DISEÑO E IMPLEMENTACIÓN DE UN ALGORITMO PARA LA IDENTIFICACIÓN DE COMUNIDADES MICROBIANAS CON POTENCIAL USO EN LA BIORREMEDIACIÓN DE RÍOS CONTAMINADOS CON CROMO (VI)

Luisa Fernanda Alzate Ruiz

Universidad el Bosque

Facultad de ingeniería

Programa académico: Bioingeniería

Bogotá D.C

2021

DISEÑO E IMPLEMENTACIÓN DE UN ALGORITMO PARA LA IDENTIFICACIÓN DE COMUNIDADES MICROBIANAS CON POTENCIAL USO EN LA BIORREMEDIACIÓN DE RÍOS CONTAMINADOS CON CROMO (VI)

Luisa Fernanda Alzate Ruiz

Proyecto de grado para optar por el título de bioingeniera

Tutor: Adriana Torres Ballesteros
Cotutor: Nuri Merchán Castellanos

Universidad el Bosque

Facultad de ingeniería

Programa académico: Bioingeniería

Bogotá D.C

2021

Tabla de contenidos

Pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

Pág.

LISTA DE ECUACIONES

pág.

RESUMEN

El cromo (VI) es un compuesto altamente contaminante debido a su naturaleza mutagénica y cancerígena, genera efectos negativos en microorganismos y plantas. La industrialización ha provocado que la concentración de estos contaminantes aumente, afectando las fuentes de agua y los suelos. Se han desarrollado nuevos enfoques biotecnológicos para el tratamiento de aguas residuales, donde el uso de microorganismos ha generado resultados prometedores en términos de eliminación, sin embargo, estos tratamientos pueden mejorarse mediante el análisis de datos de microbiomas.

El objetivo de este proyecto fue desarrollar un algoritmo para la identificación de comunidades microbianas, basado en perfiles taxonómicos y funcionales, con potencial uso para la biorremediación de ríos contaminados con cromo (VI). Para ello, se utilizó información de artículos y bases de datos que indicaban comunidades microbianas presentes en ríos contaminados con cromo (VI); luego se usaron estos datos para determinar los perfiles funcionales y taxonómicos de las comunidades microbianas. Se generaron mapas metabólicos para identificar las enzimas clave involucradas en la biorremediación del cromo. La información taxonómica y funcional se analizó en un algoritmo basado en aprendizaje automático para identificar características en comunidades microbianas con potencial para remediar el cromo (VI).

Se encontraron 50 artículos donde, el 71.2% contenían datos de amplicón 16S rRNA, el 17.3% datos metagenómicos, el 7.7 % librerías de clones, el 1.9% datos de amplicón 18S rRNA y el 1.5% de proteínas. En el análisis taxonómico realizado a través de Kbase y QIIME 2 se observó predominancia de los phylum *Proteobacteria, Firmicutes, Bacteroidetes* y *Acidobacteria* y especies como *Bacillus sp., Halomonas sp.* y *Comamonas sp.* en las muestras provenientes de efluentes de curtiduría. Estas cepas bacterianas han sido estudiadas por su capacidad de remediar metales pesados e hidrocarburos. A partir del análisis funcional se encontró la importancia de los donadores de electrones y la fuente de carbono en los procesos de remediación. Se implementaron dos clasificadores (árbol de decisión y bosques aleatorios), donde se comparó el rendimiento de cada uno usando 5 características obtenidas a través un análisis a partir de *Chi2* y 136 características obtenidas a partir de bosques aleatorios. Entre estos métodos se encontró que las características *K02227, K02232, K02233* y *K10617* estuvieron relacionadas en ambos resultados. El rendimiento en el modelo de bosques aleatorios (RF) usando las características seleccionadas a través de *Chi2* fue: exactitud del 88% y 81% (set de entrenamiento; set de validación) a comparación del análisis realizado con las características a partir del modelo de bosques aleatorios donde se obtuvo una exactitud del 100% y 63% (set de entrenamiento; set de validación). La clasificación través del modelo de RF donde, *Marinobacter hidrocarbonoclasticus* y *Bacillus paralicheniformis* se consideraron como microorganismos con alto potencial de remediación, mientras que *el* orden *Campylobacterales* tiene bajo potencial de remediación. Con la implementación del proyecto en otros ámbitos, sería posible apoyar las investigaciones y brindar soluciones en áreas como biotecnología y / o bioprocesos.

Palabras clave: Aprendizaje automático, Perfil funcional, Perfil taxonómico, Comunidad microbiana, Biorremediación

ABSTRACT

Chromium (VI) is a highly polluting compound due to its mutagenic and carcinogenic nature, it generates negative effects on microorganisms and plants. Industrialization has caused the concentration of these pollutants to increase, affecting water sources and soils. New biotechnological approaches have been developed for wastewater treatment, where the use of microorganisms has generated promising results in terms of elimination, however, these treatments can be improved by analyzing microbiome data.

The objective of this project was to develop an algorithm for the identification of microbial communities, based on taxonomic and functional profiles, with potential use for the bioremediation of rivers contaminated with chromium (VI). For this, information from articles and databases was used that indicated microbial communities present in rivers contaminated with chromium (VI); These data were then used to determine the functional and taxonomic profiles of the microbial communities. Metabolic maps were generated to identify the key enzymes involved in chromium bioremediation. Taxonomic and functional information was analyzed in an algorithm based on machine learning to identify characteristics in microbial communities with the potential to remediate chromium (VI).

Fifty articles were found where 71.2% contained 16S rRNA amplicon data, 17.3% metagenomic data, 7.7% clone libraries, 1.9% 18S rRNA amplicon data and 1.5% proteins. In the taxonomic analysis carried out through Kbase and QIIME 2, a predominance of the phylum *Proteobacteria*, *Firmicutes*, *Bacteroidetes* and *Acidobacteria* and species such as *Bacillus sp.*, *Halomonas sp*. and *Comamonas sp*. in samples from tannery effluents. These bacterial strains have been studied for their ability to remediate heavy metals and hydrocarbons. From the functional analysis, the importance of electron donors and the carbon source in remediation processes was found. Two classifiers were implemented (decision tree and random forests), where the performance of each one was compared using 5 characteristics obtained through an analysis from *Chi2* and 136 characteristics obtained from random forests. Among these methods, it was found that the characteristics *K02227*, *K02232*, *K02233* and *K10617* were related in both results. The performance in the random forest (RF) model using the characteristics selected through *Chi2* was: accuracy of 88% and 81% (training set; validation set) compared to the analysis performed with the characteristics from the model of Random forests where 100% and 63% accuracy was obtained (training set; validation set). The classification through the RF model where, *Marinobacter hydrocarbonoclasticus* and *Bacillus paralicheniformis* were considered as microorganisms with high remediation potential, while the *Campylobacterales* order has low remediation potential. With the implementation of the project in other areas, it would be possible to support research and provide solutions in areas such as biotechnology and / or bioprocesses.

Keywords: Machine learning, Functional profile, Taxonomic profile, Microbial community, Bioremediation.

INTRODUCCIÓN

El avance en el sector industrial ha hecho que la contaminación ambiental aumente a causa de los vertidos de compuestos tóxicos como el cromo (VI) que son depositados en fuentes hídricas y suelos. Debido a esta problemática se han desarrollado procesos alternativos para el tratamiento de estos recursos con el fin de disminuir el impacto ambiental. Actualmente, es posible aplicar métodos de remediación químicos, físicos o biológicos. En el caso de los métodos biológicos se pueden emplear microorganismos para la remoción de contaminantes. Se han encontrado resultados promisorios a través de estos procesos, sin embargo, en la eficiencia influyen diferentes variables como las características físico-químicas del medio y del tipo de microorganismo. Por ello, es importante la búsqueda de herramientas que permitan identificar variables para mejorar la eficiencia de la biorremediación.

El presente proyecto tiene como objetivo desarrollar un algoritmo para la identificación de comunidades microbianas, basado en perfiles taxonómicos y funcionales, con potencial uso para la biorremediación de ríos contaminados con cromo (VI). Esto con el fin de potenciar las prácticas experimentales de investigadores a través del análisis de las características funcionales de las comunidades microbianas.

Lo que busca el desarrollo del algoritmo es seleccionar los microorganismos que tengan un mayor potencial para remediar cromo (VI) teniendo en cuenta sus características funcionales y taxonómicas. Esto disminuirá costos y tiempo a la hora de buscar un microorganismo para la remediación de ambientes contaminados, además permitirá establecer relaciones para llegar a implementar consorcios microbianos para aumentar la eficiencia de la remediación.

En la primera sección de este documento se encontrarán los conocimientos base para el desarrollo del proyecto. Seguido a esto, se encuentra la metodología que se llevó a cabo, donde se explica cuál fue el procedimiento para encontrar los datos y analizarlos mediante bioinformática. Además, se especifica qué variables se analizaron para la selección de microorganismos en el algoritmo. Luego se incluyeron los resultados obtenidos donde se hizo un análisis y discusión de los mismos. Finalmente se concluye con material de referencia conformado por bibliografía y anexos.

DEFINICIÓN DEL PROBLEMA

En la última década la contaminación hídrica por metales pesados como el cromo ha incrementado debido a los procesos industriales tales como la manufactura de colorantes, el procesamiento de cueros y las actividades minero-metalúrgicas (Marino, 2006). En países en vía de desarrollo no se cuenta con las suficientes plantas de tratamiento de aguas residuales (PTAR) que garanticen la remoción total de contaminantes antes ser vertidas a los ríos, razón por la cual, estas aguas impactan de manera negativa la salud de la población (Martínez y Torres, 2018).

El cromo (VI) es un compuesto altamente contaminante debido a su naturaleza mutagénica y cancerígena, teniendo efectos negativos en diferentes microorganismos y plantas (Mayssara A. Abo Hassanin Supervised, 2014a). La exposición de este contaminante en humanos puede generar reacciones irritantes tales como úlceras en la piel, dermatitis de contacto, conjuntivitis, trastornos en las vías respiratorias y carcinoma pulmonar (Domingo-Pueyoa et al., 2014). A nivel ambiental, este compuesto altera la estructura de las comunidades microbianas, reduciendo su crecimiento y actividad enzimática, debido a su capacidad para penetrar las membranas celulares (Focardi et al., 2013).

Los métodos actualmente empleados para disminuir residuos contaminantes en ríos, como la filtración por membrana o la electrodiálisis (Caviedes Rubio et al., 2015), no son 100% eficientes en la remoción de trazas contaminantes (Gil Garzón et al., 2012). En consecuencia, la prevalencia de compuestos como el cromo (VI) en ríos genera la bioacumulación en especies acuáticas, especialmente si se alimentan del fondo del cuerpo de agua (Chávez Porras, 2010).

Nuevos enfoques biotecnológicos para el tratamiento de aguas residuales han sido desarrollados; el uso de microorganismos en procesos conocidos como biorremediación han generado resultados promisorios en términos de remoción (Jobby et al., 2018). Teniendo como base este proceso se han derivado estrategias para aumentar el rendimiento, entre las que se destacan el uso de microorganismos aislados de ambientes contaminados, microorganismos genéticamente modificados y cultivos mixtos. Aun así, una gran diversidad de organismos que podrían ser empleados para este fin no han sido considerados, por causa de la dificultad en el crecimiento de cultivos sintéticos o por las técnicas de aislamiento empleadas. De tal manera, se limita la identificación de vías metabólicas y enzimas que intervienen en la biorremediación. La escasez de información relacionada a las vías metabólicas y enzimas de microorganismos puede generar un aumento en los costos iniciales para la evaluación *in situ* y en la caracterización y evaluación de factibilidad del proceso (Garzón, Miranda y Gómez, 2017).

La secuenciación masiva de última generación (NGS) es una tecnología ampliamente usada para estudiar diversidad microbiana a nivel taxonómico (estructura de comunidades) y funcional (vías metabólicas y/o enzimas). En la actualidad cada 6 a 9 meses se duplica la capacidad de secuenciar a nivel mundial (Hernández et al.,2019) y el volumen de datos

generados incluye información de fragmentos de ADN, ARN y proteínas. La información obtenida mediante los procesos de secuenciación se deposita en bases de datos como DNA Data Bank of Japan (DDBJ), MG-Rast y MGnify (Keegan et al., 2016), NCBI, GreenGenes, SILVA y EzTaxon (Ortiz-Estrada et al., 2019). En estas bases se disponibilizan datos recolectados alrededor del mundo de diferentes ambientes, incluyendo ecosistemas impactados con contaminación con cromo (VI).

La mayoría de los análisis de datos provenientes de microbiomas han sido abordados de forma descriptiva lo que ha limitado el aprovechamiento de funciones metabólicas microbianas en procesos biotecnológicos. Por esta razón es necesario el diseño de una metodología de análisis de datos que permita identificar microorganismos y procesos metabólicos que sean reguladores en diferentes ecosistemas, esta identificación es esencial para optimizar procesos como biorremediación de cromo (VI), ya que es posible direccionar las pruebas de recuperación de ambientes impactados usando comunidades microbianas y metabolismos más eficientes.

Hasta la fecha no hay reportes bibliográficos que identifiquen comunidades microbianas con potencial de biorremediación de aguas contaminadas con cromo (VI). Aunque existen herramientas informáticas para determinar perfiles taxonómicos y/o funcionales de comunidades microbianas (Z. Liu et al., 2013), estas no están diseñadas para direccionar la selección de un potencial de actividad microbiana capaz de bioacumular, transformar o degradar cromo (VI). La implementación de una metodología que identifique los microorganismos clave asociados a un proceso metabólico en los ecosistemas tendrá múltiples aplicaciones biotecnológicas, incluyendo procesos de biorremediación.

JUSTIFICACIÓN

La contaminación hídrica generada por cromo (VI) es de gran interés dado que este compuesto ha sido considerado como uno de los 17 químicos más peligrosos para la salud humana (Ahmad, 2014). El cromo (VI) puede ingresar en varios sistemas ambientales (agua, aire o suelos) y puede afectar significativamente a los organismos presentes en fuentes de agua dulce. El cromo (VI) tiene características móviles, reactivas y tóxicas en suelos y aguas debido a su composición química, se encuentra naturalmente en el ambiente, sin embargo, cuando varios efluentes industriales liberan altas concentraciones, es absorbido por la biosfera conduciendo a la toxicidad (Jobby et al., 2018). La concentración permitida para cuerpos de agua productivos de todas las formas de cromo, incluido el cromo (VI) es de 0.1 mg L-1 (Jobby et al., 2018), concentraciones mayores a 220 µg/L y 10 µg/L pueden generar toxicidad aguda en peces e invertebrados (Initiative, 2000). Razón por la cual, es necesario implementar tecnologías que permitan disminuir la concentración de cromo (VI) en fuentes hídricas con el fin de reducir el impacto sobre la salud humana y el ambiente.

Para cumplir con los objetivos propuestos en el proyecto es importante contar con información derivada de las ciencias ómicas (genómica, transcriptómica, proteómica, metagenómica etc.) disponible en bases de datos y artículos científicos. En el 2016 la base de datos MG-RAST albergaba más de 150.000 conjuntos de datos con más de 23.000 disponibles públicamente (Keegan et al., 2016), mientras que MGnify cuenta con más de 1.9 millones de conjuntos de datos de microbiomas disponibles públicamente. El 31.5% de esos datos fueron publicados el último año y para el mes de marzo del 2020 se tenían 2,703,019,984 secuencias registradas en DNA Data Bank of Japan (*DDBJ Release Statistics*, 2020).

El análisis de datos puede potenciar prácticas experimentales de investigaciones, dado que esta información provee una descripción de los parámetros temporales y ambientales de las comunidades microbianas. Esto es fundamental para conocer la respuesta de comunidades microbianas frente a condiciones ambientales específicas, como la contaminación con cromo (VI). Es necesario integrar la información descriptiva con nuevas metodologías para identificar funciones metabólicas que puedan optimizar la biorremediación de aguas contaminadas con cromo (VI).

Con lo anterior, se apoyan los procesos investigativos y se dan soluciones en áreas como la biotecnología o los bioprocesos que pueden ser implementados por ingenieros, microbiólogos y/o científicos de datos. El beneficio a la comunidad científica también se verá reflejado en la reducción de costos y tiempo, al disminuir las pruebas de ensayo y error para encontrar el/los microorganismos más eficientes en el proceso de remoción de contaminantes de fuentes hídricas.

El reto desde la Bioingeniería con este proyecto es diseñar e implementar un algoritmo que permita determinar los perfiles taxonómicos y funcionales de comunidades microbianas con potencial para biorremediar fuentes hídricas contaminadas con cromo (VI). En este proyecto se propone la búsqueda y curación de datos registrados en artículos y bases de datos. Después de la filtración de datos, se comparan los perfiles taxonómicos y funcionales de

microorganismos presentes en aguas contaminadas y no contaminadas con cromo (VI). A Partir de esta comparación se realizará la extracción y selección de características de los datos correspondientes a comunidades microbianas con potencial de biorremediación de cromo (VI). Las características funcionales y taxonómicas serán usadas para diseñar un algoritmo de Machine Learning que permita identificar comunidades microbianas reguladoras de procesos de biorremediación de cromo (VI).

