doi:10.1186/2042-5783-2-3

Thurber, R. V. (2009). Current insights into phage biodiversity and biogeography. *Current Opinion in Microbiology*, *12*(5), 582-587. Antimicrobials Genomics. doi:https://doi.org/10.1016/j.mib.2009.08.008

Van Der Walt, S., Colbert, S. C. & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22.

Watkins, S. C. [S. C.] & Putonti, C. (2017). The use of informativity in the development of robust viromics-based examinations. *PeerJ*, *5*, e3281.

Watkins, S. C. [Siobhan C.], Kuehnle, N., Ruggeri, C. A., Malki, K., Bruder, K., Elayyan, J., … Putonti, C. (2016). Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Marine and Freshwater Research*, *67*(11), 1700-1708. Recuperado desde https://doi.org/10.1071/MF15172

Weitz, J. S., Poisot, T., Meyer, J. R., Flores, C. O., Valverde, S., Sullivan, M. B. & Hochberg, M. E. (2013). Phage–bacteria infection networks. *Trends in microbiology*, *21*(2), 82-91. doi:10.1016/j.tim.2012.11.003

Wikipedia contributors. (2020). One-hot — Wikipedia, The Free Encyclopedia.

Williams, C. (2011). Who are you calling simple? *New Scientist*, *211*(2821), 38-41.

Wimmer, E., Mueller, S., Tumpey, T. M. & Taubenberger, J. K. (2009). Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nature Biotechnology*, *27*(12), 1163-1172. doi:10.1038/nbt.1593

Wommack, K. E., Williamson, K. E., Helton, R. R., Bench, S. R. & Winget, D. M. (2009). Methods for the isolation of viruses from environmental samples. *Methods Mol. Biol. 501*, 3-14.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K. H., … Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol. 12*(9), 635-645. [DOI:10.1038/nrmicro3330] [PubMed:20531276].

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., … Giovannoni, S. J. (2013). Abundant SAR11 viruses in the ocean. *Nature*, *494*(7437), 357. doi:10.1038/nature11921

Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M.-A., Imamovic, L., … Panagiotou, G. (2019). Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome*, *7*(1), 42. doi:10.1186/s40168-019-0657-y

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic acids research*, *39*(Web Server issue), W347-W352. gkr485[PII]. doi:10.1093/nar/gkr485

Zwart, G., Crump, B. C., Agterveld, M. P. K.-v., Hagen, F. & Han, S.-K. (2002). Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology*, *28*(2), 141-155. doi:10.3354/ame028141

Anexo A

select_first_n.py

Dependencias:

- Python 3
- Pandas Statistical Computing (pandas development team, 2020)

```python
#!/usr/bin/env python3
import argparse
from pathlib import Path
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument("SRC", help="Source m8 table")
parser.add_argument("DST", help="Destination file")
parser.add_argument("N", type=int, help="N first elements to keep, assume table is sorted by BITSCORE")
args = parser.parse_args()

SRC_PATH = Path(args.SRC).resolve()
DST_PATH = Path(args.DST).resolve()
N_TOP = args.N

print(F"{SRC_PATH.name} start")
df = pd.read_csv(
    SRC_PATH,
    sep='\t',
    header=None,
    names=['qseqid', 'sseqid', 'taxids'],
    dtype={'qseqid': str, 'sseqid': str, 'taxids': pd.Int64Dtype()},
    index_col='qseqid',
    memory_map=True)\
```

```
.dropna()\
.groupby(
    level=0,
    sort=False)\
.head(N_TOP)\
.to_csv(DST_PATH,
        sep='\t',
        header=True,
        index=True)
print(F"{SRC_PATH.name} end")
```

Anexo B

filter_domains.py

Dependencias:

