Universidad Politécnica del Estado de Morelos

Desarrollo de un clasificador bayesiano para análisis masivo de secuencias ribosomales 16S

T e s i s

Que para obtener el grado de:

Maestro en Ciencias en Biotecnología

P r e s e n t a

Everardo Gutiérrez Millán

Director de tesis
Dr. Jesús Hernández Romano

Jiutepec, Morelos Abril, 2021

Upemor
Universidad Politécnica

DIRECCIÓN DE POSGRADOS Y EDUCACIÓN CONTINUA

Jiutepec, Morelos a 22 de abril de 2021.

AUTORIZACIÓN DE IMPRESIÓN DE TESIS

Una vez revisada la tesis **"Desarrollo de un clasificador bayesiano para análisis masivo de secuencias ribosomales 16S"**, presentada por el estudiante **Everardo Gutiérrez Millán,** la comisión revisora acordó que el documento cubre los requisitos de forma y fondo de una tesis de maestría, por lo tanto, se autoriza proceder con la impresión.

A continuación, se presentan las firmas de los integrantes de la comisión revisora, destacando como Director de la Tesis al **Dr. Jesús Hernández Romano** y como Codirector al **Dr. Juan Paulo Sánchez Hernández.**

DR. JESÚS HERNÁNDEZ ROMANO
DIRECTOR
Cédula Profesional: 6053275

DR. JUAN PAULO SÁNCHEZ HERNÁNDEZ
CODIRECTOR
Cédula Profesional: 8861878

DRA. SANDRA MORALES ARRIETA
VOCAL 1
Cédula Profesional: 6051484

DR. JESÚS MARTÍNEZ BARNETCHE.
VOCAL 2
Cédula Profesional: 4016227

MTRO. JONATHAN ORTUÑO TRIANA
DIRECTOR DE POSGRADO Y EDUCACIÓN CONTINUA
Cédula Profesional: 10959684

c.c.p. Dirección de Posgrado y Educación Continua

El contenido de la tesis es responsabilidad del sustentante.

AGRADECIMIENTOS

En primer lugar, quiero expresar mi agradecimiento al Dr. Jesús Romano y al Dr. Juan Paulo Sánchez por su apoyo durante la realización de esta tesis.

A mis padres por su apoyo incondicional. Por creer en mí y ser parte de este proyecto. En especial, a mi madre por luchar hasta el final.

A la Universidad Politécnica del Estado de Morelos por abrirme sus puertas para mi desarrollo profesional y personal. Siempre estaré orgulloso por ser parte de su comunidad universitaria. Lobo rojo de corazón.

A mis amigos por su paciencia y apoyo en momentos difíciles, en especial, a Gabriel Miranda.

Al Dr. Jesús Martínez y al Dr. Mario Henry Rodríguez por permitirme ser parte de sus proyectos y ayudarme a crecer en el ámbito de la bioinformática.

Por último, a Jaqueline Sánchez por su paciencia y apoyo durante los momentos más complicados. Un soporte muy importante en la parte final de este proyecto.

RESUMEN

El 16s ARNr es el marcador genómico comúnmente utilizado en los estudios de metagenómica en los últimos años para establecer las relaciones filogenéticas existentes entre los organismos procariotas. Este tipo de análisis ha tenido una enorme repercusión en la anotación taxonómica bacteriana. La secuencia del gen 16s es empleada como cronómetro molecular debido a una serie de características: a) presente en todas las bacterias, lo que lo hace un blanco molecular para su identificación; b) tiene una estructura y función que han permanecido constantes en el proceso evolutivo, por lo que las alteraciones en la secuencia reflejan probablemente cambios aleatorios (1-2% variación en la secuencia cada 50 millones de años) que contienen suficiente variabilidad para diferenciar no sólo los organismos más alejados, también los más próximos; c) el tamaño relativamente largo de los ARNr 16S (aproximadamente 1500 pb) minimiza las fluctuaciones estadísticas; d) debido a que resulta relativamente fácil de secuenciar, existen muchas bases de datos y que están en continuo crecimiento.

Para la amplificación del ARNr 16s completo se utilizan oligonucleótidos diseñados en las regiones conservadas próximas al extremo 5´y 3´del gen para amplificar un fragmento de 1500 pb. Sin embargo, se ha demostrado que para una identificación no se requiere de la secuencia completa, se puede utilizar una región de menor tamaño (aproximadamente 600 pb). Esta región en la posición 804-1392 pb del gen 16S, incluye cinco sitios variables (parte de la V4, V5-V8) y cuatro conservados, siendo una región ideal para identificar y clasificar (a nivel de género) las especies bacterianas porque presenta variabilidad y grado de conservación en su estructura.

El presente estudio tiene como objetivo desarrollar un algoritmo para el análisis masivo de secuencias ribosomales en la región de 804 a 1392 pb del gen 16S. Este algoritmo se basa en agrupar secuencias ribosomales como primer paso y con base en la formación de grupos, realiza el proceso de clasificación de secuencias bacterianas.

Las secuencias en la región de interés se agruparon mediante K-medias y DTW. El uso de estos algoritmos mejoró la agrupación con respecto a los métodos basados en Unidades Taxonómicas Operativas (OTUS). Además, la clasificación de las secuencias ribosómicas se mejoró usando una región cercana a 600 pb con respecto al uso de la secuencia completa del gen 16S.

ABSTRACT

16S rRNA is the genomic marker commonly used in metagenomics studies in recent years to establish phylogenetic relationships among prokaryotes. This type of analysis has had a huge impact on bacterial taxonomic annotation. The sequence of the 16S gene is used as a molecular marker due to a number of features: a) 16S gene is presented in all bacteria, which makes it a target for molecular identification; b) its structure and function remained constant in the evolutionary process, so the alterations in the sequence probably reflect random changes (1-2% variation in the sequence every 50 million years) that contain enough variability to distinguish not only the farthest, but also the nearest organisms (16S gene contains nine variables V1-V9 sites and ten preserved sites); c) because it is relatively easy to sequence, there are many databases and are constantly growing.

For amplification of the 16S rRNA it is used in oligonucleotides designed, conserved regions near the 5 'and 3' of gene end to amplify a fragment of 1500 bp. However, it has been shown that the identification does not require the complete sequence, it can be used a smaller region (about 600 bp). This region in the position 804-1392 bp of the 16S gene includes five variable sites (part of the V4, V5-V8) and four preserved, being an ideal region to identify and classify (at genus level) bacterial species because it presents variability and degree of conservation in its structure.

The present study aims to develop an algorithm for ribosomal sequences massive analysis in the region from 804 to 1392 bp of the 16S gene. This algorithm groups ribosomal sequences and also performs the process of identification and classification of bacterial species.

The sequences in the region of interest, were grouped by clustering algorithms (K-means and DTW). The use of these algorithms improved the grouping with respect to Operational Taxonomic Unit (OTUS) based methods.

In addition, the classification of ribosomal sequences was improved using a region close to 600 bp with respect to the use of the complete sequence of the 16S gene.

LISTA DE FIGURAS

LISTA DE TABLAS

ÍNDICE GENERAL

INTRODUCCIÓN

En la actualidad el estudio de la diversidad microbiana ha cobrado gran importancia, esta forma parte de las distintas funciones biológicas esenciales de cada uno de los ecosistemas de la Tierra. Por lo tanto, el estudio de la ecología microbiana es clave para comprender estos entornos y su efecto funcional (Lagkouvardos *et al*., 2016). Sin embargo, el estudio en esta área se ha visto complicado debido a que los procariotas son los organismos más numerosos y más diversos en el planeta. Para solucionar este problema, es trascendental analizar los múltiples ambientes en los que se detecte cualquier grupo de estos microorganismos. Para esto, la técnica más utilizada es la obtención de secuencias del gen ribosómico ARNr 16S, debido a que este gen cuenta con características específicas para la asignación taxonómica (Yang & Iwasaki, 2014). Además, es un marcador genético que está presente en todas los microorganismos procariotas, contiene regiones variables y conservadas, ampliamente analizado y estudiado para la identificación y clasificación de comunidades bacterianas, mediante la construcción de árboles filogenéticos (Vinje *et al*., 2014).

En las últimas 3 décadas, el método de Sanger ha sido el enfoque dominante y estándar la secuenciación del ADN. El lanzamiento comercial de la primera plataforma de pirosecuenciación masiva en paralelo en 2005 marcó el comienzo de la nueva era de análisis genómico de alto rendimiento, ahora se conoce como secuenciación de próxima generación (NGS por sus siglas en inglés). NGS ha alterado fundamentalmente la investigación genómica y ha permitido a los investigadores llevar a cabo experimentos que anteriormente no eran técnicamente factibles o asequibles (Voelkerding *et al*., 2009).

Con la ayuda de la tecnología de secuenciación de próxima generación, los investigadores ahora pueden obtener millones de secuencias de la firma microbiana para diversas aplicaciones que van desde los estudios epidemiológicos en humanos a las encuestas mundiales de los océanos. El desarrollo de estrategias de cálculo avanzados para extraer al máximo la información pertinente a partir de datos masivos de nucleótidos se ha convertido en un foco importante de la comunidad bioinformática (Sun *et al*., 2010).

La magnitud de la información que genera las investigaciones realizadas en las ciencias de la vida es enorme. En la actualidad existen diferentes tipos de bases de datos biológicas que son utilizadas por aplicaciones bioinformáticas para analizar y clasificar diferentes microorganismos, en las cuales se encuentran: NCBI (*National Center for Biotechnology Information*), una base de datos universal para realizar búsqueda y análisis de similitud de genes con diferentes organismos (https://www.ncbi.nlm.nih.gov/) , ITS (*Internal*

Transcribed Spacer), es una base datos específica para el estudio y clasificación de hongos, y las bases de datos para el estudio y clasificación taxonómica y filogenética de procariotas como son: RPD (*Ribosomal Database Project*) , Silva y Green Genes, en las cuales se almacena la información genética del gen 16S ribosomal (https://rdp.cme.msu.edu, http://www.arb-silva.de).

Ante la gran cantidad de información generada, uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los datos genómicos –de secuencia, de expresión, de estructura, de interacciones, etc. (Febles Rodríguez & Gonzalez-Perez, 2002). Además de evaluar la calidad de la información en los repositorios públicos. Según Ashelford y colaboradores en 2005, uno de cada 20 registros públicos de secuencias ribosomales tiene anomalías. Las quimeras formadas durante la amplificación de la reacción en cadena de la polimerasa (PCR por sus siglas en inglés) o errores producidos en el proceso de secuenciación, han existido mucho tiempo en las bases de datos públicas (Ashelford *et al*., 2005). Debido a estos problemas presentes en la gran información producida sobre secuencias ribosomales, se han creado bases de datos con secuencias ribosomales analizadas manualmente, es decir, se filtran secuencias con errores producidos en el proceso de secuenciación o quimeras generadas en la amplificación de PCR, ejemplos de estos repositorios públicos son EzTaxon y EzTaxon-e (Chun *et al*., 2007; Kim *et al*., 2012).

Las bases de datos mencionadas anteriormente, han provocado el desarrollo de una infinidad de herramientas bioinformáticas para su análisis, en especial en el área de identificación y clasificación de secuencias ribosomales. Tales como las herramientas bioinformáticas proporcionadas por RDP (J. R. Cole *et al*., 2009) y SILVA (Quast *et al*., 2013). Estas herramientas tienen una característica particular, son los creadores de las bases de datos de referencia, en este grupo también entra EzTaxon (Kim *et al*., 2012) y NCBI utilizando la Herramienta Básica de Búsqueda de Alineación Local (BLAST por sus siglas en inglés). Además existen aplicaciones bioinformáticas creadas para trabajar con bases de datos públicas o datos obtenidos por los usuarios como NAST (DeSantis *et al*., 2006), Mothur (Schloss *et al*., 2011), PyNAST (Caporaso *et al*., 2010) por mencionar algunas. Además, la gran mayoría de las herramientas desarrolladas analizan el gen 16S completo (1542 pb aproximadamente). Sin embargo, existen herramientas que tienen la particularidad de trabajar con ciertas regiones del gen 16S como RDP (J. R. Cole *et al*., 2009) y SILVA (Quast *et al*., 2013). Los estudios realizados con regiones específicas están enfocados en ciertos grupos de bacterias o *archaeas*, es decir, fragmentos del gen 16S

totalmente estudiados, son utilizados para identificar a nivel de filo, orden o familia a ciertos grupos de microorganismos (Garcia-Mazcorro *et al*., 2017).

Las herramientas bioinformáticas desarrolladas comparten una característica: el proceso de análisis es complicado. Esto es notorio desde la entrada de los archivos obtenidos por el proceso de secuenciación (diversos formatos y generación de archivos en cada paso) hasta la visualización de los resultados. Debido a esto, la interpretación de los resultados es complejo para los usuarios de estas herramientas.

En esta investigación se analizó la región 804pb-1392pb del gen 16S ribosomal como marcador para la clasificación e identificación de procariotas (Lundberg *et al*., 2012). Se tomó como base de datos control EzTaxon-e (Kim *et al*., 2012). Se agrupó EzTaxon-e con el algoritmo de *clustering* K-medias (Eiler *et al*., 2012) y Alineamiento Temporal Dinámico (DTW por sus siglas en inglés) para calcular distancias entre secuencias (Skutkova *et al*., 2015). Con esto se establece una alternativa para agrupar secuencias ribosomales con una región de aproximadamente 600pb.

Se desarrolló un algoritmo de clasificación (CTB, clasificador taxonómico bayesiano) basado en el método bayesiano simple (Wang *et al*., 2007). Se entrenó CTB con la base de datos control EzTaxon (Kim *et al*., 2012) y se realizó una prueba en la que se comparó precisión para clasificar en los diferentes niveles taxonómicos y tiempo total del proceso de clasificación frente a dos clasificadores, *RDP Classifier* (Wang *et al*., 2007) y *Classifier 16S* (Chaudhary *et al*., 2015). Para esta prueba se utilizó un *benchmark* de la base de datos de Silva *Release* 128 (https://www.mothur.org/wiki/Silva_reference_files). Los resultados demuestran que el promedio de clasificación de CTB es mayor (96.17%) con respecto a *Classifier 16S* (23.75%) y *RDP Classifier* (95.31%).

CAPÍTULO 1. MARCO TEÓRICO

1.1 Introducción

Durante décadas los investigadores han necesitado comprender la diversidad microbiana del planeta. Debido a que es enorme la cantidad de microorganismos procariotas presentes en los diversos ecosistemas y los estudios por cultivo comprenden una pequeña fracción de diversidad, los investigadores han optado por realizar estudios alternos que ofrezcan un contexto más amplio de la variedad microbiana (Klindworth *et al*., 2013)

Actualmente, se estima que la biosfera contiene 10^{30} ~10^{31} células microbianas, es decir, de 2~3 órdenes mayor que el número de células animales y vegetales. Estos microorganismos son clave en la vida de la naturaleza y de los seres vivos que se relacionan con ella, sin embargo, la estructura de las diferentes comunidades microbianas así como su diversidad son poco conocidas por parte de la sociedad científica, ya que solamente unas escasas miles de especies microbianas han sido descritas de manera formal (Sun *et al*., 2010).

Para los estudios de diversidad microbiana, el gen 16S es el principal marcador para evaluar la composición de muestras de suelos, mares o seres humanos. El alto grado de conservación de la secuencia del gen ribosomal proporciona grandes ventajas sobre los estudios filogenéticos y la identificación de taxonomías nuevas (Haas *et al*., 2011). Con el reciente desarrollo de la tecnología de secuenciación masiva en paralelo, ahora los investigadores pueden obtener material genético a partir de muestras ambientales sin necesidad de aislar o cultivar especies a nivel laboratorio y generar conocimiento sobre las grandes comunidades microbianas (Sun *et al*., 2010).

Las NGS han marcado una nueva era en el análisis de la biodiversidad, proporcionan análisis de alto rendimiento sobre comunidades microbianas complejas con amplicones cortos del gen ribosomal 16S. Estas tecnologías de nueva generación aportan una profundidad de secuenciación conveniente por muestra para caracterizar el microbioma a un 99.9% (Bokulich *et al*., 2012). Debido a la gran cantidad de información que generan las nuevas tecnologías de secuenciación masiva se han creado varios repositorios públicos (mencionados en el apartado de Introducción), sin embargo, muchas de las secuencias depositadas en estas fuentes de información es probable que sean de baja calidad aun cuando estas colecciones de datos sean revisados manualmente por “curadores” (Haas *et al*., 2011).

Mientras las tecnologías de secuenciación de próxima generación ha propiciado el aumento masivo de la disponibilidad de datos de secuencias ribosómicas, la bioinformática se enfrenta a nuevos retos informáticos que se deben afrontar para mejorar el análisis de estos metadatos generados (Scholz *et al*., 2012). Por ello, uno de los puntos importantes para la bioinformática es analizar la tasa de error generada por el secuenciador, es decir, herramientas bioinformáticas que depuren las secuencias obtenidas antes de realizar el cálculo de biodiversidad de la muestra. Con esto, se disminuye las posibilidades de sobrestimar la diversidad microbiana. Se ha comprobado que estos filtros ayudan a disminuir la cantidad de Unidades Operacionales Taxonómicas (OTUS por sus siglas en inglés) y con ello mejorar la precisión de agrupación de las secuencias analizadas (Huse *et al*., 2010).

En la actualidad, la mayoría de las herramientas bioinformáticas orientadas a la identificación recurren al algoritmo de alineamiento de pares de bases creado por Needleman y Wunsch en 1970 (Needleman & Wunsch, 1970). Este algoritmo también forma parte de la herramienta más conocida para la identificación llamada BLAST e implementada por NCBI (Altschul *et al*., 1990). En el caso de la agrupación, existen diversos métodos como la similitud de palabra corta en conjunto con árboles de decisión donde la clave es la existencia de un umbral para decidir a qué grupo pertenece la secuencia analizada (W. Li & Godzik, 2006). Otro método de agrupación es el algoritmo K-medias en conjunto con alineamientos de pares de secuencias (Edgar, 2010).

1.1.1 Ribosomas procariotas

Los ribosomas procariotas son orgánulos cuya importancia es esencial para la vida celular de los microorganismos, ya que son los encargados del complejo proceso de síntesis de proteínas (Rodicio & Mendoza, 2004). Estos orgánulos tienen un coeficiente de sedimentación de 70S, se dividen en dos subunidades: subunidad mayor y menor (Figura 1.1). La subunidad mayor tiene un coeficiente de sedimentación de 50S, además de una molécula de ARN ribosomal (ARNr 5S, 23S) y 34 proteínas. Mientras que la subunidad menor tiene un coeficiente de sedimentación del 30S. Esta subunidad tiene una molécula ARNr (16S) y 21 proteínas (Valderas Álvarez, 2012).

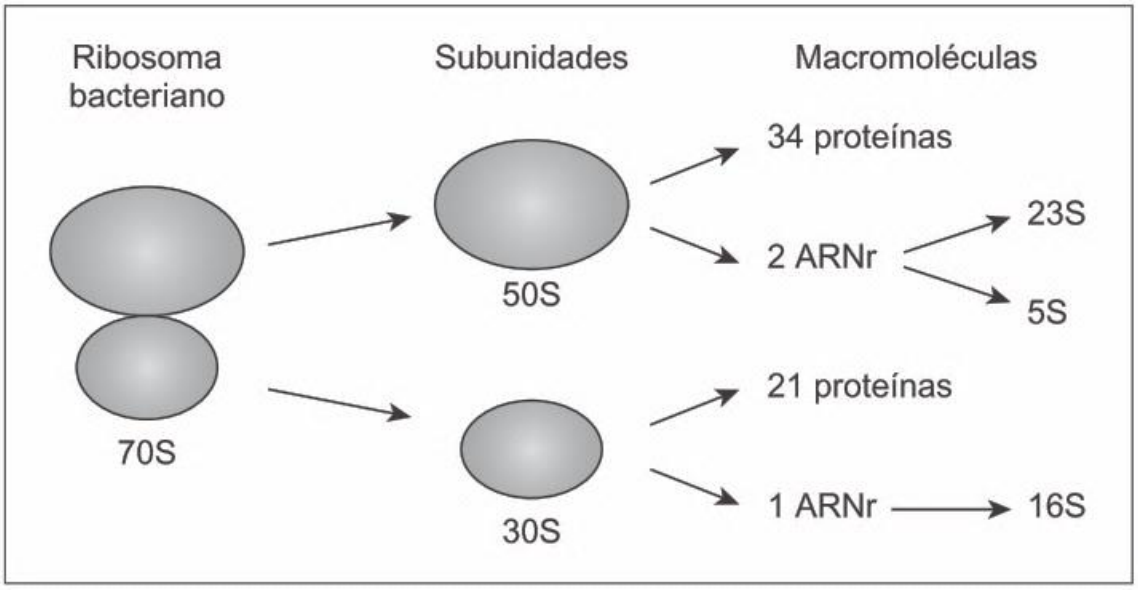


Figura 1.1 Esquema del ribosoma procariota y sus principales componentes (del Rosario Rodicio & del Carmen Mendoza, 2004).

1.1.2 Gen Ribosomal 16S

El ARNr 16S es un polirribonucleótido que se encuentra en los ribosomas de los procariotas en la región 30s. Tiene una longitud de aproximadamente 1500 nucleótidos, este ARNr es codificado por el ADN ribosomal 16S (ADNr 16S). Este gen está conformado por 9 regiones variables, las cuales están flanqueadas por regiones conservadas. Las regiones conservadas son normalmente utilizadas para realizar estudios *in silico* e *in situ* utilizando técnicas moleculares como la PCR (Van De Peer *et al*., 1994). Como se muestra en la Figura 1.2 las regiones en verde, representan secuencias que tienen un alto grado de conservación en todos los microorganismos. Estos sitios son ideales para el uso de oligonucleótidos para la amplificación por PCR de manera que todos los genes 16S en una muestra son amplificados. En cambio, las regiones grises, corresponden a secuencias con menor grado de similitud y son específicas a nivel de especie, estás permiten a los científicos ver qué especies están presentes en una comunidad (Teach the Microbiome, 2017).

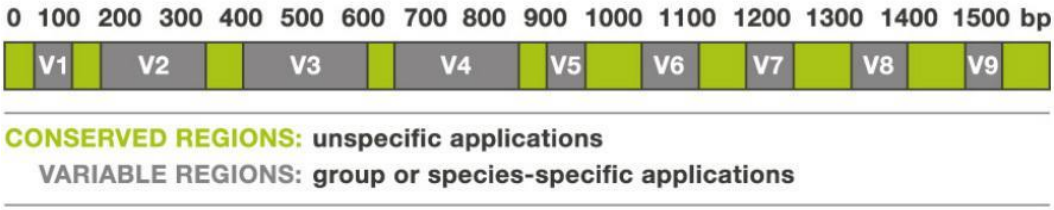


Figura 1.2 Ejemplo de un gen 16S con sus regiones variables y conservadas (Teach the Microbiome, 2017).

El método molecular de identificación bacteriana basado en el análisis de la secuenciación del gen 16S incluye tres etapas: a) amplificación del gen a partir de la muestra apropiada; b) determinación de la secuencia de nucleótidos del amplicón, y c) análisis de la secuencia (Rodicio & Mendoza, 2004). La tercera etapa se refiere al proceso bioinformático que se realiza para la identificación y clasificación bacteriana a partir de secuenciación masiva (Figura 1.3) (Regier *et al*., 2019).

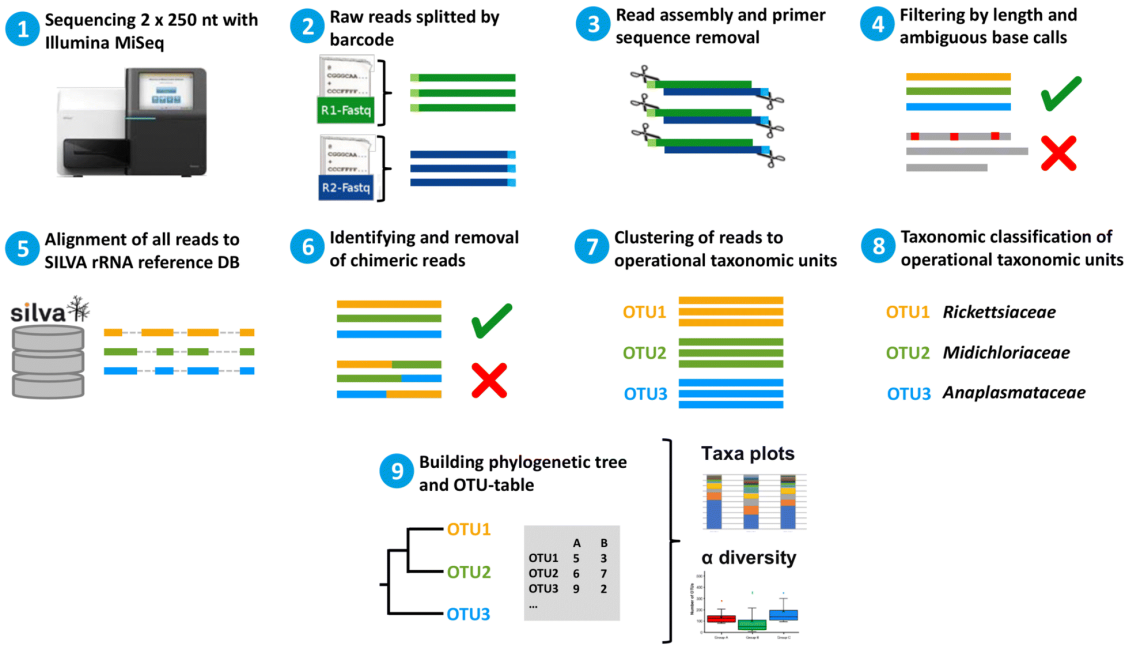


Figura 1.3 Etapas a seguir en el proceso bioinformático para la identificación bacteriana mediante el gen 16S (Regier *et al*., 2019).

El gen 16s ha sido ampliamente utilizado para identificar especies de bacterias y estudios taxonómicos (Choi *et al*., 1996). Y es considerado como un marcador genético bacteriano por las siguientes características (Woese, 1987):

1. Es una molécula muy antigua, presente en todos los microorganismos procariotas. Por lo tanto, es un marcador universal para su identificación.
2. Su conservación en cuanto a estructura y función es muy alta, de tal modo que las mutaciones en su secuencia manifiestan solo cambios aleatorios.
3. La tasa de mutación es muy baja y además muy lenta, por lo cual aportan gran información acerca de los procariotas. Los ARNr de la subunidad pequeña contienen, sin embargo, suficiente variabilidad para diferenciar no sólo los organismos más alejados, sino también los más próximos.

4. El tamaño relativamente largo del gen 16S (1.500 nt) minimiza las oscilaciones estadísticas.
5. La conservación en estructura secundaria puede servir de ayuda en las comparaciones, aportando una base para el alineamiento preciso.
6. Dado que resulta relativamente fácil secuenciar el gen 16S existen bases de datos amplias, en continuo crecimiento.