Los resultados del presente proyecto serán un aporte fundamental para investigaciones realizadas por el instituto Rothamsted Research, debido a que actualmente se busca la optimización de modelos de disponibilidad de carbono en suelos para uso agrícola, donde uno de los componentes que debe ser integrado al modelo "Rothc", es la actividad microbiana y su relación con variables físicas del suelo. Esperamos que el algoritmo diseñado en este proyecto pueda ser aplicado en el futuro a diversos ambientes y que sea una guía para la identificación de características microbianas reguladoras de procesos metabólicos en suelos. Igualmente, la información derivada de este proyecto también es de interés para el grupo de investigación de Rothamsted Research, debido a que permitirá conocer cómo la contaminación de cromo (VI) en agua puede tener un impacto en los sistemas agrícolas alrededor del mundo (Ertani et al., 2017) además de un efecto directo en el ciclo de carbono. Lo anterior está relacionado a la inhibición causada por contaminación con cromo (VI) en funciones microbianas en el suelo, por ejemplo, mineralización de carbono y otras actividades enzimáticas (Dotaniya et al., 2017).

Con el perfil de un bioingeniero se puede llevar a cabo este tipo de proyectos porque enlaza conocimientos biológicos, matemáticos y de programación. Así se aportarán soluciones, mediante un modelo computacional, para el análisis de datos resultantes de un proceso bioinformático. El proyecto se alinea en el foco misional de relación hombre-agua, ya que está orientado a generar estrategias que permitan reducir el impacto ambiental que se genera sobre los ríos por las actividades antropogénicas y/o industriales. Teniendo en cuenta que los recursos hídricos son importantes para la salud humana, la sostenibilidad del ambiente y el desarrollo económico, el planteamiento de proyectos encaminados a la preservación de este recurso representa la principal herramienta para cumplir este propósito. El presente trabajo aporta soluciones dirigidas al objetivo de desarrollo sostenible número 6 declarado por la Organización de Naciones Unidas, donde se plantea "*Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos/as*".

OBJETIVOS

OBJETIVO GENERAL

Desarrollar un algoritmo para la identificación de comunidades microbianas, a partir de perfiles taxonómicos y funcionales, con potencial uso para biorremediación de ríos contaminados con cromo (VI).

OBJETIVOS ESPECÍFICOS

Construir una base de datos curada con secuencias de ADN, ARN y/o proteínas de comunidades microbianas en ríos contaminados con cromo (VI).

Determinar abundancia y mapas metabólicos de comunidades microbianas presentes en ríos contaminados con cromo (VI), basado en perfiles taxonómicos y funcionales.

Diseñar un algoritmo basado en Machine Learning, a partir de perfiles taxonómicos y funcionales, para la identificación de comunidades microbianas con potencial de biorremediación en ríos contaminados con cromo (VI).

Implementar un algoritmo que identifique factores microbianos, a nivel taxonómico y funcional con potencial de biorremediación de cromo (VI).

MARCO TEÓRICO

1. Contaminación Hídrica

La contaminación del agua consiste en la presencia de componentes químicos, físicos o biológicos que producen una condición de deterioro en un cuerpo de agua. La capacidad de una fuente hídrica de soportar la intervención de este tipo de sustancias depende del tipo de cuerpo de agua, su ubicación y beneficios que aporta (Ninla Elmawati Falabiba, 2019a). La contaminación puede ser producto de desechos domésticos, insecticidas y herbicidas, desechos de procesamiento de alimentos, metales pesados, desechos químicos y demás que son descartados en cuerpos de agua. De acuerdo con el contaminante que se encuentre en el agua se pueden generar diferentes efectos sobre el ecosistema, por ejemplo, cuando el agua tiene una alta concentración de nutrientes (fosfatos, nitrógeno, etc.) comienza a existir un crecimiento excesivo de algas tóxicas que otros animales acuáticos consumirán ocasionándoles la muerte (Muralikrishna V., I., 2017).
Las fuentes de contaminación pueden dividirse en fuentes puntuales y fuentes no puntuales. Las fuentes puntuales son aquellas que son identificables como las refinerías, fábricas, plantas de tratamiento de aguas residuales, etc. Mientras que las fuentes no puntuales son aquellas distribuidas en un área geográfica amplia como una cuenca hidrográfica, también pueden considerarse a aquellas fuentes móviles (Mayssara A. Abo Hassanin Supervised, 2014b). El sector industrial ha sido considerado como uno de los principales responsables del alto consumo hídrico, con ello se deriva la contaminación de mares, ríos y lagos ya que se han convertido en vertederos de los residuos generados por la actividad industrial (Gamba & Pedraza, 2017).

1.1. Contaminación hídrica en Colombia

La problemática actual de la calidad del agua está derivada principalmente por las descargas de residuos que resultan de las actividades humanas, lo que interfiere con el uso adecuado del recurso. La contaminación de cuerpos de agua en Colombia se debe a efectos naturales y antropogénicos (Gualdrón Durán, 2018). Estos factores han generado el impacto y presión sobre las fuentes de agua, por ejemplo, el oxígeno disuelto en los ríos Bogotá, Medellín, Alto Cauca, Chicamocha y Sogamoso ha descendido con el transcurso de los años debido a vertimientos domésticos e industriales (Valencia et al., 2009).

1.2. Contaminación por cromo (VI) en ríos

Uno de los problemas ambientales que más impacta en Colombia está relacionado con el uso excesivo de sustancias químicas en actividades ilícitas, el uso de metales pesados, el vertimiento de aguas con compuestos contaminantes relacionados con las actividades industriales y agrícolas, lo que ha causado daños irreparables a los ecosistemas acuáticos (Coelho et al., 2015). La contaminación generada por metales pesados está agrupada en sustancias como cadmio, mercurio, cromo, cobalto, cobre, níquel, plomo, estaño, vanadio

zinc o plata. Cada uno de estos componentes constituye un riesgo ambiental alto ya que son sustancias que difícilmente pueden ser degradadas debido a su estabilidad química, por lo que no pueden ser metabolizadas por los seres vivos, generando sobre ellos la bioacumulación del contaminante (Mancera & Álvarez Ricardo, 2006).
El cromo y los compuestos derivados son importantes para diferentes aplicaciones como: el procesamiento del curtido del cuero, conservación de la madera, la cerámica, la pirotécnica, la fabricación de aleaciones metálicas, etc. Por ello es importante tener en cuenta la toxicidad que este compuesto tiene ambientalmente (Guevara, 2010).

2. Tratamientos de agua mediante microorganismos

Actualmente existe una variedad de métodos para la descontaminación de aguas residuales, entre ellos se encuentra la utilización de microorganismos que pueden usar los componentes contaminantes presentes en las aguas como una fuente de energía para su metabolismo y crecimiento. La importancia del uso de microorganismos para el tratamiento de aguas resulta en que su presencia no genera subproductos contaminantes lo que permite que se mantenga un equilibrio natural entre los microorganismos que habitan el entorno, trayendo efectos positivos sobre el ecosistema (Romero López & Vargas Mato, 2017).

2.1. Biorremediación

La biorremediación es el uso de especies microbianas con el fin de limpiar suelos o aguas que han sido contaminados por productos químicos descargados. Este proceso estimula el crecimiento de organismos específicos que utilizan los contaminantes como fuente de alimento y energía. Cuando se trata de un ambiente no contaminado, las bacterias, hongos, protistas y otros microorganismos trabajan constantemente para la descomposición de la materia orgánica. Esta tecnología proporciona fertilizantes a estos organismos que basan su alimentación en compuestos contaminantes para fomentar su crecimiento, lo que haría que la descomposición del contaminante se realice con mayor velocidad (Speigth G., 2018). La biorremediación de las aguas residuales puede dividirse en tres tecnologías principales, estas son: La depuración natural, la bioestimulación y por último, la bioaumentación (Ome et al., 2018). Es importante diferenciar que la biorremediación es una intervención humana, mientras que la biodegradación es una propiedad biológica de los microorganismos (Chen B., Ye X., Zhang B., Jing L., 2018). La bioaumentación anteriormente mencionada consiste en la adición de cultivos microbianos en la fuente contaminada con el fin de mejorar limpieza de los contaminantes reduciendo el tiempo requerido para ello. Los microorganismos nativos están presentes en cantidades muy pequeñas, por lo que es posible que no puedan evitar la propagación del contaminante. Por lo tanto, la bioaumentación ofrece una manera de proporcionar un microorganismo específico en las cantidades suficientes para completar la biodegradación (Speigth G., 2017).

2.1.1. Interacción de microorganismos con cromo (VI)

Los microorganismos eucariotas y procariotas responden a la influencia de Cr (VI) mediante la combinación de redes celulares que actúan en varios niveles, como el poder de reductor que se genera por el metabolismo de energía basal, la adquisición de hierro y azufre, la homeostasis, la protección del estrés oxidativo de proteínas, la reparación de ADN y enzimas de desintoxicación (Viti et al., 2014). En la figura 1 se pueden observar los mecanismos de interacción previamente mencionados en las células procariotas.

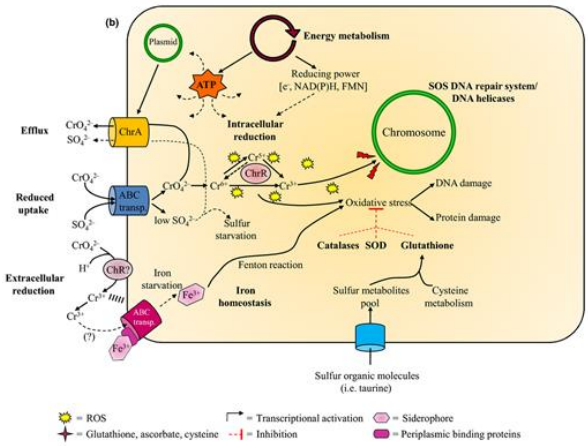

Figura 1. *interacción de microorganismos procariotas con cromo (VI). (Viti et al., 2014)*

2.1.1. Mecanismos de reducción microbiana de cromo (VI)

Los microorganismos al encontrarse en ambientes contaminados pueden comenzar a desarrollar mecanismos de resistencia para soportar condiciones tóxicas. Al encontrarse en ambientes contaminados se desarrollan mecanismos de reducción microbiana de cromo (VI) a cromo (III), este puede ser considerado como un mecanismo de desintoxicación de cromato. Se han descrito dos mecanismos de reducción directa. El primero donde el cromo (VI) se reduce en condiciones aeróbicas asociadas con reductasas de cromato solubles y el segundo donde este compuesto puede ser usado como aceptor de electrones en condiciones anaeróbicas por algunas bacterias (Viti et al., 2014).

3. Multi Ómica

Los estudios en el campo de la ómica apuntan a la abundancia y a la caracterización estructural de una variedad de moléculas biológicas de organismos en diferentes escenarios y campos como el clínico, ambiental o de la ciencia de los alimentos. Entre las diferentes ciencias ómicas se encuentran las relacionadas con el ADN (genómica, epigenómica), el ARN (transcriptómica), las proteínas (proteómica) y los metabolitos (metabolómica) cada una de estas ramas son de gran interés debido a su presencia en diferentes etapas de la información biológica. El estudio de estas moléculas proporciona nuevos conocimientos de

los procesos biológicos ya que cada una de ellas está interconectada y permite el análisis de datos a partir de ellas (Lamichhane et al., 2018).

3.1. Genómica

La genómica consiste en el estudio de la estructura y función de los ácidos desoxirribonucleicos (ADN) dentro del contexto genético y genómico. Esta ciencia investiga cómo la estructura molecular, la variación del haplotipo y la complejidad del gen en su relación con el ambiente puede alterar el genotipo y fenotipo (Ninla Elmawati Falabiba, 2019).

3.2. Proteómica

La proteómica involucra la aplicación de tecnologías para la identificación y cuantificación del contenido presente de proteínas de una célula, tejido y organismo, esto nos permite exponer la identidad de las proteínas de un organismo, conocer la estructura y las funciones de una proteína en específico (B. Aslam et al., 2017).

3.3. Metagenómica

La metagenómica es la ciencia que se encarga de realizar el análisis genómico de diferentes comunidades de microorganismos mediante la extracción directa y clonación de su ADN (Handelsman, 2005).
Actualmente, la metagenómica se puede dividir en dos enfoques principales que están dirigidos a diferentes aspectos de una comunidad microbiana asociada a un entorno en específico. El primero es el enfoque metagenómico estructural, donde se busca estudiar la estructura de una población microbiana no cultivada y esta puede expandirse hacia otras propiedades, como a la construcción de una red metabólica compleja entre los miembros de la comunidad, en este enfoque es posible conocer la composición de una población y su dinámica en un ecosistema en respuesta a presiones selectivas y parámetros espaciotemporales. Mientras que el enfoque metagenómico funcional tiene como objetivo identificar genes que codifican una función de interés (Alves et al., 2018). Mediante la metagenómica se puede conocer información sobre las capacidades funcionales de un microbioma al perfilar las abundancias relativas de los genes dentro de la comunidad microbiana (Langille, 2018).

4. Bioinformática

La bioinformática es un campo interdisciplinario que involucra principalmente la biología molecular y la genética, las ciencias de la computación, matemáticas y la estadística. Con ello se busca solucionar los problemas biológicos que tienen un uso intensivo de datos abordándolos desde un punto de vista computacional. De los problemas más comunes que se

pueden encontrar son la modelación de procesos biológicos a nivel molecular y hacer inferencias a partir de los datos recopilados (Can T., 2014).

La bioinformática se encarga de integrar muchos campos científicos y su tarea con la información está relacionada con la comprensión y la traducción de términos y conceptos de estas disciplinas, comenzando con un problema biológico a solucionar. Para cumplir con los objetivos se usan programas informáticos que incluyen algoritmos formulados por matemáticos y que son implementados por científicos informáticos con el fin de analizar esos datos e identificar genes, proteínas o enzimas según se requiera (Christensen Editor, 2018).

4.1. Base de datos biológica

Una base de datos biológica es un cuerpo extenso y organizado de datos, generalmente está asociado con un software diseñado para actualizar, consultar y recuperar componentes de los datos almacenados en el sistema. Una base de datos sencilla podría ser un archivo con diferentes registros que incluya un mismo conjunto de información (Jung, 2020). Entre las bases de datos populares se encuentra el GenBank del NCBI, MGnify, MG-Rast, etc.

4.2. Metadatos

Los metadatos hacen referencia a la información asociada a la secuencia de ADN que se está analizando, esta incluye la información descrita en los objetos de estudio, la muestra, el experimento, abarca el contexto de muestreo, la descripción del procesamiento de la muestra y la configuración del secuenciador. Mediante esta información se realizan los análisis genéticos con el fin de tener la mayor información posible de la procedencia de una muestra microbiana (Ten Hoopen et al., 2017).

4.3. QIIME 2

QIIME es una plataforma bioinformática robusta que permite combinar conjuntos de datos experimentales heterogéneos para obtener nuevos conocimientos acerca de comunidades microbianas (Caporaso et al., 2010).
Este software ha sido utilizado para analizar e interpretar datos de secuencias de ácido nucleico de comunidades fúngicas, virales, bacterianas y arqueales. Un análisis QIIME estándar comienza con datos de secuencia de una o más tecnologías de secuenciación, como Sanger, Roche/454, Illumina u otras. El uso de la plataforma para el análisis de datos de comunidades microbianas consiste en escribir una serie de comandos en una ventana terminar y finalmente ver la salida gráfica y textual, su interfaz de línea de comando tiene el estilo de Linux. Los protocolos muestran el uso del software para el procesamiento de datos de un estudio de secuenciación de ARNr 16S de alto rendimiento, donde se comienza con lecturas de secuencia multiplexadas de un instrumento de secuenciación y se finaliza con los perfiles

taxonómicos y filogenéticos y una comparación de las muestras en el estudio (Kuczynski et al., 2011).

5. Modelos para el análisis predictivo

El análisis predictivo de datos consiste en construir y usar modelos que generan predicciones basado en patrones provenientes de datos históricos. En el análisis de datos, una predicción es la asignación de un valor a cualquier variable desconocida (Kelleher et al., 2015).

5.1. Predicción Funcional

El aumento de volumen de datos genómicos y proteómicos ha generado el interés en la extracción automática de información funcional de estos conjuntos de datos. Un enfoque importante es la predicción *in silico* de la función de los genes y proteínas, donde se puede ir desde el papel bioquímico hasta su impacto en el fenotipo. Los métodos de predicción pueden dividirse en dos grandes categorías: los métodos que buscan predecir la función a partir de las propiedades intrínsecas de un gen y los enfoques realizados por la asociación que predicen rasgos funcionales basados en la similitud de un gen con funcionalidades ya existentes (Lehtinen et al., 2015).