- Python 3
- Pandas Statistical Computing (pandas development team, 2020)
- Numpy Númerical Computing (Oliphant, 2006; Van Der Walt, Colbert & Varoquaux, 2011)
- ETE3 Phylogenetic DataStructures (Huerta-Cepas, Serra & Bork, 2016)

```
#!/usr/bin/env python3
import typing
import argparse
import operator as op
from functools import reduce, lru_cache
from pathlib import Path
from enum import IntFlag, auto
import pandas as pd
import numpy as np
from ete3 import NCBITaxa


parser = argparse.ArgumentParser()
parser.add_argument("ICTVRANK", type=str, help="ICTV database prepared CSV")
parser.add_argument("SRC", help="Source m8 table")
parser.add_argument("DST", help="Destination m8 table")
args = parser.parse_args()

ICTVRANK_PATH = Path(args.ICTVRANK).resolve()
SRC_PATH = Path(args.SRC).resolve()
```

```
21 DST_PATH = Path(args.DST).resolve()
22
23 KEEP_HOST_DOMAINS = ["Bacteria", "Archaea"]
24 KEEP_VIRAL_DOMAIN = ["Viruses"]
25
26 class Partition(IntFlag):
27     DISCARDABLE = int(0)
28     HOSTLIKE = auto()
29     VIRAL = auto()
30
31 # -- Loads NCBI taxonomy database wrapper
32 NCBI = NCBITaxa()
33
34 # -- Loads prepared ICTV species list
35 ICTVrank = pd.read_csv(ICTVRANK_PATH,
36                        sep='\t',
37                        usecols=['species'],
38                        dtype=int)
39
40 class Utilities:
41     ''' Define Utility functions '''
42     HOSTLIKE_TAXID = set(reduce(op.add,
       ↪  NCBI.get_name_translator(KEEP_HOST_DOMAINS).values(), []))
43     VIRAL_TAXID = set(reduce(op.add,
       ↪  NCBI.get_name_translator(KEEP_VIRAL_DOMAIN).values(), []))
44
45     @staticmethod
46     @lru_cache(maxsize=1024, typed=False)
47     def get_partition(taxid: int) -> Partition:
48         ''' Returns the partition to assign taxid
49         '''
50         try:
51             lineage = set(NCBI.get_lineage(taxid))
52         except ValueError:
53             return Partition.DISCARDABLE
54         except UserWarning:
55             pass # Ignore renaming warnings
56         if bool(set(Utilities.HOSTLIKE_TAXID) & lineage):
57             return Partition.HOSTLIKE
58         if bool(set(Utilities.VIRAL_TAXID) & lineage):
59             return Partition.VIRAL
60         return Partition.DISCARDABLE
61
62     @staticmethod
```

```
63      def species_or_discard(taxids: typing.Sequence[int]) ->
        ↪ typing.List[int]:
64          ''' Takes sequence of taxids and returns them at the species
            ↪ rank-level
65              * Discards those above and those which can't be changed due to
    ↪ absent information
66          '''
67          uniqs = np.unique(taxids).astype(int).tolist()
68          translator = NCBI.get_lineage_translator(uniqs)
69          # re-adds taxids that were implicitly renamed
70          for k in {t for t in uniqs if t not in translator}:
71              translator[k] = NCBI.get_lineage(k)
72          # leaves only species, in-place
73          keys = list(translator.keys())
74          for k in keys:
75              translator[k] = next((x for x, y in
                ↪ NCBI.get_rank(translator[k]).items() if y == 'species'), -1)
76          return [translator.get(x, -1) for x in taxids]
77
78  #
    ↪ ------------------------------------------------------------------------
    ↪ #
79
80
81  # Loads Source Table
82  df = pd.read_csv(SRC_PATH,
83                   sep='\t',
84                   dtype={'qseqid': str, 'sseqid': str, 'taxids': int},
85                   memory_map=True)
86  print(F"\r{SRC_PATH.name} [ ][ ][ ][ ]", end='')
87
88  # Classify and remove those which aren't neither host nor viral
89  df['partition'] = df['taxids'].apply(Utilities.get_partition).astype(int)
90  df = df.loc[df['partition'] != 0, :]
91  print(F"\r{SRC_PATH.name} [x][ ][ ][ ]", end='')
92
93  # Discards anything that can't be annotated as 'species'
94  df['correction'] = Utilities.species_or_discard(df['taxids'])
95  df = df.loc[df['correction'] > 0, :]
96  print(F"\r{SRC_PATH.name} [x][x][ ][ ]", end='')
97
98  df.drop(columns=['taxids'], inplace=True)
99  df.sort_values(by=['partition', 'correction', 'sseqid'], inplace=True)
100 df.rename(columns={'correction': 'taxids'}, inplace=True)
101 print(F"\r{SRC_PATH.name} [x][x][x][ ]", end='')
```

```
102
103 # Discards anything that's not in ICTVrank db
104 df = df.loc[(
105         df['partition'] == Partition.HOSTLIKE) | (
106             df.loc[df['partition'] == Partition.VIRAL,
                ↪  'taxids'].isin(ICTVrank['species'])
107         ), :]
108 df.drop(columns=['partition'], inplace=True)
109 print(F"\r{SRC_PATH.name} [x][x][x][x]", end='')
110
111 df.to_csv(DST_PATH, sep='\t', header=True, index=False)
112 print(F"\r{SRC_PATH.name} [x][x][x][x] ok", end='\n')
```