1.1.3 Bases de datos biológicas

Actualmente existen descripciones de 179 bases de datos, de las cuales 95 son nuevas en la revista británica *Nucleic Acids Research*. Estas bases de datos, en conjunto con otras reportadas en diferentes revistas, han sido incluidas en la *Molecular Biology Database Collection* (Colección de Bases de Datos de Biología Molecular) de *Nucleic Acids Research*, registrando 1170 bases de datos en total (Galperin & Cochrane, 2009).

En la Tabla 1.1 se muestran las principales bases de datos públicas disponibles actualmente.

Tabla 1-1 Principales bases de datos disponibles en la actualidad.

Base de Datos	Breve descripción	Liga
AceBD	Base de datos del genoma del *Caenorhabditis elegans*	www.acebd.org
DDBJ	Principal base de datos de secuencias de nucleótidos en Japón	www.ddbj.nig.ac.jp
EMBL	Principal base de datos de secuencias de nucleótidos en Europa	www.ebi.ac.uk/embl/index.html
ExPASY	Base de datos de la proteómica	http://us.expasy.org/
FlyBase	Base de datos del genoma de Drosophila	http://flybase.bio.indiana.edu/
GenBank	Principal base de datos de secuencias de nucleótidos en la NCBI	www.ncbi.nlm.nih.gov/Genbank
HIV databases	Datos de la secuencia del VIH y la información inmunológica relacionada	www.hiv.lanl.gov/content/index
TAIR	Base de datos de información del	www.arabidopsis.org

	Arabidopsis	
Ribosomal database Project	Secuencias de ARN ribosomal y arboles filogenéticos derivados de las secuencias	http://rdp.cme.msu.edu
SILVA	Bases de datos alineadas (16S / 18S, SSU) (23S / 28S, LSU) de secuencias de ARN ribosomal (ARNr)	http://www.arb-silva.de
Greengenes	Bases de datos de 16S ARNr y herramientas bioinformáticas	http://greengenes.secondgenome.com

De las cuales, los repositorios con información del gen 16S son:

- RDP: Base de datos del Proyecto Ribosomal (Ribosomal Database Project, por sus siglas en inglés). Base de datos que contiene alineamientos y anotaciones sobre datos de secuenciación de ARNr, junto con herramientas para permitir a los investigadores analizar sus propias secuencias de rRNA en el marco de RDP. Contiene colecciones de genes de la subunidad pequeña del rRNA de bacteria y arquea así como el gen de la subunidad larga rRNA de hongos (James R Cole *et al*., 2014). (http://rdp.cme.msu.edu/)
- NCBI: base de datos compuesta ya que alberga un conjunto de bases de datos de secuencia, taxonomía, genomas, mutaciones, entre otras y además herramientas como BLAST para búsquedas por similitud de secuencia (www.ncbi.nlm.nih.gov/nuccore).
- Greengenes: es una base de datos 16S ARNr que proporciona a los usuarios una taxonomía curada basada en la inferencia *de novo* árbol (McDonald *et al*., 2012). Además provee monitoreo de quimeras, alineamiento estándar y clasificación taxonómica usando múltiples taxonomías publicadas (DeSantis *et al*., 2006) (http://greengenes.secondgenome.com/).
- SILVA: sitio web que contiene bases de datos de calidad controlada de secuencias del gen 16S alineadas de dominios de Bacteria*, Archea* y *Eukaryota* y servicios en línea complementarios. Contiene secuencias de genes de 3 194 778 de la subunidad pequeña y 288 717 de la subunidad grande (Quast *et al*., 2013) (http://www.arb-silva.de).
- Eztaxon: EzBiocloud es el portal público de datos y análisis de ChunLab que se centra en taxonomía, ecología, genómica, metagenómica y microbioma de bacterias y arqueas. Ofrecen servicios en la nube que incluyen herramientas de

bioinformática y funcionan con bases de datos anteriores publicadas, que incluyen EzTaxon, EzTaxon-e, y EzGenome. La red EzBioCloud ha sido utilizada en más de 50 países diferentes por más de 22.000 usuarios, desde académicos, organizaciones de investigación sin fines de lucro, comunidades médicas, agencias gubernamentales y compañías globales (Chun *et al*., 2007; Kim *et al*., 2012) (http://www.ezbiocloud.net/).

Se cuenta con una gran cantidad de datos de donde se puede elegir para analizar, sin embargo, en algunas situaciones, una sola base de datos no puede dar respuestas a los complejos problemas de los biólogos. La integración o la recopilación de información de varias bases de datos para resolver problemas y descubrir nuevos conocimientos son otros retos importantes en bioinformática. La transformación de datos biológicos voluminosos en información útil y en conocimiento valioso es un reto a la hora de descubrir conocimientos. La identificación e interpretación de patrones interesantes que están escondidos en miles de millones de datos biológicos genéticos es una meta clave de la bioinformática. Este objetivo abarca la identificación de las estructuras de genes útiles en secuencias biológicas, la derivación de los conocimientos de diagnóstico a partir de datos experimentales, y la extracción científica de la información de la literatura.

Por ello, las bases de datos biológicas están bajo constante estudio para hacerlas más eficientes y facilitar su uso. Esto ayuda a la integración de diferentes fuentes de información y por consecuencia aporta ventajas a los recursos bioinformáticos disponibles para la comunidad científica (Baker & Brass, 1998).

1.1.4 Secuenciación masiva

La diversidad de los microorganismos de la tierra sigue siendo poco conocida a pesar de un estimado de 1.5 millones de especies de bacterias y hongos. Estos tienen funciones vitales como descomponedores, simbiontes y patógenos en los ecosistemas. Hasta la fecha, sólo el 5% del número estimado de especies bacterianas se ha documentado. En los últimos años, los estudios metagenómicos han mejorado la comprensión de la diversidad de microbios en diversos hábitats. Esto incluye los microbios asociados con las plantas, microbioma intestinal animal y humana, agua y aire (Akinsanya *et al*., 2015).

La secuenciación de próxima generación (NGS por sus siglas en inglés) ha permitido el desarrollo de plataformas de secuenciación masiva que han dado lugar al aumento de los esfuerzos de investigación para comprender la composición y la función de las poblaciones bacterianas (Jovel *et al*., 2016). El principio básico de este tipo se secuenciación surge en los años 70’s, cuando se publicaron 2 artículos que describen métodos para la secuenciación

del ADN. Estos métodos son los de Allan Maxam y Walter Gilbert, Frederick Sanger y colaboradores. El refinamiento y comercialización de este último método condujo a su amplia difusión en toda la comunidad de investigación y, en última instancia, en el diagnóstico clínico. La tecnología Sanger fue utilizada en la secuenciación del primer genoma humano, que se completó en el año 2003 a través del Proyecto Genoma Humano, un esfuerzo de 13 años con un costo estimado de $ 2,7 mil millones. En 2008, en comparación, un genoma humano fue secuenciado en un período de 5 meses por aproximadamente $ 1.5 millones. Este último logro pone de relieve la capacidad del campo en rápida evolución de la NGS, tecnologías que han surgido durante los últimos 5 años (Voelkerding *et al*., 2009). Por tanto, las nuevas plataformas se distinguen por su capacidad de secuenciar millones de fragmentos de ADN de forma paralela a un precio mucho más barato por base (Figura 1.4). Además la secuenciación masiva tiene el potencial de detectar todos los tipos de variación genómica en un único experimento, incluyendo variantes de nucleótido único o mutaciones puntuales, pequeñas inserciones y deleciones, y también variantes estructurales tanto equilibradas (inversiones y traslocaciones) como desequilibradas (deleciones o duplicaciones) (Febles Rodríguez & Gonzalez-Perez, 2002).

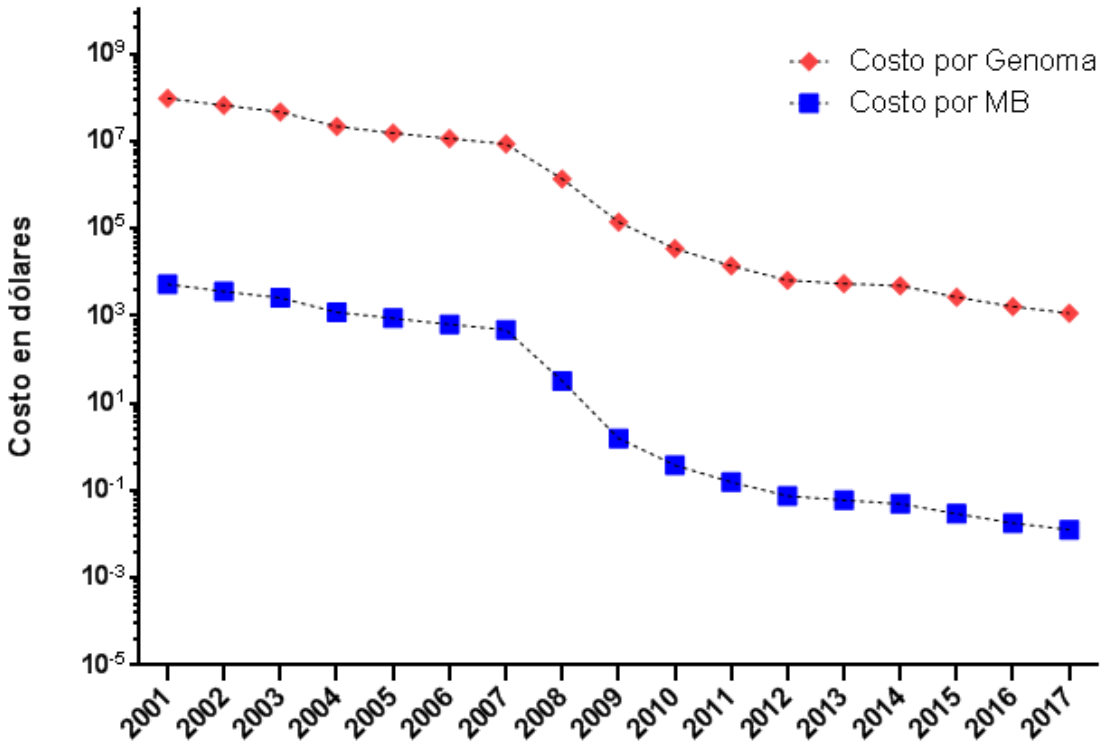


Figura 1.4 Costos de la secuenciación a lo largo de los últimos 17 años por MB y por genoma. Los costos han disminuido significativamente en los últimos diez años (https://www.genome.gov/sequencingcostsdata).

En este contexto, la NGS de genomas es una tecnología ideal debido a su rapidez y economía, pero adicionalmente porque permite obtener datos confiables y de buena calidad para el entendimiento de sistemas biológicos a nivel de genoma, transcriptoma, metaboloma o epigenoma. La NGS de genomas implica la confluencia de equipos automatizados, en los que los avances experimentales en química, ingeniería, biología molecular, nanotecnología se entrelazan con la computación de alto rendimiento, para incrementar la velocidad con la cual se obtienen los datos, su almacenamiento, así como su análisis y procesamiento (Tabla 1-2) (Rodríguez-Santiago & Armengol, 2012).

Tabla 1-2 Tipos de secuenciación disponibles por las principales empresas (Pillai *et al*., 2017)

Compañía	**Plataforma**	**Tipo de secuenciación**	**Longitud de Lectura/Salida**	**Ventajas**	**Desventajas**
Roche	454 GS junior	Pirosecuenciación	400-500 pb/35Mb	Lecturas con longitud largas Corridas rápidas	Caro Baja profundidad Errores con homopolímeros
Illumina	MiSeq MiSeq Dx MiSeq FGx NextSeq 500 NextSeq 550	Secuenciación por síntesis	2 x 300pb/15G 2 x 300pb/15G 2 x 300pb/15G 2 x 500pb/120G 2 x 500pb/120G	Menos cantidad de errores Alta profundidad Bajo costo Flujo de trabajo completamente automático	Corridas lentas Lecturas cortas
Life Technologies	Ion PG Ion Proton	Basado en semiconductores Secuenciación por síntesis	200-400 pb/20Mb-1Gb Dependiendo del chip 200 pb/10Gb	Corridas cortas Alta escalabilidad Diferentes chips disponibles Bajo costo	Alta tasa de error

1.1.5 Bioinformática

La bioinformática es la aplicación de las tecnologías informáticas en las ciencias biológicas. De acuerdo al Centro Nacional para la Información Biotecnológica, NCBI (National Center for Biotechnology Information por sus siglas en inglés), la bioinformática se define como un campo de la ciencia en el que confluyen varias disciplinas como son: la biología, la computación y las tecnologías de la información; con el fin de facilitar el descubrimiento de nuevos conocimientos y el desarrollo de perspectivas globales a partir de las cuales puedan discernirse principios unificadores en el campo de la biología (Cañedo & Arencibia, 2004).

Los estímulos principales para el desarrollo de la bioinformática son (Febles Rodríguez & Gonzalez-Perez, 2002):

- El enorme volumen de datos generados por los distintos proyectos denominados genoma (humano y de otros organismos).
- Los nuevos enfoques experimentales, basados en biochips, que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos) de enfoques celulares (expresión génica).
- El desarrollo de Internet, que permite el acceso universal a las bases de datos de información biológica.

La bioinformática es un área del espacio que representa la biología molecular computacional, que incluye la aplicación de las computadoras y de las ciencias de la información en áreas como la geonómica, el mapeo, la secuencia y determinación de las secuencias y estructuras por métodos clásicos. Muchos de los métodos de la computación y de las ciencias de la información sirven para estos fines, incluyendo el aprendizaje de las máquinas, las teorías de la información, la estadística, la teoría de los grafos, los algoritmos, la inteligencia artificial, los métodos estocásticos, la simulación, la lógica, etc. (Febles Rodríguez & Gonzalez-Perez, 2002).

La bioinformática involucra la solución de problemas complejos biológicos usando herramientas y sistemas computacionales que incluyen la recolección, organización, almacenamiento y recuperación de información biológica a partir de la base de datos (Abd-Elsalam, 2003).

El desarrollo de la técnica de reacción en cadena de la polimerasa (PCR, por sus siglas en inglés) en 1986 ha hecho que el crecimiento de los datos biológicos pase de 606 secuencias a más de 82 millones. El desarrollo de la tecnología de secuenciación de nueva generación ha hecho posible secuenciar con una simple corrida más de un millón de millones

de bases. Las aplicaciones de estas metodologías incluyen análisis de genómica comparativa, genómica de microorganismos, detección de SNP de alto rendimiento, secuenciación y análisis de micro ARN, identificación de mutaciones de genes en rutas metabólicas asociadas a enfermedades, análisis de transcriptomas de organismos (transcriptómica), determinación de genes con baja tasa de exposición en sus ambientes naturales, y obtención de información sobre la variación genética a nivel de especies, poblaciones y ecosistemas. La bioinformática tendrá que desarrollar formas distribuidas de almacenamiento para sus bases de datos cada vez más eficientes y veloces, que permitan la actualización, sincronización y consulta permanente de las bases de datos (Barreto Hernández, 2008).

La bioinformática con el tiempo se ha vuelto parte importante para la genómica básica y los estudios en biología molecular, también está siendo aplicada en muchas áreas de la biotecnología y las ciencias biomédicas, influyendo de manera directa en los resultados obtenidos. Ejemplos de estas aplicaciones son los conocimientos de diseño de fármacos, estudios en el área forense, genómica humana e investigaciones agrícolas. Uno de los grandes avances en cuanto a la relación bioinformática-genómica, es la medicina personalizada. Esto permitirá a los médicos detectar posibles mutaciones en el genoma de un paciente y aplicar tratamientos eficaces contra enfermedades. Desde este punto de vista, el genoma humano se convierte en el principal protagonista en el diagnóstico y tratamiento de enfermedades de manera personalizada. Por otra parte, los estudios en la agricultura se han favorecido por las herramientas bioinformáticas desarrolladas, las bases de datos de plantas así como el estudio de expresión génica ha propiciado el desarrollo de nuevos cultivos con características especiales como mayor resistencia a enfermedades y mayor crecimiento en menor tiempo, aumentando los números de producción (Escobar *et al*., 2011).

1.1.5.1 DTW en el análisis de distancias entre secuencias de ADN

DTW es un algoritmo muy utilizado en la comparación de dos series de tiempo. El objetivo de esta comparación es encontrar la mínima distancia entre estas dos series de tiempo (vectores numéricos), para ello, una serie se toma como referencia y la segunda serie (serie de entrada) se alarga y se deforma de tal modo que se encuentre la mayor similitud, es por ello que las longitudes de ambas series pueden ser desiguales (Figura 1.5) (Giorgino, 1996).

La aplicación de DTW en el área biológica se ha dado en el análisis de redes regulatorias de genes (Figura 1.6) (Lee *et al*., 2012), en la secuenciación selectiva en tiempo

real (Loose *et al*., 2017) y en la evaluación de similitud de diferentes genes en estudios filogenéticos (Skutkova *et al*., 2015), por mencionar algunos estudios.

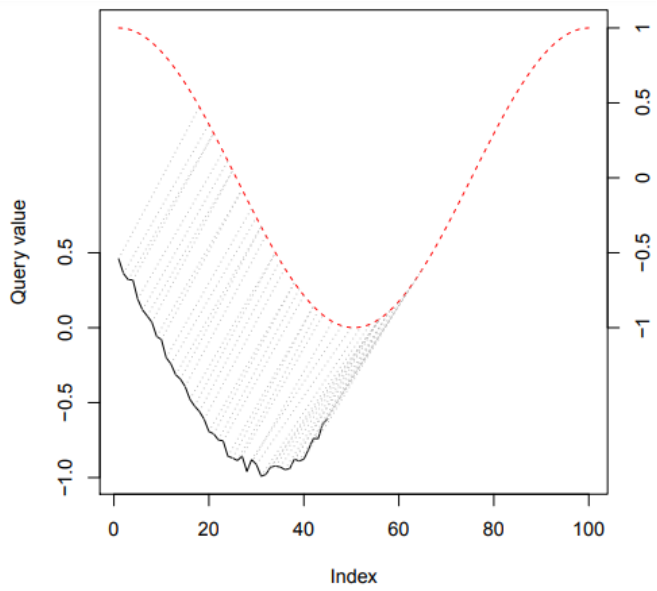


Figura 1.5 Alineamiento parcial entre una serie de entrada (línea sólida) y una serie de referencia (línea punteada). La serie de entrada se deforma para encontrar la mayor similitud con la serie de referencia (Giorgino, 1996).

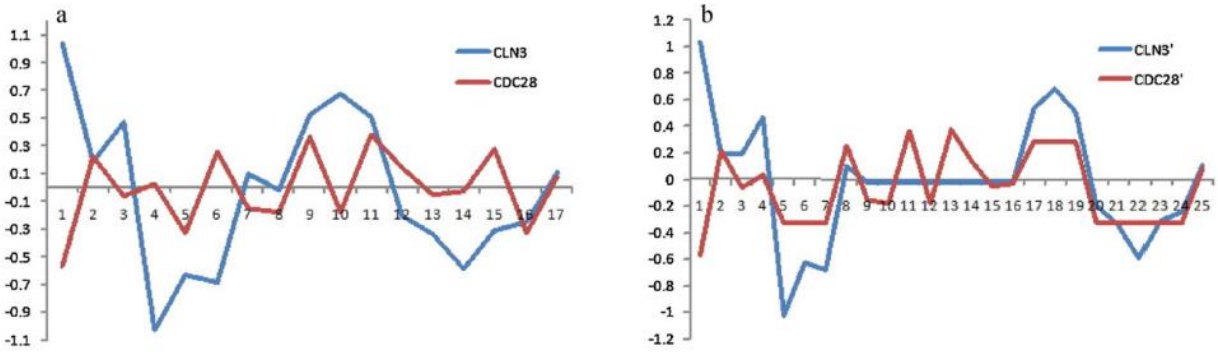


Figura 1.6 Aplicación de DTW en la categorización de diversos genes (activación, inhibición o no regulación) basado en la relación de expresión entre pares de genes en diferentes tiempos. a) expresión de dos pares de genes (CLN3, CDC28) y b) aplicación de DTW en los dos pares de genes analizados (CLN3´, CDC28´).

1.1.5.2 Clasificador *Naive* Bayes

Como se ha mencionado en apartados anteriores, ante la enorme cantidad de datos que se generan a partir de las nuevas tecnologías de secuenciación masiva, surgen necesidades tecnológicas que ayuden en el descubrimiento de información en temas biológicos. Debido a la complejidad que conlleva los análisis biológicos, es necesaria la implementación de herramientas y métodos propios de la inteligencia artificial (Orozco Arias

& Arango López, 2016). Existen métodos de clasificación de datos, como los algoritmos de aprendizaje supervisado, que han sido utilizados en cuestiones biológicas como son los árboles aleatorios (*Random forest*), regresión logística, vecino más cercano (*K-Nearest Neighbor* o KNN por sus siglas en inglés) y *Naive* Bayes. Este último, es un clasificador que se basa en la probabilidad, en el cual, con el conteo de frecuencias y combinaciones de valores de un conjunto de datos, se calcula el conjunto de probabilidades con las que se determinará la probabilidad condicional de que ocurra el evento X, el evento Y, del evento X dado evento Y y viceversa. Con esto, se determina la categoría en la cual se clasificará cada muestra estudiada. Este tipo de algoritmo bajo problemas de clasificación controlados tiende a aprender rápidamente (Villagrana-Bañuelos *et al*., 2020).

En 2019 Alaff y colaboradores realizaron un estudio sobre la detección de casos positivos de personas infectadas por el virus SARS-CoV-2 con base en la clasificación de texto publicado en la red social Twitter en Turquía. Evaluaron 8 algoritmos de clasificación (*Naive Bayes, Generalized linear model, Logistic regression, Fast large margin, Deep learning, Decision tree, Random forest, Gradient boosted trees*) con los siguientes atributos: tos, fiebre, dolor de garganta, dificultad para respirar y dolor de cabeza. Además, incluyeron palabras relacionadas con estos atributos como falta de aire, asfixia. Realizaron 30 corridas independientes para cada algoritmo con 1000 datos sin procesar respectivamente.

Los resultados reflejan una alta precisión por parte del algoritmo *Naive* Bayes con una tasa de error del 6.4% y un tiempo de entrenamiento menor (2.2 segundos) con respecto a los algoritmos evaluados. Además, arrojó que los atributos con mayor impacto en los casos positivos detectados (69%) son tos y falta de aire.

1.2 Planteamiento del problema

El análisis de los datos procedentes de estudios de secuenciación masiva es complejo no solo por la enorme cantidad de información que se genera por muestra sino también por las peculiaridades asociadas a cada una de las combinaciones posibles entre plataforma de secuenciación y sistema de captura. Actualmente, existe una amplia gama de posibles piezas de software orientadas a resolver pasos muy concretos del extenso flujo de análisis necesario para el procesamiento de este tipo de datos. Sin embargo, la combinación de diferentes herramientas, así como la elección de diferentes parámetros en su ejecución producen resultados muy dispares ofreciendo distintos grados de incertidumbre que repercuten en los resultados y su interpretación. Aunado a esto, las representaciones de los resultados provenientes de los exhaustivos análisis por parte de las diferentes herramientas

bioinformáticas son poco claros, difíciles de interpretar y en diversas ocasiones es necesaria otra herramienta para facilitar la interpretación por parte del usuario.

CAPÍTULO 2. ANTECEDENTES

2.1 Características particulares del gen 16S

Las regiones variables y conservadas del gen 16S ARNr son potenciales blancos de estudio, dado que su papel principal es la identificación y clasificación taxonómica de bacterias. Diferentes grupos de investigación se han enfocado en elucidar si existen regiones de menor tamaño que pudieran igualar o mejorar los resultados reportados al utilizar la secuencia completa del gen. Con esto se puede disminuir la carga computacional requerida para los análisis masivos, ya que realizar alineamientos y análisis de secuencias más grandes suele ser complejo.

En ese sentido Vinje *et al*., 2014, realizaron un estudio con el objetivo de investigar dónde se encuentran los sitios más discriminativos en el gen marcador 16S y si estos sitios se encuentran en regiones conservadas o variables, en la Figura 2.1 se muestra las regiones variables y conservadas que se identificaron después de calcular la entropía de los alineamientos de secuencias con las tres bases de datos de referencia (Greengenes, RDP y SILVA) (Vinje *et al*., 2014).

Los resultados del estudio muestran que los sitios más variables se ubican en V1-V3, esta región tiene el sitio con mayor conservación ubicado entre V2 y V3. El sitio V4 más el fragmento que lo limita con V5 tienen un nivel equilibrado entre variabilidad y conservación, además con el sitio de mayor número de huecos resultado de los alineamientos y con el sitio más discriminativo según los puntajes de relación de selectividad utilizados como criterio de discriminación de secuencias como se muestra en la Figura 2.2.

La región V5-V9 (situada en 900pb-1500pb) está compuesta por sitios variables y conservados bien definidos en los tres alineamientos con las diferentes librerías. Esta región contiene el mayor número de sitios conservados, en los cuales existe cierta variabilidad que puede ser una ventaja para los análisis taxonómicos, así como de identificación. Es decir, comprende los sitios más discriminativos según los puntajes de relación de selectividad (véase Figura 2.2).

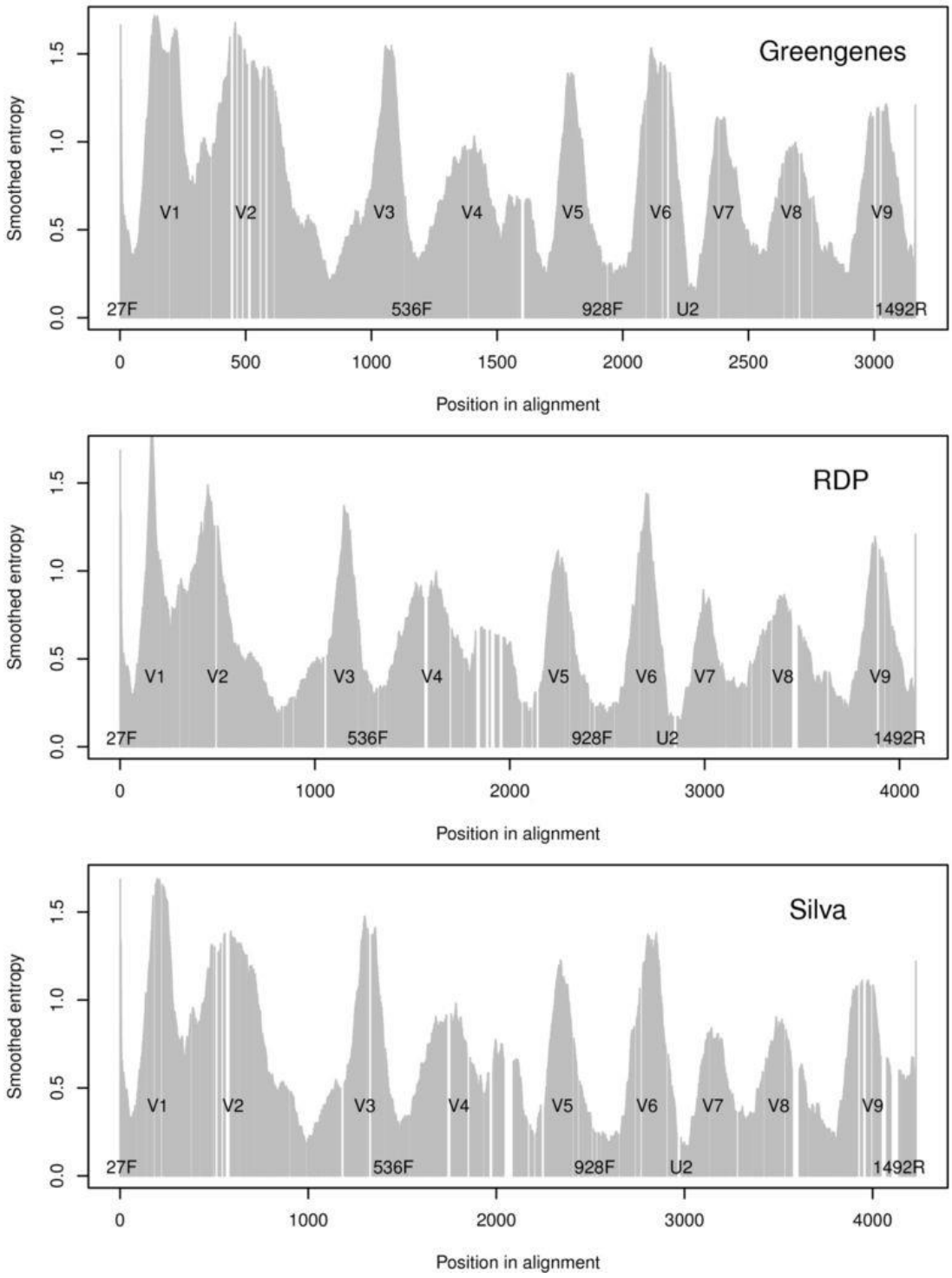


Figura 2.1 Sitios variables y conservados identificados en el análisis del gen 16S (Vinje *et al*., 2014).