5.1.1. Basada en Homología

La predicción basada en la homología se caracteriza por determinar la similitud entre secuencias, donde se supone que las secuencia que son similares están funcionalmente relacionadas. Por lo tanto, si se tiene una proteína desconocida, puede atribuirse aspectos de su función de acuerdo con la similitud de esta con otra proteína conocida (Kleine, 2012).

5.1.2. No-homóloga

Para una aproximación no-homóloga se deben tener en cuenta una serie de eventos, como la fusión o división génica y la ubicación relativa de los genes en el genoma (Kleine, 2012).

5.1.3 Basada en perfiles filogenéticos

Debido a la cantidad de información de genomas secuenciados que es almacenada en bases de datos, es posible buscar genes homólogos en diferentes organismos. De esta manera para todos los genes de un organismo puede determinarse la presencia o ausencia de un gen homólogo de un grupo de organismos previamente secuenciados completamente (Kleine, 2012).

5.2 Modelos basados en Machine Learning

Una de las características que define los sistemas de aprendizaje automático (ML) es que estos sistemas pueden ser mejorados a partir de la experiencia. Un sistema típico de ML requiere de al menos tres componentes diferentes que se mencionan a continuación (Soueidan & Nikolski, 2016).

1. *Experiencia*, representada en forma de datos.
2. *Tarea*, representada como la salida del algoritmo.
3. *Objetivo*, representado como la medición del rendimiento de un producto.

Para generar un modelo, es necesario partir de unos datos denominados datos de entrenamiento, estos son introducidos en el algoritmo de aprendizaje con el fin de identificar un modelo óptimo a partir de una hipótesis. Es importante tener en cuenta la extracción de características ya que este es el componente más importante de un sistema de aprendizaje (Soueidan & Nikolski, 2016).

5.2.1 Modelo de Árbol de Decisión

Los modelos basados en árboles de decisión (DT) son una estrategia de aprendizaje supervisado que se emplea para clasificación y regresión. Este sistema tiene como objetivo predecir el valor de una variable específica mediante el aprendizaje de una serie de reglas de decisión que se infieren a partir de las características de los datos (Pedregosa, F. \emphet al. et al., 2011).

Fórmulas matemáticas

Los vectores de entrenamiento $x_i \in R^n, i = 1$ y el vector de etiqueta $y \in R^l$, el árbol de decisión divide el espacio en dos para comenzar a tomar las respectivas decisiones. Por cada individuo dividido $\theta = (j, t_m)$, donde j hace referencia a una característica y t_m a un umbral. A partir de esta declaración se debe subdividir $Q_{left}(\theta)$ y $Q_{right}(\theta)$ de la siguiente manera,

$$Q_{left}(\theta) = (x, y)|x_j \leq t_m \qquad \textit{Ecuación 1}$$

$$Q_{right}(\theta) = Q \qquad \textit{Ecuación 2}$$

$$Q_{left}(\theta) \qquad \textit{Ecuación 3}$$

De esta manera se irán subdividiendo las decisiones que toma el modelo DT (Pedregosa, F. \emphet al. et al., 2011).

5.2.2 Modelo de Bosques Aleatorios (RF)

Este tipo de modelo consiste en construir un conjunto (bosque) de árboles de decisión a partir de una serie de datos. Este modelo de machine learning es ideal, ya que permite tener un sesgo bajo y una gran varianza entre los datos (Louppe, 2014).

6. SITUACIÓN ACTUAL EN EL ÁREA DE INVESTIGACIÓN – ESTADO DEL ARTE

Diferentes investigaciones se encargan de realizar estudios experimentales con muestras biológicas de diversos ambientes, para su posterior análisis y almacenamiento en bases de datos destinadas para tal fin. La información genómica es obtenida mediante secuenciación o microarreglos; en un estudio realizado (Woloszynek et al., 2019) se empleó secuenciación 16s rRNA con el fin de comparar perfiles funcionales con un enfoque basado en la abundancia relativa de las Unidades Taxonómicas Operativas (OTU), esta investigación muestra un modelo de donde puede basarse la predicción funcional microbiana. Permite observar qué aspectos se deben tener en cuenta a la hora de abordar una serie de datos de procedencia genómica.

En el Río Yamuna ubicado en India se realizó un estudio metagenómico para identificar las comunidades microbianas presentes en el mismo, teniendo en cuenta que es uno de los ríos más contaminados de la indica ya que recibe diversos efluentes con productos químicos tóxicos y metales pesados (Mittal et al., 2019), esto con el fin de estudiar el impacto ambiental de la contaminación en la microbiota del río. Otro estudio realizado en India se enfocó en el análisis de la diversidad microbiana de los vertederos de desechos de las curtiembres Jajmau y Unnao, ya que representan una gran amenaza a nivel ambiental (Verma & Sharma, 2020). En Pakistán también se han registrado estudios relacionados con curtiembres, en uno de ellos se realizó la secuenciación completa del genoma (WGS) donde se aislaron bacterias de efluentes de curtiembres con el fin de convertirlas en herramientas para la biorremediación, a partir del análisis de la diversidad taxonómica (Muccee & Ejaz, 2020).

El estudio enfocado a la predicción funcional es muy útil porque permite distinguir entre comunidades microbianas, por ejemplo, en una investigación a cargo de (Z. Ren et al., 2017) se realizó esta distinción entre diferentes comunidades a lo largo del río Qinghai y sus corrientes de entrada. Otras investigaciones que guían el estudio hacia los perfiles funcionales microbianos se enfocan en los cambios que ocurren en las comunidades microbianas por la interacción con diferentes contaminantes, se ha observado que los patrones funcionales microbianos cambian de acuerdo al gradiente de contaminación por metales pesados en diferentes muestras (Y. Ren et al., 2016)

Algunos estudios han demostrado las diferencias estructurales de comunidades microbianas al analizar muestras de diferentes zonas contaminadas y cómo estas comunidades se ven afectadas por la contaminación en su hábitat. En un estudio se observó los cambios funcionales de los microorganismos metabólicamente adaptados en ambientes contaminados con cromo (VI) y su capacidad para reducir este elemento. El estudio presentó datos derivados de secuenciación Illumina, MiSeq de ADNr 16s (Pei et al., 2018). El estudio de estas características permite identificar las diversas respuestas que presentan los microorganismos frente a la contaminación de los entornos por ejemplo con hidrocarburos (Mukherjee et al., 2017).

La predicción funcional de microorganismos también ha sido desarrollada para su aplicación sobre los suelos, donde se puede observar el perfil funcional de los microorganismos, de acuerdo a las características presentes en el lugar de estudio, de tal forma que se ha demostrado la respuesta de estas comunidades a la contaminación por cromo y arsénico en suelo de arroz (Kuang et al., 2016).

Mediante la bioinformática es posible realizar diversos análisis a partir de la información recolectada en bases de datos proveniente de la secuenciación genómica de diversos microorganismos. Estudios se han enfocado en la reconstrucción de la red metabólica a nivel metagenómico, lo que permite revelar información acerca de las características funcionales de una comunidad microbiana, estas características han sido estudiadas en organismos presentes en el río Bogotá (Ruiz-Moreno et al., 2019).

Para abordar la asignación taxonómica y funcional se han implementado algoritmos de ensamblaje y el análisis de lecturas de secuencias buscando homologías entre las mismas. En un estudio se implementó un modelo de predicción basado en ensambles microbianos, realizado mediante redes neuronales artificiales, donde se buscaba capturar y modelar las interacciones de taxones microbianos mediante inferencias bayesianas y diferentes métodos estadísticos (Kuang et al., 2016).

Uno de los métodos que se abordan con frecuencia en la actualidad es machine learning, la predicción de las funciones de una comunidad microbiana donde se integran diversos conocimientos como las regresiones lineales, redes neuronales y la selección de características con el fin de proporcionar una respuesta de mayor confianza. Un algoritmo basado en Machine Learning según investigaciones fue empleado para el análisis predictivo de carbono orgánico disuelto en la descomposición de la basura (Thompson et al., 2019).

El aprendizaje automático ha sido implementado en el área de la microbiología con diferentes enfoques, algunos de ellos guiados a la predicción de fenotipos ambientales del huésped aplicado al área de la salud, para ser de utilidad como apoyo a los brotes de enfermedad; además de esta aplicación también ha sido empleado para comprender la interacción entre microorganismos, permitiendo conocer el intercambio de metabolitos, los procesos de señalización, la inhibición de crecimiento y muerte (Qu et al., 2019). A partir de estos estudios se puede observar la aplicabilidad que tiene este método en temas micro ambientales.

7. METODOLOGÍA

7.1 Desarrollo del objetivo n°1

- Construir una base de datos curada con secuencias de ADN, ARN y/o proteínas de comunidades microbianas en ríos contaminados con cromo (VI).

7.1.1 Recolección de información genómica de microorganismos en ambientes contaminados con Cr (VI).

En esta fase se realizó la recolección de información con el fin de generar una base de datos de microorganismos presentes en ambientes contaminados con cromo (VI) para poder usarla en el desarrollo del proyecto. La información fue recolectada de bases de datos como: Scopus, Plos One, Science Direct, PudMed, Web of Science y PudMed.

La información que se recolectó se organizó teniendo en cuenta el año de publicación, los autores y el título del artículo de donde provenía. Cada artículo contenía los datos genómicos a través de números de acceso que permitían hacer una búsqueda en el NCBI para obtener las secuencias que finalmente serían usadas para el desarrollo del modelo de Machine Learning. Además de ello, se clasificó la información teniendo en cuenta el tipo de dato recolectado (16s rRNA, 18s rRNA, proteína, metagenoma) y la estrategia usada para obtenerlos (WGS, Reacción en Cadena de la Polimerasa; PCR, Electroforesis en gel con gradiente de desnaturalización; DGGE, pirosecuenciación, secuenciación de alto rendimiento). En la figura 2 se explica el proceso que se siguió.

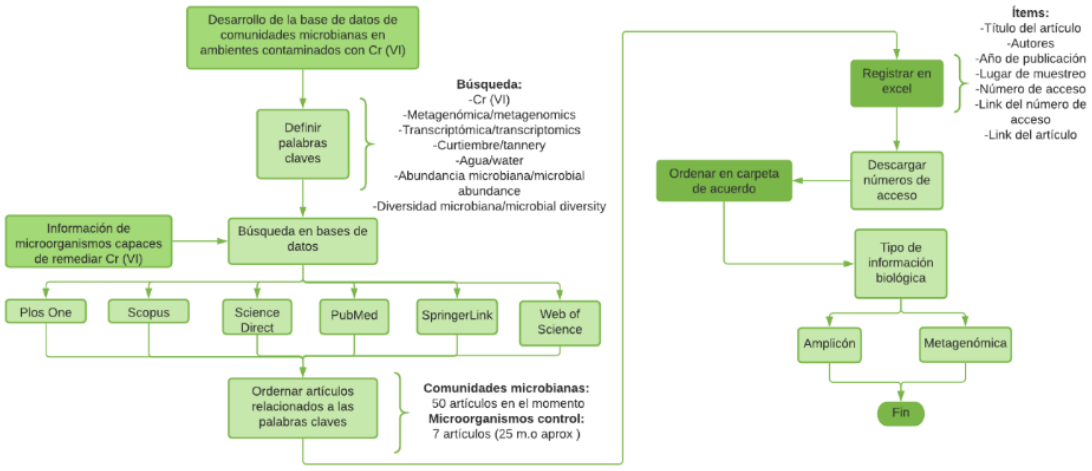

Figura 2. *Metodología del objetivo n°1. Autoría propia (2021)*

Adicionalmente, se buscaron genomas de microorganismos que tienen la capacidad de remediar ambientes contaminados con cromo (VI), esto con el fin de saber qué características se podrían emplear para identificar el potencial de remediación. Para generar un contraste entre los datos, se consultó también acerca microorganismos presentes en ambientes

contaminados con cromo (VI), pero que tuvieran un menor porcentaje de remoción registrado en bibliografía.

Se obtuvieron 3 grupos de datos a tener en cuenta para el desarrollo del algoritmo, estos se nombran a continuación:

1. Datos de bibliografía asociada a microorganismos en ambientes contaminados (Efluentes de curtiembres, residuos sólidos de curtiembres, lodos activados).
2. Datos de bibliografía asociada a microorganismos conocidos por su capacidad para remediar cromo (VI).
3. Datos de bibliografía asociada a microorganismos no reconocidos por su capacidad para remediar cromo (VI), se determina de esta manera ya que presentan porcentajes de remoción menores al 75%, sin embargo, no se descarta que puedan tener potencial de remediación.

Los datos correspondientes a microorganismos conocidos o no reconocidos por remediar cromo (VI), fueron organizados en un mismo archivo de Excel. En este se ordenó la información teniendo en cuenta lo siguiente: Título de la publicación, año de publicación, ambiente del aislado, microorganismo y link de donde fue recolectada la información.

7.2 Desarrollo del objetivo n°2

- Determinar abundancia y mapas metabólicos de comunidades microbianas presentes en ríos contaminados con cromo (VI), basado en perfiles taxonómicos y funcionales.

Una vez organizados los datos, se comenzó a hacer el análisis bioinformático para obtener la información taxonómica y funcional a emplear en el algoritmo. El análisis se realizó teniendo en cuenta los 3 grupos de datos mencionados anteriormente. Se emplearon datos metagenómicos y de datos correspondientes a amplicón 16 rRNA que fueron analizados usando las plataformas Kbase y QIIME2, respectivamente.

7.2.1 Análisis bioinformático a través de Kbase

La metodología para el análisis bioinformático en Kbase se basó en una de las narrativas que han sido generadas para la extracción de datos WGS (Chivian et al., 2020). La metodología empleada se puede encontrar en la página de Kbase en la sección de búsqueda de datos de referencia.

7.2.1.1 Importación de datos a Kbase para el análisis bioinformático

El análisis se realizó a través de Kbase, una plataforma bioinformática web que integra diferentes aplicaciones para el análisis de datos de comunidades microbianas.

Para obtener los datos metagenómicos para analizar en Kbase se realizó el siguiente proceso:

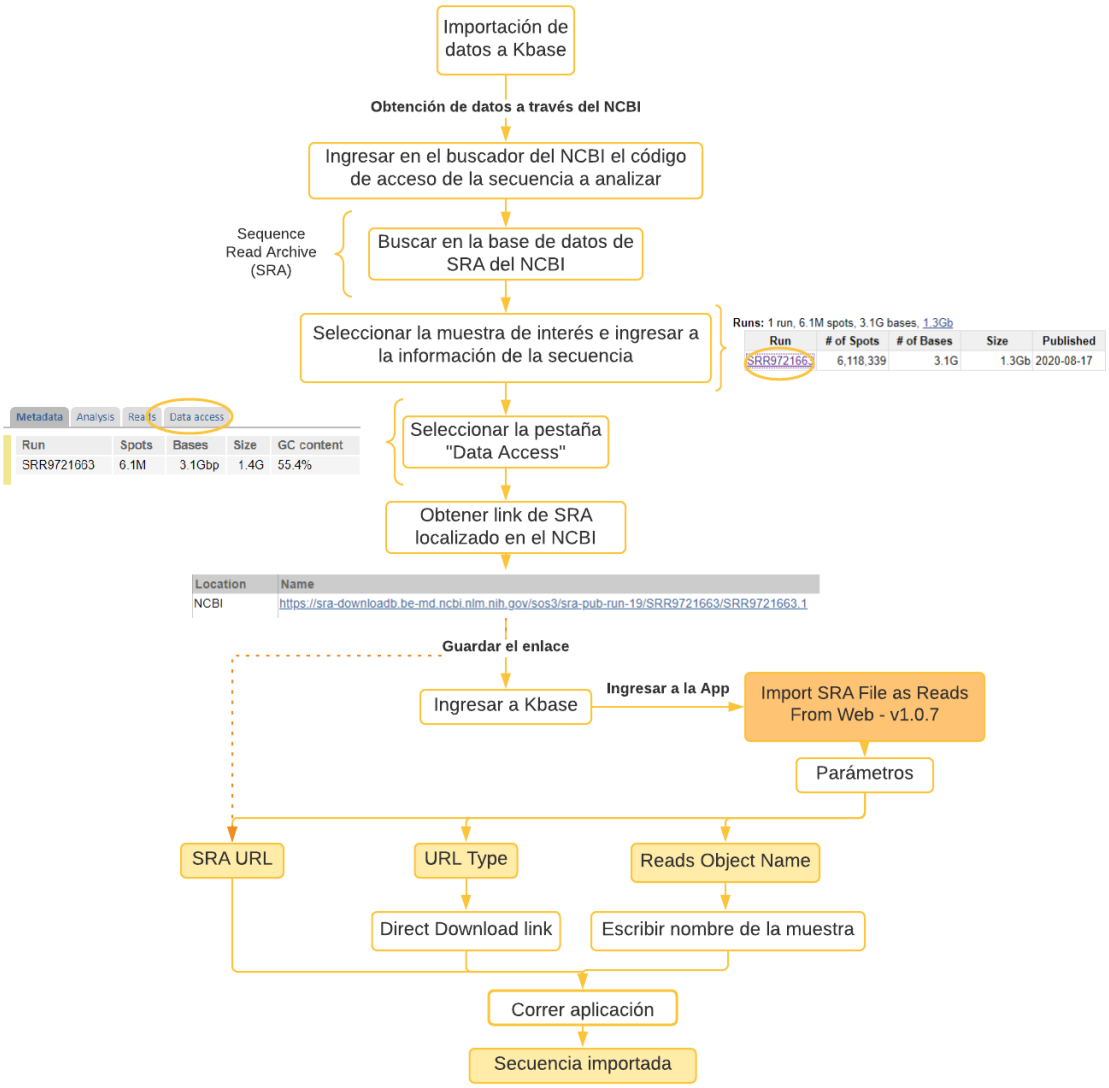

Figura 3.*Metodología para importar datos a Kbase. Autoría propia (2021)*

7.2.1.2 Preprocesamiento de los datos recolectados para su análisis funcional y taxonómico

Una vez se obtuvieron los enlaces de Sequence Read Archive (SRA), se comenzaron a cargar en Kbase. En la figura 4 se pueden observar cada uno de los pasos que se tuvieron en cuenta para el análisis funcional y taxonómico.