Anexo C

remove_low_freq.py

Dependencias:

- Python 3
- Pandas Statistical Computing (pandas development team, 2020)
- Numpy Númerical Computing (Oliphant, 2006; Van Der Walt y col., 2011)
- ETE3 Phylogenetic DataStructures (Huerta-Cepas y col., 2016)

```
#!/usr/bin/env python3
import argparse
import operator as op
from functools import reduce, lru_cache
from enum import IntFlag, auto
from pathlib import Path
import pandas as pd
import numpy as np
from ete3 import NCBITaxa


parser = argparse.ArgumentParser()
parser.add_argument("SRC", help="Source m8 table")
parser.add_argument("DST", help="Destination file")
parser.add_argument("min_vir", type=int,
                    help="Minimum count of protein matches per VIRAL species assignment to keep that match")
parser.add_argument("min_host", type=int,
                    help="Minimum count of protein matches per HOST species assignment to keep that match")
args = parser.parse_args()
```

```
20
21  SRC_PATH = Path(args.SRC).resolve()
22  DST_PATH = Path(args.DST).resolve()
23  MINIMUM_VIRA = args.min_vir
24  MINIMUM_HOST = args.min_host
25
26  KEEP_HOST_DOMAINS = ["Bacteria", "Archaea"]
27  KEEP_VIRAL_DOMAIN = ["Viruses"]
28
29  # -- Loads NCBI taxonomy database wrapper
30  NCBI = NCBITaxa()
31
32
33  class Utilities:
34      ''' Define Utility functions '''
35      HOSTLIKE_TAXID = set(
36          reduce(op.add, NCBI.get_name_translator(KEEP_HOST_DOMAINS).values(),
            ↪  []))
37      VIRAL_TAXID = set(
38          reduce(op.add, NCBI.get_name_translator(KEEP_VIRAL_DOMAIN).values(),
            ↪  []))
39
40      @staticmethod
41      @lru_cache(maxsize=1024, typed=False)
42      def get_partition(taxid: int) -> Partition:
43          ''' Returns the partition to assign taxid
44          '''
45          lineage = set(NCBI.get_lineage(taxid))
46          if bool(set(Utilities.HOSTLIKE_TAXID) & lineage):
47              return "h"
48          if bool(set(Utilities.VIRAL_TAXID) & lineage):
49              return "v"
50
51
52  print(F"{SRC_PATH.name} starts", end='\n')
53  df = pd.read_csv(SRC_PATH,
54                   sep='\t',
55                   dtype={'sseqid': str, 'taxids': int},
56                   index_col='taxids',
57                   memory_map=True)
58  print(F"{SRC_PATH.name} [ ][ ][ ][ ]", end='')
59
60  # -- Counts frequency of matches per protein per assigment
61  count = df.groupby('taxids', sort=False)['sseqid'].nunique()
62  print(F"\r{SRC_PATH.name} [x][ ][ ][ ]", end='')
```

```
63
64 # --Paritionate
65 df['part'] =
   ↪  df['taxids'].apply(Utilities.get_partition).astype('categorical')
66 print(F"\r{SRC_PATH.name} [x][x][ ][ ]", end='')
67
68
69 host_min = np.array(count[count > MINIMUM_HOST].index)
70 vira_min = np.array(count[count > MINIMUM_VIRA].index)
71 print(F"\r{SRC_PATH.name} [x][x][x][ ]", end='')
72
73 # If protein was matched less than MINIMUM_COUNT times per assigned
   ↪  species,
74 # .. it will be removed
75 df = pd.concat([
76     df[(df.index.isin(host_min)) & (df['part']=="h")],
77     df[(df.index.isin(vira_min)) & (df['part']=="v")]
78 ]).drop(columns=["part"])
79 print(F"\r{SRC_PATH.name} [x][x][x][x]", end='')
80 df.to_csv(DST_PATH,
81           sep='\t',
82           header=True,
83           index=True)
84 print(F"\r{SRC_PATH.name} [x][x][x][x] ok", end='\n')
```