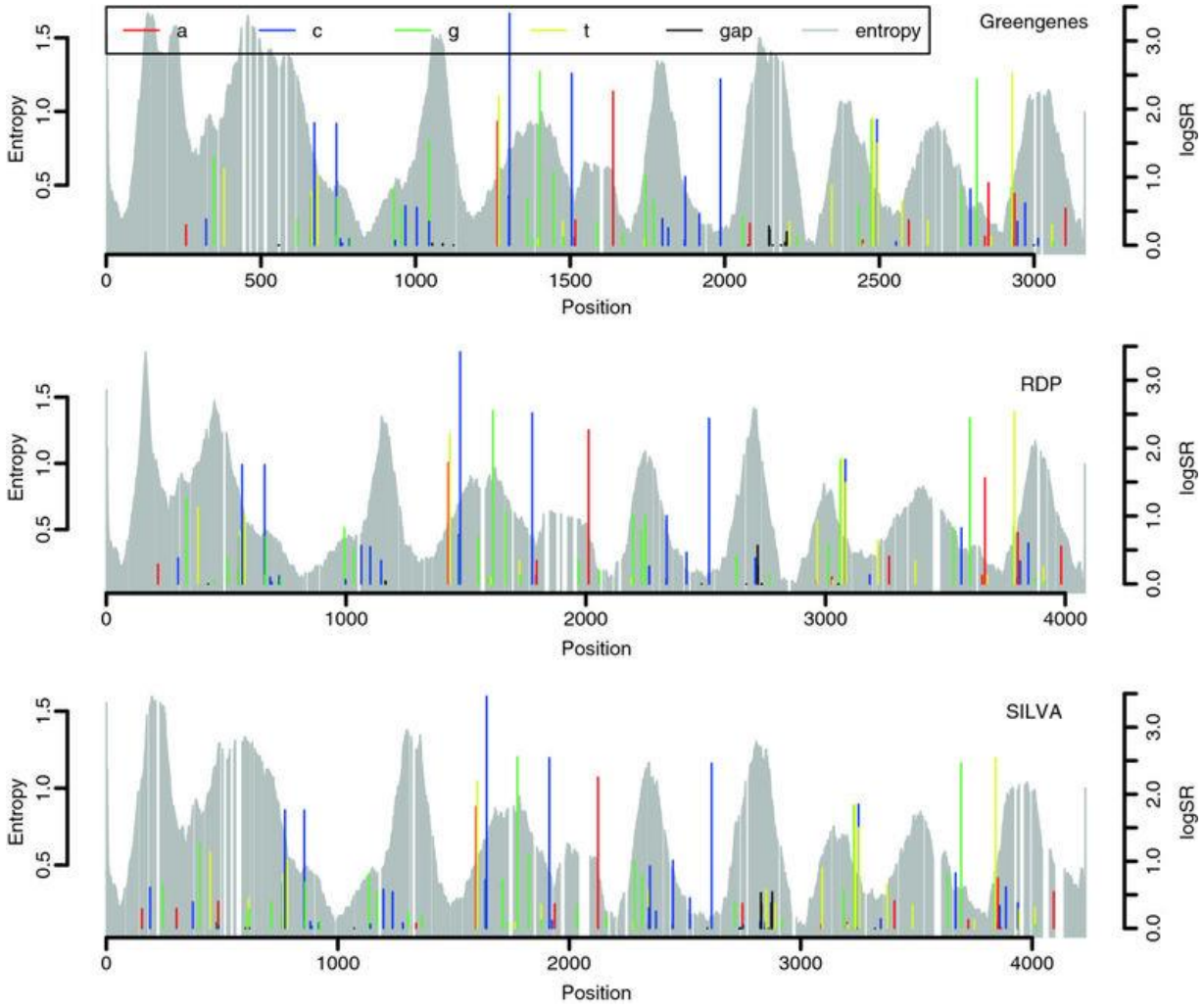


Figura 2.2 Sitios más discriminativos detectados (barras de diferentes colores) según relación de selectividad (Vinje *et al*., 2014).

Otro estudio en el que se desarrolló una herramienta bioinformática llamada V-Xtractor para localizar, verificar y extraer regiones hipervariables (V-regiones) del gen ribosomal 16S de procariotas y secuencias de hongos fue realizado en 2010 por Hartmann y colaboradores. Este instrumento de *software* tiene una fiabilidad del 99. 6% y una gran sensibilidad para identificar falsos positivos. Esta herramienta es útil en el pretratamiento de los datos para un análisis comunitario (Hartmann *et al*., 2010).

En este estudio evaluaron la base de datos SILVA de referencia, amplificando mediante oligonucleótidos en la ubicación 28F y 1491R según las posiciones de *E. Coli*. Encontraron sitios consenso para flanquear las V-regiones utilizando fragmentos de 20, 30, 40 y 50 nucleótidos en dirección río arriba y río abajo respectivamente. En la Figura 2.3 se muestra en la parte superior el análisis de secuencias bacterianas, en la parte de en medio los resultados de secuencias de *archeas* y en la gráfica inferior sobre secuencias de hongos.

Las diferentes posiciones de inicio se pueden observar en la parte superior de cada gráfica, río arriba en las barras de color verde y río abajo en color rojo. Las barras en color negro son posiciones de *primers* utilizados comúnmente en la amplificación del gen ribosomal. El área negra es el promedio de variabilidad dada los resultados calculados por los Modelos Ocultos de Markov y estos se pueden observan como líneas verdes (río arriba) y rojas (río abajo) a lo largo de los alineamientos (Hartmann *et al*., 2010). Con este estudio se identifican las nueve regiones hipervariables en las secuencias ribosomales en procariotas y en hongos.

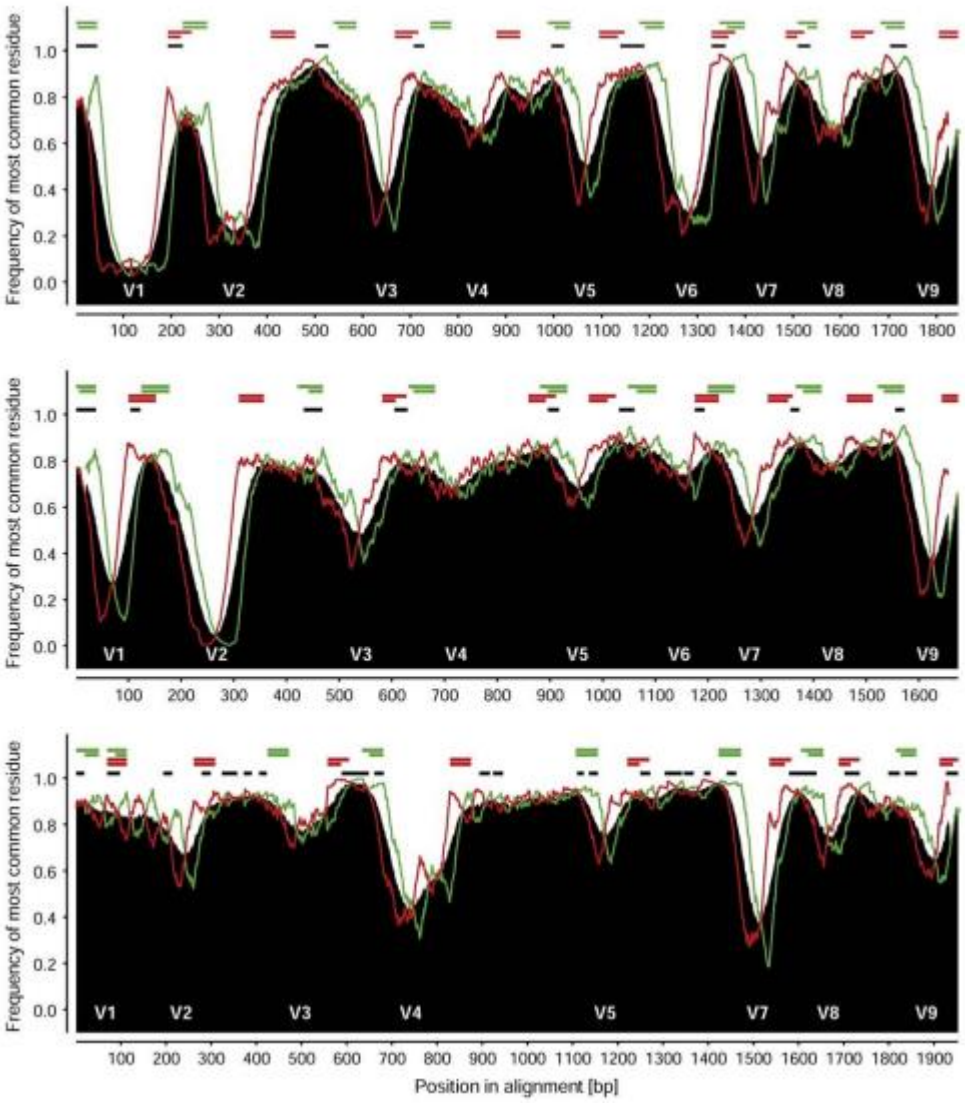


Figura 2.3 Análisis de secuencias ribosomales para localizar, verificar y extraer regiones hipervariables (Hartmann *et al*., 2010).

Este mismo proceso lo aplicaron a tres bases de datos RDP, Greengenes y SILVA para medir la eficiencia en el proceso de extracción de regiones variables, la Figura 2.4 muestra que los porcentajes en la mayoría de las regiones analizadas son del 100%. Se

puede observar que la región con menor eficiencia es la V9 en las secuencias procariotas, sin embargo, la disminución en la eficiencia no es significativa con respecto a las otras regiones analizadas.

	Reference	Extended[a]		
	Silva	Silva[b]	RDP	Greengenes
Bacteria	13104 (100.0)	90972 (100.0)	121975 (100.0)	87981 (100.0)
V1	13102 (100.0)	90750 (99.8)	121792 (100.0)	87922 (100.0)
V2	13104 (100.0)	90880 (99.9)	121841 (99.9)	87832 (99.9)
V3	13104 (100.0)	90911 (99.9)	121909 (100.0)	87896 (100.0)
V4	13103 (100.0)	90822 (99.8)	121813 (99.9)	87860 (99.9)
V5	13103 (100.0)	90855 (99.9)	121817 (100.0)	87897 (100.0)
V6	13104 (100.0)	90886 (99.9)	121882 (100.0)	87908 (100.0)
V7	13104 (100.0)	90902 (99.9)	121906 (100.0)	87929 (100.0)
V8	13104 (100.0)	90927 (100.0)	121929 (100.0)	87945 (100.0)
V9	13071 (99.7)	90507 (99.5)	119997 (98.5)	87265 (99.8)
Processing time[c]	46 min	372 min	433 min	313 min
Archaea	247 (100.0)	1841 (100.0)	2471 (100.0)	1626 (100.0)
V1	247 (100.0)	1838 (99.8)	2464 (99.7)	1623 (99.8)
V2	247 (100.0)	1839 (99.9)	2468 (99.9)	1625 (99.9)
V3	247 (100.0)	1839 (99.9)	2468 (99.9)	1624 (99.9)
V4	247 (100.0)	1834 (99.6)	2464 (99.7)	1624 (99.9)
V5	247 (100.0)	1832 (99.5)	2462 (99.6)	1626 (100.0)
V6	247 (100.0)	1841 (100.0)	2471 (100.0)	1626 (100.0)
V7	247 (100.0)	1839 (99.9)	2468 (99.9)	1626 (100.0)
V8	247 (100.0)	1839 (99.9)	2468 (99.9)	1626 (100.0)
V9	245 (99.2)	1835 (99.7)	2319 (93.8)	1623 (99.8)
Processing time	1 min	6 min	9 min	8 min
Fungi	357 (100.0)	2349 (100.0)	NA	NA
V1	357 (100.0)	2296 (97.7)	NA	NA
V2	357 (100.0)	2251 (95.8)	NA	NA
V3	357 (100.0)	2344 (99.8)	NA	NA
V4	357 (100.0)	2272 (96.7)	NA	NA
V5	357 (100.0)	2340 (99.6)	NA	NA
V7	357 (100.0)	2240 (95.4)	NA	NA
V8	357 (100.0)	2252 (95.9)	NA	NA
V9	357 (100.0)	2247 (95.7)	NA	NA
Processing time	1 min	9 min		

[a] Datasets were not de-replicated and individual entries may exist in multiple datasets.
[b] The extended SILVA datasets do not contain the sequences already included in the reference set.
[c] Time (minutes) required to extract all V-regions from the respective dataset (3 GHz dual-core, 8 GB memory).

Figura 2.4 Porcentaje de eficiencia en la extracción de regiones hipervariables (Hartmann *et al*., 2010).

Los resultados sobre la localización de regiones variables y conservadas obtenidas por Vinje y colaboradores son similares a los expuestos por Hartmann y colaboradores. Las regiones variables en procariotas en ambos estudios están bien definidas, así como las regiones conservadas, además se observa una consistencia en las tres bases de datos analizadas (RDP, SILVA y Greengenes).

Otro estudio realizado por Vasileiadis *et al*., 2012, cuyo objetivo es analizar las secuencias ribosomales de las regiones variables V3, V4, V5 y V6 obtenidas por la secuenciación masiva de Illumina tomadas de muestras de suelo, localizaron las regiones

conservadas y variables de 42109 secuencias. Para ubicar los fragmentos conservados calcularon entropía de Shannon y además los sitios conservados que limitaban las regiones hipervariables fueron estudiados para el diseño de cebadores. En la Figura 2.5 se muestran los resultados del cálculo de entropía de Shannon de secuencias casi completas. Las regiones que presentaron mayor variabilidad fueron V3 y V6, las regiones con secuencias con una longitud mayor (105pb) fueron identificadas en V3 y V4, por el contrario, V5 y V6 tienen una longitud de secuencia corta entre 27pb-35pb.

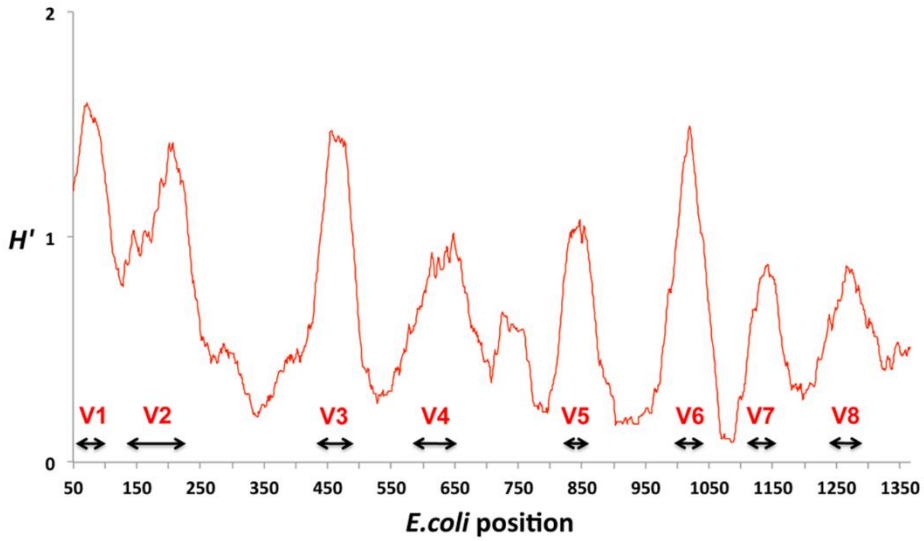


Figura 2.5 Cálculo de entropía de secuencias casi completas de muestras de suelo (Vasileiadis *et al*., 2012).

El estudio concluye que la región V3 en el proceso de identificación muestra mejores resultados con datos experimentales de suelo que con datos de referencia. La región V4 mostró una baja conservación en comparación con las otras regiones variables, lo cual puede bajar el rendimiento en el estudio de diversidad. V5 manifestó mejores resultados en la selección de diversidad con datos de referencia. El rendimiento de V6 fue muy bajo, es por ello que los estudios se enfocaron a las otras regiones. Los análisis de estas regiones fueron comparados con los mismos estudios con secuencias completas de referencia (Figura 2.7).

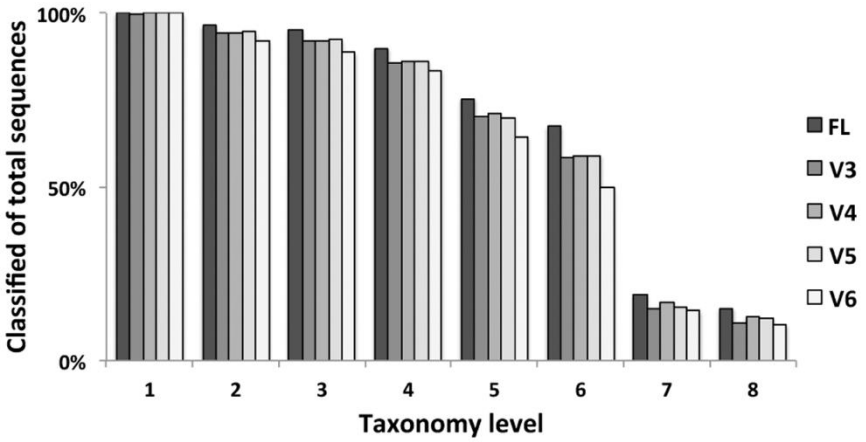


Figura 2.6 Nivel de clasificación con diferentes regiones variables y secuencia completa (Vasileiadis *et al*., 2012).

Por último, Vasileiadis *et al*., 2012 sugiere que las lecturas de regiones variables presentan errores en los procesos de clasificación e identificación en la relación con el uso de secuencias completas. Además, V3 es una región con propiedades que mejoran los resultados en estudios de análisis de suelos. V5 es una región relativamente eficaz dentro de las regiones variables cortas.

2.2 El uso de regiones cercanas al 804pb-1392pb del gen 16S

Diversos estudios se han desarrollado acerca de regiones del gen ribosomal 16S, tal es el caso de Lundberg *et al*., 2012, quienes diseñaron oligonucleótidos sentido que abarcaban las posiciones 804pbF, 926pbF 1114pbF y un oligonucleótido universal en la posición 1392R, donde demostraron por medio de secuenciación que el oligonucleótido 804pbF tiene el menor número de lecturas con respecto a los otros oligonucleótidos, sin embargo, el porcentaje de lecturas que reconocen OTUs de plantas o quiméricos es mínimo. Es decir, las lecturas del amplicón 804pbF identifican cerca del 100% a especies procariotas con un error bastante bajo. Con esto, es evidente que la diversidad calculada con Shannon-Wiener y Chao a nivel de especie con muestras mayores a 1000 lecturas de los amplicones 926pbF y 1114pbF es mucho mayor con respecto al 804pbF (Figura 2.7).

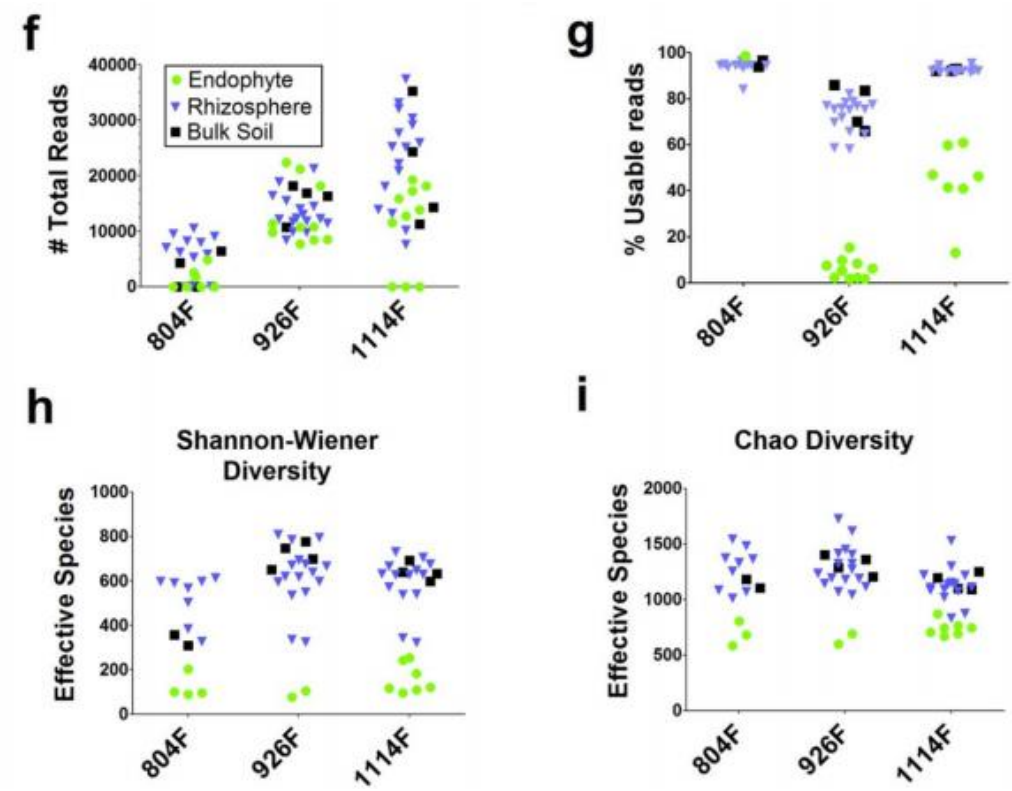


Figura 2.7 Muestra la homogeneidad en el comportamiento de las lecturas con los diferentes oligos diseñados (Lundberg *et al*., 2012).

Chelius & Triplett en 2001 desarrollaron el primer oligonucleótido de desapareamiento (799F), el diseño del oligonucleótido se centró en torno a dos desajustes de pares de bases en las posiciones 798-799 y dos desajustes de pares de bases adicionales en las posiciones 783 y 784 en el ADN del cloroplasto. Cabe mencionar que estos estudios están orientados al estudio de comunidades bacterianas en plantas, en los cuales utilizan oligonucleótidos que amplifican secuencias del gen 16S bacterianas y evitan la amplificación de secuencias de ADN de cloroplasto.

Los estudios basados en evitar la amplificación de mitocondrias y cloroplastos se han hecho *in silico*, es por ello, que Beckers *et al*., 2016 realizaron un estudio experimental con un conjunto de oligonucleótidos (Figura 2.8) para el analizar de las comunidades bacterianas asociadas a plantas en la rizosfera, las raíces, los tallos y las hojas de los álamos híbridos (*Populus tremula x P. alba*).

Primer pairs	Primer sequence (5′–3′)	A	B	C	D	References
799F	AACMGGATTAGATACCCKG	79.7	0.29	4	V5-V6-V7	Chelius and Triplett, 2001
1391R	GACGGGCGGTGWGTRCA	84.6	1.44	0		Walker and Pace, 2007
967F	CAACGCGAAGAACCTTACC	80.9	0.34	0	V6-V7	Sogin et al., 2006
1391R	GACGGGCGGTGWGTRCA	84.6	1.44	0		Walker and Pace, 2007
799F	AACMGGATTAGATACCCKG	79.7	0.29	4	V5-V6-V7	Chelius and Triplett, 2001
1193R	ACGTCATCCCCACCTTCC	78.1	0.20	0		Bodenhausen et al., 2013
341F	CCTACGGGNGGCWGCAG	91.2	0.05	0	V3-V4	Klindworth et al., 2013
785R	GACTACHVGGGTATCT AATCC	86.2	0.09	0		Klindworth et al., 2013
68F	TNANACATGCAAGTCGRRCG	72.5	0.60	0	V1-V4	McAllister et al., 2011
783Rabc	CTACC*AGGGTATCTAATCC*TG	70.9	5.05	3		Sakai et al., 2004
68F	TNANACATGCAAGTCGRRCG	72.5	0.60	0	V1-V3	McAllister et al., 2011
518R	WTTACCGCGGCTGCTG G	87.6	0.09	0		Lee et al., 2010
341F	CCTACGGGNGGCWGCAG	91.2	0.05	0	V3-V4	Klindworth et al., 2013
783Rabc	CTACC*AGGGTATCTAATCC*TG	70.9	5.05	3		Sakai et al., 2004

Primers are indicated as forward (F) or reverse (R).
**Primer 783Rabc is a primer mix (Sakai et al., 2004):*
(a) 5′-CTACCAGGGTATCTAATCCTG-3′,
(b) 5′-CTACCGGGGTATCTAATCCCG-3′,
(c) 5′- CTACCCGGGTATCTAATCCGG-3′.
(A) Primer coverage (%) for Bacteria using Silva (Quast et al., 2013) and probeBase (Loy et al., 2007), (B) Primer Score tested with PrimerProspector 1.0.1 (Walters et al., 2011) (Table S1), (C) Number of mismatches with poplar chloroplast DNA, (D) Hypervariable region of the 16S rRNA operon targeted by primer pairs.

Figura 2.8 Conjunto de cebadores seleccionados en el estudio de Beckers *et al*., 2016.

Las diferentes cantidades de contenido de ADN de plásmidos de estos compartimentos vegetales, que van desde prácticamente ningún contenido plastidial (suelo de la rizosfera) a muy altas (cloroplastos y hojas), les permitió evaluar el rendimiento de los conjuntos de oligonucleótidos seleccionados en condiciones específicas (Figura 2.9), concluyendo que el mejor par de cebadores que evita la amplificación de plásmidos son el 799F y 1391R.