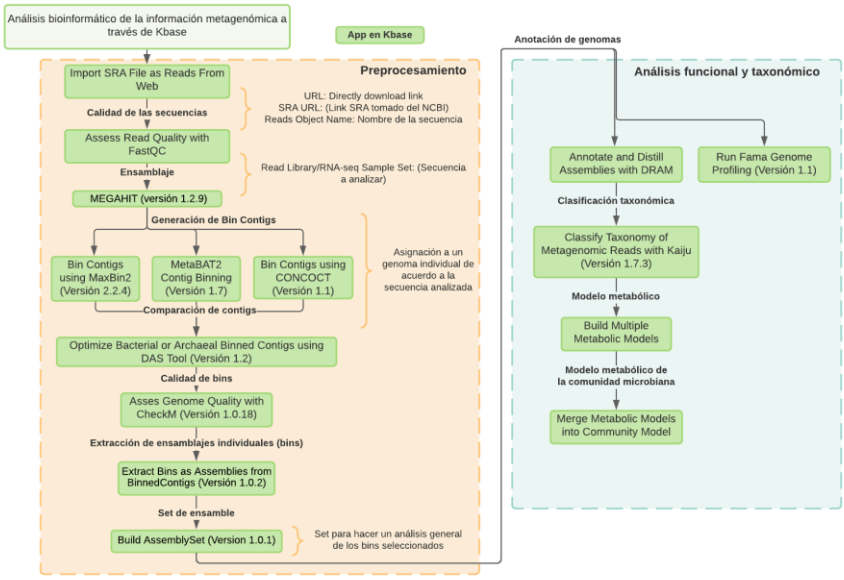

Figura 4. *Metodología del objetivo n°2. Autoría propia (2021)*

El análisis bioinformático en Kbase comenzó a través del preprocesamiento de los datos, lo que permite controlar la calidad en los datos de secuencia provenientes de las bases de datos, ya que estos no han sido manipulados previamente. Este procedimiento se realiza a través de la aplicación "*Assess Read Quality with FastQC*" que permite la obtención de una serie de estadísticas básicas para identificar la calidad de los datos de secuenciación (Andrews, 2020). Entre las métricas que se obtienen se encuentra la calidad por secuencia por base, el contenido de guanina-citocina por secuencia y los niveles de duplicación de secuencia.

En el análisis que se llevó a cabo se tuvo en cuenta la calidad de secuencia por base. El resultado que arroja este tipo de análisis es una serie de gráficos de caja y bigotes que muestran la calidad en cada posición a lo largo de todas las lecturas del archivo de secuenciación (Andrews, 2020). En el eje **Y** de estos diagramas se encuentra el índice de Phred que es un estándar para establecer el grado de fiabilidad de las bases nitrogenadas de la lectura de secuenciación. Esto nos permite identificar la probabilidad de error para cada nucleótido que ha sido secuenciado (Zúñiga Trejos, 2014).

Este coeficiente de calidad se puede determinar a partir de la siguiente ecuación.

$$Q = -10logE \qquad \textit{Ecuación 4.}$$

Donde la puntuación de calidad está representada por Q y la estimación del error por E (GATK team, 2021). Por lo tanto, lo que se busca es tener un mayor valor de Phred para tener la certeza de que se cuenta con una mayor precisión en los datos que se analizan. En la Tabla 1 se puede observar la relación que tiene el coeficiente de Phred con la precisión de las bases nitrogenadas en una secuencia.

Tabla 1. *Coeficiente de Phred*

Nivel de calidad Phred	Error	Precisión (1-error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%

Fuente: GATK team (2021)

De acuerdo a lo anteriormente mencionado se determinó realizar un filtrado de las bases en las secuencias que presentaran un coeficiente de Phred menor a 15. Este proceso se realizó a través de la aplicación "*Trim Reads with Trimmomatic"* en los casos que era necesario, allí se modificaron los parámetros *Sliding window size* y *Sliding window minimum quality* para obtener la secuencia filtrada (Bolger et al., 2014).

Posterior a filtrar la parte de las secuencias que no cumplían con el criterio de calidad que se estableció se realizó un ensamblaje de las secuencias. Este proceso se llevó a cabo con la aplicación "*MEGAHIT*" y consiste en organizar la secuencia de nucleótidos en el orden correcto, ya que lo que se recibe de la secuenciación son fragmentos y es necesario unificarlos para obtener una secuencia continua y ordenada. Es importante tener en cuenta que al trabajar con metagenomas se tiene información acerca de diferentes microorganismos, por lo que se obtienen diferentes secuencias.

Una vez se realiza el ensamble, estas deben ser agrupadas en Bin Contigs que almacenan bines. Estos son una asignación de un genoma individual, es decir, son una secuencia que representa un microorganismo y varía dependiendo al método de agrupación de Bin Contigs que se use.

En Kbase existen 3 aplicaciones para realizar la agrupación que son las siguientes: "*Bin Contigs using MaxBin2*","*MetaBAT2 Contig Binning*" y "*Bin Contig using CONCOCT*". Cada una de ellas puede generar una agrupación diferente por lo que se determinó usar las 3 y realizar una comparación de los resultados a través de la aplicación "*Optimize Bacterial or Archael Binned Contigs using DAS Tool*".

DAS Tool nos permite obtener un gráfico que indica la cantidad de bines que se obtienen y lo completo que se encuentra cada uno de ellos representado a través de porcentajes (Sieber et al., 2018). Al obtener el resultado se debe seleccionar el que contenga los bines más completos con el fin de que sea posible identificar el microorganismo que está siendo manipulado.

Posterior a este proceso se debe estimar la integridad y contaminación de los genomas (bines), para ello se usa la aplicación "*Assess Genome Quality with CheckM*". Este proceso se debe realizar para prevenir errores en análisis posteriores por contaminación de los datos o la falta de información en los mismos (Parks et al., 2015).

Para finalizar con el preprocesamiento de los datos se deben extraer los bines escogidos a través de la aplicación "*Extract Bins as Assemblies from BinnedContigs*" teniendo en cuenta que la contaminación en los bines no supere la mitad del genoma registrado.

Una vez se extraen los bines se debe generar un set de datos con todos los bines seleccionados y se realiza a través de la aplicación "*Build AssemblySet"* que permite obtener un solo archivo continuar con el análisis funcional y taxonómico.

El análisis funcional se realizó a través de 2 aplicaciones que tiene disponible Kbase, estas son: "*Annotate and Distill Assemblies with DRAM"* y "*Run Fama Genome Profiling".* Para hacer uso de estos aplicativos es necesario ingresar el archivo previamente ensamblado.

DRAM es una aplicación que nos ayudará a identificar el resumen de los metabolismos que cada bin es capaz de realizar a través de un archivo .zip que recopila lo siguiente (Shaffer et al., 2020).

1. Tabla de estadísticas del genoma.
2. Resumen del metabolismo donde se indica el recuento de genes funcionales en diversos metabolismos.
3. Mapa de calor que indica la presencia o ausencia de los componentes de transporte de electrones y la presencia de las diferentes funciones metabólicas en cada muestra.

Fama Genome Profiling es una aplicación que permite identificar perfiles funcionales y el perfil taxonómico de los genes funcionales relacionados a los siguientes datos de referencia (Kazakov A, 2019):

- Conjunto de datos de enzimas relacionadas al ciclo del nitrógeno para el perfil funcional y taxonómico de genes metabólicos de nitrato/nitrito y amoniaco.
- Marcadores universales para la verificación de contaminación.

En este caso se tuvieron en cuenta los datos de referencia asociados al ciclo del nitrógeno.

Posterior a este proceso se generó un modelo metabólico que permite observar las reacciones químicas que se llevan a cabo en la comunidad microbiana, para ello se usaron las aplicaciones "*Build multiple Metabolic Models" y "Merge Metabolic Models into Community Model".*

Para finalizar se realizó el análisis taxonómico a través de la aplicación "*Classify Taxonomy of Metagenomic Reads with Kaiju"* que permite obtener una serie de gráficos de barras organizando los datos de acuerdo a la clasificación de filo, clase, orden, familia, género y especie (Menzel et al., 2016).

7.2.2 Análisis bioinformático a través de QIIME 2

La metodología para el análisis bioinformático en QIIME 2 se basó en un código que fue otorgado por el área *Sustainable Agriculture Sistems* del Instituto Rothamsted Research. Además de ello, el servidor para hacer uso del programa bioinformático también fue concedido por el instituto.

En la figura 5 se puede observar el procedimiento que se llevó a cabo para el análisis. En los recuadros de color amarillo se describen los archivos de salida luego de ejecutar el código correspondiente, estos archivos tienen un formato ". qza" que es usado para poder analizar los datos en QIIME 2. Los recuadros azules representan el nombre de la función usada para generar un archivo de visualización, que es el usado para saber cómo modificar la línea de código del paso a seguir. Finalmente, en los recuadros verdes se muestra el nombre de archivos de visualización mediante formato ". qzv"; estos archivos se deben visualizar a través de QIIME 2 View en algún buscador Web.

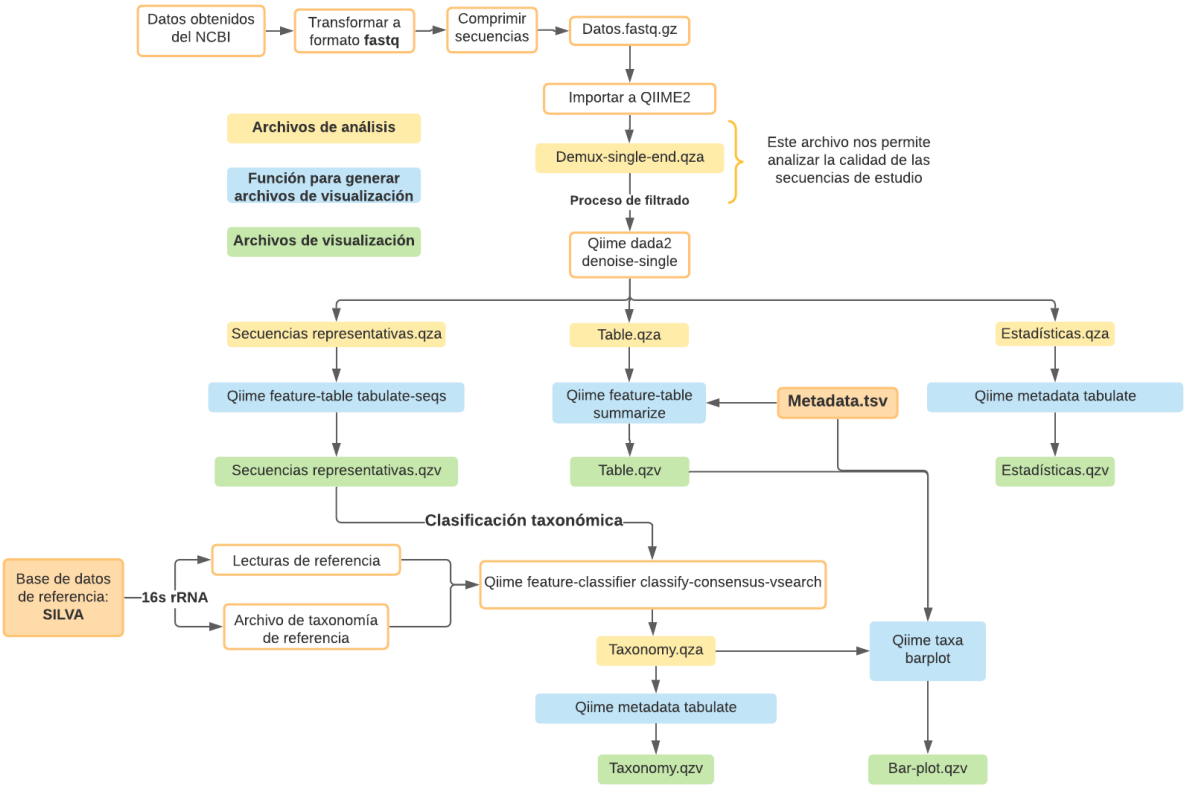

Figura 5. *Flujo para la clasificación taxonómica con Qiime 2. Autoría propia (2021)*

7.2.2.1 Transformación de los datos de la Web para su análisis mediante QIIME2

Inicialmente se descargaron los datos correspondientes a amplicón 16s rRNA de la base de datos desarrollada en el objetivo 1.

Es necesario que los datos estén en formato *fastq* y que el directorio donde se encuentran comprima cada uno de las muestras; lo que nos permite obtener un archivo *fastq.gz*. Posterior a este proceso se deben nombrar las muestras de acuerdo a la convención de nomenclatura de Ilumina (Jiménez & Ph, 2021). Por ejemplo, **NombreDeLaMuestra_S1_L001_R1_001.fastq.gz** donde,

S1, corresponde al identificador del código de barras.

L001, corresponde al número de carril.

R1, corresponde a la dirección de la lectura.

001, corresponde al número de muestra.

7.2.2.2 Creación de archivo de metadata

Luego de obtener un directorio con cada uno de los archivos organizados y renombrados se debe hacer un documento de metadata. Este es un archivo que proporciona la descripción de los sitios de muestreo, lo que permite tener un contexto de la información de cada secuencia (Committee, 2007). Para la creación de este archivo se realizó una tabla en Excel teniendo indicando en cada una de las columnas lo siguiente: *Sample ID* (por ejemplo, NombreDeLaMuestra_S1_L001_R1_001), *ubicación* y *año*, se tuvo en cuenta para cada una de las muestras. Finalmente, el archivo se debe guardar en formato *.tsv* en el servidor para poder usarlo en los análisis de QIIME 2.

7.2.2.3 Importación de secuencias al entorno de QIIME 2

En esta sección se ejecuta una línea de código con la función *Qiime tools import* con el fin de importar las secuencias al entorno bioinformático. Este proceso permite obtener un resumen de las secuencias donde se proporcionará información visual acerca de la distribución de las muestras donde se podrá observar el coeficiente Phred por base (Jiménez & Ph, 2021). Esto con el fin de saber cómo se deben filtrar los datos en el paso posterior.

7.2.2.4 Eliminación de ruido de secuencia

QIIME 2 ofrece la eliminación de ruido través de diferentes softwares bioinformáticos. En este caso se hizo uso de DADA2, ya que fue el que se dio en la línea de código entregada por el instituto. DADA2 es un paquete de software que corrige errores de amplicones que han sido secuenciados mediante Ilumina. Este software nos permite obtener 3 archivos que se mencionan a continuación (Callahan et al., 2016):

Table.qza: Es un artefacto de QIIME 2 que contiene la frecuencia de cada secuencia única en cada muestra del conjunto de datos.

Secuencias representativas.qza: Es un artefacto que asigna identificadores de características en cada una de las secuencias que se encuentran con mayor abundancia. En este caso cada una de las muestras recibe un hipervínculo que dirige a la página del NCBI para poder realizar un alineamiento informático a través de BLAST con el fin de identificar a qué microorganismo o parte de la célula corresponde la muestra.

Estadísticas.qza: Es un artefacto que nos permite saber la cantidad de secuencias existentes antes del proceso de filtrado y la cantidad de secuencias que fueron filtradas. Nos permite saber qué datos son los que quedan para ser analizados.

7.2.2.5 Clasificación taxonómica

Para asignar la taxonomía a cada una de las secuencias presentes se debe tener una base de datos de referencia. En este caso se hizo uso de SILVA ya que es una base de datos que nos provee información de ARN ribosómico y se usó en este caso para análisis de 16S. En este caso se obtuvo un archivo que contiene la asignación taxonómica de cada uno de los microorganismos respecto a cada muestra y adicional se generó un archivo de visualización en gráfico de barras a través de la función *qiime taxa barplot*. Este proceso se hizo adicionalmente comparando los datos con las bases de datos PR2 y con SILVA dirigida a 18S. Con el fin de comparar los resultados y observar la variación de las comunidades microbianas en las muestras.