Anexo D

to_encoding.py

Dependencias:

- Python 3
- Pandas Statistical Computing (pandas development team, 2020)
- Numpy Númerical Computing (Oliphant, 2006; Van Der Walt y col., 2011)
- ETE3 Phylogenetic DataStructures (Huerta-Cepas y col., 2016)

```
#!/usr/bin/env python3
"""
to_encoding.py
python3 to_encoding.py SRC_DIR ENC.tsv LINEAGE.tsv
"""
import argparse
import typing
import operator as op
from functools import reduce
from pathlib import Path
import pandas as pd
import numpy as np
from ete3 import NCBITaxa

parser = argparse.ArgumentParser()
parser.add_argument("SRC_DIR", help="Directory where m8 table are")
parser.add_argument("DST_DS", help="Destination for encoding")
parser.add_argument("DST_LN", help="Destination for lineage metadata")
args = parser.parse_args()

SRC_DIR = Path(args.SRC_DIR).resolve()
```

```
22 DST_DS = Path(args.DST_DS).resolve()
23 DST_LN = Path(args.DST_LN).resolve()
24
25 KEEP_HOST_DOMAINS = ["Bacteria", "Archaea"]
26 KEEP_VIRAL_DOMAIN = ["Viruses"]
27
28 # -- Loads NCBI taxonomy database wrapper
29 NCBI = NCBITaxa()
30
31 class Utilities:
32     ''' Define Utility functions '''
33     HOSTLIKE_TAXID = set(reduce(op.add,
       ↪  NCBI.get_name_translator(KEEP_HOST_DOMAINS).values(), []))
34     VIRAL_TAXID = set(reduce(op.add,
       ↪  NCBI.get_name_translator(KEEP_VIRAL_DOMAIN).values(), []))
35     LINEAGE_COLUMNS = ["superkingdom", "kingdom", "subkingdom", "phylum",
       ↪  "subphylum", "class", "subclass", "order", "suborder", "family",
       ↪  "subfamily", "genus", "subgenus", "species"]
36
37     @staticmethod
38     def get_lineage_columns(taxid: int) -> typing.List[int]:
39         records = {v:k
40                    for k, v in
                       ↪  NCBI.get_rank(NCBI.get_lineage(taxid)).items()
41                    if v in Utilities.LINEAGE_COLUMNS}
42         return [
43             records.get(x, None)
44             for x in Utilities.LINEAGE_COLUMNS
45         ]
46
47 print(F"{SRC_DIR.name} starts", end='\n')
48 _FILES = [x.resolve() for x in SRC_DIR.iterdir()]
49 _DF = [
50     pd.read_csv(_F, sep='\t', usecols=['taxids'], dtype={'taxids': int},
       ↪  memory_map=True)
51     for _F in _FILES
52 ]
53 print(F"{SRC_DIR.name} [x][][][][]", end='')
54
55 _TAXIDS = [
56     np.unique(df['taxids']).astype(int)
57     for df in _DF
58 ]
59 print(F"\r{SRC_DIR.name} [x][x][][][]", end='')
60
```

```
61  _ALL_TAXIDS = np.unique(np.concatenate(_TAXIDS))
62  print(F"\r{SRC_DIR.name} [x][x][x][][]", end='')
63
64  ENCODING = np.array([
65      np.isin(_ALL_TAXIDS, m, assume_unique=True)
66      for m in _TAXIDS
67  ]).astype(bool)
68  print(F"\r{SRC_DIR.name} [x][x][x][x][]", end='')
69
70  _RECORDS = {k: Utilities.get_lineage_columns(k) for k in _ALL_TAXIDS}
71  print(F"\r{SRC_DIR.name} [x][x][x][x][x]", end='')
72
73  pd.DataFrame(
74      data=ENCODING,
75      index=[x.name for x in _FILES],
76      columns=[str(x) for x in _ALL_TAXIDS],
77      dtype=int
78  ).to_csv(DST_DS, sep='\t', index=True, header=True)
79
80  pd.DataFrame.from_dict(
81      data=_RECORDS,
82      orient='index',
83      dtype=pd.Int64Dtype(),
84      columns=Utilities.LINEAGE_COLUMNS
85  ).to_csv(DST_LN, sep='\t', index=True, header=True)
86
87  print(F"\r{SRC_DIR.name} [x][x][x][x][x] ok", end='\n')
```