Otro estudio realizado por Yarza *et al*., 2014, demostraron que la existencia de umbrales en los análisis taxonómicos proporcionan ventajas en la clasificación de especies cultivadas así como los microorganismos de muestras ambientales. Esto con base en que para una asignación taxonómica confiable es necesario que el tamaño de la secuencia analizada sea mayor a 1300 nucleótidos, lo que actualmente es imposible con las tecnologías de secuenciación de próxima generación, ya que el tamaño de lectura es de alrededor de ~300 pb. Para demostrar esto, evaluaron regiones de aproximadamente ~250 nucleótidos y regiones combinadas. Los resultados que obtuvieron demuestran que entre mayor sea la secuencia analizada, el nivel de confiabilidad con respecto a identificación y taxonomía crece significativamente, sin embargo, el uso de la secuencia completa del gen ribosomal tiende a estimaciones exactas de riqueza y aumenta la precisión sobre clasificaciones taxonómicas (véase Figura 2.10). Además, concluyen que el análisis realizado facilitará el desarrollo de una taxonomía común entre la comunidad científica (Yarza *et al*., 2014).

A. Total reads	799F-1391R	967F-1391R	799F-1193R	341F-785R	68F-783Rabc	68F-518R	341F-783Rabc
Rhizosphere soil	2235 ± 165	2550 ± 673	956 ± 285	1961 ± 119	3346 ± 454	1519 ± 217	2196 ± 317
Root	2728 ± 74	2577 ± 56	1943 ± 129	2916 ± 438	2484 ± 155	3488 ± 532	2548 ± 403
Stem	2811 ± 117	2502 ± 159	2456 ± 486	2199 ± 350	2068 ± 384	3412 ± 632	1386 ± 18
Leaf	2665 ± 100	2402 ± 231	2410 ± 197	2621 ± 134	1961 ± 64	3257 ± 367	1678 ± 81
Read length before QC	405 ± 96	401 ± 101	364 ± 105	392 ± 105	348 ± 139	349 ± 105	361 ± 129
Read length after QC	207 ± 4	208 ± 4	217 ± 5	233 ± 5	222 ± 5	200 ± 4	205 ± 4
NORMALIZATION TO 1000 READS							
B. Chloroplast DNA	**799F-1391R**	**967F-1391R**	**799F-1193R**	**341F-785R**	**68F-783Rabc**	**68F-518R**	**341F-783Rabc**
Rhizosphere soil	0 [a]	0.2 ± 0.3 (< 0.1)[a]	0 [a]	1 ± 2 (0.1)[a]	0 [a]	0 [a]	0.2 ± 0.3 (< 0.1)[a]
Root	0 [a]	786 ± 79 (79)[b]	0 [a]	863 ± 54 (86)[b]	736 ± 90 (74)[b]	975 ± 8 (97)[c]	270 ± 87 (26)[d]
Stem	2 ± 3(0.2)[a]	997 ± 3 (99)[b]	0 [a]	962 ± 1 (96)[b]	993 ± 4 (99)[b]	998 ± 1 (99)[b]	804 ± 36 (80)[c]
Leaf	0 [a]	907 ± 35 (91)[b]	0 [a]	910 ± 29 (91)[b]	894 ± 12 (89)[b]	985 ± 4 (98)[c]	518 ± 71 (52)[d]
C. Mitochondrial DNA	**799F-1391R**	**967F-1391R**	**799F-1193R**	**341F-785R**	**68F-783Rabc**	**68F-518R**	**341F-783Rabc**
Rhizosphere soil	0 [a]	0 [a]	0.5 ± 0.5 (< 0.1)[a]	0 [a]	0 [a]	0 [a]	0 [a]
Root	0 [a]	0 [a]	9 ± 1 (1)[b]	45 ± 17 (5)[c]	15 ± 5 (1)[b]	4 ± 1 (0.5)[b]	136 ± 17 (14)[d]
Stem	0 [a]	0 [a]	19 ± 11 (2)[b]	35 ± 1 (4)[b]	6 ± 3 (0.5)[a]	1 ± 1 (0.1)[a]	173 ± 25 (17)[c]
Leaf	0 [a]	0 [a]	11 ± 2.5 (1)[b]	69 ± 16 (7)[c]	20 ± 13 (2)[b]	6 ± 3 (0.5)[b]	196 ± 53 (20)[d]
D. Bacterial rDNA	**799F-1391R**	**967F-1391R**	**799F-1193R**	**341F-785R**	**68F-783Rabc**	**68F-518R**	**341F-783Rabc**
Rhizosphere soil	1000 ± 0 (100)[a]	999 ± 0.26 (99)[a]	999 ± 0.3 (99)[a]	998 ± 3 (99)[a]	1000 ± 0 (100)[a]	1000 ± 0 (100)[a]	999 ± 0.52 (99)[a]
Root	1000 ± 0 (100)[a]	414 ± 79 (21)[b]	992 ± 1 (99)[a]	92 ± 41 (9)[b]	250 ± 88 (25)[b]	22 ± 7 (2)[c]	594 ± 72 (60)[d]
Stem	997 ± 3 (99)[a]	2 ± 3 (0.2)[b]	982 ± 11 (98)[a]	4 ± 2 (0.3)[b]	1 ± 1 (0.1)[b]	1 ± 2 (< 0.1)[b]	25 ± 12 (3)[b]
Leaf	1000 ± 0 (100)[a]	93 ± 35 (9)[b]	989 ± 3 (98)[a]	22 ± 15 (2)[b]	85 ± 37 (9)[b]	10 ± 6 (1)[b]	278 ± 25 (28)[c]

(A) Total number of reads (± standard deviation) obtained per plant compartment for each primer pair and average read length (± standard deviation) before and after quality control (QC). Average number of chloroplast (B) and mitochondrial (C) sequences (non-target DNA) obtained from each plant compartment by the selected primer pairs. (D) Amplification of bacterial rDNA reads. Values were normalized to 1000 reads and are averages of three biologically independent replicates ± standard deviation. Values between brackets represent the average percentage (%) of reads. Sequence counts were statistically analyzed using a one-way ANOVA within each plant compartment to compare primer pairs. Differences at the 95% significance level are indicated with lower case letters (P < 0.05).

Figura 2.9 Resultados de la amplificación del conjunto de oligonucleótidos (Beckers *et al.*, 2016).

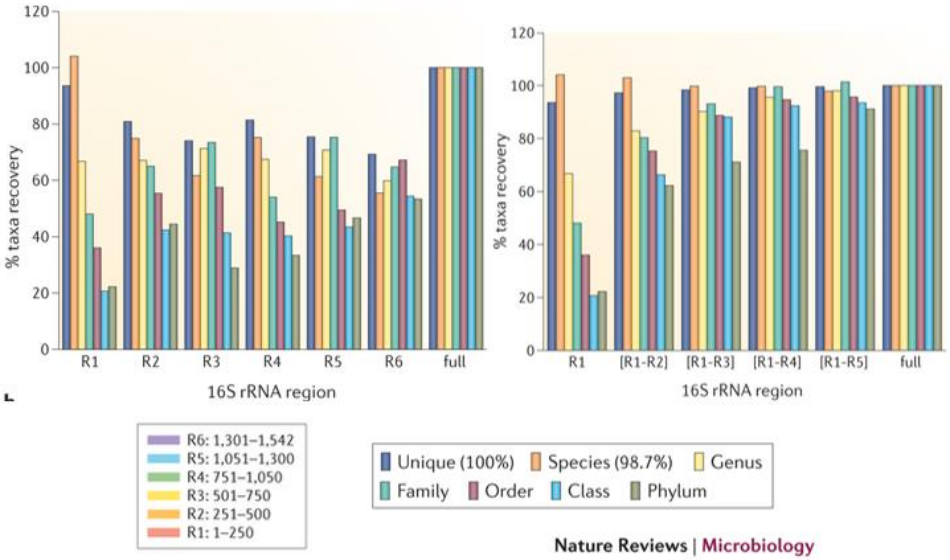


Figura 2.10 Análisis de distintas regiones del gen ribosomal (Yarza *et al.*, 2014).

2.3 Herramientas bioinformáticas para clasificar e identificar secuencias ribosomales

Actualmente, la metagenómica ofrece a los investigadores la imagen más completa de la taxonomía; es decir, ¿qué organismos están allí?, y las relaciones funcionales; es decir, ¿cuáles son los organismos que hacen?; de la composición de las comunidades microbianas de forma nativa en la muestra, por lo que es posible llevar a cabo las investigaciones que incluyen organismos que antes eran intratables para el cultivo en laboratorio controlado (Keegan *et al*., 2016). Es por ello, que se han desarrollado diversas herramientas bioinformáticas para analizar la gran cantidad de información que se genera con las nuevas tecnologías de secuenciación.

myPhyloDB es un ejemplo de ello, esta herramienta es una base de datos personal fácil de usar con una interfaz de navegador diseñado para facilitar el almacenamiento, procesamiento, análisis y distribución de las poblaciones microbianas de una muestra (por ejemplo datos de metagenómica 16S), es decir, análisis comparativo de la abundancia taxonómica, riqueza de especies y la diversidad de especies (Figura 2.10) para proyectos de diversos tipos (por ejemplo, asociada a la humana, microbioma intestinal humana, aire, suelo y agua) para cualquier nivel taxonómico deseado. MyPhyloDB fue diseñado para ser una aplicación basada en web personal, fácil de usar para proveer el almacenamiento de bases de datos para un mejor manejo, procesamiento, análisis y distribución de datos de metagenómica. El objetivo de myPhyloDB no es necesariamente el desarrollo de nuevas herramientas analíticas; más bien, el objetivo es complementar las herramientas de la metagenómica disponibles en la actualidad con capacidades de base de datos y una interfaz gráfica de usuario (GUI) para facilitar el análisis y la comparación de los estudios de metagenómica de diferentes proyectos y áreas de enfoque (por ejemplo, humanos, aire, suelo y agua) (Manter *et al*., 2016).

Chun y colaboradores en 2007, crearon una base de datos de secuencias ribosomales con nombres de cepas válidos publicados y fue revisada manualmente, es decir, filtraron secuencias que contenían errores de secuenciación o quimeras surgidas en la amplificación de PCR (Chun *et al*., 2007). Posteriormente, esta base de datos se amplió en 2012, incorporando secuencias ribosomales de muestras ambientales revisadas manualmente. Este repositorio público contiene representantes de filotipos aún no cultivados. Las asignaciones tanto de especies como de filotipos fueron dadas por análisis filogenéticos basados en el gen 16S ribosomal (Kim *et al*., 2012). Esta base de datos es utilizada por las herramientas basadas en la web, desarrolladas en EzBioCloud (http://www.ezbiocloud.net) y proporcionan diversas funciones, incluyendo un motor de búsqueda de similitud, cálculo de

similitud por pares, la alineación de secuencias múltiples y algoritmos para generar árboles filogenéticos en el lado del servidor. También podría ser útil en diversos niveles de clases para propósitos educativos (Figura 2.12).

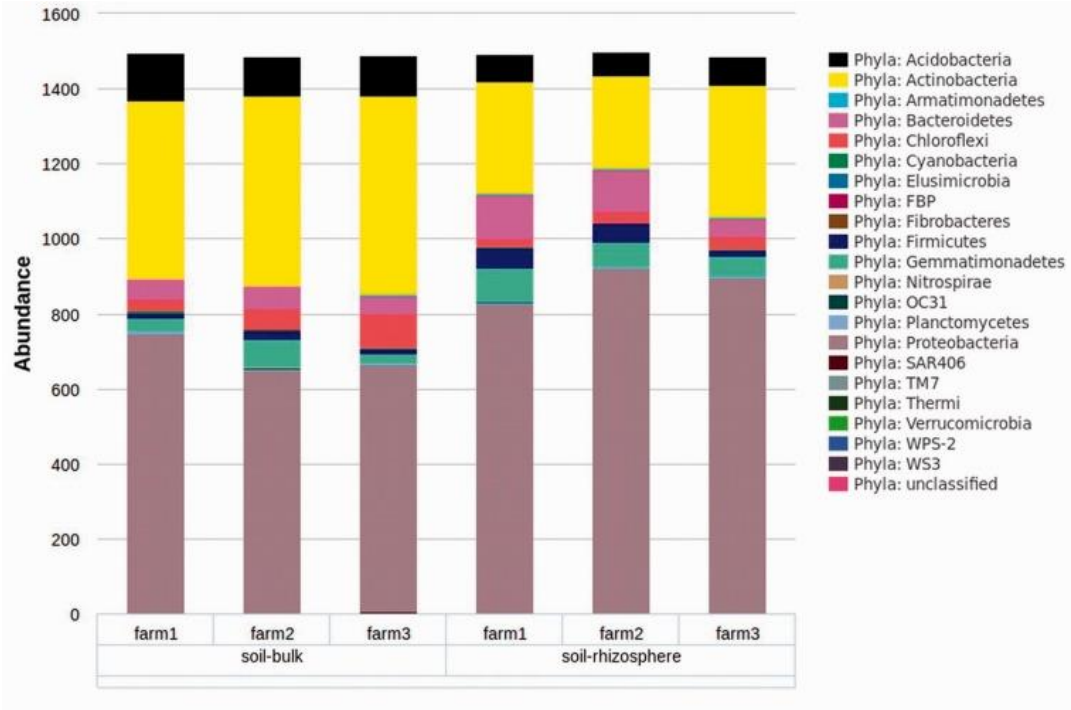


Figura 2.11 Ejemplo de salida del gráfico de un análisis de ANcOVA. Los gráficos de barras son la abundancia taxonómica promedio para cada muestra analizada (Manter *et al*., 2016).

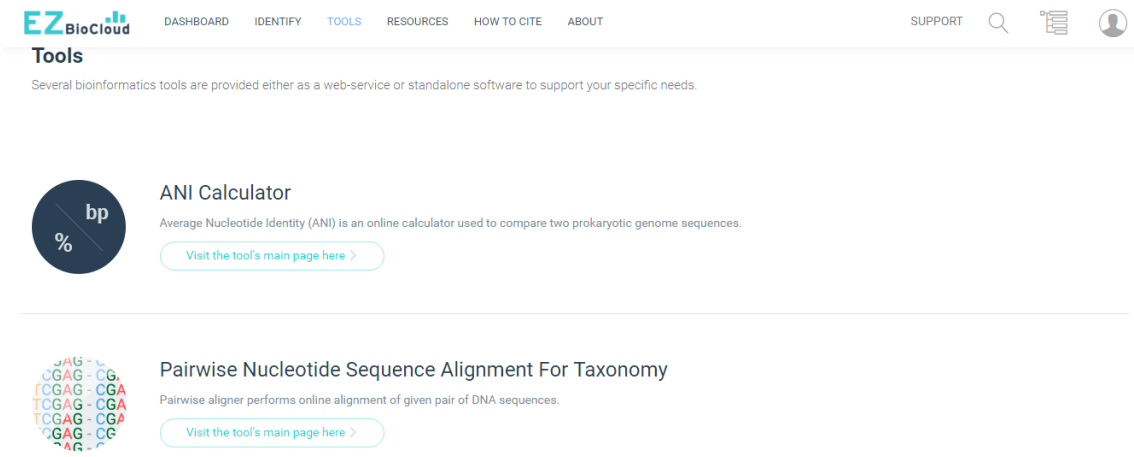


Figura 2.12 Interfaz de EzBioCloud, se muestran dos de sus herramientas que ofrecen.

SILVA es otro proyecto que su principal objetivo es centralizar secuencias ribosomales de los genes de la subunidad menor y mayor, además de secuencias eucariotas. Las secuencias seleccionadas para la base de datos que se lanza de forma libre a la comunidad

se someten a un proceso riguroso que se resume en el *Alignment Quality*. Esta “Calidad de Alineamiento” por su traducción al español, consta de tres valores medidos por el *software* SINA: la puntuación de alineamiento, la puntuación de par de bases y, como de liberación, la identidad de alineación. Además, las secuencias de bacterias y eucariotas deben tener una longitud mayor a las 1200pb, mientras las secuencias de archeas al menos una longitud de 900pb. SILVA ofrece diferentes herramientas bioinformáticas para diferentes objetivos como un alineador de secuencias para la identificación con la capacidad de analizar 1000 secuencias, de un tamaño mayor a las 600 bases, proporcionadas por el usuario y otra herramienta de prueba de cebadores sobre sus bases de datos con la posibilidad de elegir el número máximo de desajustes por base permitidos (véase Figura 2.13). La base de datos de SILVA está basada en las anotaciones de EMBL-BANK (Quast *et al*., 2013).

MetaMetaDB es una base de datos que contiene secuencias de 16S ARNr obtenida de un conjunto de datos de gran tamaño. La información fue extraída de tal manera que para hacer más fácil el análisis de una gran cantidad de datos, la base de datos es compacta y la información está asociada con diversos entornos estudiados (Figura 2.14). El análisis de las secuencias por MetaMetaDB de ciertos procariotas proporciona al usuario ventajas sobre el estudio de la diversidad microbiana y su evolución (Yang & Iwasaki, 2014).

Figura 2.13 Resultados del análisis de un par de cebadores universales en la herramienta TestPrime (Quast *et al*., 2013).

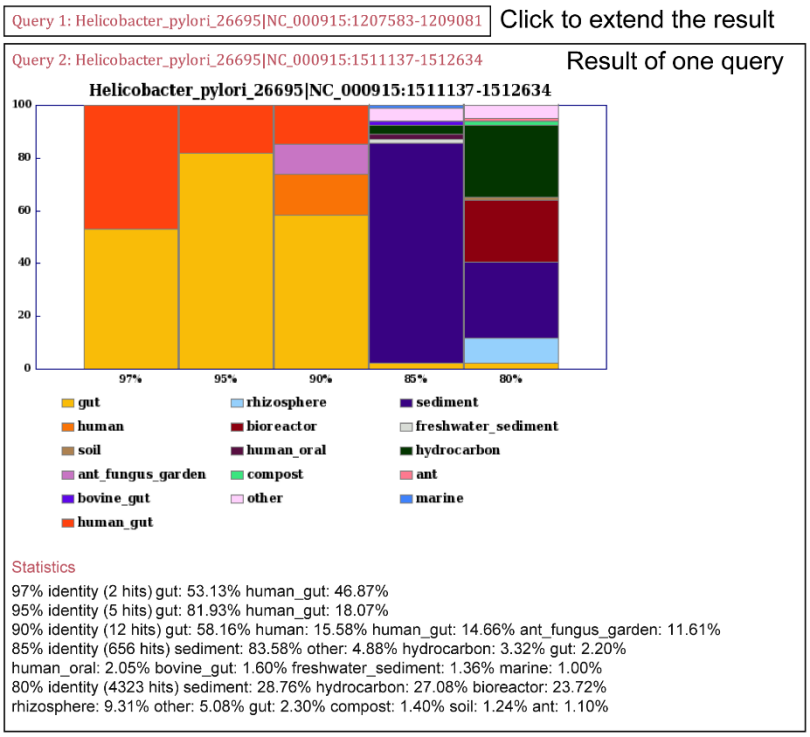


Figura 2.14 Ejemplo de análisis con la base de datos MetaMetaDB (Yang & Iwasaki, 2014).

En 2007 Wang y colaboradores desarrollaron una herramienta de clasificación utilizando el algoritmo Naive Bayes simple. Este clasificador forma parte de las múltiples aplicaciones bioinformáticas ofrecidas por RDP (Wang *et al*., 2007).

El estudio evaluó dos diferentes esquemas taxonómicos, el primero fue el Esquema Taxonómico de los Procariotas de Bergey (*Taxonomic Outline of the Prokaryotes)* (Garrity *et al*., 2004) con 5014 cepas de y el segundo asignado por NCBI con 23095 secuencias del gen ARNr tomadas de la liberación de la base de datos RDP en 2004 y la taxonomía fue tomada de GenBank en la liberación del mismo año (Figura 2.15). Los resultados demostraron que el 98% de las clasificaciones fueron con un nivel de confianza mayor al 95% y una precisión del 98% (Wang *et al*., 2007).

Taxonomy	No. of sequences in corpus	No. of:					
		Domains	Phyla	Classes	Orders	Families	Genera
Bergey's	5,014	1	24	33	79	211	988
NCBI	23,095	1	24	31	82	209	1,187

Figura 2.15 Taxonomía de los diferentes esquemas estudiados por el clasificador RDP (Wang *et al*., 2007).

Además, evaluaron diferentes regiones tomadas de V2-V4 con tamaños de 50, 100, 200 y 400 bases. Los resultados para los dos esquemas (Figuras 2.16-2.17) demuestran que con la secuencia completa y con el segmento de 400 bases se tiene una precisión mayor al 88.7% a nivel de género. Para los segmentos de 200 bases, la precisión estaba por arriba del 92.1% en el nivel de familia, mientras que a nivel de género fue del 83.2%. Para los segmentos de 50 bases fue de sólo 94.1% a nivel de filo y disminuyó abruptamente a 51.5% a nivel de género (Wang *et al*., 2007).

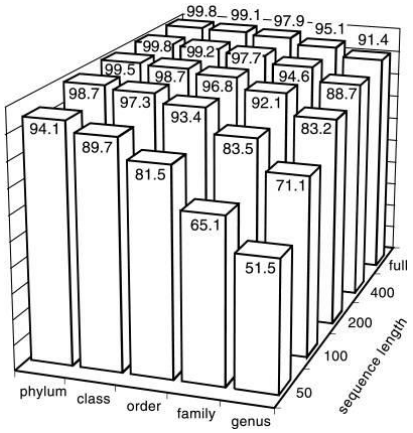


Figura 2.16 Precisión de las asignaciones taxonómicas con el esquema de Bergey (Wang *et al*., 2007).

Length of segment (bases)	% of segments accurately identified in:				
	Phylum	Class	Order	Family	Genus
Full	99.8	99.3	98.6	97.1	92.1
400	99.7	99.3	98.5	97.0	90.4
200	99.7	99.2	98.1	95.7	86.6
100	99.2	98.4	95.7	88.9	74.9
50	94.6	90.9	81.6	69.2	52.8

Figura 2.17 Precisión del clasificador en los diferentes segmentos con el esquema de NCBI (Wang *et al*., 2007).

Es importante mencionar que algunas herramientas bioinformáticas implementan algoritmos para los procesos de identificación o clasificación y algunas otras han desarrollado sus propios algoritmos que se adapten a problemas específicos. En la implementación intervienen a menudo la combinación de algoritmos para mejorar los resultados finales o el análisis pretratamiento. En consecuencia, Chen y colaboradores en 2013 realizaron un estudio para evaluar los diversos métodos de *clustering* en secuencias

del gen 16S usando como criterio los OTUs con diferentes herramientas bioinformáticas. En este estudio, se usaron datos reales con secuencias de la región V6 de una comunidad de 43 especies microbianas conocidas, el total de lecturas fue de 202340 con un tamaño de 57 a 145 bases. Además, se utilizaron datos simulados obtenidos con la herramienta 454Sim. El conjunto de datos simulados estaba constituido por 10 grupos como se muestra en la Figura 2.18. Las herramientas analizadas fueron de tres clases amplias: *clustering* jerárquico (Mothur, ESPRIT, ESPRIT-Tree, SLP, Muscle+Mothur), *clustering* heurístico (CD-HIT, Uclust, GramCluster, DNAClust) y *clustering* basado en modelos (CROP). Los resultados demuestran que la mayoría de las herramientas para calcular la diversidad microbiana sobrestiman o subestiman la cantidad de OTUs debido a que se establece un umbral inapropiado (Figura 2.19). En ese sentido, el cálculo del umbral depende de la herramienta utilizada además de la complejidad del conjunto de datos estudiados. Para datos con una baja complejidad y un tamaño de secuencia corta se puede establecer un umbral mayor en la definición de OTUs. Para este caso, SLP y CROP tuvieron un mejor desempeño en el cálculo de OTUs. Para el caso contrario, alta complejidad y longitud larga, los métodos de agrupamiento jerárquico necesitan un umbral menor. CD-HIT mostró un rendimiento parecido a los métodos de agrupación jerárquica (Chen *et al*., 2013).

	Num. species	Species length	Speices similarity arrange (%)	Total reads	Initial Abundance ratio (%)
Simclone10_1	10	226~252	70.00~93.00	69958	[6.25,3.89,5.06,11.31,7.74,29.77,8.04,12.50,5.95,9.52]
Simclone10_2	10	218~255	70.59~92.94	148374	[6.13,8.02,4.25,9.91,9.43,14.62,10.38,12.26,11.79,13.21]
Simclone15_1	15	59~71	50.68~83.04	63616	[3.57,1.78,1.78,5.35,3.57,17.86,3.57,7.14,1.79,1.79,5.36,3.57,10.71,14.29,17.86]
Simclone15_2	15	59~82	50.68~82.54	134092	[7.58,5.30,12.12, 18.94, 3.79, 8.71, 17.05, 3.79, 8.71,14.02]
Simclone20	20	64~261	25.58~94.19	115654	[2.87,1.97,2.12,4.54,4.84,9.53,6.05,6.05,3.02,2.42,3.18,5.45,3.18,7.57,8.62,6.81,5.45,4.84,8.62,2.87]
Simclone30	30	212~241	69.72~94.44	128308	[4.41,5.26,1.07,4.21,2.98,0.96,1.73,3.07,4.56,5.45,1.12,4.92,4.76,2.66,3.93,1.02,2.10,4.90,4.56,4.48,3.46,0.85,4.62,4.77,3.31,4.01,3.55,2.25,3.87,1.14,]
Simclone50	50	210~242	69.86~95.85	152373	[1.22,3.10,2.78,0.68,0.58,1.70,3.50,1.46,2.19,0.90,3.22,1.09,1.87,3.08,3.41,3.72,2.18,0.68,0.81,0.97,3.21,0.99,3.15,1.10,4.06,1.27,0.85,1.19,2.20,1.86,1.66,3.53,2.16,1.94,3.72,1.19,3.21,3.39,1.42,2.07,0.44,0.38,2.38,2.95,3.93,0.62,2.47,1.91,0.21,1.38]
Simclone100	100	212~276	67.03~95.93	248968	[0.14,0.94,1.55,0.72,1.44,0.89,0.62,0.91,1.29,0.97,0.37,1.76,0.70,0.43,1.02,0.48,0.32,1.12,1.25,0.81,1.04,0.36,0.94,0.42,0.57,0.60,0.32,1.44,0.45,0.89,1.15,0.81,1.02,1.47,1.38,0.70,0.65,1.01,0.46,1.62,0.41,0.40,1.48,1.20,1.68,0.86,0.72,0.80,0.38,1.46,1.80,1.02,1.25,1.77,1.29,1.09,1.54,0.80,1.74,1.62,1.06,1.70,1.28,1.56,1.13,1.50,0.38,1.58,0.57,1.01,1.13,0.68,1.27,1.76,0.99,0.75,0.37,1.39,0.34,0.77,1.25,0.44,0.94,0.48,1.48,1.23,1.24,0.97,1.58,1.17,1.41,1.58,0.51,0.91,0.58,1.47,0.42,1.16,0.91,0.68]
Simclone150	150	211~276	67.03~96.97	359153	[0.09,0.64,1.03,0.50,1.00,0.61,0.40,0.64,0.89,0.68,0.25,1.17,0.47,0.29,0.76,0.32,0.21,0.75,0.90,0.55,0.73,0.24,0.65,0.30,0.41,0.41,0.23,1.00,0.31,0.62,0.82,0.57,0.71,1.01,0.97,0.52,0.47,0.69,0.33,1.09,0.28,0.27,1.02,0.82,1.11,0.61,0.51,0.56,0.28,0.97,1.20,0.72,0.86,1.18,0.87,0.76,1.09,0.56,1.19,1.11,0.72,1.16,0.91,1.06,0.78,1.04,0.25,1.05,0.40,0.68,0.77,0.48,0.85,1.17,0.71,0.50,0.26,0.92,0.21,0.53,0.83,0.30,0.65,0.33,0.99,0.84,0.86,0.69,1.07,0.77,0.93,1.06,0.34,0.62,0.43,0.99,0.29,0.82,0.64,0.48,0.72,0.34,0.38,1.32,0.62,0.36,0.25,1.20,1.27,0.60,0.27,1.10,0.59,0.34,0.85,0.25,0.22,0.77,1.24,0.38,0.30,0.21,0.88,0.95,1.17,0.46,0.86,0.27,0.32,0.65,0.72,0.32,0.91,0.70,0.75,0.88,0.31,0.84,1.20,0.60,0.29,0.24,0.19,0.54,0.41,0.55,0.86,0.82,0.49,0.65]
Simclone200	200	190~276	64.26~96.97	484404	[0.06,0.47,0.77,0.37,0.75,0.46,0.31,0.48,0.65,0.50,0.19,0.87,0.35,0.22,0.54,0.24,0.16,0.57,0.65,0.42,0.54,0.19,0.49,0.22,0.29,0.30,0.18,0.74,0.22,0.46,0.61,0.43,0.52,0.75,0.71,0.38,0.35,0.51,0.24,0.80,0.19,0.20,0.76,0.61,0.83,0.45,0.38,0.40,0.19,0.73,0.88,0.52,0.63,0.86,0.64,0.55,0.82,0.40,0.86,0.82,0.53,0.86,0.65,0.78,0.58,0.76,0.19,0.79,0.30,0.51,0.58,0.35,0.63,0.87,0.53,0.38,0.20,0.70,0.16,0.41,0.60,0.23,0.47,0.24,0.75,0.62,0.64,0.50,0.80,0.59,0.69,0.77,0.25,0.46,0.32,0.71,0.21,0.59,0.47,0.35,0.53,0.26,0.28,0.98,0.47,0.26,0.18,0.88,0.93,0.44,0.18,0.81,0.42,0.25,0.61,0.18,0.17,0.59,0.92,0.28,0.23,0.14,0.66,0.70,0.87,0.33,0.63,0.19,0.24,0.49,0.53,0.23,0.66,0.52,0.56,0.63,0.25,0.62,0.90,0.45,0.21,0.16,0.14,0.39,0.31,0.41,0.62,0.60,0.37,0.48,0.45,0.77,0.46,0.43,0.56,0.56,0.16,0.27,0.47,0.83,0.34,0.65,0.16,0.14,0.62,0.68,0.43,0.51,0.46,0.76,0.62,0.31,0.91,0.48,0.40,0.41,0.86,0.56,0.51,0.81,0.28,0.27,0.37,0.14,0.89,0.18,0.15,0.84,0.76,0.82,0.82,0.18,0.33,0.63,0.92,0.60,0.54,0.75,0.62,0.44]

doi:10.1371/journal.pone.0070837.t001

Figura 2.18 Datos simulados generados con la herramienta 454Sim (Chen *et al*., 2013).