7.2.3 Modelado metabólico para el análisis de la remediación de cromo (VI)

El modelado metabólico se realizó a partir de una representación de diferentes vías metabólicas intracelulares o extracelulares que usan los microorganismos para la remediación de cromo (VI). Este proceso se realizó a través de la plataforma online BioRender, que permite realizar ilustraciones de diferentes procesos metabólicos. Este proceso se realizó con el fin de comprender los factores que influyen en la remediación.

7.3 Desarrollo del objetivo n°3

- Diseñar un algoritmo basado en Machine Learning, a partir de perfiles taxonómicos y funcionales, para la identificación de comunidades microbianas con potencial de biorremediación en ríos contaminados con cromo (VI).

En esta sección se llevó a cabo el diseño del algoritmo de machine learning donde se buscó información sobre este tipo de modelos y se generó un proceso de selección para aplicar en este proyecto.

7.3.1 Búsqueda bibliográfica sobre algoritmos de machine learning empleados para el análisis de comunidades microbianas

Se realizó una búsqueda bibliográfica de las aplicaciones que han tenido los algoritmos de machine learning sobre diferentes comunidades microbianas. Algunas de las bases de datos consultadas fueron: Science Direct, Springer Link, Plos One, Nature y PubMed.

La información recolectada se fue almacenando en un archivo Excel teniendo en cada columna lo siguiente: *Publicación, año de publicación, tipo de muestra, descripción y link de consulta*.

7.3.2 Selección de modelo de machine learning a emplear

Para la selección del modelo de machine learning se realizó una matriz de decisión, con el fin de determinar de acuerdo a ponderación el algoritmo tuviera mayor cumplimiento de las siguientes condiciones:

- Capacidad de clasificar datos binarios.
- Capacidad para trabajar con datos de alta dimensión.
- Complejidad de implementación.
- Obtención de resultados en términos probabilísticos.

7.3.3 Etiquetado de datos de entrada

Se realizó un proceso de etiquetado con los datos que se usaron para el entrenamiento del modelo. Estos datos corresponden a microorganismos capaces de remediar cromo (VI) identificados en bibliografía.

Se determinaron las etiquetas con un uno (1) y un cero (0) donde se encontrarían los microorganismos con un mayor potencial y un menor potencial para remediar, respectivamente. Teniendo en cuenta el porcentaje de remediación de cada microorganismo se decidió etiquetar como microorganismos con bajo potencial los que contaron con un porcentaje de remediación menor a 75%. Para tomar esta decisión se tuvo en cuenta la cantidad de datos presentes para hacer el análisis y los porcentajes reportados en bibliografía (Ayele, A., & Godeto, 2009; Ayele & Godeto, 2021), ya que la mayoría de ellos se encuentra en un rango de 30% (Zhang, 2012) a 99% (Ayele & Godeto, 2021).

7.3.4 Desarrollo del modelo de machine learning basado en árboles de decisión

7.3.4.1 Requerimientos del diseño

- El algoritmo debe tener como datos de entrada un archivo *.xslx* con la asignación taxonómica de las muestras y la descripción metabólica (funcional) de los genes presentes o ausentes en cada una de las muestras.
- La salida del algoritmo debe ser un archivo *.xlsx* con la asignación taxonómica de cada muestra y la predicción otorgada por el modelo.
- La cantidad de muestras de entrada con la información taxonómica y funcional debe ser la misma.
- Los datos a usar en el algoritmo deben provenir del mismo proceso bioinformático en Kbase para la obtención de las características funcionales y taxonómicas.

7.3.4.2 Condiciones para el desarrollo del árbol de decisión

En esta sección se determinaron los parámetros a tener en cuenta para el modelo, además de ello se estructuró como se realizaría el entrenamiento y validación del modelo. La explicación se puede observar con mayor detalle en la figura 6.

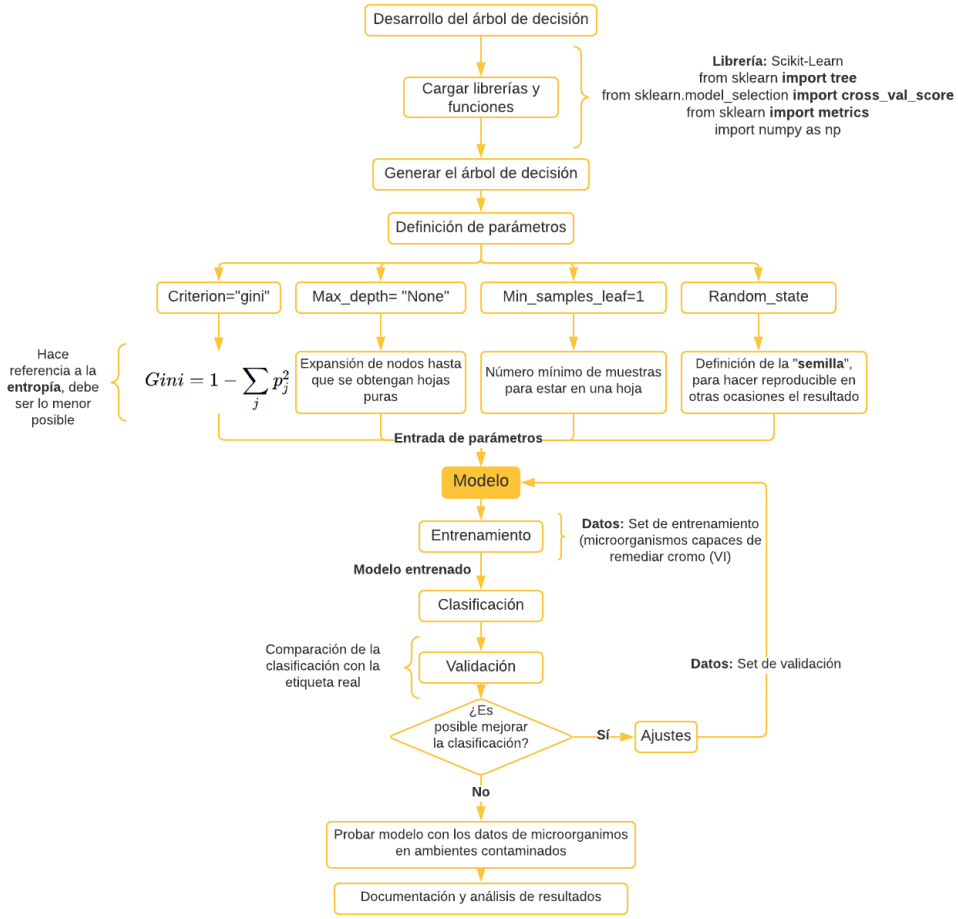

Figura 6. *Metodología para el desarrollo del árbol de decisión. Autoría propia (2021)*

7.4 Desarrollo del objetivo n°4

- Implementar un algoritmo que identifique factores microbianos, a nivel taxonómico y funcional con potencial de biorremediación de cromo (VI).

7.4.1 Implementación del algoritmo de árbol de decisión

Una vez se determinó el modelo de machine learning a usar se comenzó a implementar el código a través de Google Colabolatory. Esta es una plataforma de software libre que permite la creación de modelos de machine learning a través de cuadernos o "notebooks" basados en el lenguaje de programación de Python (Bisong, 2019). Se realizó de esta manera para que la programación pudiera ser compartida fácilmente, ya que no es necesario tener los programas cargados en el computador.

Adicional, la programación en notebooks permite organizar las líneas de código a través de bloques de ejecución. Lo que hace que se pueda explicar el proceso que se está haciendo en cada sección, de esta manera se explicó el paso a paso en el código.

7.4.2 Preparación y análisis de datos

Posterior al análisis bioinformático se obtiene una serie de tablas con información referente al metabolismo de los microorganismos que son analizados. Los datos usados para el desarrollo del algoritmo fueron:

- Tabla con los genes funcionales asociados al metabolismo de los microorganismos.
- Tabla con la asignación taxonómica de cada microorganismo.
- Tabla con las etiquetas para el entrenamiento del modelo.

Esta información tuvo que pasar por un preprocesamiento donde se eliminaron los datos duplicados y se traspusieron con el fin de que los genes se ubicaran en las columnas y los microorganismos en las filas.

El procedimiento se puede observar de manera detallada en la figura 7.

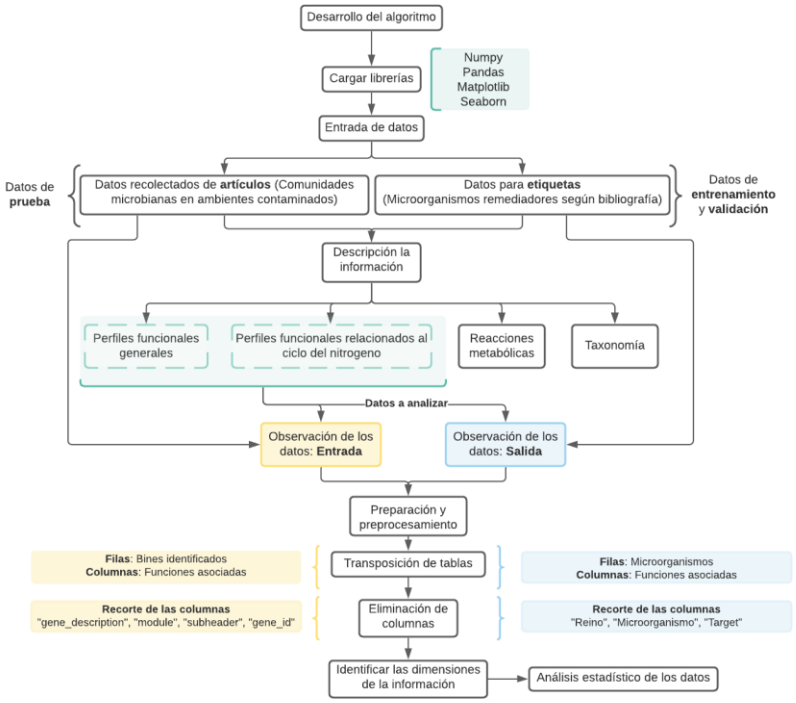

Figura 7. *Metodología para el desarrollo del objetivo n ° 4. Autoría propia (2021)*

Al finalizar el preprocesamiento de los datos, se realizó la observación de ellos con el fin de saber si la información seguía algún patrón que nos pudiera ayudar en el desarrollo del algoritmo. Este proceso se realizó a través de la función "*.describe()*" de pandas en Python que nos permite obtener un resumen estadístico de los datos. En este caso, al contar con datos

categóricos lo que se esperaba es que esta función nos permitiera obtener los valores más comunes entre las muestras y la frecuencia de los valores más comunes acuerdo a la documentación de pandas.

7.4.3 Selección y evaluación de características

La selección de características se realizó teniendo en cuenta los perfiles funcionales obtenidos a partir del análisis bioinformático. Para la selección de características se usaron dos métodos: *i)* Selección de características a partir del coeficiente *Chi2*. *ii)* Selección de características usando bosques aleatorios en Scikit-Learn.

7.4.3.1 Selección de características usando Chi2

Para generar una selección de características de manera objetiva se decidió usar la librería *feature selection* de Scikit-learn, que nos permite obtener una serie de características de acuerdo a la relación de los datos de entrada con los datos de salida (Pedregosa et al., 2011).

Se realizó una selección de características basada en pruebas estadísticas univariadas, esto permite seleccionar las mejores para implementarlas en un modelo. Existen diversos métodos para la selección de características, cada uno se usa de acuerdo a naturaleza de los datos.

En este caso se trabaja con datos categóricos (1 y 0) en un modelo de clasificación, por lo que se decidió usar el método: *Chi2* (Pedregosa et al., 2011). Este método permite saber cuál es la relación entre una categoría independiente y una dependiente, donde *si* la dependencia es mayor se cuenta con una característica que se podría usar para el entrenamiento de un modelo (Cardona et al., 2020).

El coeficiente de Chi busca encontrar la desviación entre el valor real y el valor teórico, mientras mayor sea el valor de Chi, más significativa es su relación con la salida (Cai et al., 2021). Esta estadística se puede definir como la siguiente ecuación:

$$X^2 = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad \text{Ecuación 5}$$

Donde N corresponde al número total de muestras; A es el número de muestras que contienen una característica (presencia de gen) y pertenecen a la clase remediadora; B es el número de muestras que contienen una característica (presencia de gen), pero no pertenecen a la clase remediadora; C es el número de muestras que no cuentan con una característica, pero pertenecen a la clase remediadora; D es el número de muestras que no contienen la característica específica y tampoco pertenecen a la clase remediadora.

Posterior a la selección de características, realizó la evaluación de ellas a través de una matriz de covarianza. La covarianza es una medida de cuánto varían dos variables aleatorias juntas y puede ser expresada como se observa en la siguiente ecuación (Nikolai, 2018):

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \qquad \text{Ecuación 6}$$

Donde el set de datos pertenece a números reales. En las diagonales de la matriz se obtendrá como resultado la variación de una característica en específico respecto a las muestras presentes.

7.4.3.1 Selección de características usando bosques aleatorios (Scikit-learn)

El modelo de bosques aleatorios además de brindar la solución a problemas de clasificación y/o regresión, permite generar la selección de características. Los bosques aleatorios están formados por un rango de 4 a 1200 árboles de decisión y cada uno de ellos realiza una extracción aleatoria de características, donde no todos tienen en cuenta las mismas características. Por lo tanto, este modelo permite seleccionar características de acuerdo a la pureza en cada uno de los grupos (Linusson & Olausson, 2012). Para usar este modelo se deben seguir estos pasos:

1. Importar librerías de *Scikit-learn*

Librerías: *Pandas, RandomForestClassfier, SelectFromModel*.

2. Implementar el modelo de bosques aleatorios con un set de entrenamiento.
3. Usar la librería *SelectFromModel* para seleccionar automáticamente las características de acuerdo al resultado del modelo.
4. Identificar cuáles son las características que tienen un mayor peso sobre la clasificación a través de la función *get_support ()*.

Esta función permite obtener una matriz con booleanos, donde se representarán con la etiqueta "*True*" y "*False*" las características que podrían tenerse en cuenta o no para la clasificación.

5. Realizar una lista con los nombres de las características identificadas con la etiqueta "*True*".

De esta manera se realizó el proceso para la obtención de características e identificar cuáles son relevantes teniendo en cuenta el modelo de bosques aleatorios.

7.4.4 Creación del set de datos (Entrenamiento y prueba)

La construcción del set de datos se realizó teniendo en cuenta los datos provenientes de los microorganismos capaces de remediar cromo (VI). Teniendo en cuenta que se contó con 27 datos para el entrenamiento del algoritmo, se usó una validación cruzada de tipo *Leave-One-Out* (LOOCV). Este método es recomendado cuando se cuenta con un set de datos pequeño (Sammut & Webb, 2010).

En la figura 8 se pueden apreciar los pasos que se siguieron para generar la validación cruzada.

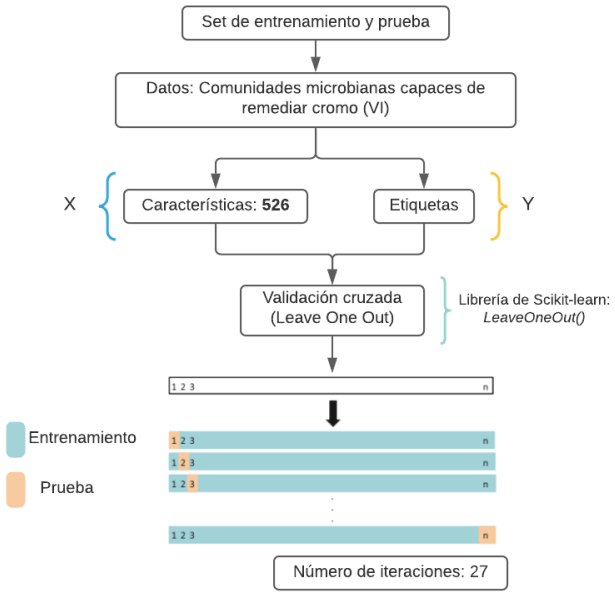

***Figura 8**. Set de entrenamiento y prueba mediante validación cruzada. Autoría propia (2021)*

Este proceso se realizó usando la función de Scikit-learn *LeaveOneOut()* que permitió obtener 27 conjuntos de prueba y validación, en donde cada uno de ellos va dejando una de las muestras afuera para validar el modelo y usa las 26 restantes para entrenarlo.

7.4.5 Entrenamiento del árbol de decisión

Una vez se genera el código del árbol de decisión a través de la función *tree.DecisionTreeClassifier* en Python se debe entrenar el modelo con los datos referentes a las características y etiquetas de los microorganismos. Este procedimiento se realiza por cada iteración del y va ingresando al modelo a través de la función "*dt.fit"* (Pedregosa et al., 2011).