Anexo E

Implementación NODF en numpy

Dependencias:

- Python 3
- Numpy Númerical Computing (Oliphant, 2006; Van Der Walt y col., 2011)

```
from itertools import combinations
import numpy as np

def NODF(M):
    m,n = M.shape
    Fr = np.sum(M, axis=1)
    Fc = np.sum(M, axis=0)
    Npair = 0
    for i,j in combinations(range(n), 2):
        if Fc[i] > Fc[j]:
            mask = M[:,j]
            Npair += np.sum(np.logical_and(M[mask,i],
            ↪  M[mask,j]))/np.sum(mask)
    for i,j in combinations(range(m), 2):
        if Fr[i] > Fr[j]:
            mask = M[j]
            Npair += np.sum(np.logical_and(M[i,mask],
            ↪  M[j,mask]))/np.sum(mask)
    return 2*Npair/(m*(m-1) + n*(n-1))
```

Anexo F

Implementación WNODF en numpy

Dependencias:

- Python 3
- Numpy Númerical Computing (Oliphant, 2006; Van Der Walt y col., 2011)

```
from itertools import combinations
import numpy as np

def WNODF(M):
    m,n = M.shape
    Fc = np.sum(M > 0,axis=0)
    Fr = np.sum(M > 0,axis=1)
    Npair = 0
    for i,j in combinations(range(n), 2):
        if Fc[i] > Fc[j]:
            mask = M[:,j] > 0
            Npair += np.sum(M[mask,j] < M[mask,i])/np.sum(mask)
    for i,j in combinations(range(m), 2):
        if Fr[i] > Fr[j]:
            mask = M[j] > 0
            Npair += np.sum(M[j,mask] < M[i,mask])/np.sum(mask)
    return 2*Npair/(m*(m-1) + n*(n-1))
```

Anexo G

Metagenomas de Agua Dulce

metagenome name	length (bp)	reads	size (MB)	latitude	longitude
49247	3143505549	14582269	3404.2 MB	43,1998	-86,5698
50239	4443417385	21063651	4823.3 MB	43,1998	-86,5698
50240	4203379499	19634419	4555.3 MB	43,1998	-86,5698
50241	1114565803	5201101	1207.8 MB	43,1998	-86,5698
50250	1117565272	5235126	1211.6 MB	43,1998	-86,5698
50251	1183883598	5398073	1279.7 MB	43,1998	-86,5698
50252	1379730813	6361793	1493.2 MB	43,1998	-86,5698
50253	1253920933	5810864	1357.7 MB	43,1998	-86,5698
50271	2287109190	10622223	2477.1 MB	43,1998	-86,5698
50272	2379543906	10861389	2572.4 MB	43,1998	-86,5698
50563	4783846154	22211909	5181 MB	43,1998	-86,5698
50564	1916756797	8943020	2077 MB	43,1998	-86,5698
51410	923769962	4458913	1004.8 MB	43,1998	-86,5698
51411	1016760170	4833128	1104 MB	43,1998	-86,5698
51412	963559343	4622247	1047.3 MB	43,1998	-86,5698
51413	130125969	619759	141.3 MB	43,1998	-86,5698
51415	1481238335	7049593	1608.6 MB	43,1998	-86,5698
51416	1209237111	5760983	1313.4 MB	43,1998	-86,5698
51866	881476992	4213566	957.7 MB	43,1998	-86,5698
54966	612642485	2894971	648.2 MB	46,0072	-89,6063
54967	565271016	2704598	598.8 MB	46,0072	-89,6063
54968	477762584	2286310	506.1 MB	46,0072	-89,6063
54970	636177172	3028562	673.6 MB	46,0072	-89,6063
54973	497702245	2349505	526.6 MB	46,0072	-89,6063
54974	492839008	2332595	521.6 MB	46,0072	-89,6063
54976	406880008	1910869	430.3 MB	46,0072	-89,6063
54978	551607110	2665429	584.8 MB	46,0072	-89,6063