	Clone43				Simclone15_1			
	Expected OTUs	Inferred* OTUs(2%)	inferred OTUs(3%)	inferred OTUs(4%)	Expected OTUs	inferred OTUs(2%)	inferred OTUs(3%)	inferred OTUs(4%)
Mothur	43	1882	720	369	15	63	41	20
Muscle+Mothur		2478	1418	784		117	89	54
ESPRIT		4474	4397	1733		131	131	55
ESPRIT-Tree		2301	1096	279		96	29	16
SLP		286	245	227		17	17	15
Uclust		2177	1883	597		80	75	51
CD-HIT		1473	1464	481		50	49	32
DNAClust		3768	3658	1103		239	225	53
GramCluster		2119	2071	2071		70	70	70
CROP		339	133	62		21	15	15

*: all the listed numbers of OTU are the average numbers over xx simulations.
doi:10.1371/journal.pone.0070837.t002

Figura 2.19 Número de OTUs calculados con diferentes umbrales (Chen *et al*., 2013).

CAPÍTULO 3. OBJETIVOS

3.1 Justificación

Las herramientas bioinformáticas son utilizadas en el análisis de datos biológicos como procesos de alineamientos múltiples de secuencias de ADN y secuenciación masiva. Sin embargo, el tiempo de análisis depende del tamaño de la secuencia utilizada y las características de los algoritmos que utilicen, no obstante, el uso y aplicación de la bioinformática en el análisis y depuración de bases de datos ha sido poco aplicada, por lo tanto, la implementación de un algoritmo para el análisis de metadatos biológicos optimizados facilitará la clasificación e identificación de bacterias y *archaeas*.

3.2 Hipótesis

El diseño de una base de datos cercana a la región 804-1392 del gen 16S, permitirá analizar y filtrar grandes volúmenes de datos para la identificación y clasificación de bacterias mediante la implementación de un clasificador bayesiano de análisis masivo de secuencias 16S.

3.3 Objetivos

3.3.1 Objetivo General

Desarrollar un clasificador bayesiano para el análisis de datos de secuenciación masiva del gen 16s, en la región 804-1392 para identificar y clasificar bacterias.

3.3.2 Objetivos Específicos

- Diseñar una base de datos 16s en la región 804-1392 para agrupar y clasificar secuencias ribosomales de bacterias.
- Agrupar las secuencias 16s de la base datos diseñada en la región 804-1392 para clasificar secuencias ribosomales.
- Desarrollar un clasificador bayesiano para el análisis de secuencias 16s en la región 804-1392.
- Implementar el flujo de trabajo para análisis masivo de secuencias 16s en la región 804-1392.

CAPÍTULO 4. METODOLOGÍA

Como parte de este trabajo de investigación, se desarrolló una serie de pasos para obtener una base de datos organizada y consistente. Esto con el objetivo de que el clasificador bayesiano desarrollado realice los procesos de clasificación e identificación con mayor precisión.

La metodología está compuesta por dos etapas, la primera etapa abarca el análisis y agrupación de tres bases de datos RDP, SILVA y Greengenes, además de la base de datos control EzTaxon (Kim *et al*., 2012). La segunda etapa tuvo como objetivo el desarrollo del clasificador bayesiano, el cual hace uso de la base de datos obtenida en la primera etapa para analizar los datos de los archivos de secuenciación masiva proporcionados por los usuarios (véase Figura 4.1).

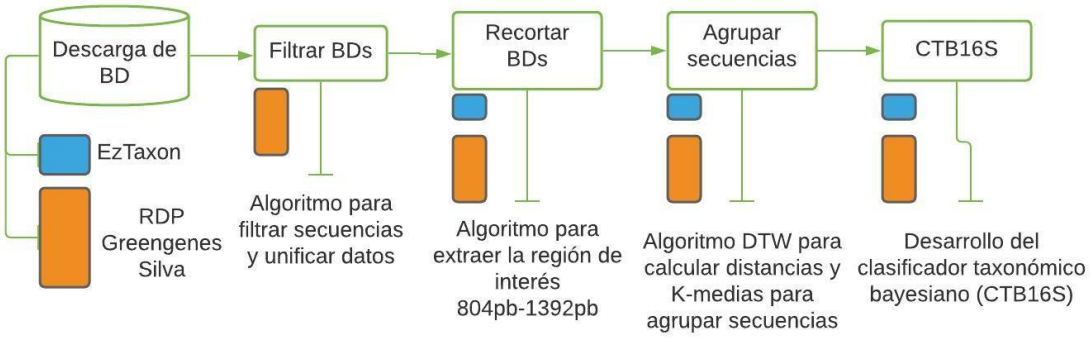


Figura 4.1 Metodología de trabajo para el análisis masivo de secuencias ribosomales.

4.1 Materiales

4.1.1 Fuente de Datos

Para crear un conjunto amplio de secuencias ribosomales 16S y además confiable, se eligieron fuentes de datos previamente analizados como son RDP, Greengenes, SILVA y EzTaxon por diversos estudios (véase el apartado de Marco Teórico, Antecedentes).

Las bases de datos se descargaron desde sus páginas oficiales, SILVA ref de bacterias, *archaeas* y no clasificadas disponible en http://www.arb-silva.de, Greengenes desde http://greengenes.secondgenome.com, RDP de http://rdp.cme.msu.edu y EzTaxon del sitio http://www.ezbiocloud.net, este proceso de descarga se realizó en el mes de septiembre del 2016. En la Tabla 4.1 se muestran las secuencias totales descargadas por cada repositorio.

Tabla 4-1 Número de secuencias descargadas por base de datos.

Base de datos	Número de secuencias descargadas
RDP	2948785
SILVA	4620004
Greengenes	406997
EzTaxon	63240

La taxonomía adoptada para los posteriores procesos de identificación y clasificación fue la de EzTaxon por ser una base de datos "curada" manualmente, además por los cálculos y asignación de filotipos para las secuencias bacterianas y *archaeas* de muestras ambientales (Kim *et al*., 2012).

4.1.2 Hardware y software utilizado para el análisis de secuencias

Para realizar los procesos de análisis de secuencias ribosomales se utilizó una computadora portátil modelo DELL Latitude E6430, procesador Intel® Core™ i5-3320M con frecuencia básica a 2.6 GHz y máxima a 3.3 GHz, memoria RAM de 12GB y como sistema operativo Ubuntu 16.04.2 LTS de 64 bits.

El software para la filtración de datos que se usó fue Bourne again shell (Bash por sus siglas en inglés) y para el proceso de *clustering* y clasificación se utilizó Java en la versión 1.8 con la IDE NetBeans 8.2.

Los gráficos de los distintos estudios elaborados se realizaron en GraphPad Prism 6.0.1 en la plataforma Windows 7, mientras que los cladogramas circulares para la validación de los grupos formados se obtuvieron con la herramienta CLC Sequence Viewer 7 versión 7.8.1 en la plataforma Ubuntu 16.04.2 LTS.

4.2 Métodos

4.2.1 Primera etapa

Para obtener todas aquellas secuencias únicas de SILVA, RDP y Greengenes, así como de la base de datos control EzTaxon, se realizó el filtrado mediante un *script* en *Bash*, en el cual las secuencias se compararon como "palabras", es decir, las secuencias que tuvieran al menos una base diferente se consideraban como diferentes y por consiguiente,

en este paso no se alinearon las secuencias descargadas, esto con el fin de obtener mayor diversidad de secuencias en el conjunto de datos para análisis posteriores.

El siguiente paso fue el recorte de los datos filtrados, en este paso se alinearon el juego de oligonucleótidos diseñados por Lundberg y colaboradores en 2012 (oligonucleótido sentido: AGATTAGATACCCDRGTAGT y antisentido: ACGGGCGGTGTGTRC) con las secuencias filtradas. Las secuencias que reconocieron los dos cebadores, se les aplicó el *script* en *BASH* para obtener las secuencias únicas de la región de interés (804pb-1392pb).

Se analizó la variabilidad de la base de datos control recortada a lo largo de las 600 bases aproximadamente de longitud de sus secuencias, con el objetivo de localizar las regiones conservadas y variables dentro la de región 804pb-1392pb (Vinje *et al*., 2014). Para ello, se dividió las secuencias recortadas en diez segmentos, cada segmento de sesenta bases cada uno aproximadamente. Se calculó la entropía de cada uno de los segmentos con la fórmula $H_k = -\sum_{i=1}^{i=5} p_i \log(p_i)$, donde *H* es la entropía y *k* es el segmento evaluado. P_1, P_2, P_3, P_4 y P_5 son los estados posibles de las bases A=1, T=2, G=3, C=4, N=5.

4.2.2 Agrupar secuencias recortadas

La agrupación de secuencias se realizó mediante el algoritmo de *clustering* K-medias. Este algoritmo requiere del cálculo de distancia entre los diferentes centroides, en este proceso se utilizó el algoritmo Alineamiento Temporal Dinámico (DTW por sus siglas en inglés). Se agrupan las secuencias más cercanas a los centroides, DTW busca el mejor patrón para cada secuencia entrada respecto a cada centroide.

Se comenzó con 3099 géneros que se encontraron en la unificación de las secuencias obtenidas de los diferentes repositorios antes mencionados. Para la selección de centroides, se eligió aquella secuencia que mantuvo un patrón alto, en cuestión de similitud, con las de su mismo género, esto calculado por DTW. Posteriormente, la formación de grupos finalizó en dos ciclos, en cada uno de ellos se recalcularon los centroides.

La validación de los grupos formados se realizó mediante la construcción de cladogramas con el software CLC Sequence Viewer. Se realizaron diversas pruebas con grupos del mismo filo, grupos con diferente filo y secuencias ajenas a los grupos formados. Además, se crearon grupos con la herramienta Cd-hit (W. Li & Godzik, 2006) con diferentes porcentajes de identidad (90% hasta 97%). Estos *clusters* se compararon con los grupos antes formados mediante DTW y K-medias.

4.2.3 Clasificador CTB16S

El clasificador taxonómico bayesiano 16s (CTB16S) está basado en la clasificación de texto mediante Bayes ingenuo o simple. La idea principal de este método, es calcular las probabilidades de cada una de las palabras en el conjunto de texto objetivo, este cálculo toma como medida *n* veces las palabras superpuestas en el conjunto de texto. Las diferentes muestras que componen el conjunto objetivo se le llaman clases (Y. H. Li, 1998). Partiendo de este principio, los diversos grupos formados en el proceso de *clustering* forman las clases del conjunto texto objetivo. Es decir, se tienen 3099 clases y, por otra parte, el grupo de palabras para llevar a cabo el cálculo de las probabilidades de cada uno de ellas en las diversas clases fueron tomadas de la siguiente forma: las palabras posibles de ocho letras formadas por A, T, C, G.

Se calcularon dos probabilidades para las 65536 posibles palabras, una probabilidad global y una por grupo o clase.

Siendo *W= {w1, w2…wi}* el conjunto de palabras a evaluar y *N* las secuencias de la base de datos, se calculó las *n(wi)* secuencias que presentaban la palabra *wi* en el conjunto *N*. La probabilidad global se calculó de la siguiente manera: $PG = [n(wi) + 0.5]/(N + 1)$

Para el conjunto de clases *C*, las cuales contienen *S* secuencias; se calculó las *s(wi)* secuencias que contienen la palabra *wi*. La fórmula para obtener la probabilidad de cada palabra por clase es la siguiente: $PC(wi|C) = [s(wi) + PGi]/(S + 1)$

La probabilidad conjunta de que una secuencia que contiene *wi* palabras forme parte de una clase se calculó con la fórmula $P(X|C) = \Pi P(vi|C)$

Donde *vi* forma parte del conjunto *V* de palabras que contiene la secuencia a evaluar.

El fundamento de tomar ocho como tamaño de palabra está basado en el estudio del clasificador de RDP (Wang *et al*., 2007). En este estudio se evaluaron diferentes tamaños de palabra (7, 8 y 9 respectivamente) y la mejor que resultó fue 8 en su caso. Para nuestro estudio, tomamos el tamaño de palabra de ocho solapando la última base.

Para la clasificación por el método bayesiano simple de una nueva secuencia *X*, se realizó el cálculo de la mejor probabilidad según la probabilidad de palabra por grupo y global con la fórmula $P(C|X) = P(X|C) \cdot P(C) / P(G)$

Donde *P(X|C)* es la probabilidad de que una secuencia nueva forme parte de una clase, *P(C)* es la probabilidad previa de que la secuencia *X* forme parte de *C* y *P(G)* es la probabilidad de que *X* forme parte de cualquier clase.

La asignación bayesiana de clase es dada según la probabilidad más alta. Para calcular esta asignación, se construye la serie de palabras únicas que contiene cada

secuencia nueva a clasificar. Este mismo proceso se realiza para cada secuencia previamente clasificada de la base de datos de referencia. La probabilidad más alta se calcula según la fórmula de la Figura 4.6 para las coincidencias entre las dos series de palabras (nueva y referencia). Además, es importante el orden de estas palabras en ambas secuencias comparadas. Aquellas palabras que tienen la misma posición, tanto en la referencia como la secuencia por clasificar, se toman como similares y se realiza el cálculo de la probabilidad antes mencionada.

CAPÍTULO 5. RESULTADOS

5.1 Descarga de bases de datos

Se obtuvieron más de 8 millones de secuencias (8039026 secuencias) y la base de datos que aportó el mayor porcentaje fue SILVA con cerca del 58%. El tamaño total en GB de los datos descargados fue de 8.79 GB (Figura 5.1).

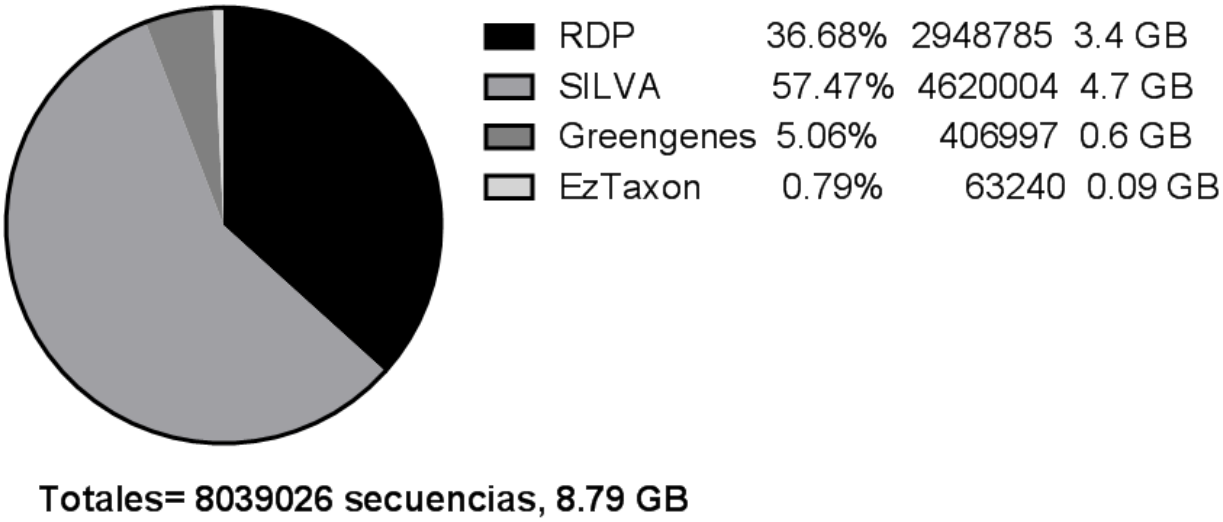


Figura 5.1 Secuencias descargadas de RDP, SILVA, Greengenes y EzTaxon y tamaño que ocupan en GB.

5.2 Filtrado de las bases de datos descargada

Las secuencias descargadas de RDP, SILVA y Greengenes fueron analizadas por un *script* en *Bash*, esto con el objetivo de excluir secuencias con al menos una base diferente comparadas como “palabras”. Este proceso no se aplicó a la base de datos control ya que cuenta con secuencias únicas, previamente analizadas por los creadores de la base de datos EzTaxon (Kim *et al*., 2012).

El 44% de las secuencias totales de la unión de RDP, Greengenes y SILVA fueron filtradas como únicas, mientras el resto fueron secuencias duplicadas. La Figura 5.2 muestra el total de secuencias únicas (3736415) y el tamaño en GB (3.8 GB) y las secuencias redundantes encontradas (4239371) y su tamaño (4.9 GB).

La riqueza de secuencias que se obtuvo con respecto a las secuencias únicas es mayor en comparación con la base de datos control. Con esto, los análisis posteriores son de suma importancia para evaluar la calidad de las secuencias únicas obtenidas.

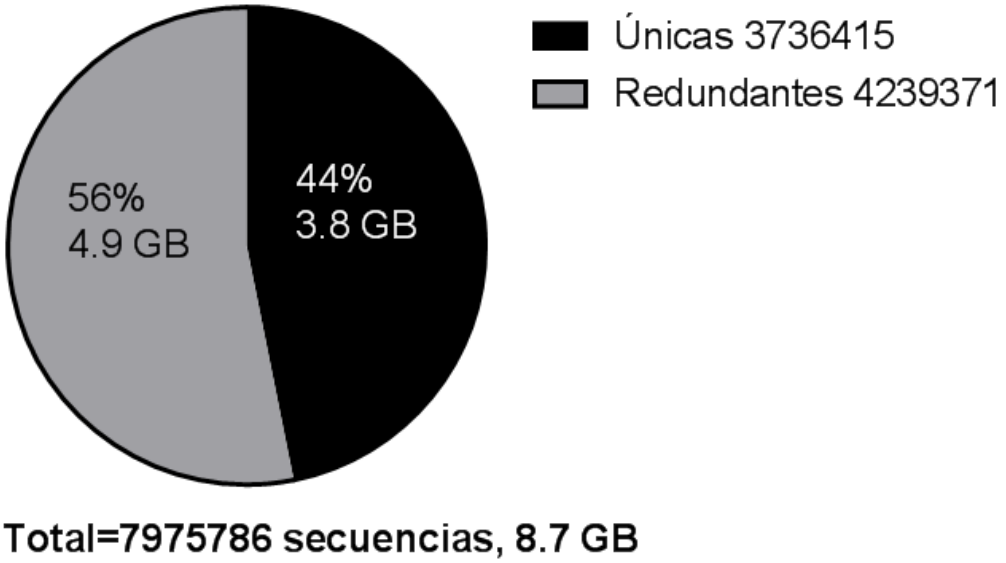


Figura 5.2 Secuencias únicas y redundantes de RDP, SILVA y Greengenes.

5.3 Pruebas de la base de datos filtrada

Se realizaron pruebas a la base de datos filtrada con Sequenceserver (Priyam *et al*., 2015) y BLAST con la base de datos 16S de NCBI. Para las pruebas se seleccionaron secuencias de la base de datos filtrada de manera aleatoria. La primera prueba se realizó con *Salmonella Bongori* cepa NCTC 12419. Se obtuvieron resultados similares a nivel de género por parte de las dos herramientas utilizadas (Figura 5.3 con Blast y Figura 5.4 con Sequenceserver) con los primeros alineamientos.

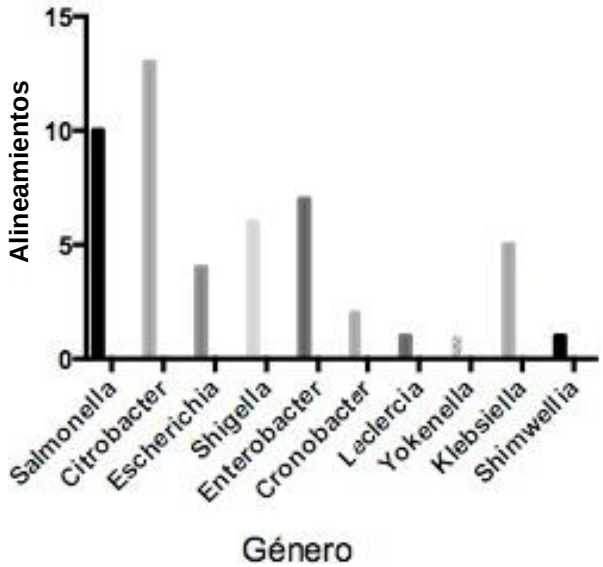


Figura 5.3 *Salmonella Bongori* cepa NCTC 12419 con BLAST.

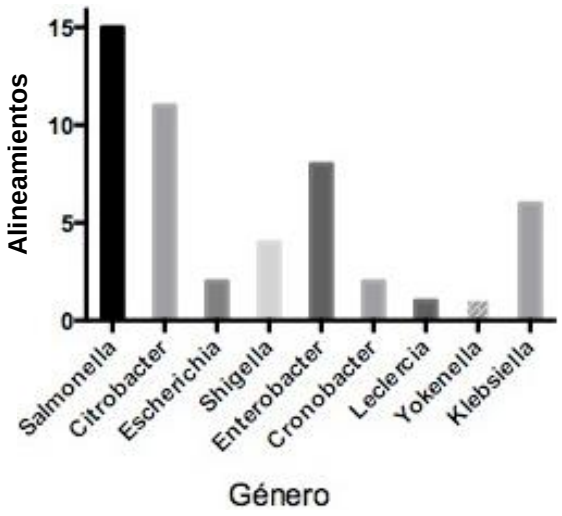


Figura 5.4 *Salmonella Bongori* cepa NCTC 12419 con Sequenceserver.

Además, se identificó la secuencia de *Salmonella Bongori* cepa NCTC 12419 con un porcentaje de identidad del 100% con la base de datos unificada (Figura 5.5).

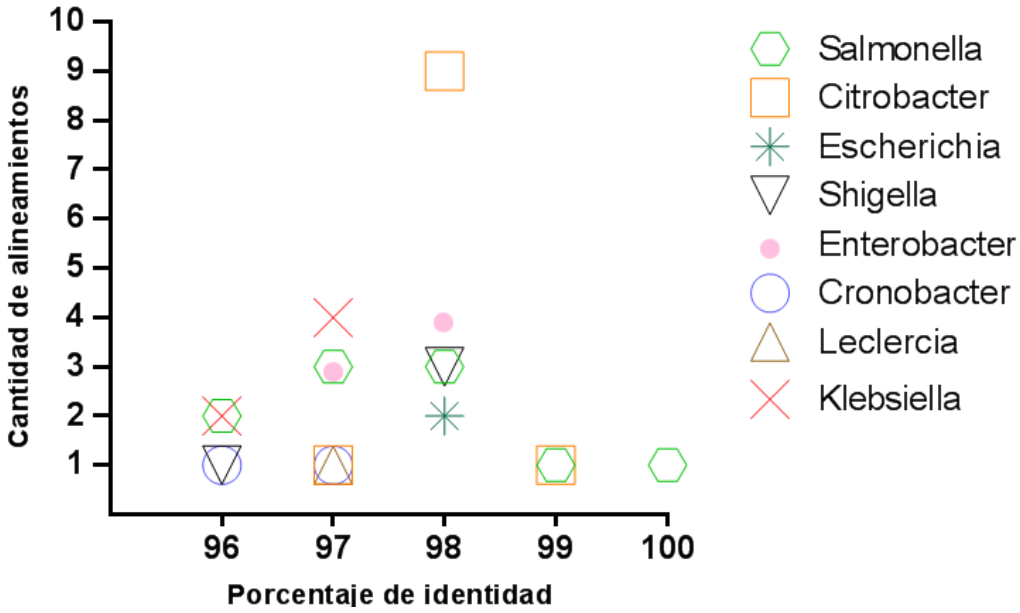


Figura 5.5 Identificación de *Salmonella Bongori* cepa NCTC 12419.

La segunda prueba se realizó con *Pseudomonas syringae pv. phaseolicola* cepa Psp-1, lo interesante de esta prueba fue la identificación al 100% con la base de datos unificada (Figura 5.6) mientras que con la herramienta BLAST tuvo un porcentaje de identidad del 99% (Figura 5.7).