7.4.6 Validación del modelo

Una vez se entrena el modelo es necesario probarlo y realizar una comparación entre el resultado obtenido y el resultado esperado. Este proceso es llamado validación y se realiza a través de la comparación métodos como las matrices de decisión. De esta manera es posible saber el rendimiento del modelo entrenado. En la figura 9 se muestra con detalle el proceso que se siguió.

***Figura 9**. Metodología para la validación del modelo. Autoría propia (2021)*

Para medir el desempeño del modelo a través de las matrices de confusión se usaron las métricas de exactitud, precisión y sensibilidad, para ello se tuvieron en cuenta las siguientes ecuaciones (Ferreira, 2018).

$$precisión = \frac{VP}{VP+FP} \qquad \textit{Ecuación 7}$$

$$Exactitud = \frac{VP+VN}{VP+FP+FN+VN} \qquad \textit{Ecuación 8}$$

$$Sensibilidad = \frac{VP}{VP+FN} \qquad \textit{Ecuación 9}$$

Teniendo en cuenta que: Verdadero positivo (VP); Verdadero negativo (VN); Falso positivo (FP); Falso negativo (FN)

Adicional, se tuvo en cuenta la métrica *F1-Score* nos indica el promedio ponderado entre la precisión y la sensibilidad (Solutions, 2016). Por lo tanto, en ella se tienen en cuenta los falsos positivos y los falsos negativos. La fórmula que la describe se muestra a continuación,

$$FScore = \frac{2*Precisión*Sensibilidad}{Precisión+Sensibilidad} \qquad \textit{Ecuación 10}$$

Esta métrica suele de ser de gran ayuda especialmente si se tiene una distribución de clases desbalanceada.

7.4.7 Comparación del modelo de árbol de decisión con otros modelos de machine learning

Se realizó una comparación del modelo de árbol de decisión diseñado con un modelo bosques aleatorios con el fin de identificar y comparar si se obtienen mejores resultados con alguno

de estos métodos. La implementación de cada uno de los modelos se realizó empleando las características previamente seleccionadas.

- *Bosques aleatorios (RF)*

La implementación del modelo de bosques aleatorios se realizó a través de la función *RandomForestClassifier* de Scikit-learn. Este método se basa en los árboles de decisión, sin embargo, los RF usan varios árboles de decisión en su código con el fin de mejorar la precisión predictiva y evitar el sobreajuste (Pedregosa et al., 2011).

7.4.7 Implementación del modelo entrenado con la información de comunidades microbianas presentes en ríos contaminados con cromo (VI)

La implementación del modelo seleccionado se realizó con la información perteneciente a comunidades microbianas en ríos contaminados con cromo (VI). Estos datos son nuevos para el algoritmo ya que no intervinieron en el proceso de entrenamiento.

Se generó un archivo de Excel a través de la programación con el fin de poder observar cada especie con su respectiva predicción del potencial de remediación.

8. RESULTADOS Y DISCUSIÓN

8.1 Resultados del objetivo n°1

8.1.1 Recolección de información de genomas de microorganismos en ambientes contaminados con Cr (VI).

Para la selección de los artículos que fueron usados en el proyecto se tuvo en cuenta que en ellos estuvieran los códigos de acceso del NCBI o MG-Rast para poder descargar las secuencias. Además de ello, se escogieron los artículos que presentaban una relación con la contaminación de metales pesados en suelos y fuentes hídricas o que se relacionaran con muestras tomadas de efluentes de curtiembres.

Finalmente, en la búsqueda se encontraron 50 artículos relacionados con comunidades microbianas presentes en ríos y suelos contaminados con cromo (VI) con datos disponibles para su uso libre.

En la figura 10 se puede apreciar la distribución de la información obtenida, donde el 71.2% de la información corresponde a información de amplicón 16s rRNA; el 17.3% a información metagenómica; el 7.7% a información de librería de clones y un 1.9% a información de proteínas y amplicón 18s rRNA. La documentación se puede encontrar en el *Anexo 1*.

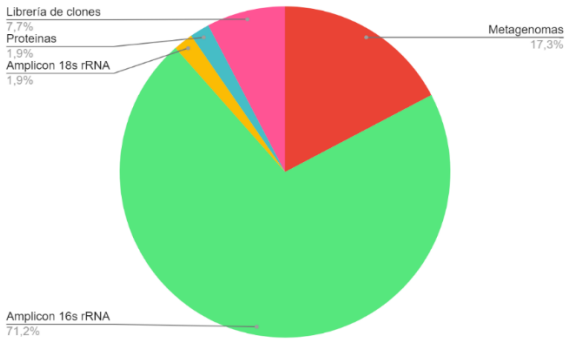

Figura 10. *Distribución de la información recolectada. Autoría propia (2021)*

La información que se empleó para hacer el análisis bioinformático mediante Kbase se recuperó a partir de una serie de artículos titulados: i) *Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes* (Mittal et al., 2019) ii) *Wheel genome shotgun sequencing of POPs degrading bacterial community dwelling tannery effluents and petrol contaminated soil* (Muccee & Ejaz, 2020) iii) *NGS-based characterization of microbial diversity and functional profiling of solid tannery waste metagenomes* (Verma & Sharma, 2020). De estos artículos se tomaron las siguientes muestras de datos metagenómicos: SRR9721663-1, SRR9721664-1, SRR10083574-1, SRR8870488-1.

Entre los microorganismos remediadores que se encontraron en bibliografía están: *Ochrobactrum anthropi* (Vélez et al., 2021), *Pseudomonas aeruginosa, Desulfovibrio vulgaris, Klebsiella pneumoniae* (F. Aslam et al., 2020), *Schizosaccharomyces pombe, Desulfovibrio desulfuricans* (Joo et al., 2015), *Enterobacter cloacae* (Bhattacharya et al., 2019), *Bacillus cereus, Escherichia coli, Bacillus subtilis* (Jin et al., 2017), *Bacillus megaterium, Pseudomonas putida, Staphylococcus aureus* (Kalsoom et al., 2021; Tariq et al., 2019), *Bacillus sp.* (Bhattacharya et al., 2019), *Acinetobacter haemolyticus* (Zakaria et al., 2007), *Staphylococcus capitis, Pseudomonas fluorescens, Aspergillus nidulans, Enterococcus casseliflavus, Serratia marcescens* (Campos et al., 2005), *Enterococcus gallinarum, Aspergillus flavus, Ochrobactrum intermedium, Penicillium rubens wisconsin, Aspergillus niger* (Y. Gu et al., 2015) *y Trichoderma harzianum* (Mala et al., 2020).

8.2 Resultados del objetivo n°2

8.2.1 Análisis bioinformático de los datos metagenómicos mediante Kbase

En esta sección, se realizó el análisis bioinformático de los datos obtenidos previamente con el fin de determinar las características funcionales y taxonómicas de los microorganismos.

En total se analizaron 4 muestras de metagenoma (SRR9721663-1, SRR9721664-1, SRR10083574-1, SRR8870488-1) a través de Kbase y 37 muestras de amplicón 16s rRNA a través de QIIME2.

8.2.2 Preprocesamiento de los datos recolectados.

8.2.2.1 Calidad de las secuencias a analizar

De este proceso se obtuvieron gráficas de calidad como por ejemplo la que se aprecia en la figura 11. En este caso no se filtraron los datos de entrada porque la calidad de secuencias por base se encontraba en un rango de 32 y 40 lo que significa que se trataba con secuencias con una precisión del 99.9%.

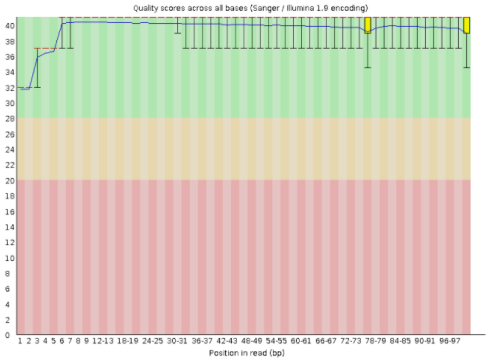

Figura 11. *Resultado de calidad con FastQC. Autoría propia (2021)*

A comparación de la figura 12 se puede observar que la calidad de secuencias por base es variable y llega incluso a un coeficiente de Phred de 14. A pesar de que este valor puede significar una precisión del 93% (GATK team, 2021) se decidió filtrar estas secuencias con el fin de disminuir el error.

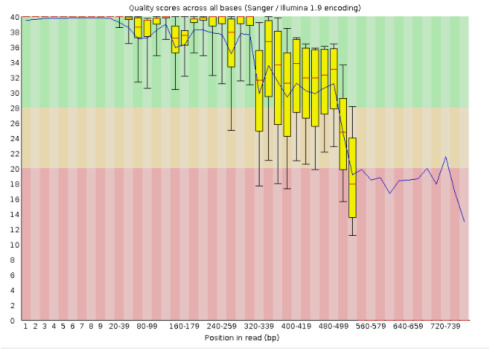

Figura 12. *Resultado de FastQC antes de recortar las secuencias. Autoría propia (2021)*

Para este proceso se determinó promediar la lectura de 4 bases nitrogenadas y filtrar en caso de que se tuviera un coeficiente de Phred menor a 15. Se obtuvo la gráfica que se aprecia en la figura 13.

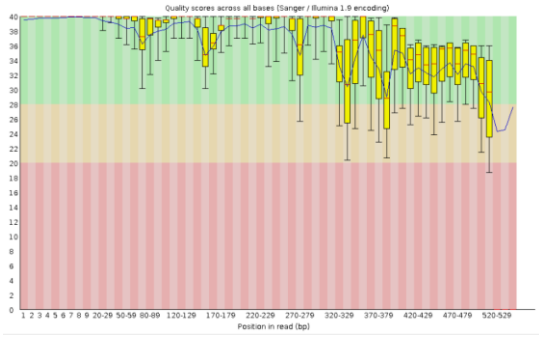

Figura 13. *Resultado de FastQC después de recortar las secuencias. Autoría propia (2021)*

Este procedimiento se llevó a cabo únicamente con la secuencia SRR10083574-1, ya que las demás presentaron una calidad de secuencias por base superior a 15.

8.2.2.2 Generación de Bin Contigs

Luego de agrupar y optimizar los bin contigs se obtuvo el diagrama que se observa en la figura 14.

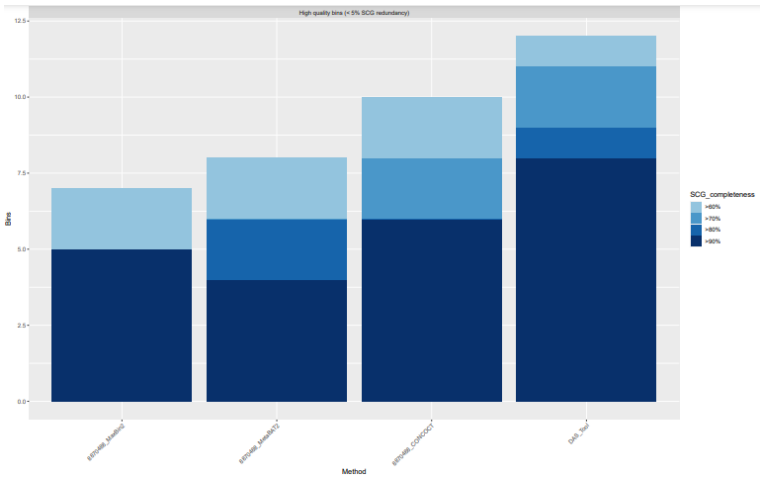

Figura 14. *Optimización de bin contigs mediante DAStool (Versión 1.2). Autoría propia (2021)*

Una vez se obtuvieron los genomas de alta calidad con baja contaminación (< 5%) (Sieber et al., 2018) se observó el resultado. Teniendo en cuenta la gráfica de barras se seleccionaron los contigs agrupados a través de *DAS Tool* porque nos permitió obtener 11 bines de los cuales 7 contaban con más del 90% de información completa para analizar, 1 con más del 80% y los 3 restantes con menos del 70%.

8.2.2.3 Prueba de calidad en los bins

El resultado de calidad de los bines agrupados por *DAS Tool* se puede observar en las figuras 15 y 16.

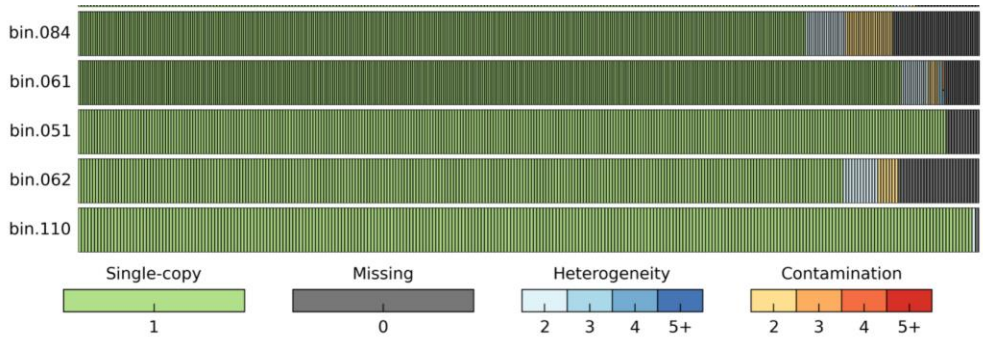

Figura 15. *Resultado de Assess Genome Quality with CheckM (Versión 1.0.18). Autoría propia (2021)*

En la figura 15 se pueden observar las 5 muestras con el menor porcentaje de contaminación. Las muestras se indican a continuación con el porcentaje mencionado: Bin.084. (10.03%), Bin. 061 (5.41%), Bin.051(0%), Bin.062 (8.2%), Bin.110. (0.27%). Además de esto, las estadísticas de *CheckM* indican que las secuencias se encuentran completas en un 92.73%, 94.74%, 95.08%, 95.16% y 99.46%, respectivamente. Este valor se determina a partir de una serie de genes marcadores que permiten identificar la integridad de las secuencias. Las barras verdes representan los marcadores que fueron identificados correctamente una vez, mientras que las barras en gris representan los marcadores faltantes (Parks et al., 2015). Teniendo en cuenta estos datos, es posible decir que la muestra del Bin.110 es pura ya que contiene un alto porcentaje de marcadores identificados y bajo porcentaje de contaminación.

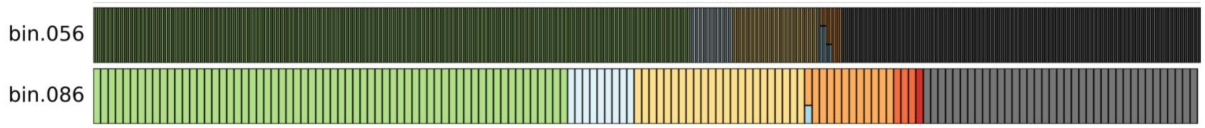

Figura 16. *Resultado de Asses Genome Quality: Contaminación. Autoría propia (2021)*

A diferencia de la figura anterior, en la figura 16 se pueden observar dos de las muestras con mayor contaminación. Las muestras se indican a continuación con mencionado: Bin.056 (17.19 %) y el Bin.0.86 (31.03 %), además los porcentajes de la cantidad de marcadores detectados es de 68.49% y 69.55%, respectivamente. Las barras rojas o de color naranja hacen referencia a la contaminación en la secuencia y se da cuando se encuentra un marcador de genes tienen una menor similitud de aminoácidos en la muestra (AAI) (Medlar et al., 2018). En caso contrario, cuando la AAI ≥ 90% se representa a través de barras azules. Esto quiere decir que en la muestra se encuentran marcadores repetidos lo que puede significar que fragmentos de genoma de otro microorganismo se encuentren en la muestra (Parks et al., 2015). Por lo tanto, si se observa el Bin.086 se puede identificar que tiene un alto porcentaje de contaminación y de heterogeneidad. Esto puede hacer que se clasifique en un microorganismo de manera errónea.

8.2.2.4 Extracción de bines

De acuerdo a la evaluación de calidad en los bines se determinó conservar los que no presentaran un porcentaje de contaminación mayor al 10% ni un valor de heterogeneidad mayor a 3. Además de esto, se tuvo en cuenta que la detección de marcadores de genes en la muestra fuera mayor al 70%. Teniendo en cuenta estos criterios los bines extraídos y usados para el análisis funcional y taxonómico fueron: Bin.001.10083574, Bin.003.10083574, Bin.004.10083574, Bin.001.9721664, Bin.003. 9721664, Bin.004.8870488, Bin.005.8870488, Bin.012.9721663, Bin.012.8870488, Bin.051.8870488, Bin.059.8870488, Bin.061.8870488, Bin.104.8870488 y Bin.110.8870488. Las secuencias extraídas se identifican a través del número de bin seguido de la muestra metagenómica a la que pertenecen.

8.2.3 Análisis funcional y taxonómico: microorganismos en ambientes contaminados con cromo (VI)

En esta sección se realizó un análisis de los perfiles funcionales y taxonómicos obtenidos a partir de la información de artículos asociada a comunidades microbianas en ambientes contaminados con cromo (VI).