Continua en siguiente página

metagenome name	length (bp)	reads	size (MB)	latitude	longitude
54980	433400770	2073239	459.1 MB	46,0072	-89,6063
54981	697843262	3367891	739.8 MB	46,0072	-89,6063
54982	563950326	2706899	597.5 MB	46,0072	-89,6063
54984	517233409	2505961	548.5 MB	46,0072	-89,6063
54986	544801377	2597646	576.9 MB	46,0072	-89,6063
54987	521394927	2516227	552.7 MB	46,0072	-89,6063
54994	469419617	2211521	496.6 MB	46,0072	-89,6063
54995	576314767	2734657	610 MB	46,0072	-89,6063
54996	664414417	3149681	703.2 MB	46,0072	-89,6063
54997	534065672	2597486	566.5 MB	46,0072	-89,6063
54999	781453818	3722303	827.4 MB	46,0072	-89,6063
55000	848284642	4005029	897.5 MB	46,0072	-89,6063
55003	716199060	3351216	757.2 MB	46,0072	-89,6063
55004	685907925	3208857	725.1 MB	46,0072	-89,6063
55005	680752272	3188317	719.8 MB	46,0072	-89,6063
56244	467523285	2353824	497.5 MB	46,0072	-89,6063
57364	741248967	3496868	787.6 MB	43,1998	-86,5698
57919	919960599	4424687	979.2 MB	43,1998	-86,5698
57987	288155733	1404731	307.1 MB	46,0072	-89,6063
59164	1637313135	7809181	1741.4 MB	43,1998	-86,5698
59212	674521435	3203441	714 MB	46,0072	-89,6063
59214	638021046	3046560	675.7 MB	46,0072	-89,6063
59216	431249297	2047817	456.5 MB	46,0072	-89,6063
59217	393000051	1839047	415.5 MB	46,0072	-89,6063
59220	381941751	1821503	404.5 MB	46,0072	-89,6063
59307	895348798	4308657	953 MB	43,1998	-86,5698
59308	729917032	3499083	776.7 MB	43,1998	-86,5698
59309	699102032	3321311	743.3 MB	43,1998	-86,5698
59310	818310692	3927291	870.8 MB	43,1998	-86,5698
59338	1345584695	6398401	1430.7 MB	43,1998	-86,5698
59500	1554837506	6386507	1632.6 MB	43,1998	-86,5698
60364	381290375	1674534	401 MB	46,0072	-89,6063
60368	392521990	1810458	414.5 MB	46,0072	-89,6063
60369	421851418	1955781	445.6 MB	46,0072	-89,6063
60372	393879651	1743155	414.5 MB	46,0072	-89,6063
60383	1575360304	6538836	1655.5 MB	46,0072	-89,6063
65607	386026983	1605629	404.2 MB		
66374	621545671	3009081	659.1 MB	43,1998	-86,5698
66437	1455698086	6107686	1531 MB	41,69957	-83,2941
66442	1652564675	7066847	1740.8 MB	41,69957	-83,2941