Number	Sequences producing significant alignments	Total score	E value	Length
1.	JX876900.1.1445	2607.15	0.00	1445
2.	137936	2542.23	0.00	1540
3.	137935	2542.23	0.00	1540
4.	137934	2542.23	0.00	1540
5.	137933	2542.23	0.00	1540
6.	137932	2542.23	0.00	1540
7.	137383	2542.23	0.00	1540
8.	137382	2542.23	0.00	1540

JX876900.1.1445 1 / 50

Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas;Pseudomonas syringae pv. phaseolicola

Hit length: 1445 Select | Sequence | FASTA

1. Score	E value	Identities	Gaps	Strand
2607.15 (2890)	$0.00 \times 10^{+00}$	1445/1445 (100.00)	0/1445 (0.00)	+/+

Figura 5.6 Identificación de *Pseudomonas syringae pv. Phaseolicola* cepa Psp-1 con Sequenceserver.

Pseudomonas syringae pv. phaseolicola 1448A strain 1448A; BAA-978 16S ribosomal RNA, complete sequence

Sequence ID: ref|NR_074598.1| Length: 1518 Number of Matches: 1

Range 1: 23 to 1466 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
2542 bits(2818)	0.0	1429/1436(99%)	6/1436(0%)	Plus/Plus

Figura 5.7 Mejor resultado en la identificación de *Pseudomonas syringae pv. Phaseolicola* cepa Psp-1 con BLAST.

5.4 Evaluación de cebadores 804pb-1392pb

Los oligonucleótidos descritos por Lundberg *et al*., 2012 se evaluaron a través del cálculo del porcentaje de identidad contra la base de datos filtrada. Para los seis cebadores sentido se probaron con dos porcentajes de identidad, el primero con el 100% de identidad (20pb) y 85% (17pb). Los resultados manifiestan que muy pocas secuencias son reconocidas para el porcentaje de 100% (Figura 5.8). Para el porcentaje de identidad de 85%, gran parte de las secuencias fueron reconocidas, por lo que este porcentaje se utilizó para los distintos oligonucleótidos sentido en el proceso de recorte (Figura 5.9).

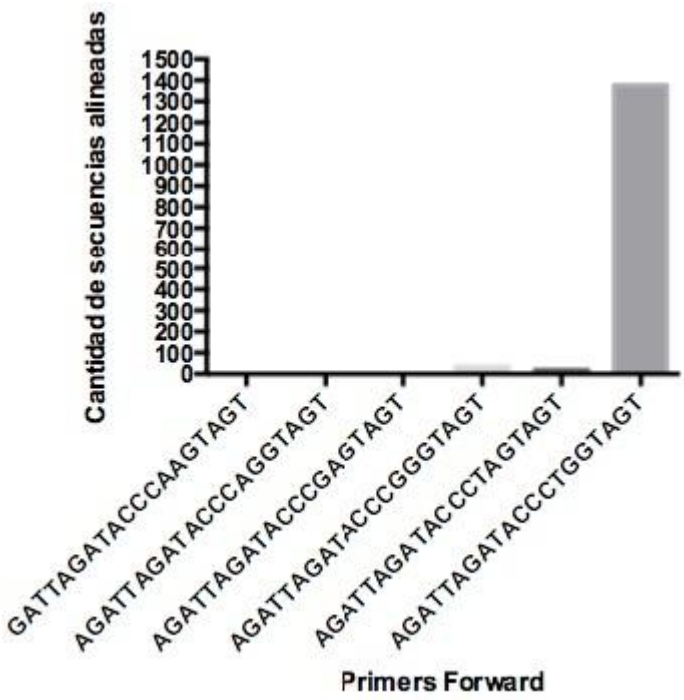


Figura 5.8 Cebadores alineados con un porcentaje de identidad 100%.

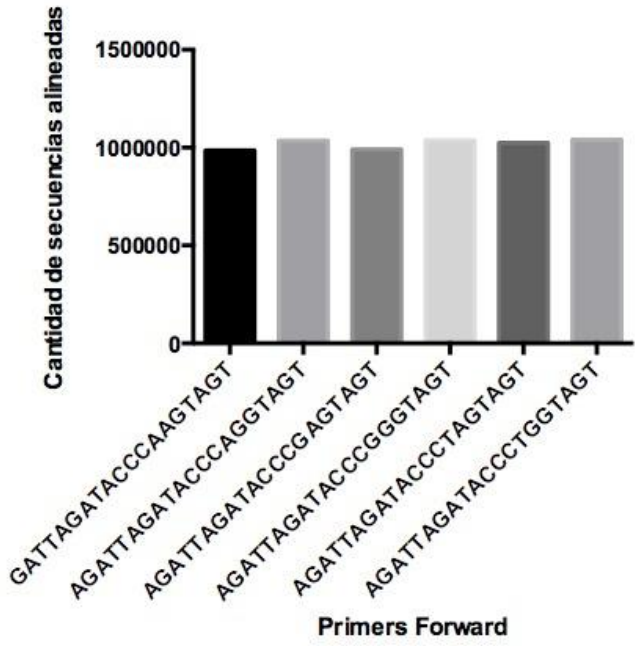


Figura 5.9 Alineamiento al 85% de los cebadores sentido.

Para el par de oligonucleótidos antisentido, se utilizaron dos porcentajes de identidad, uno del 100% y otro del 93% (15pb y 14pb respectivamente). El porcentaje de identidad que

tuvo mejores resultados fue el de 93%, por lo que se concluyó que este porcentaje fuera elegido para ser utilizado en el recorte de la región de interés (Figura 5.10).

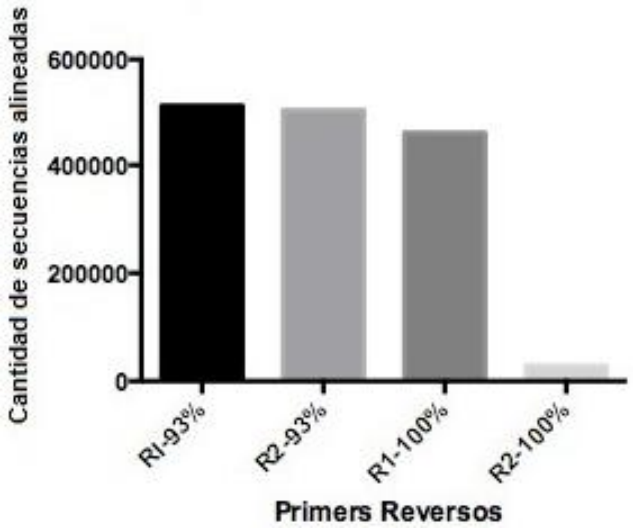


Figura 5.10 Alineamiento del par de *primers reverse* con los dos tipos de porcentajes.

5.5 Extracción de la región de interés

Con la finalidad de obtener la región 804pb-1392pb, se alinearon los oligonucleótidos diseñados por Lundberg *et al*., 2012 con la base de datos unificada obtenida del proceso de filtración y la base de datos control.

De la base de datos unificada se obtuvieron 562711 (13%) secuencias recortadas de un total de 3736415 secuencias como se muestra en la Figura 5.11. Mientras que la base de datos control fue reconocida por los cebadores al 100%, sin embargo, el 24.9% de las secuencias tienen una nomenclatura difícil de identificar, es decir, nomenclatura con códigos de acceso a niveles de filo hasta especie.

Figura 5.11 Secuencias recortadas de la base de datos unificada.

En la Figura 5.12 se muestra un ejemplo de una secuencia alineada en EzBioCloud con nomenclatura no validada con un porcentaje de identidad del 100%, en la siguiente figura se muestran los resultados con nombres válidos con la herramienta que ofrece el mismo sitio (Figura 5.13), además se realizó la misma prueba con la herramienta Blast de NCBI (5.14) y los porcentajes de identidad son muy bajos (80% en ambas herramientas).

Similarity	Diff/Total nt	Hit taxonomy	Completeness
100.00	0/1458	Bacteria;TM6;TM6_c;Babela_o;FJ264771_f;FJ264771_g	100.0
96.77	47/1457	Bacteria;TM6;TM6_c;Babela_o;FJ264771_f;FJ264771_g	100.0
96.16	56/1458	Bacteria;TM6;TM6_c;Babela_o;FJ264771_f;FJ264771_g	100.0
94.99	73/1458	Bacteria;TM6;TM6_c;Babela_o;FJ264771_f;FJ264771_g	100.0
93.90	89/1458	Bacteria;TM6;TM6_c;Babela_o;FJ264771_f;FJ264771_g	100.0

Figura 5.12 Taxonomía de la base de datos EzTaxon con nombres no validados.

Similarity	Diff/Total nt	Hit taxonomy	Completeness
80.55	274/1409	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobulbaceae_o;Desulfobulbaceae;Desulfobulbus	97.1
80.04	275/1378	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Phenylobacterium	98.4
79.85	293/1454	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobulbaceae_o;Desulfobulbaceae;Desulfobulbus	100.0

Figura 5.13 Taxonomía de la base de datos EzTaxon con nombres validados.

	Max score	Total score	Query cover	E value	Ident	Accession
Desulfosalsimonas propionicica strain PropA 16S ribosomal RNA gene, partial sequence	1243	1243	98%	0.0	79%	NR_115678.1
Desulfonatronobacter acetoxydans strain APT3 16S ribosomal RNA, partial sequence	1232	1232	100%	0.0	79%	NR_145936.1
Desulfonatronobacter acidivorans strain APT2 16S ribosomal RNA gene, partial sequence	1220	1220	100%	0.0	78%	NR_117486.1
Desulfobulbus mediterraneus strain 86FS1 16S ribosomal RNA gene, partial sequence	1211	1211	95%	0.0	79%	NR_025150.1
Desulfobulbus alkaliphilus strain APS1 16S ribosomal RNA gene, partial sequence	1178	1178	100%	0.0	78%	NR_117882.1
Desulfatiglans parachlorophenolica strain DS 16S ribosomal RNA gene, partial sequence	1162	1162	100%	0.0	78%	NR_126176.1

Figura 5.14 Taxonomía de NCBI *versus* secuencia con anotación no válida de EzTaxon.

Es por ello, se decidió filtrar aquellas secuencias que presentaban este tipo de nomenclatura hasta nivel de género, consiguiendo un total de 47524 secuencias (75.1%) de un total de 63240 secuencias (5.15).

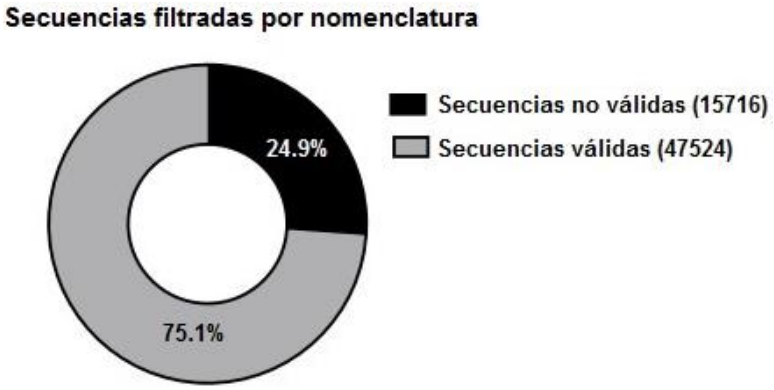


Figura 5.15 Secuencias filtradas por nomenclatura.

5.6 Agrupación de la base de datos control

Se obtuvieron 3099 grupos a nivel de género, a partir de la base de datos control (EzTaxon). La Figura 5.16 muestra los grupos formados desde el punto de vista de taxonomía procariota.

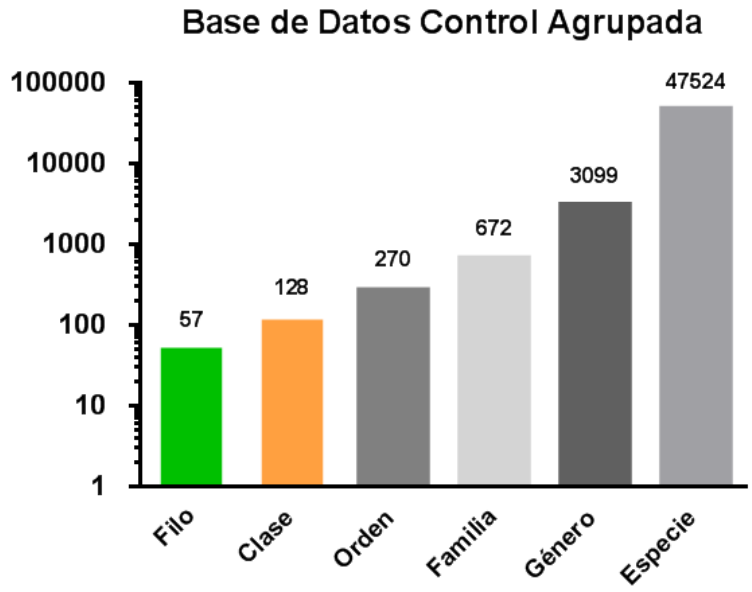


Figura 5.16 Taxonomía de la base de datos agrupada.

A partir del agrupamiento de las secuencias, se analizó la composición de la base de datos con el fin de comparar con el estudio de Kim *et al*., 2012. Los resultados muestran que el filo con mayor abundancia relativa con relación al dominio de bacteria fue *Proteobacteria* y *Euryarchaeota* como el filo con mayor presencia con respecto al dominio *Archaea* (Figura 5.17). Además, se analizaron los filos de ambos dominios con menor presencia como se muestra en la Figura 5.18. El porcentaje relativo de los filos con baja abundancia son uniformes en el dominio de bacterias y por el contrario en el dominio de *Archaea*, el filo con mayor presencia fue *Thaumarchaeota*. En ambos dominios existen más filos con baja abundancia, en el dominio bacteriano es muy notable la diversidad a nivel de filo.

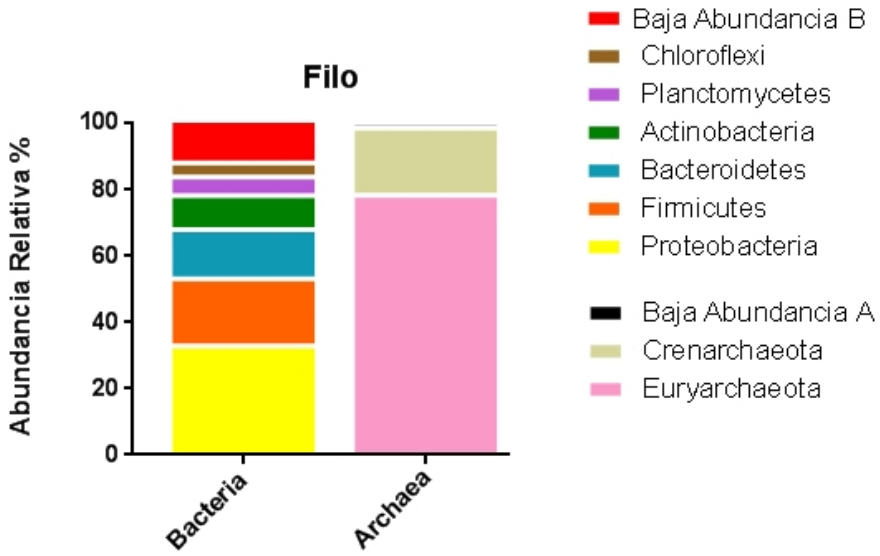


Figura 5.17 Composición de la base de datos control agrupada.

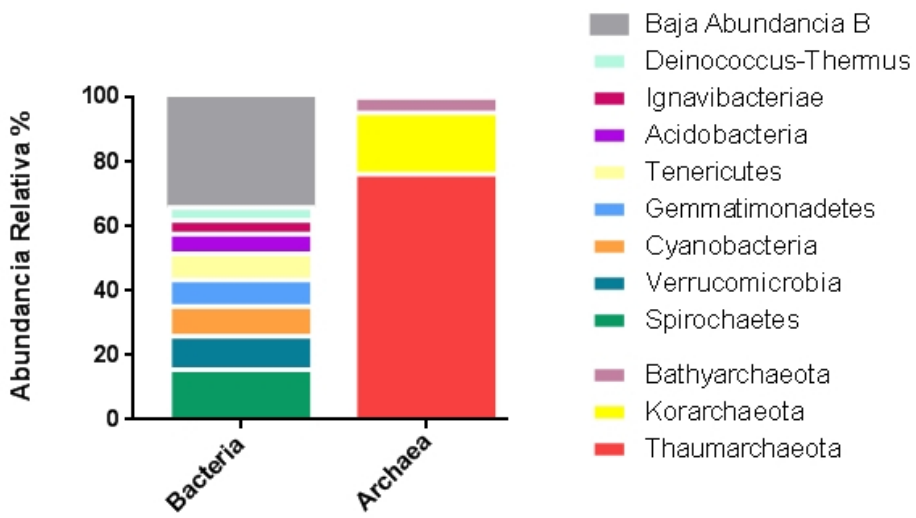


Figura 5.18 Filos con menor presencia en la composición de la base de datos agrupada.

Se analizó la correlación de los grupos formados por el método DTW-K-medias con los géneros reportados en la base de datos control (Kim *et al*., 2012) (Figura 5.19). Para esto, se calculó el coeficiente de correlación de Pearson (**r**) obteniendo una correlación positiva **r=0.9026** y una **R^2=0.8146**.

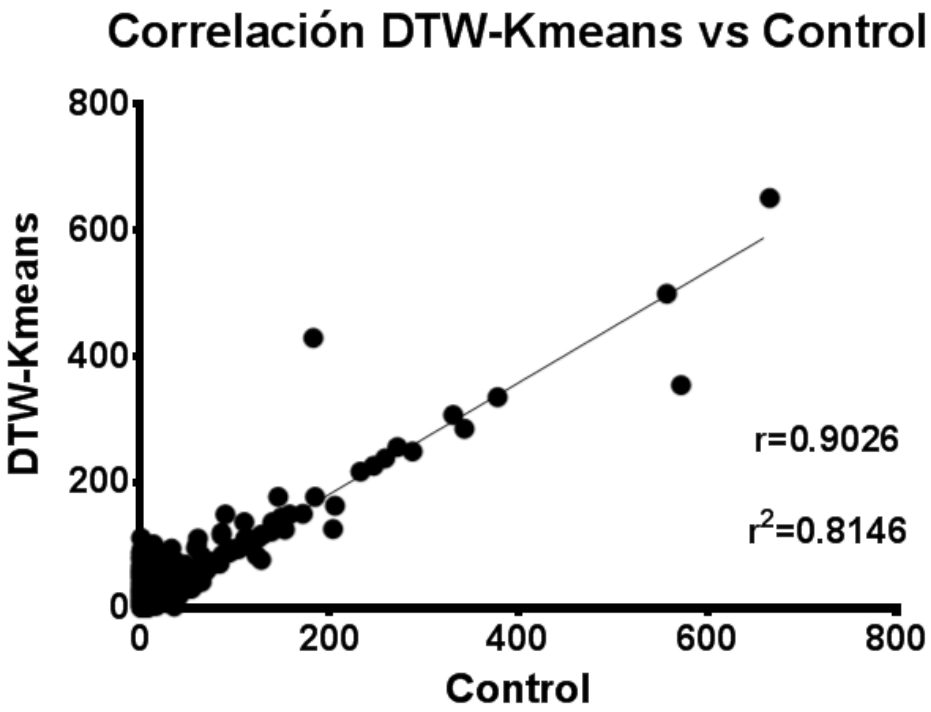


Figura 5.19 Correlación positiva con base en el cálculo del coeficiente de correlación (r) de Pearson entre los grupos formados por el método DTW-K-medias y los géneros reportados en la base de datos de Kim *et al*., 2012.

Para comprobar que las secuencias que forman parte de cada uno de los grupos obtenidos mediante el algoritmo de *clustering* K-medias son afines, se calculó la entropía de cada una de ellas por grupo. Para ello, cada secuencia se dividió en 10 segmentos, cada segmento de 60 bases. En la Figura 5.20 se muestran los resultados del cálculo de entropía con sus respectivas desviaciones estándar de tres de los grupos formados.

Los resultados muestran una afinidad de las secuencias agrupadas en la mayoría de sus segmentos, esto con base en la baja dispersión de información que muestra la desviación estándar en al menos la mitad de los segmentos evaluados. También se pueden observar segmentos con una alta conservación, pero con una cierta variabilidad que es notorio debido a una desviación estándar alta. En los tres grupos mostrados, el comportamiento a lo largo de los diez segmentos es similar, a pesar de que el grupo 57 no pertenece al mismo filo (*Firmicutes*) que los dos grupos restantes (*Proteobacteria*). Además, los segmentos S3 y S7 de los 3 grupos analizados presentan un alto grado de conservación, mientras que S4 y S8 manifiestan una mayor variabilidad.

La localización de los segmentos con mayor conservación, así como variabilidad, aportan información para el siguiente proceso, la clasificación de secuencias por Bayes simple.

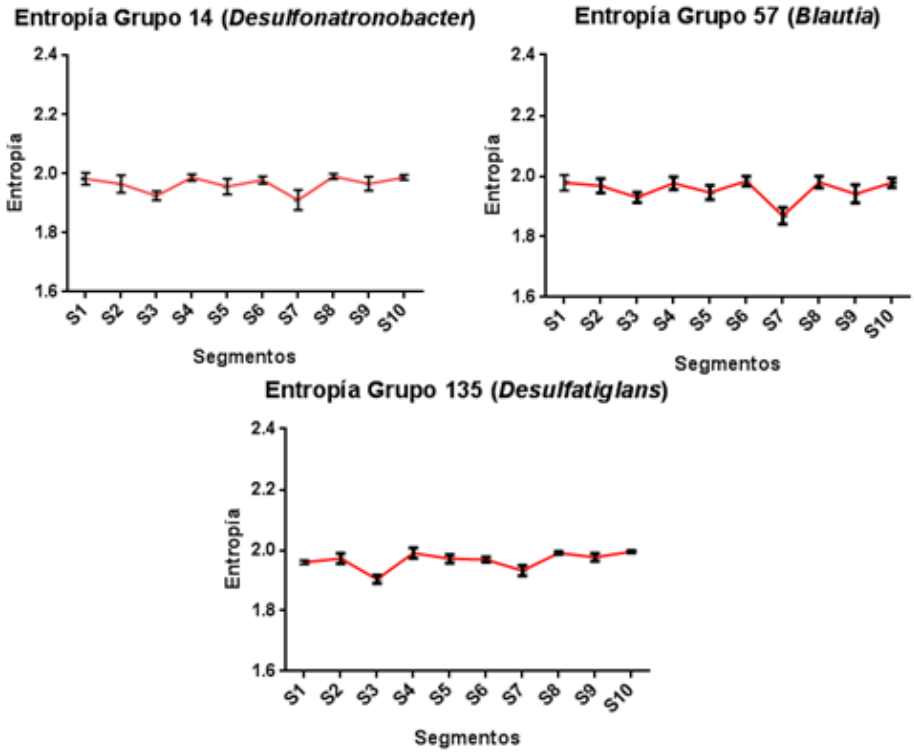


Figura 5.20 Entropía calculada de tres grupos formados.

5.7 Validación de los grupos formados

Se evaluaron los grupos formados mediante la construcción de cladogramas circulares. Para ello, se realizaron varias pruebas para valorar la calidad del proceso de agrupación de secuencias. La construcción del cladograma circular se realizó mediante la herramienta CLC Sequence Viewer 7. Se eligió un costo de hueco de 10 y por extensión de 4, y alineamiento muy exacto, esto para el proceso de alineamiento. Para la creación del cladograma circular se utilizó el método UPGMA para construir el árbol, la medida de distancia entre nucleótidos se utilizó Jukes-Cantor (Jukes & Cantor, 1969) y con 100 replicaciones.

Para las cuatro pruebas se utilizaron los grupos y su taxonomía proporcionada por la base de datos control (Tabla 5.1).

La primera prueba consistió en el análisis de dos grupos del mismo filo (*Proteobacteria*), el grupo bacteriano número 5 que tiene como género asignado *Syntrophus* y el grupo número 221 con el género asignado *Gallionella*. El objetivo de esta prueba fue estudiar el comportamiento de secuencias del mismo filo y si era posible observar la separación de los grupos en el cladograma, ya que pertenecen al mismo filo, pero a diferentes clases.

Tabla 5-1 Taxonomía de las diversas secuencias utilizadas en las pruebas de validación de grupos formados.

Grupo	Dominio	Filo	Clase	Orden	Familia	Género
3099	*Bacteria*	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Macellibacteroides*
1365	*Bacteria*	*Armatimonadetes*	*Chthonomonadetes*	*Chthonomonadales*	*Chthonomonadaceae*	*Chthonomonas*
808	*Bacteria*	*Firmicutes*	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Faecalibacterium*
1569	*Archaea*	*Euryarchaeota*	*Halobacteria*	*Halobacteriales*	*Halobacteriaceae*	*Halobacterium*
272	*Bacteria*	*Actinobacteria*	*Rubrobacteria*	*Rubrobacterales*	*Rubrobacteraceae*	*Rubrobacter*
5	*Bacteria*	*Proteobacteria*	*Deltaproteobacteria*	*Syntrophobacterales*	*Syntrophaceae*	*Syntrophus*
221	*Bacteria*	*Proteobacteria*	*Betaproteobacteria*	*Gallionellales*	*Gallionellaceae*	*Gallionella*
88	*Bacteria*	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	*Acholeplasmataceae*	*Acholeplasma*

En la Figura 5.21, se muestra el cladograma circular de la primera prueba, los grupos analizados muestran una relación cercana, es decir, la distancia entre las ramas de cada grupo que conectan hacia la raíz es pequeña, sin embargo, la separación es notable en el cladograma como se esperaba.

La etiqueta de cada hoja está formada por dos identificadores separados por un punto, la primera cifra identifica el número de secuencia y la siguiente cifra, posterior al punto, el número de grupo.

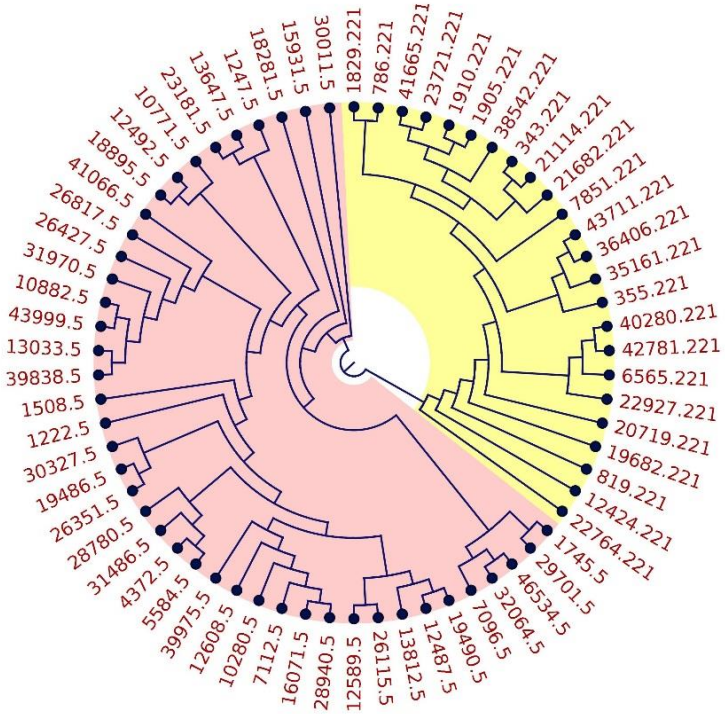


Figura 5.21 Análisis de dos grupos del mismo filo (*Proteobacteria*).