8.2.3.1 Análisis funcional y taxonómico general

A través del análisis bioinformático realizado mediante Kbase se obtuvo la clasificación taxonómica de los microorganismos presentes en ambientes contaminados. En la plataforma fue posible obtener el: filo, clase, orden, familia, género y especie de algunos microorganismos presentes en las muestras de estudio.

De los artículos encontrados se seleccionaron 4 que permitieron obtener la información metagenómica para el análisis. Cada uno de los artículos tenía diferentes muestras (SRR9721663-1, SRR9721664-1, SRR10083574-1, SRR8870488-1), a continuación, se explica de manera detallada la información sobre cada una de ellas. En figura 17 se pueden observar las muestras y el phylum de las especies analizadas.

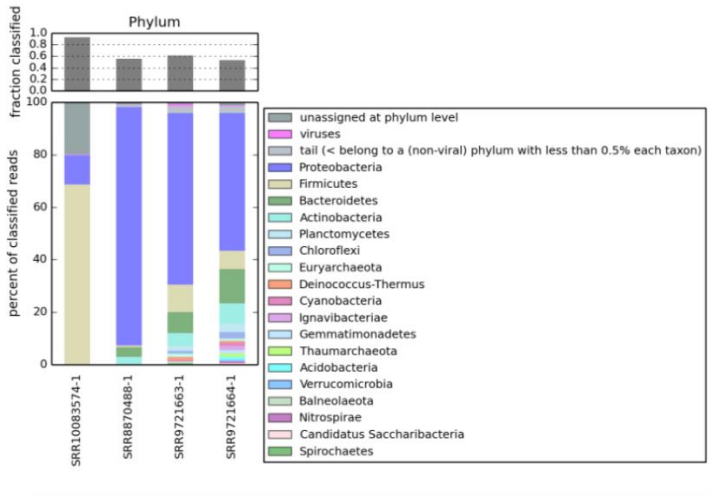

Figura 17.*Perfiles taxonómicos (phylum) de los microorganismos presentes en la información metagenómica seleccionada. Autoría propia (2021)*

Las muestras SRR9721663-1 y SRR9721664-1 pertenecen al mismo estudio, donde se secuenciaron dos metagenomas pertenecientes a desechos de curtiembres de Jajmau y Unnao en India. De acuerdo al análisis realizado en Kbase se puede observar que predominan los phylum *Proteobacteria*, *Firmicutes*, *Bacteroidetes* y *Acidobacteria*. Algunos de estos phylum se caracterizan por tener bacterias pertenecientes a la microbiota de animales (Thomas et al., 2011) y por permitir el desarrollo de los ciclos biogeoquímicos como por ejemplo, las Acidobacterias (Kalam et al., 2020).

Se encontraron especies que tienen la capacidad de remediar, como lo son *Brevibacillus sp.* (Ruiz-Lozano & Azcón, 2011), *Bacillus licheniformis* (Jamil et al., 2014), *Marinobacter sp.* (Chernikova et al., 2020), *Dechloromonas aromática* (Salinero et al., 2009), *Halomonas sp* (Abdel-Razik et al., 2020), *Methylophilus sp*. (Zhang, SY, Wang, QF, Wan, R. y Xie, 2011) y *Comamonas sp*. (Qurbani & Hamzah, 2020; Rudakiya, 2013). Las especies anteriormente mencionadas tienen la capacidad de remediar metales pesados e hidrocarburos y algunas de ellas son empleadas mediante consorcios microbianos para mejorar la eficiencia de la remediación.

Para la muestra SRR10083574-1 se usaron bacterias aisladas de curtiembres, este estudio tuvo lugar en Pakistán y fue publicado en el año 2020. En el análisis taxonómico se encontró que en esta muestra predominan los phylum *Firmicutes* y *Proteobacteria*. En el phylum *Firmicutes* podemos encontrar miembros del género *Bacillus*, estos han sido microorganismos ampliamente estudiados por su aplicación en la agroecología. Sin embargo, otros miembros de este phylum tienen la capacidad de promover el crecimiento de plantas, controlar patógenos en plantas y fitorremediar metales pesados (Hashmi, I., Bindschedler, S. and Junier, 2020). Los microorganismos con mayor abundancia en este estudio fueron *Bacillus paralincheniformis*, *Brevibacillus agri* y *Burkholderia pseudomultivorans*. Este último género puede llegar a desarrollar relaciones simbiótico-mutualistas, así como relaciones simbiótico-patogénicas. Cuando se desarrolla una relación simbiótico-mutualista por *Burkholderia sp*. puede existir la promoción del crecimiento vegetal, fijación biológica

de nitrógeno, solubilización de fosfatos, nodulación y biorremediación (Espinosa-Victoria et al., 2020).

Finalmente, la muestra SRR8870488-1 fue extraída de un estudio metagenómico realizado en el Río Yamuna. Este es uno de los ríos más contaminados de India producto de químicos tóxicos y metales pesados que son desechados en el río por parte de las industrias (Mittal et al., 2019). En el análisis realizado mediante Kbase de los datos disponibles de este estudio, se encontró que el phylum que tiene mayor abundancia en la muestra es Proteobacteria. Este phylum se caracteriza por ser uno de los más diversos ya que contiene desde patógenos oportunistas como *Escherichia coli*, *Salmonella* y *Campylobacter* (Moon et al., 2018) que afectan la salud hasta microorganismos importantes para la mineralización del suelo como *Acinetobacter* (Adewoyin & Okoh, 2018).

Adicional, en el análisis se observó la abundancia de los phylum Bacteroidetes y Acidobacteria. Entre los phylum anteriormente mencionados se encontraron géneros abundantes como *Acinetobacter sp.*, *Brevibacillus sp.*, *Pseudomonas sp.*, *Aeromonas sp*. y *Bacillus sp*. Estos cuatro últimos géneros se han relacionado en diversos estudios por su resistencia a metales pesados (Fakhar et al., 2020). En la figura 18 se encuentran las especies que se obtuvieron en el análisis taxonómico.

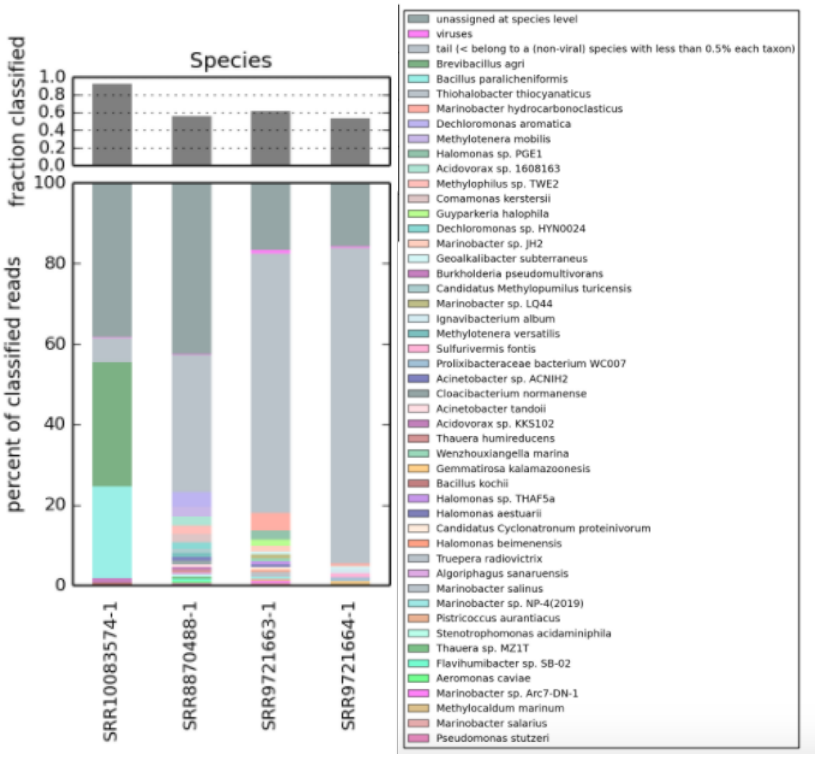

Figura 18. *Microorganismos identificados en la información metagenómica recolectada. Autoría propia (2021)*

Posterior al análisis taxonómico de las muestras, se identificaron los perfiles funcionales de las especies encontradas, esto con el fin de saber que comportamiento tienen y si muestran algún mecanismo que pueda asociarse a la remediación de cromo (VI).

La figura 19 nos permite observar qué papel metabólico están cumpliendo los diferentes microorganismos de las muestras. Los tonos van de azul oscuro hasta amarillo pálido, indicando la cantidad de pasos involucrados en el proceso metabólico, siendo el 100% de estos hasta el 0%, respectivamente. Se puede apreciar que se presentaron 2 muestras (Bin. 002.Fasta9721664 y Bin. 003.Fasta9721664), que no mostraron el comportamiento de las demás. Esto pudo ocurrir porque esos bins no tenían una información tan completa como en los otros casos, por lo que no se puede hacer un análisis tan riguroso en ellos. Exceptuando las muestras mencionadas anteriormente, en este resultado podemos apreciar que se presentan vías metabólicas en común como la glucólisis, vía pentosa-fosfato y el ciclo de Calvin. La glucólisis representa una de las diversas vías por las cuales las bacterias degradan la glucosa y está asociada comúnmente con el metabolismo anaeróbico o fermentativo de bacterias o levaduras (Wolfe, 2015) por lo que se puede inferir que en estas muestras hay presencia de este tipo de microorganismos. Adicional a ello, la presencia de procesos relacionados con el ciclo de Calvin nos permite identificar que los microorganismos presentes requieren de la fijación de carbono para la obtención de energía (Dijkhuizen & Harder, 1984).

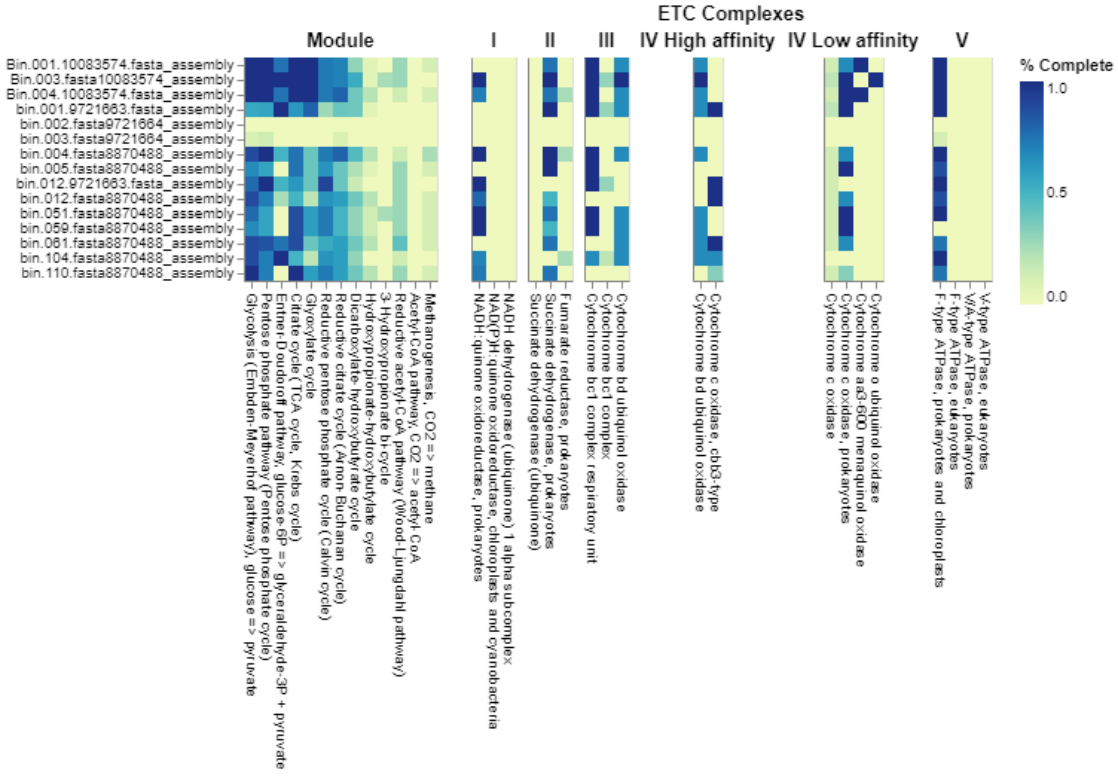

Figura 19. *Perfiles funcionales relacionados con las cadenas de transporte de electrones de las especies presentes. Autoría propia (2021)*

En los resultados se presentan ATPasas de tipo F, siendo estas las ATP sintasas de translocación de H+ más comunes y pueden encontrarse en las membranas de bacterias, cloroplastos y mitocondrias (Ozawa et al., 2000). Estas son importantes para la generación de ATP usando energía a partir del gradiente de protones, se crea a partir de la cadena respiratoria o complejos fotosintéticos (Doucet, 2008).

Existen además las ATPasas de tipo P, a pesar de que no se pueden visualizar en el resultado, son importantes ya que permiten controlar funciones vitales de la célula como la contracción muscular, el potencial de membrana y la señalización (Mathur, 2017). Además, se ha encontrado que tienen relación con los mecanismos de amortiguación a la concentración de metales pesados. Este mecanismo se presenta en *Staphylococcus aureus* lo que le confiere resistencia al cadmio (Ueno et al., 2000) y podría relacionarse también a la capacidad que tiene esta especie para la remediación de cromo (VI) (Kalsoom et al., 2021; Mathur, 2017) .

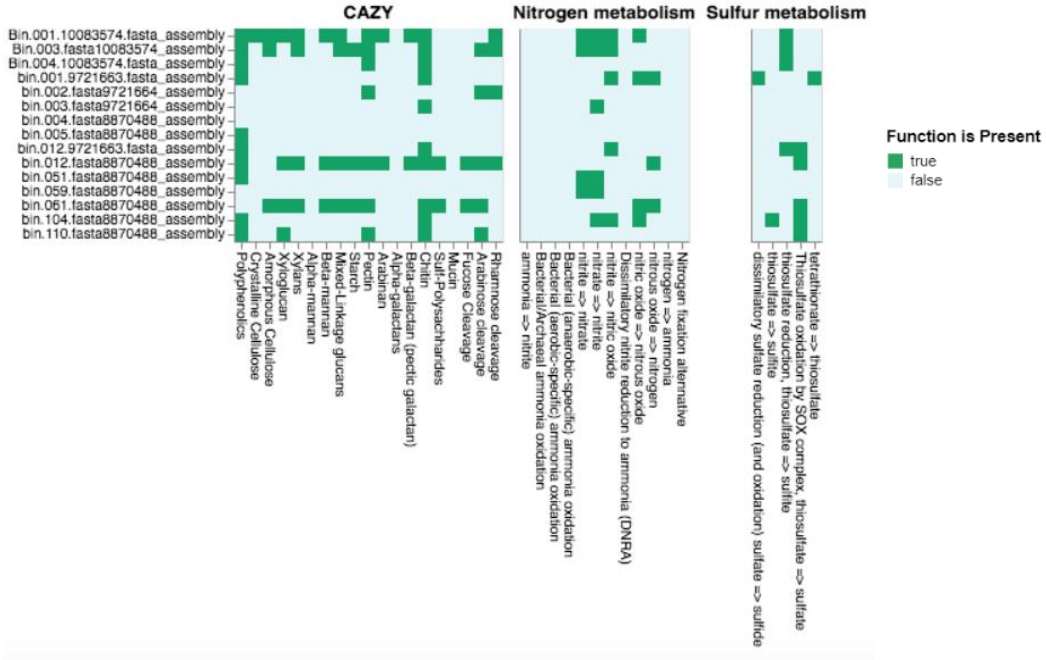

Figura 20. *Perfiles funcionales relacionados con la presencia de enzimas en diferentes procesos metabólicos. Autoría propia (2021)*

En el caso de la figura 20 se observa el resultado de las enzimas presentes en las muestras, allí se aprecian cuadros de color verde que indican la presencia de genes asociados a enzimas activas en carbohidratos (CAZY); la descripción de ellas se encuentra en la parte inferior del diagrama. Las enzimas que se presentan con mayor frecuencia entre las muestras están relacionadas con la degradación de quitina y polifenoles. Se ha encontrado a través de estudios que los productos de degradación de fenoles pueden ser usados como donantes de electrones para la reducción de cromo (VI) (Chen & Tian, 2021) por lo que los microorganismos que presentan esta característica podrían asociarse como potenciales remediadores. No se ha encontrado que la degradación de quitina se asocie a la remediación;

sin embargo, las bacterias pueden producir estas enzimas para suministrarse de nitrógeno y carbono como fuente de nutrientes. Además, juegan un papel importante en la patogénesis bacteriana si su huésped contiene quitina como escudo protector (Rathore & Gupta, 2015). Una de las bacterias quitinolíticas mejor estudiadas es *Serratia marcescens* que también se ha estudiado por su capacidad para degradar cromo (VI) (Campos et al., 2005; Rathore & Gupta, 2015).