Continua en siguiente página

metagenome name	length (bp)	reads	size (MB)	latitude	longitude
66716	349636476	1488959	366.8 MB	44,504638	-83,045851
66717	332501609	1402119	348.5 MB	44,701094	-82,854085
66718	374815256	1588128	393 MB	44,701094	-82,854085
66719	319140165	1333384	334.3 MB	44,504638	-83,045851
66720	379126061	1611936	397.6 MB	44,701094	-82,854085
66721	343903824	1468336	360.8 MB	44,5046	-83,045851
67506	1297054124	6121605	1378.1 MB	43,1998	-86,5698
70958	426192842	1807569	446.9 MB	46,0072	-89,6063
70959	462674667	1980492	485.5 MB	46,0072	-89,6063
71192	286073246	1219220	300.1 MB	46,0072	-89,6063
71193	287454134	1235085	301.7 MB	46,0072	-89,6063
71203	287058233	1230802	301.3 MB	46,0072	-89,6063
71204	589857183	2567630	619.8 MB	46,0072	-89,6063
71206	415493588	1787616	436.2 MB	46,0072	-89,6063
71208	404698779	1763917	425.3 MB	46,0072	-89,6063
71209	400847242	1738857	421.1 MB	46,0072	-89,6063
71212	301530629	1274823	316.1 MB	46,0072	-89,6063
71213	311609734	1326438	326.8 MB	46,0072	-89,6063
71222	522084787	2223087	547.6 MB	46,0072	-89,6063
71223	504421687	2161724	529.4 MB	46,0072	-89,6063
71224	419359066	1802249	440.2 MB	46,0072	-89,6063
71227	338156471	1443628	354.8 MB	46,0072	-89,6063
71228	441238101	1892695	463.1 MB	46,0072	-89,6063
71229	260327227	1098878	272.9 MB	46,0072	-89,6063
71233	354809095	1507659	372.1 MB	46,0072	-89,6063
78087	1036360697	4355058	1089.9 MB	-77605	163163
78088	640730940	2781069	675.6 MB	-77605	163163
78089	646093290	2684515	678.9 MB	-77605	163163
78090	561605452	2363221	590.7 MB	-77714	162445
78091	478537891	1992503	502.9 MB	-77714	162445
78092	1250361832	5201770	1313.9 MB	-77605	163163
78093	500442691	2097978	526.2 MB	-77714	162445
78382	1303549407	5423545	1369.8 MB	-77605	163163
78383	636973212	2831602	672.9 MB	-77605	163163
78384	1241959047	5146457	1304.7 MB	-77605	163163
78385	1127495356	4664595	1184.3 MB	-77605	163163
87017	1070161572	4412466	1123.8 MB	46,8319	-72,5
87018	1492771219	6192863	1574.2 MB	48,2311	-71,2508
87019	1172576248	4788908	1230.4 MB	46,8319	-72,5
87020	888642069	3603223	932 MB	45,4091	-72,0994

Continua en siguiente página

metagenome name	length (bp)	reads	size (MB)	latitude	longitude
87027	1295387770	5269948	1363.9 MB	48,2311	-71,2508
87031	1063512423	4333798	1115.8 MB	48,2311	-71,2508
87040	9208189074	42927262	9807.4 MB	61,5637	22044
87041	1398664491	7255121	1504.8 MB	61,5637	22044
87318	963316162	3919175	1010.6 MB	48,2311	-71,2508
87319	949209731	3834840	995.2 MB	48,2311	-71,2508
92642	1888936422	9205585	2019.9 MB	41,69957	-83,2941
mgm4441590.3	315151139	296355	516.1 MB	9,1644	-79,83611
mgm4453064.3	19066746	45902	21.8 MB	-35,18	138,46
mgm4453083.3	121619467	288786	139 MB	-35,18	138,46
mgm4481964.3	5896590700	58965907	8830.1 MB	45,139111	106,767658
mgm4481966.3	1458532200	14585322	2184.2 MB	45,139111	106,767658
mgm4516288.3	82832106	236932	82.8 MB	39,3037632	-0,3226403
mgm4516289.3	96008868	236172	95.4 MB	39,3037632	-0,3226403
mgm4516290.3	60567270	149835	60.2 MB	39,3037632	-0,3226403
mgm4534328.3	1888723700	18887237	2612.1 MB		
mgm4534330.3	1847964400	18479644	2555.8 MB		
mgm4534332.3	812416500	8124165	1123.6 MB		
mgm4534334.3	841902400	8419024	1164.4 MB		
mgm4534339.3	1599106200	15991062	2211.6 MB		
mgm4534340.3	1427643200	14276432	1974.5 MB		
mgm4534342.3	1647353500	16473535	2278.4 MB		
mgm4534344.3	1606029200	16060292	2221.2 MB		
mgm4534346.3	1395406900	13954069	1929.9 MB		
mgm4534348.3	1031183000	10311830	1426.1 MB		
mgm4534350.3	1623038100	16230381	2244.7 MB		
mgm4673358.3	3226945593	22707545	3956.7 MB		
mgm4673360.3	6749988239	38221444	8177.7 MB		
mgm4673644.3	3226945593	22707545	3956.7 MB		
mgm4673646.3	6749988239	38221444	8177.7 MB		
mgm4680904.3	7077078057	45414929	8610.7 MB	22,572646	88,363895
mgm4680906.3	4411781846	31246455	5447.3 MB		
totals:	1.191.553.167	6.714.311	189664.0 MB		