La Tabla 5.1 muestra la taxonomía de los dos grupos examinados de la base de datos control. La taxonomía de los grupos analizados es similar a la reportada por NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) (Figura 5.22 y 5.23).

Syntrophus

Taxonomy ID: 43773
Inherited blast name: **d-proteobacteria**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Syntrophus Mountfort et al. 1984**

Lineage(abbreviated)
Bacteria; Proteobacteria; Deltaproteobacteria; Syntrophobacterales; Syntrophaceae

Figura 5.22 Taxonomía reportada en NCBI para el del grupo 5

Gallionella

Taxonomy ID: 96
Inherited blast name: **b-proteobacteria**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Gallionella Ehrenberg 1838**

Lineage(abbreviated)
Bacteria; Proteobacteria; Betaproteobacteria; Nitrosomonadales; Gallionellaceae

Figura 5.23 Taxonomía reportada en NCBI Taxonomy para el grupo 221.

La segunda prueba fue realizada para examinar la distancia de secuencias de diferentes filos, en la prueba anterior es notorio la corta longitud de las ramas que dividen los grupos que pertenecen al mismo filo, por lo que se espera que la distancia entre los grupos analizados sea mayor con respecto a la primera prueba. Los géneros analizados fueron *Syntrophus* del filo *Proteobacteria* como grupo 5 con 42 secuencias y como grupo 88 *Acholeplasma* del filo *Tenericutes* con 5 secuencias.

La Figura 5.24 muestra los resultados de la segunda prueba con grupos de diferentes filos, las ramas que dividen los grupos analizados tienen una longitud mayor en comparación con géneros del mismo filo.

La taxonomía de los grupos analizados se muestra en la Tabla 5.1. Esta taxonomía se comparó con la reportada en NCBI Taxonomy (Figura 5.22 para el grupo 5 y 5.25 para el grupo 88), los resultados fueron similares.

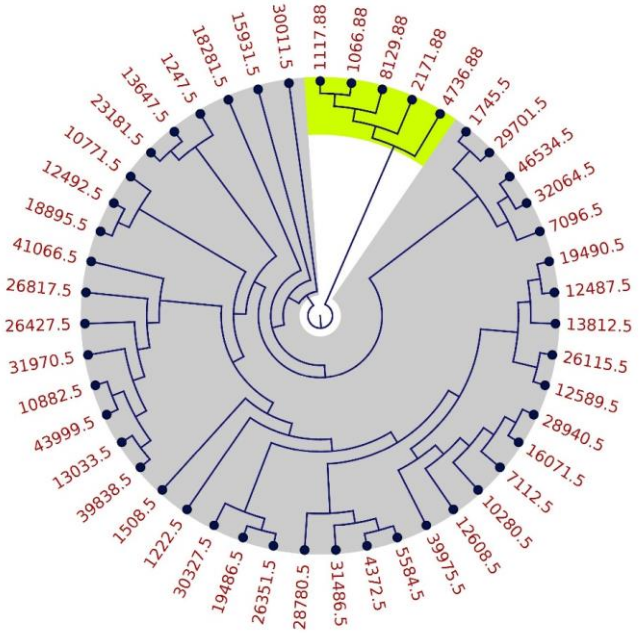


Figura 5.24 Análisis de dos grupos con diferentes filos.

Acholeplasma

Taxonomy ID: 2147
Inherited blast name: **mycoplasmas**
Rank: genus
Genetic code: Translation table 4 (Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate Mitochondrial; Mycoplasma; Spiroplasma)
Other names:
synonym: **Sapromyces**
authority: **Acholeplasma Edward and Freundt 1970**
authority: **"Sapromyces" Sabin 1941**

Lineage (full)
cellular organisms; Bacteria; Terrabacteria group; Tenericutes; Mollicutes; Acholeplasmatales; Acholeplasmataceae

Figura 5.25 Taxonomía reportada de NCBI Taxonomy para el grupo 88.

Para la siguiente prueba se eligieron cinco grupos, cuatro del dominio Bacteria con diferentes filos y un grupo del dominio *Archaea*. De los cuatro conjuntos bacterianos se tomaron 20 secuencias por cada uno y ocho secuencias por el grupo de *Archaea*. En la Tabla 5.4 se muestra el número de grupo y la taxonomía para esta tercera prueba.

La Figura 5.26 muestra los resultados de la tercera prueba con los cuatro conjuntos bacterianos y el grupo del dominio *Archaea*.

El análisis de esta tercera prueba muestra la separación evidente de cada uno de los grupos examinados. El conjunto de datos bacterianos muestra una distancia menor entre ellos y una clara separación hacia las secuencias pertenecientes al dominio *Archaea* (Figura 5.26). La rama que conecta la raíz con las secuencias de *Archaea* tiene una longitud mayor en relación con las secuencias bacterianas.

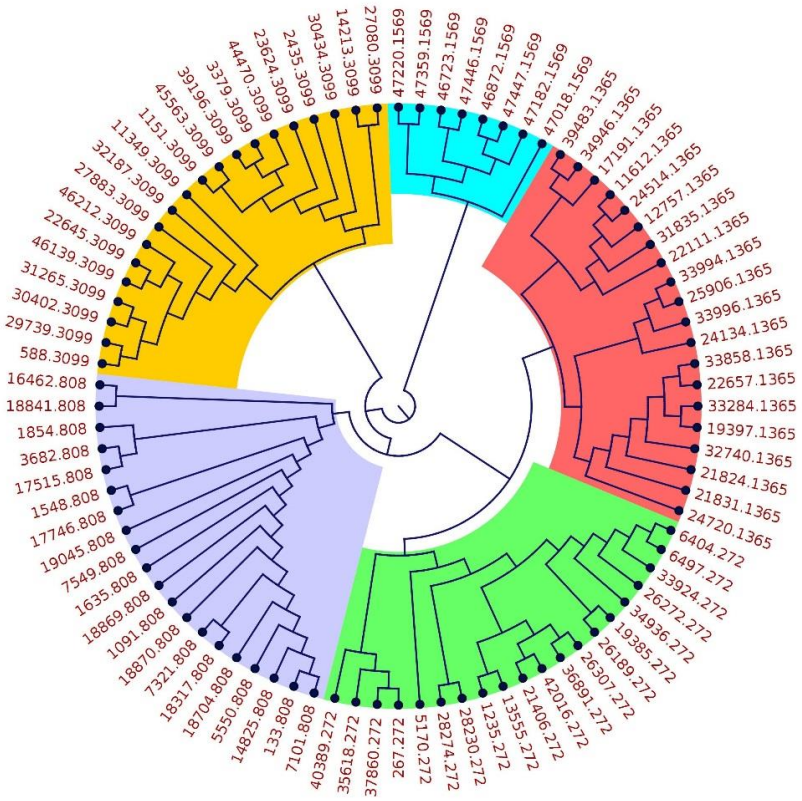


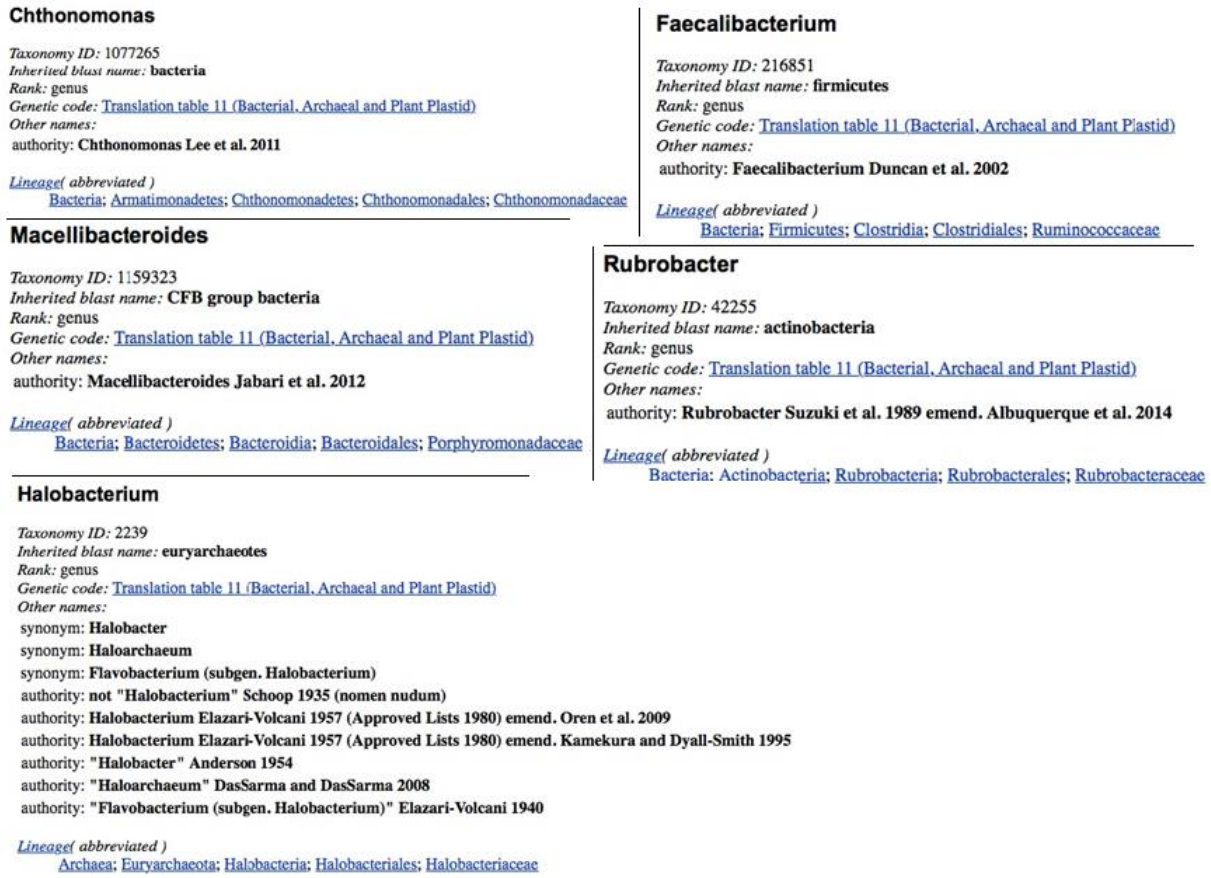
Figura 5.26 Cladograma circular de la tercera prueba con grupos bacterianos *y Archaea*.

La taxonomía de los grupos analizados de esta prueba (Tabla 5.1) se comparó con la que proporciona NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy). Los resultados demuestran que la taxonomía de la base de datos control con los grupos examinados es similar con la reportada en NCBI Taxonomy (Figura 5.26).

La última prueba se realizó con cinco grupos bacterianos, dos de ellos del mismo filo: grupo 5 del género *Syntrophus* y 221 del género *Gallionella*. Los grupos restantes son de distintos filos, grupo 88 del género *Acholeplasma*, grupo 808 del género *Faecalibacterium*, grupo 1365 del género *Chthonomonas* (taxonomía completa en la Tabla 5.1). La prueba tiene como objetivo evaluar grupos asignados al mismo filo y comparar las distancias entre ellos en una muestra con grupos del dominio bacteriano.

Los resultados de esta prueba indican que los grupos del mismo filo tienen una distancia menor con relación a los grupos con filos diferentes. En la Figura 5.28 se muestra el cladograma construido con los cinco grupos estudiados. Las ramas que conectan (color naranja) a los grupos 5 y 221 tienen una longitud corta, algunas secuencias muestran mayor similitud entre estos dos grupos. El grupo 88 muestra una mayor similitud con respecto a los

otros dos grupos (grupo 808 y 1365). El grupo 1365 es el conjunto más alejado, los cuatro grupos sobrantes mantienen una cierta similitud.



Chthonomonas

Taxonomy ID: 1077265
Inherited blast name: **bacteria**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Chthonomonas Lee et al. 2011**

Lineage(abbreviated)
Bacteria; Armatimonadetes; Chthonomonadetes; Chthonomonadales; Chthonomonadaceae

Faecalibacterium

Taxonomy ID: 216851
Inherited blast name: **firmicutes**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Faecalibacterium Duncan et al. 2002**

Lineage(abbreviated)
Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcaceae

Macellibacteroides

Taxonomy ID: 1159323
Inherited blast name: **CFB group bacteria**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Macellibacteroides Jabari et al. 2012**

Lineage(abbreviated)
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Porphyromonadaceae

Rubrobacter

Taxonomy ID: 42255
Inherited blast name: **actinobacteria**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
authority: **Rubrobacter Suzuki et al. 1989 emend. Albuquerque et al. 2014**

Lineage(abbreviated)
Bacteria: Actinobacteria; Rubrobacteria; Rubrobacterales; Rubrobacteraceae

Halobacterium

Taxonomy ID: 2239
Inherited blast name: **euryarchaeotes**
Rank: genus
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
synonym: **Halobacter**
synonym: **Haloarchaeum**
synonym: **Flavobacterium (subgen. Halobacterium)**
authority: **not "Halobacterium" Schoop 1935 (nomen nudum)**
authority: **Halobacterium Elazari-Volcani 1957 (Approved Lists 1980) emend. Oren et al. 2009**
authority: **Halobacterium Elazari-Volcani 1957 (Approved Lists 1980) emend. Kamekura and Dyall-Smith 1995**
authority: **"Halobacter" Anderson 1954**
authority: **"Haloarchaeum" DasSarma and DasSarma 2008**
authority: **"Flavobacterium (subgen. Halobacterium)" Elazari-Volcani 1940**

Lineage(abbreviated)
Archaea; Euryarchaeota; Halobacteria; Halobacteriales; Halobacteriaceae

Figura 5.27 Taxonomía reportada de NCBI Taxonomy similar a la prueba realizada con grupos formados.

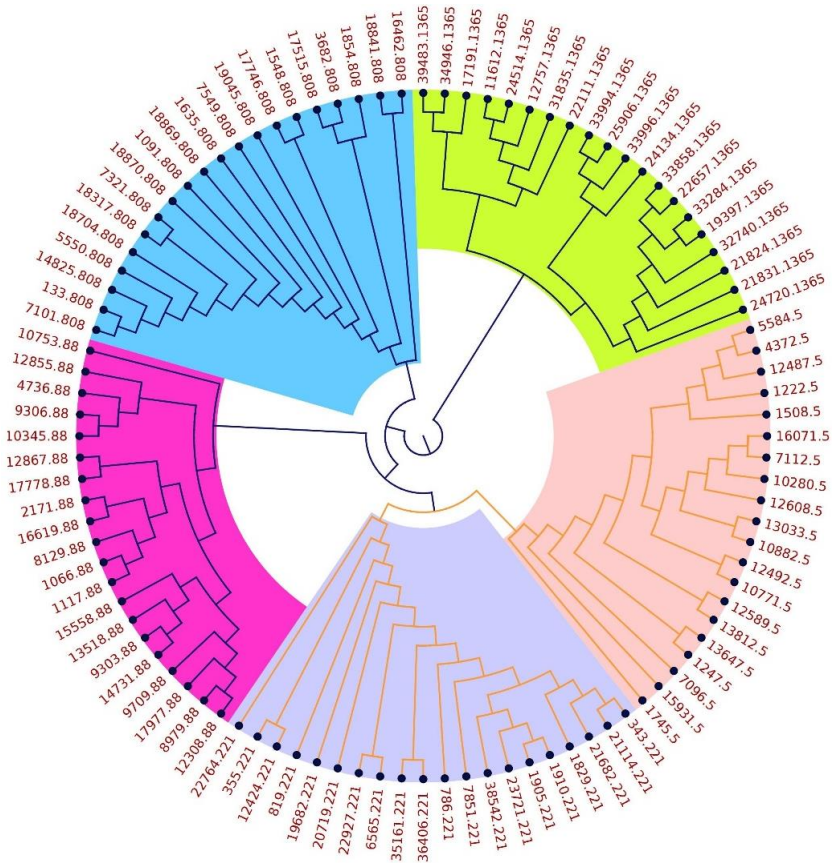


Figura 5.28 Cladograma construido con cinco grupos bacterianos, dos grupos del mismo filo.

5.7.1 *Clustering* de base de datos control con Cd-hit

Para evaluar la calidad de los grupos formados por DTW y K-medias, se crearon grupos con la herramienta Cd-hit (W. Li & Godzik, 2006) Esta herramienta bioinformática se caracteriza por tener una gran capacidad de manejar grandes conjuntos de datos.

Cd-hit agrupa secuencias de nucleótidos en grupos dado un umbral de similitud. Se utilizaron diferentes umbrales de similitud, de 90% hasta 97% de similitud.

Los datos de entrada para Cd-hit fueron etiquetados según el grupo creado con DTW y K-medias. Es decir, los datos están en formato fasta y los identificadores de cada secuencia tiene desde el número de especie hasta número de filo, según los grupos creados por los algoritmos antes mencionados.

Los resultados del *clustering* por parte de Cd-hit, muestran un elevado número de grupos al aumentar el umbral de similitud como se esperaba, sin embargo, con el umbral más bajo (90%) Cd-hit formó un número mayor de grupos comparado con los grupos formados por DTW y K-medias (3639 grupos vs 3099 grupos) (Figura 5.29).

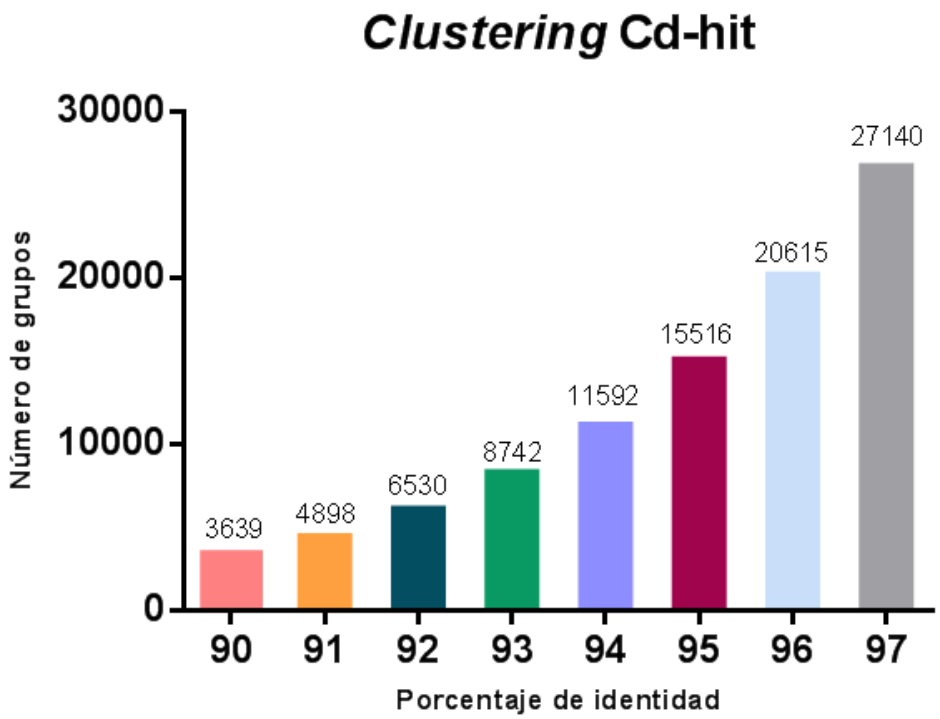


Figura 5.29 Grupos formados por la herramienta Cd-hit (W. Li & Godzik, 2006) con distintos umbrales de similitud (90-97%).

Por otra parte, se tomaron cinco grupos (20 secuencias de cada uno) de los 3639 creados por CD-hit con un porcentaje de similitud del 90%. Se etiquetó cada secuencia según el grupo formado por CD-hit separado por una coma, seguido por el grupo y número de secuencia dado por DTW y K-medias separados por un guion medio. Se creó un cladograma circular para validar los grupos formados por CD-hit (Figura 5.30). Se marcaron las secuencias de los grupos analizados por un color distinto (714-amarillo, 169-gris, 11-azul, 477-rosa, 591-verde). Las secuencias que presentaron menor dispersión fueron las del grupo 11 y 714. Sin embargo, el grupo 714 muestra una gran cercanía con más de la mitad de las secuencias del grupo 169. La estructura del árbol creado muestra que las distancias entre secuencias de diferentes grupos son muy cortas a tal grado que forman nuevos grupos.

Se tomó el porcentaje de similitud del 90% para la formación del árbol circular debido a la poca diferencia en cuanto a número de grupos formados por CD-hit y nuestro método. Se tomó la decisión de no evaluar los diferentes umbrales estudiados ya que es evidente que a mayor número de grupos mayor dispersión de secuencias y por ende formación de nuevos grupos.

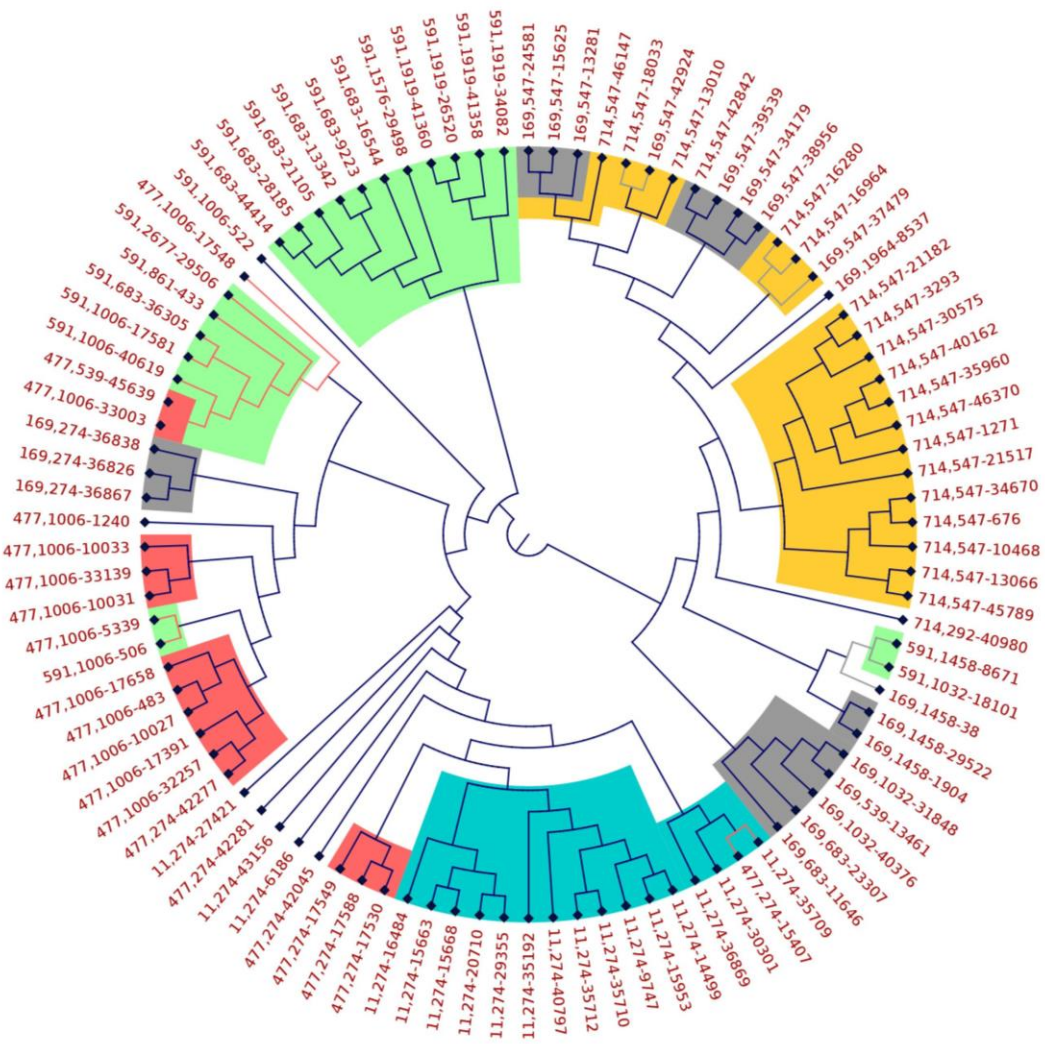


Figura 5.30 Análisis de grupos formados por CD-hit con un umbral de disimilitud del 10%. Cinco grupos fueron evaluados con 20 secuencias cada uno, cada grupo marcado con un color diferente. Se observa una gran dispersión entre grupos, además de formación de nuevos grupos.

5.8 Clasificador CTB

Se evaluaron dos herramientas de clasificación: *RDP classifier* (Wang *et al*., 2007) y *16S Classifier* (Chaudhary *et al*., 2015). Se descargó la base de datos SILVA *release 102*, la cual contiene 17385 secuencias del dominio Bacteria y 656 secuencias de *Archaea*. Se tomó esta información descargada de https://www.mothur.org/wiki/Silva_reference_files como *benchmark* para evaluar las diferentes herramientas de clasificación.

Se recortaron las secuencias del dominio Bacteria y *Archaea*, se obtuvieron más del 94% del dominio Bacteria y más del 95% en dominio *Archaea* (Figura 5.31).

Se evaluaron dos aspectos en el proceso de clasificación de las herramientas *RDP classifier, 16S classifier* y el clasificador CTB: tiempo y precisión.

Para CTB se utilizó las secuencias recortadas en la región de interés 804pb-1392pb del *benchmark*. Para las herramientas *RDP classifier y 16S classifier* se utilizaron las secuencias que reconocieron los oligonucleótidos 804F y 1392R, pero se tomaron sin recortar, esto para analizar el mismo número de secuencias en las 3 herramientas a evaluar.

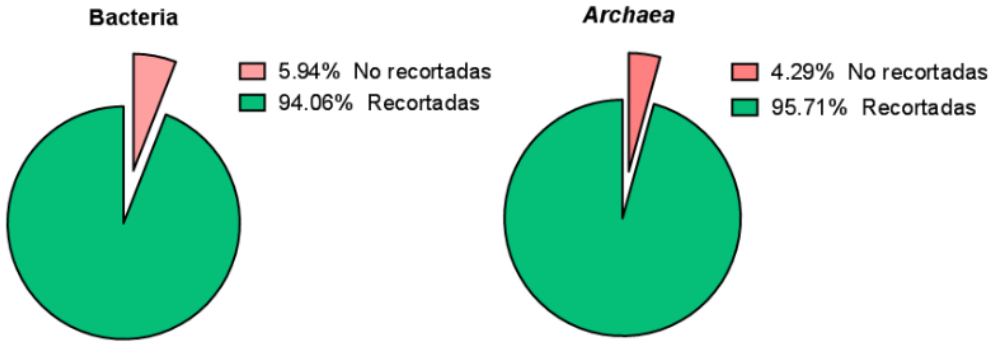


Figura 5.31 Porcentaje de secuencias recortadas en los diferentes dominios del *benchmark* en la región 804-pb-1392pb.

RDP classifier fue la herramienta con mayor rapidez en el proceso de entrenamiento y clasificación. CTB fue la herramienta con mayor precisión en la clasificación de secuencias. Mientras que *16S Classifier* fue la herramienta con el mayor tiempo en el proceso de clasificación y menor precisión respectivamente (véase Tabla 5-2 y Figura 5.32).

Tabla 5-2 Evaluación de dos aspectos en la clasificación de secuencias ribosomales 16S: tiempo y precisión para las herramientas *RDP classifier*, *16S classifier* y CTB.

Herramienta	**Tiempo de entrenamiento (horas: minutos: segundos)**	**Tiempo de clasificación (horas: minutos: segundos)**	**Porcentaje de precisión**
RDP classifier	00:20:16	00:06:03	95.31%
16S Classifier	----	16:21:49	23.75%
CTB	00:04:30	03:25:16	96.17%

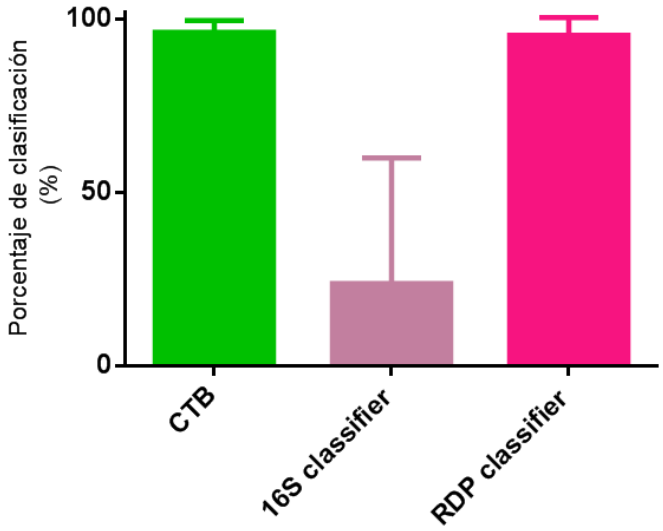


Figura 5.32 Porcentaje promedio según la precisión de asignación para cada nivel taxonómico, donde CTB tuvo el mejor promedio de clasificación (96.17%), seguido de *RDP classifier* con el 95.31% y por último *16S classifier* con el 23.7%.

Para evaluar la profundidad de la clasificación de secuencias nuevas con las diferentes herramientas, se separaron las secuencias provenientes del *benchmark* de acuerdo a su asignación en cada nivel taxonómico (desde dominio hasta género).

Las tres herramientas muestran una precisión de clasificación similar a nivel de dominio. Sin embargo, en los siguientes niveles, *16S classifier* baja significativamente la precisión para clasificar información (Figura 5.33). Esto se puede observar en la gran desviación estándar que tiene *16S classifier* con respecto a *RDP classifier* y CTB (Figura 5.32).

La herramienta *RDP classifier* mantiene un porcentaje de clasificación por encima del 98% en los tres primeros niveles taxonómicos, sin embargo, tiene una disminución significativa a nivel de orden (87.2%). En los dos niveles taxonómicos siguientes se mantiene por arriba del 91% (Figura 5.33).

Por otra parte, CTB tiende a tener un porcentaje de clasificación por arriba del 97% en los 3 primeros niveles taxonómicos. A nivel de orden, nivel en el cual *RDP classifier* presenta una baja evidente, y niveles posteriores mantiene un porcentaje de clasificación por encima del 90% (Figura 5.33).

A niveles más específicos, como es familia y género, las dos herramientas con mejores resultados son similares. Por un lado, *RDP classifier* a nivel de familia mejora su clasificación en 3.24% con respecto a CTB. Por otra parte, CTB mejora la clasificación a nivel de género en un 2.2% con respecto al clasificador *RDP classifier* (Figura 5.33).

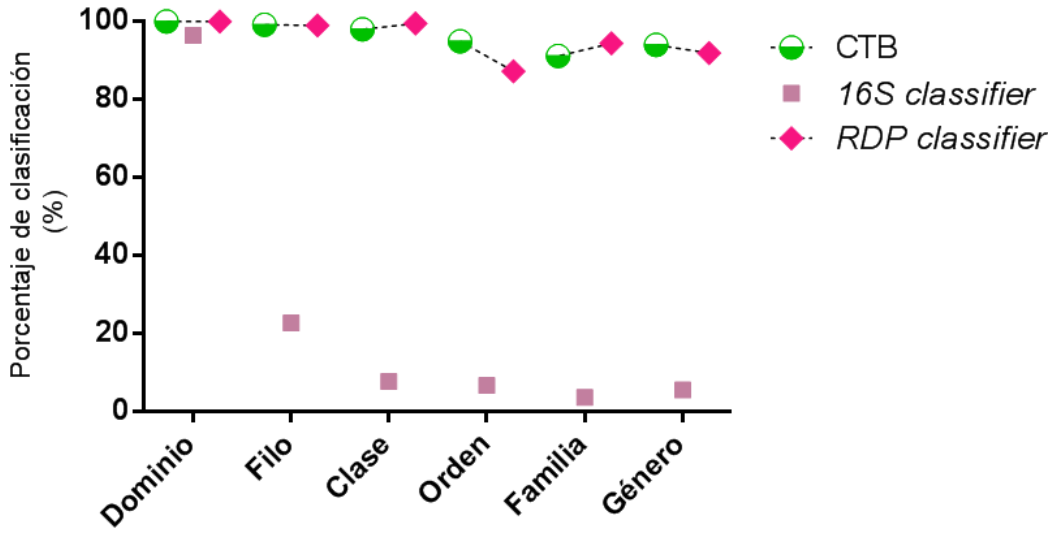


Figura 5.33 Porcentaje de clasificación del *benchmark* para cada nivel taxonómico evaluado en las tres herramientas de clasificación (*RDP classifier, 16S classifier* y CTB).

CAPÍTULO 6. DISCUSIÓN

Los estudios metagenómicos sobre comunidades microbianas basados en el análisis del gen 16S se han convertido en una gran herramienta para la investigación ambiental y humana. La gran cantidad de información producida por las nuevas tecnologías de secuenciación masiva se guarda en diferentes repositorios públicos como SILVA (Quast *et al*., 2013), RDP (J. R. Cole *et al*., 2009) y Greengenes (http://greengenes.secondgenome.com/). Estos repositorios contienen una gran cantidad de información por lo que el análisis de calidad y asignación taxonómica es complicado, cabe mencionar que nuestros análisis fueron con datos "curados" y sin "curar" con los tres repositorios. Los resultados reflejan una alta redundancia con las colecciones mezcladas (44% redundante) y, además solo el 13% de datos no redundantes reconoció los oligonucleótidos para la extracción de la región de interés. Con esto, se demuestra que un bajo porcentaje de la gran cantidad de información disponible tiene una calidad fiable. Esta calidad con base en la región analizada, cerca de 600 pb y ubicada en 804pb-1392pb del gen 16S.

Por otra parte, se analizó la base de datos "curada" de EzBioCloud (Kim *et al*., 2012) con organismos provenientes de muestras ambientales, los cuales son difíciles de clasificar e identificar y que además proporcionan diversidad en estudios y taxonómicos de agrupación. El reconocimiento de los oligonucleótidos de la región de interés (Lundberg *et al*., 2012) fue al 100% con las secuencias de EzBioCloud. Sin embargo, se encontraron nomenclaturas con códigos de acceso, lo que hace una asignación taxonómica no confiable. Estas nomenclaturas principalmente son de especies ambientales. Más del 75% de las secuencias de EzBioCloud tienen una nomenclatura clara hasta nivel de género, esto representa mayor diversidad con respecto a la base de datos 16S disponible en NCBI en el formato de referencia.

Por otro lado, los algoritmos computacionales son clave en la extracción eficiente de información de las grandes cantidades de datos generadas por las nuevas tecnologías de secuenciación masiva, es decir, juegan un papel muy importante en la generación de nuevo conocimiento. Existen diversas herramientas bioinformáticas, que integran algoritmos computacionales, enfocadas en el análisis de diversidad microbiana, estudios independientes de la taxonomía microbiana, entre las que se encuentran USEARCH (Edgar, 2010) y CD-hit (Huang *et al*., 2010). En la literatura basada en estudios de comunidades microbianas, se utiliza el 3% y 5% para diferenciar entre especies y géneros respectivamente (Cai & Sun, 2011). USEARCH y CD-hit son dos herramientas en las que se debe definir el

umbral de disimilitud para el proceso de *clustering* o creación de OTUS. Sin embargo, definir un umbral de disimilitud para la formación de OTUS y delimitar un nivel taxonómico especifico es complicado, esto debido a la complejidad de los datos. Por ello, es importante tomar en cuenta la complejidad de los datos para no subestimar o sobrestimar el número de OTUS sobre los esperados en un estudio (Chen *et al*., 2013). Este es un problema frecuente en el análisis de comunidades microbianas. Por consecuencia, en este estudio se realizó el proceso de *clustering* mediante dos algoritmos, el primero K-medias utilizado por USEARCH y el segundo, DTW para obtener distancias entre pares de secuencia. Con esto, se evitó el problema de definir un umbral de disimilitud. La diferencia radicó en que se utilizó los géneros reportados en la base de datos de EzBioCloud para elegir secuencias lideres o centroides y la región analizada (804pb-1392pb) presentó un comportamiento favorable en cuanto a la anotación, dado que en los ciclos de K-medias*,* pocas secuencias presentaron cambio constante de grupo. En cambio, CD-hit sobrestimó el número de grupos con el mismo umbral de disimilitud del 3% (27140 grupos vs 3099 grupos). Con un umbral elevado de disimilitud se acercó al número de grupos creados por nuestro método (3639 grupos vs 3099 grupos). Por otra parte, el cálculo de coeficiente de correlación de Pearson (0.9026) demuestra una correlación positiva fuerte entre los grupos creados por DTW-K-medias y los géneros reportados por Kim *et al*., 2012 y una R^2 del 0.8146, determina que nuestro método de agrupamiento y la anotación reportada es adecuado para la asignación de géneros en la región de interés con un intervalo de confianza del 95%. Además, la validación de los grupos por medio de cladogramas circulares demuestra que los grupos creados por nuestro método mantienen una distancia regular capaz de diferenciar a nivel de género y, además secuencias pertenecientes a un mismo grupo forman subgrupos y algunas otras se aíslan, pero no presentan reagrupación con grupos ajenos.

En la parte de clasificación de secuencias ribosomales, Pablo Yarza y colaboradores en su trabajo evaluaron diversas regiones del gen 16S, cada región de 250 pb (1-250, 251-500, 501-750…) (Yarza *et al*., 2014). Sus resultados demuestran que segmentos de 250 pb sobrestima la diversidad de OTUS y, por lo tanto, la asignación taxonómica tiene una alta tasa de error. También evaluaron el comportamiento con segmentos de 500pb, 750pb, 1000pb y 1250pb. Conforme la longitud del segmento aumenta, la asignación taxonómica mejora considerablemente. Al final, los resultados obtenidos por este estudio sugieren que el uso de la secuencia completa del gen mejora la asignación taxonómica en todos los niveles taxonómicos. Sin embargo, nuestros resultados demuestran que la región 804pb-1392pb es un segmento útil en la clasificación taxonómica hasta niveles de género. A pesar de ser un

fragmento cercano a las 600pb, la región estudiada mejoró la clasificación en diversos niveles taxonómicos comparado con el clasificador RDP (Wang *et al*., 2007), esta última herramienta utilizando la secuencias completas del gen 16S.

CAPÍTULO 7. CONCLUSIONES Y PERSPECTIVAS

La presente investigación se ha dedicado al estudio del gen 16S, en especial, la región 804pb-1392pb con el objetivo de crear una base de datos que contenga esta región de aproximadamente 600pb. Esta región contiene segmentos altamente conservados con cierta variabilidad, además de segmentos con gran variabilidad. Estas características benefician los resultados del proceso de identificación y clasificación de secuencias ribosomales.

El objetivo principal de la creación de una base de datos que contenga secuencias con las características de la región 804pb-1392pb es para identificar y clasificar secuencias ribosomales provenientes de archivos de secuenciación masiva mediante la implementación de un clasificador bayesiano.

En primera instancia, la creación de la base de datos era a partir de la unificación de tres bases de datos (SILVA, RDP y Greengenes), sin embargo, la base de datos que se obtuvo en el proceso de unificación fue muy grande, poco más de medio millón de secuencias. Por lo tanto, el proceso de agrupamiento de secuencias necesitaba gran poder de cómputo. Es por ello, se optó por una base de datos control para disminuir la necesidad de recursos computacionales, así como tener punto de comparación en cuanto a los resultados de agrupación obtenidos. Se utilizó la base de datos EzTaxon de Kim *et al*., 2012 como control. Esta base de datos esta analizada manualmente y contiene cerca de 60000 secuencias curadas.

En esta tesis se estudia la agrupación de secuencias ribosomales mediante un algoritmo utilizado en otras áreas como reconocimiento de voz y movimiento. Dicho algoritmo es el de Alineamiento Temporal Dinámico (DTW por sus siglas en inglés). Además de un algoritmo de *clustering* K-medias que es utilizado en diversas áreas como minería de datos, bioinformática, etc.

Con los resultados obtenidos de la agrupación de secuencias por medio de DTW y K-medias, se concluye que estos dos algoritmos en conjunto son una alternativa para el análisis de secuencias ribosomales, en especial, para el proceso de *clustering*.

Los resultados de correlación de Pearson obtenidos, combinando los algoritmos DTW con K-medias y la anotación reportada en la base de datos EzTaxon, mejoran la agrupación de secuencias ribosomales en la región de interés (804pb-1392pb) con respecto a los métodos de *clustering* en los que es necesario definir un umbral de disimilitud. Así mismo, los análisis de filogenia plasmados en los distintos cladogramas refuerzan la idea de que los algoritmos utilizados en este estudio tienden a mejorar la agrupación.

Por otra parte, a pesar de que la identificación y clasificación de secuencias ribosomales se encuentran bastante estudiadas, la región 804pb-1392pb contiene características particulares para una clasificación hasta nivel de género. Los resultados reportados en este estudio en cuanto a clasificación de secuencias ribosomales, utilizando Bayes simple y un diseño de datos con secuencias de aproximadamente 600 bases de nucleótidos, son similares con aquellos en los que se utiliza la secuencia ribosomal completa con el clasificador *RDP*. Con esto, se comprueba la hipótesis de que el diseño de una base de datos en la región 804-1392 del gen ribosomal 16S permite agrupar y clasificar secuencias ribosomales de bacterias.

REFERENCIAS BIBLIOGRÁFICAS

Abd-Elsalam, K. A. (2003). Bioinformatic tools and guideline for PCR primer design. *African Journal of Biotechnology*, *2*(5), 91–95. https://doi.org/10.4314/ajb.v2i5.14794

Akinsanya, M. A., Goh, J. K., Lim, S. P., & Ting, A. S. Y. (2015). Metagenomics study of endophytic bacteria in Aloe vera using next-generation technology. *Genomics Data*, *6*, 159–163. https://doi.org/10.1016/j.gdata.2015.09.004

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, *71*(12), 7724–7736. https://doi.org/10.1128/AEM.71.12.7724-7736.2005

Baker, P. G., & Brass, A. (1998). Recent developments in biological sequence databases. In *Current Opinion in Biotechnology* (Vol. 9, Issue 1, pp. 54–58). https://doi.org/10.1016/S0958-1669(98)80084-0

Barreto Hernández, E. (2008). Bioinformática: una oportunidad y un desafío Bioinformatics presents both an opportunity and a challenge. *Rev. Colomb. Biotecnol*, *X*(1), 132–138. http://dialnet.unirioja.es/servlet/articulo?codigo=2731617

Beckers, B., Op De Beeck, M., Thijs, S., Truyens, S., Weyens, N., Boerjan, W., & Vangronsveld, J. (2016). Performance of 16s rDNA primer pairs in the study of rhizosphere and endosphere bacterial microbiomes in metabarcoding studies. *Frontiers in Microbiology*, *7*(MAY). https://doi.org/10.3389/fmicb.2016.00650

Bokulich, N. A., Joseph, C. M. L., Allen, G., Benson, A. K., & Mills, D. A. (2012). Next-generation sequencing reveals significant bacterial diversity of botrytized wine. *PLoS ONE*, *7*(5), 3–12. https://doi.org/10.1371/journal.pone.0036357

Cai, Y., & Sun, Y. (2011). ESPRIT-Tree: Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research*, *39*(14), 1–10. https://doi.org/10.1093/nar/gkr349

Cañedo, R., & Arencibia, R. (2004). Bioinformática: en busca de los secretos moleculares de la vida. *Medicina*, *12*, 1–21. http://scielo.sld.cu/pdf/aci/v12n6/aci02604.pdf

Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics*, *26*(2), 266–267. https://doi.org/10.1093/bioinformatics/btp636

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., & Sharma, V. K. (2015). 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS ONE*, *10*(2), 1–13. https://doi.org/10.1371/journal.pone.0116106

Chelius, M. K., & Triplett, E. W. (2001). The diversity of archaea and bacteria in association with the roots of Zea mays L. *Microbial Ecology*, *41*(3), 252–263. https://doi.org/10.1007/s002480000087

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., & Zhao, H. (2013). A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS ONE*, *8*(8), e70837. https://doi.org/10.1371/journal.pone.0070837

Choi, J., Chen, J., Schreiber, S. L., & Clardyt, J. (1996). Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP. *Science*, *273*(July), 239–242. https://doi.org/10.1126/science.273.5272.239

Chun, J., Lee, J.-H., Jung, Y., Kim, M., Kim, S., Kwon Kim, B., Lim, Y.-W., & Jongsik Chun jchun, C. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, *57*(10), 2259–2261. https://doi.org/10.1099/ijs.0.64915-0

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., & Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(SUPPL. 1), 141–145. https://doi.org/10.1093/nar/gkn879

Cole, James R, Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633-42. https://doi.org/10.1093/nar/gkt1244

del Rosario Rodicio, M., & del Carmen Mendoza, M. (2004). Identificación bacteriana mediante secuenciación del ARNr 16S: fundamento, metodología y aplicaciones en microbiología clínica. *Enfermedades Infecciosas y Microbiología Clínica*, *22*(4), 238–245. https://doi.org/10.1016/S0213-005X(04)73073-6

DeSantis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R., & Andersen, G. L. (2006). NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.), 394–399. https://doi.org/10.1093/nar/gkl244

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST.

Bioinformatics, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Eiler, A., Heinrich, F., & Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal*, *6*(2), 330–342. https://doi.org/10.1038/ismej.2011.113

Escobar, C. A. M., Murillo, L. V. R., & Soto, J. F. (2011). Tecnologías bioinformáticas para el análisis de secuencias de ADN. *Scientia et Technica*, *3*(49), 116–121. https://doi.org/10.1017/CBO9781107415324.004

Febles Rodríguez, J. P., & Gonzalez-Perez, A. (2002). Aplicación de la minería de datos en la bioinformática. *ACIMED*, *10*(July), 69–76. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000200003&lng=es&nrm=iso

Galperin, M. Y., & Cochrane, G. R. (2009, January 1). Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Research*, *37*(SUPPL. 1), D1–D4. https://doi.org/10.1093/nar/gkn942

Garcia-Mazcorro, J. F., Castillo-Carranza, S. A., Guard, B., Gomez-Vazquez, J. P., Dowd, S. E., & Brigthsmith, D. J. (2017). Comprehensive Molecular Characterization of Bacterial Communities in Feces of Pet Birds Using 16S Marker Sequencing. *Microbial Ecology*, *73*(1), 224–235. https://doi.org/10.1007/s00248-016-0840-7

Garrity, G. M., Bell, J. A., Lilburn, T. G., & Lansing, E. (2004). Taxonomic outline of the prokaryotes. *Bergey's Manual of Systematic Bacteriology*, *2*(May), 1–399. https://doi.org/10.1007/bergeysoutline200405

Giorgino, T. (1996). *Computing and Visualizing Dynamic Time Warping Alignments in R : The dtw Package*.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Petrosino, J. F., Knight, R., & Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, *21*(3), 494–504. https://doi.org/10.1101/gr.112730.110

Hartmann, M., Howes, C. G., Abarenkov, K., Mohn, W. W., & Nilsson, R. H. (2010). V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, *83*(2), 250–253. https://doi.org/10.1016/j.mimet.2010.08.008

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). *CD-HIT Suite : a web server for clustering*

and comparing biological sequences. *26*(5), 680–682. https://doi.org/10.1093/bioinformatics/btq003

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, *12*(7), 1889–1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., & Wong, G. K.-S. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, *7*, 459. https://doi.org/10.3389/fmicb.2016.00459

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, 21–123. https://doi.org/citeulike-article-id:768582

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). *MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function* (pp. 207–233). Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-3369-3_13

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S. C., Jeon, Y. S., Lee, J. H., Yi, H., Won, S., & Chun, J. (2012). Introducing EzTaxon-e: A prokaryotic 16s rRNA gene sequence database with phylotypes that represent uncultured species. *International Journal of Systematic and Evolutionary Microbiology*, *62*(PART 3), 716–721. https://doi.org/10.1099/ijs.0.038075-0

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Gl??ckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, *41*(1), 1–11. https://doi.org/10.1093/nar/gks808

Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., & Clavel, T. (2016). IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Scientific Reports*, *6*(1), 33721. https://doi.org/10.1038/srep33721

Lee, C. P., Leu, Y., & Yang, W. N. (2012). Constructing gene regulatory networks from microarray data using GA/PSO with DTW. *Applied Soft Computing Journal*, *12*(3), 1115–1124. https://doi.org/10.1016/j.asoc.2011.11.013

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Y. H. (1998). Classification of Text Documents. *The Computer Journal*, *41*(8), 537–546.

https://doi.org/10.1093/comjnl/41.8.537

Loose, M., Malla, S., & Stout, M. (2017). *Europe PMC Funders Group Real-time selective sequencing using nanopore technology*. *13*(9), 751–754. https://doi.org/10.1038/nmeth.3930.Real-time

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Rio, T. G. del, Edgar, R. C., Eickhorst, T., Ley, R. E., Hugenholtz, P., Tringe, S. G., & Dangl, J. L. (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature*, *488*(7409), 86–90. https://doi.org/10.1038/nature11237

Manter, D. K., Korsa, M., Tebbe, C., & Delgado, J. A. (2016). myPhyloDB: a local web server for the storage and analysis of metagenomic data. *Database*, *2016*, baw037. https://doi.org/10.1093/database/baw037

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. https://doi.org/10.1038/ismej.2011.139

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Orozco Arias, S., & Arango López, J. (2016). Aplicación de la Inteligencia Artificial en la Bioinformática, avances, definiciones y herramientas. *UGCiencia*, *22*(1), 159. https://doi.org/10.18634/ugcj.22v.1i.494

Pillai, S., Gopalan, V., & Lam, A. K. Y. (2017). Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. *Critical Reviews in Oncology/Hematology*, *116*, 58–67. https://doi.org/10.1016/j.critrevonc.2017.05.005

Priyam, A., Woodcroft, B. J., Rai, V., Munagala, A., Moghul, I., Ter, F., Gibbins, M. A., Moon, H., Leonard, G., Rumpf, W., & Wurm, Y. (2015). Sequenceserver: a modern graphical user interface for custom BLAST databases. *BioRxiv*, 033142. https://doi.org/10.1101/033142

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219

Regier, Y., Komma, K., Weigel, M., Kraiczy, P., Laisi, A., Pulliainen, A. T., Hain, T., & Kempf,

V. A. J. (2019). Combination of microbiome analysis and serodiagnostics to assess the risk of pathogen transmission by ticks to humans and animals in central Germany 11 Medical and Health Sciences 1108 Medical Microbiology. *Parasites and Vectors*, *12*(1), 11. https://doi.org/10.1186/s13071-018-3240-7

Rodicio, M. D. R., & Mendoza, M. D. C. (2004). Identificación bacteriana mediante secuenciación del ARNr 16S: fundamento, metodología y aplicaciones en microbiología clínica. *Enfermedades Infecciosas y Microbiología Clínica*, *22*(4), 238–245. https://doi.org/10.1157/13059055

Rodríguez-Santiago, B., & Armengol, L. (2012). Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal. *Diagnostico Prenatal*, *23*(2), 56–66. https://doi.org/10.1016/j.diapre.2012.02.001

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing Artifacts on 16s rRNA-based studies. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0027310

Scholz, M. B., Lo, C. C., & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. *Current Opinion in Biotechnology*, *23*(1), 9–15. https://doi.org/10.1016/j.copbio.2011.11.013

Skutkova, H., Vitek, M., Sedlar, K., & Provaznik, I. (2015). Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of Theoretical Biology*, *385*, 20–30. https://doi.org/10.1016/j.jtbi.2015.08.007

Sun, Y., Cai, Y., Mai, V., Farmerie, W., Yu, F., Li, J., & Goodison, S. (2010). Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Research*, *38*(22), 1–10. https://doi.org/10.1093/nar/gkq872

Teach the Microbiome. (2017). *Sequencing the microbiome*. http://teachthemicrobiome.weebly.com/sequencing-the-microbiome.html

Valderas Álvarez, K. A. (2012). *Caracterización de Cepas de Bacillus Aisladas de Muestras de Miel y de Colmena Mediante la Secuenciación del Gen Ribosomal 16S* [Universidad Austral de Chile]. http://cybertesis.uach.cl/tesis/uach/2014/bmfciu.41i/doc/bmfciu.41i.pdf

Van De Peer, Y., Robbrecht, E., De Hoog, S., Caers, a, De Rijk, P., & De Wachter, R. (1994). Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Research*, *22*(1), 111–116. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148128&tool=pmcentrez&rendertype=abstract

Vasileiadis, S., Puglisi, E., Arena, M., Cappa, F., Cocconcelli, P. S., & Trevisan, M. (2012). Soil Bacterial Diversity Screening Using Single 16S rRNA Gene V Regions Coupled with Multi-Million Read Generating Sequencing Technologies. *PLoS ONE*, *7*(8), e42671. https://doi.org/10.1371/journal.pone.0042671

Villagrana-Bañuelos, K. E., Zanella-Calzada, L. A., Galván-Tejada, C. E., Gamboa-Rosales, H., & Galván-Tejada, J. I. (2020). *Evaluación de cuatro clasificadores para el reconocimiento de s ´ ındrome de muerte s ´ ubita del lactante utilizando ´ acidos grasos de cadena corta como fuente de informaci ´ on Evaluation of Four Classifiers for the Recognition of Sudden Infant Deat*. *149*(8), 1061–1071.

Vinje, H., Almøy, T., Liland, K. H., & Snipen, L. (2014). A systematic search for discriminating sites in the 16S ribosomal RNA gene. *Microbial Informatics and Experimentation*, *4*(1), 2. https://doi.org/10.1186/2042-5783-4-2

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-generation sequencing:from basic research to diagnostics. *Clinical Chemistry*, *55*(4), 641–658. https://doi.org/10.1373/clinchem.2008.112789

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Woese, C. R. (1987). Bacterial Evolution. *Microbiology*, *51*(2), 221–271. https://doi.org/10.1139/m88-093

Yang, C. C., & Iwasaki, W. (2014). MetaMetaDB: A database and analytic system for investigating microbial habitability. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0087126

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. https://doi.org/10.1038/nrmicro3330

ANEXO