En el segundo cuadro de la figura 21 se encuentra la presencia de enzimas asociadas al metabolismo del nitrógeno. Se puede observar que seis de las muestras realizan desnitrificación a través de una enzima denominada nitrato reductasa periplásmica (Nap) que permite catalizar la transformación de nitrato a nitrito en procariotas (Sparacino-Watkins et al., 2014). En el caso del metabolismo del azufre se observa que tres muestras no cumplen esta función, sin embargo, la que se repite en las otras muestras es la oxidación de tiosulfato por medio del complejo SOX. No se encontró información directamente asociada con esta vía metabólica y la remediación. Sin embargo, el metabolismo del azufre es de gran importancia para la disimilación de sulfonatos como fuente de carbono y energía. Algunos microorganismos que también se asocian a la remediación de cromo (VI) usan esta vía metabólica, como *Acinetobacter* y *Pseudomonas*(Wu et al., 2021).

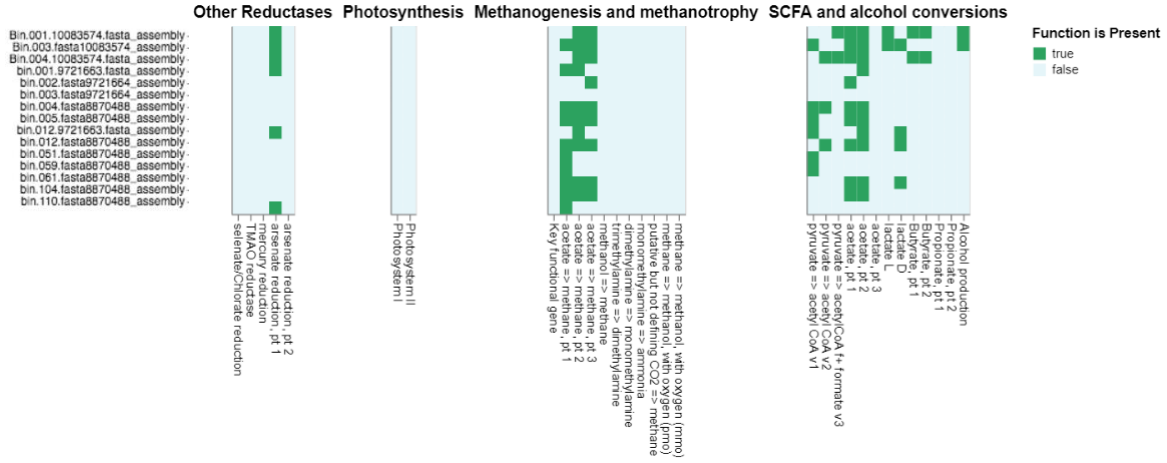

Figura 21. *Perfiles funcionales relacionados con la presencia de enzimas en diferentes procesos metabólicos. Autoría propia (2021)*

En la figura 21 se pueden observar otras características metabólicas como: Reductasas relacionadas a la cadena de transporte de electrones, la presencia de enzimas asociadas a la fotosíntesis, metanogésis y metanotropía y enzimas asociadas a la degradación de ácidos grasos de cadena corta (SCFA) y conversiones alcohólicas.

En los resultados se observa que la mayoría de las muestras cumplen un proceso de reducción de arsenato a través de arsenato reductasas como la glutaredoxina. Este mecanismo se presenta en bacterias resistentes al arsénico como *B. licheniformis, Micrococcus luteus, Pseudomonas fluorescens* y hongos como *Thricoderma sp*. (Upadhyay et al., 2018). Algunos metales pesados como el Cu, Fe (III) y As (III) actúan como mediadores para la aceleración

enzimática para la reducción de cromo (VI) (Z. Rahman & Thomas, 2021). *Micrococcus luteus* ha sido estudiada por su capacidad para remediar cromo (VI) (Katyal & Kaur, 2018); sin embargo, puede cumplir este proceso gracias a la actividad de sus cromato reductasas. Esto demuestra la diversidad de procesos que puede llevar a cabo un solo microorganismo, ya que se adaptan de acuerdo al ambiente en el que se encuentran (Jaiswal & Shukla, 2020).

En el caso de las vías que se siguen para la metanogénesis los resultados nos indican que se sintetiza metano a partir de acetato, estas reacciones son mediadas por fosfato acetiltransferasa, acetato quinasa y Acetil CoA sintetasa. Finalmente, la mayoría de las muestras sintetizan acetato con ayuda del acetato quinasa y fosfato acetiltransferasa. El acetato es conocido por ser un donante de electrones importante para la reducción de cromo (VI); por ejemplo, en microorganismos como *Pannonibacter phragmitetus* aumentó la reducción de cromo (VI) en presencia de acetato, lactato y piruvato (Z. Rahman & Thomas, 2021).

8.2.3.2 Análisis funcional relacionado al ciclo del nitrógeno

Se obtuvieron los perfiles funcionales asociados al ciclo del nitrógeno a partir de la aplicación bioinformática FAMA. En la figura 22 se puede apreciar un gráfico que se realizó a partir de los resultados obtenidos.

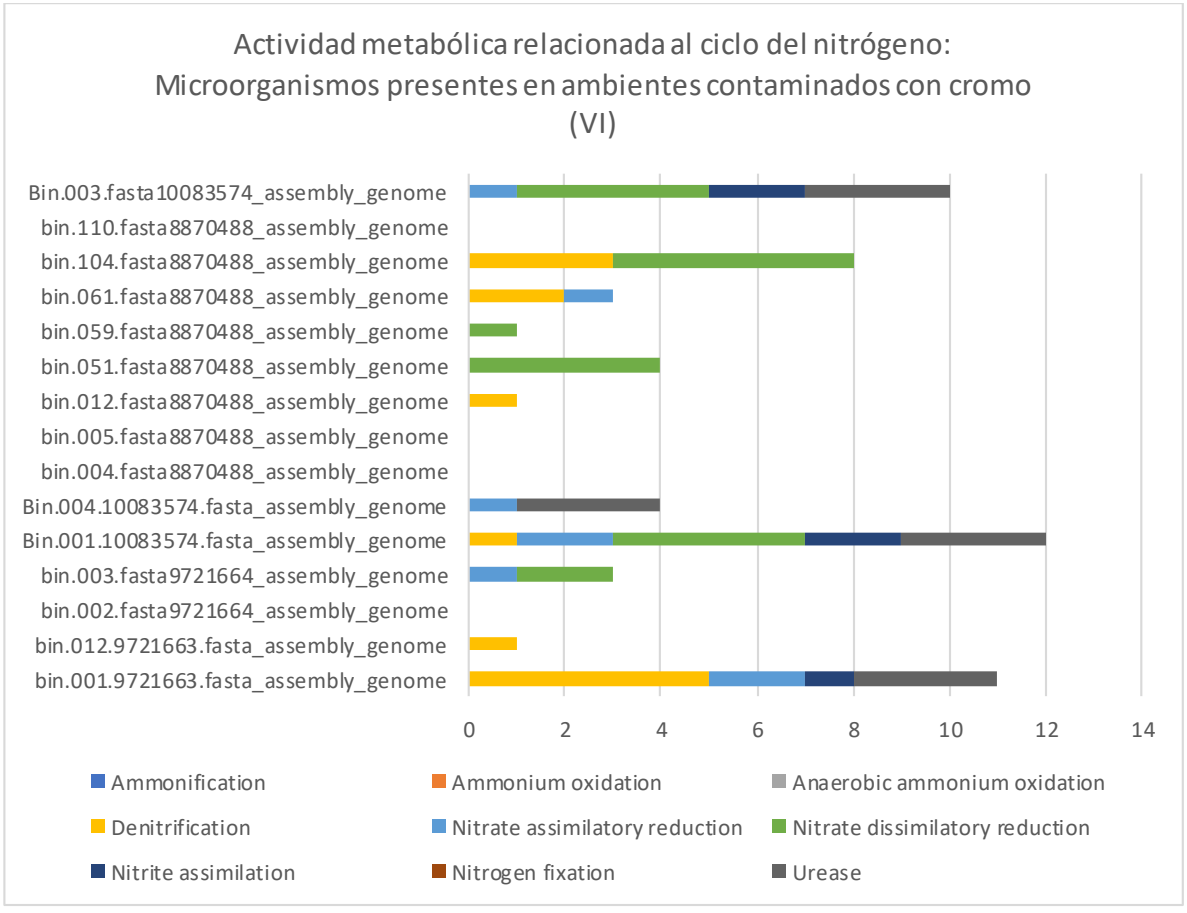

Figura 22. *Vías metabólicas asociadas al ciclo del nitrógeno. Autoría propia (2021)*

Se encontró que en varias muestras se presentan vías metabólicas asociadas a la reducción de nitrato y a la desnitrificación. Esto puede deberse a que las muestras provienen de efluentes de curtiembres. Estas aguas además de contener cromo (VI), tienen altas concentraciones de nitrógeno, especialmente de amonio (Wang et al., 2014).

La desnitrificación se relaciona con las vías catabólicas del carbono, como la degradación de aminoácidos, ácidos grasos y carbohidratos. Esto porque pueden generar donantes para la reducción de nitratos y cromo (VI) (Emmanuel et al., 2019; Z. Rahman & Thomas, 2021; Sul et al., 2016). De acuerdo a un estudio realizado, los productos generados por la degradación de aminoácidos, ácidos grasos y carbohidratos ingresa al ciclo de Krebs donde se genera un agente reductor NADH que se emplea en el proceso de desnitrificación. Teniendo en cuenta esto, es posible que este producto pueda influir en la reducción de cromo (VI), ya que los microorganismos hacen uso de enzimas como: cromato reductasas (ChrA y YieF) (Baldiris et al., 2018), nitroreductasas y flavin reductasas (O'Neill et al., 2020) que son dependientes de la acción del donador NADH.

8.2.4 Análisis funcional y taxonómico: microorganismos capaces de remediar cromo (VI)

8.2.4.1 Análisis funcional y taxonómico general

Se realizó un análisis funcional con los microorganismos que son conocidos por remediar cromo (VI). En la figura 23 se puede observar la información relacionada a las cadenas de transporte de electrones. En este caso se conocen los nombres de los microorganismos y se puede saber qué función cumple cada uno.

En el mapa de calor se definen los colores de acuerdo al porcentaje de cumplimiento de cada vía metabólica que se menciona en la parte inferior. Por ejemplo, la glucólisis en *Staphylococcus capitis* se aprecia un tono azul fuerte, lo que quiere decir que cumple el 100% (9) de las reacciones de esta vía.

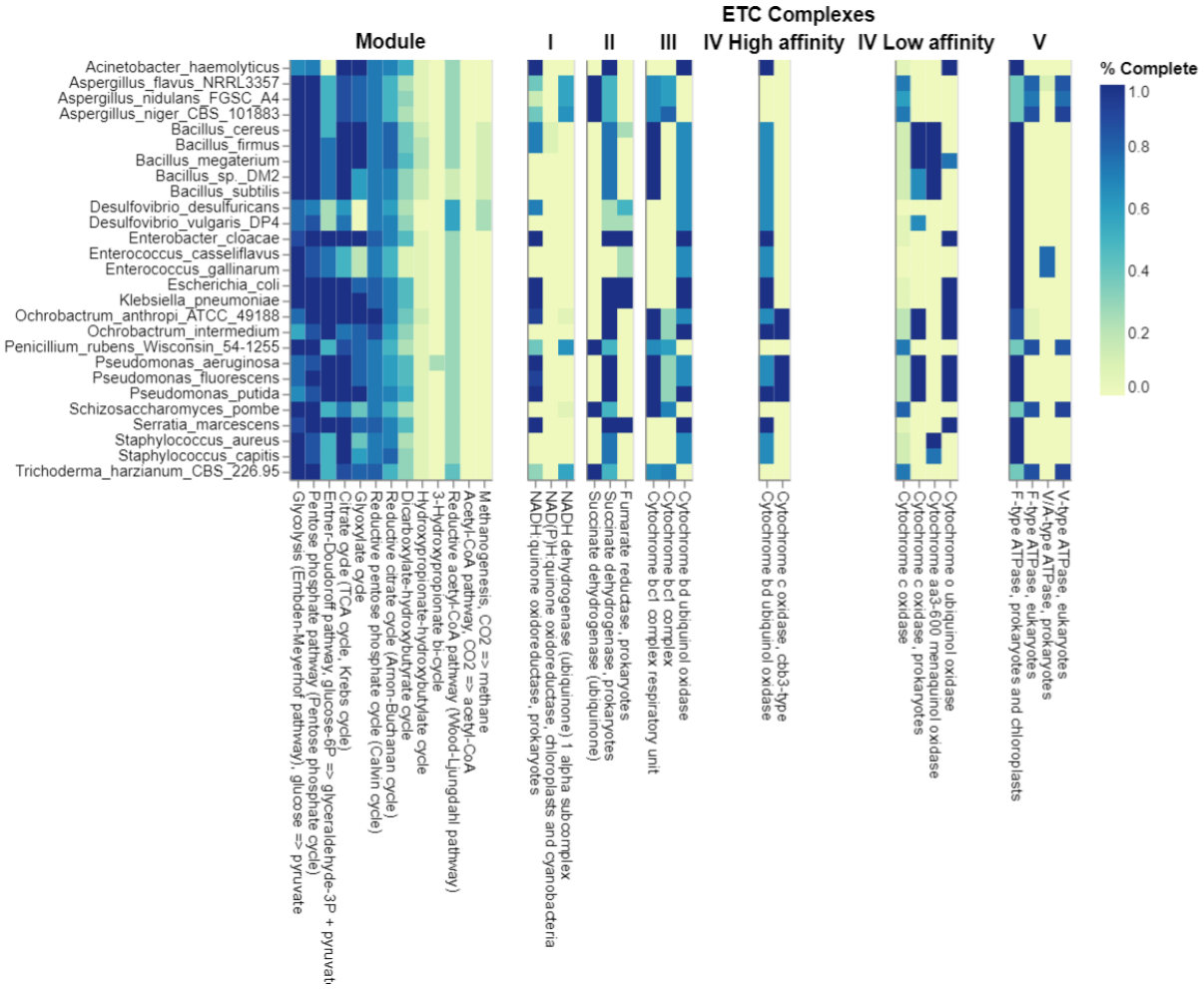

Figura 23*.Perfiles funcionales relacionados con las cadenas de transporte de electrones de los microorganismos conocidos por remediar cromo (VI). Autoría propia (2021)*

Se puede identificar que microorganismos como *Enterobacter cloacae, Klebsiella pneumonioae, Ochrobactrum anthropi* y *Pseudomonas putida* presentan las vías metabólicas del ciclo de Krebs, glioxilato y Entder-Doudoroff. El ciclo de Krebs es una vía catabólica, mientras que el ciclo del glioxilato es anabólico y permite la síntesis de glucosa a partir de ácidos grasos (Berg et al., 2002). A pesar de que estas vías son similares, se generan en compartimentos diferentes de la célula. El ciclo del ácido acético se genera en la mitocondria y el del glioxilato en glioxisomas, pero son procesos que actúan de manera coordinada para la producción de energía (Kevin Ahern et al., 2021). La vía de Entder-Doudoroff se usa por algunas bacterias que no cuentan con enzimas claves para poder realizar la glucólisis, adicional el rendimiento de esta vía es de 1 ATP por cada molécula de glucosa procesada, mientras que en la glucólisis se generan 2 ATP y 2 NADH por cada molécula de glucosa (Vegge et al., 2016). Es posible que estas especies presenten una gran capacidad metabólica ya que son caracterizadas por ser patógenos oportunistas lo que hace necesario la búsqueda de diversas fuentes energéticas cuando no poseen un huésped (Richardson et al., 2015; Ryan & Tony Pembroke, 2020).

En la mayoría de las muestras se presenta el citocromo *bd*, es una oxidasa que se presenta en varios procariotas, incluido en patógenos como *Klebsiella pneumoniae* (Giuffrè et al., 2014). Esta oxidasa permite que los microorganismos puedan conservar su energía y en las bacterias patógenas les permite colonizar en ambientes pobres en O_2 (Borisov et al., 2011). Este factor es importante ya que puede mejorar la tolerancia bacteriana a condiciones de estrés oxidativo (Giuffrè et al., 2014) que se pueden generar por ambientes contaminados como lo son los efluentes de curtiembre.

Así como en la figura 23 la mayoría de los microorganismos presentan ATPasas de tipo F que se presentan en procariotas y cloroplastos, como fue mencionado anteriormente. A diferencia de las bacterias los hongos presentaron ATPasas de tipo V pertenecientes a eucariotas. Las V-ATPasas están presentes en las membranas vacuolares de los hongos y están involucradas en procesos como la absorción de nutrientes (Portillo, 2000) y el mantenimiento de la homeostasis iónica intracelular ya que ayudan a mantener el pH dentro de la célula (Minematsu et al., 2019).

El análisis enzimático se indica en la figura 24. Así como en el caso anterior, se pueden apreciar las enzimas involucradas representadas a través de tonos verdes y en la zona inferior se especifica a cuál se hace referencia.

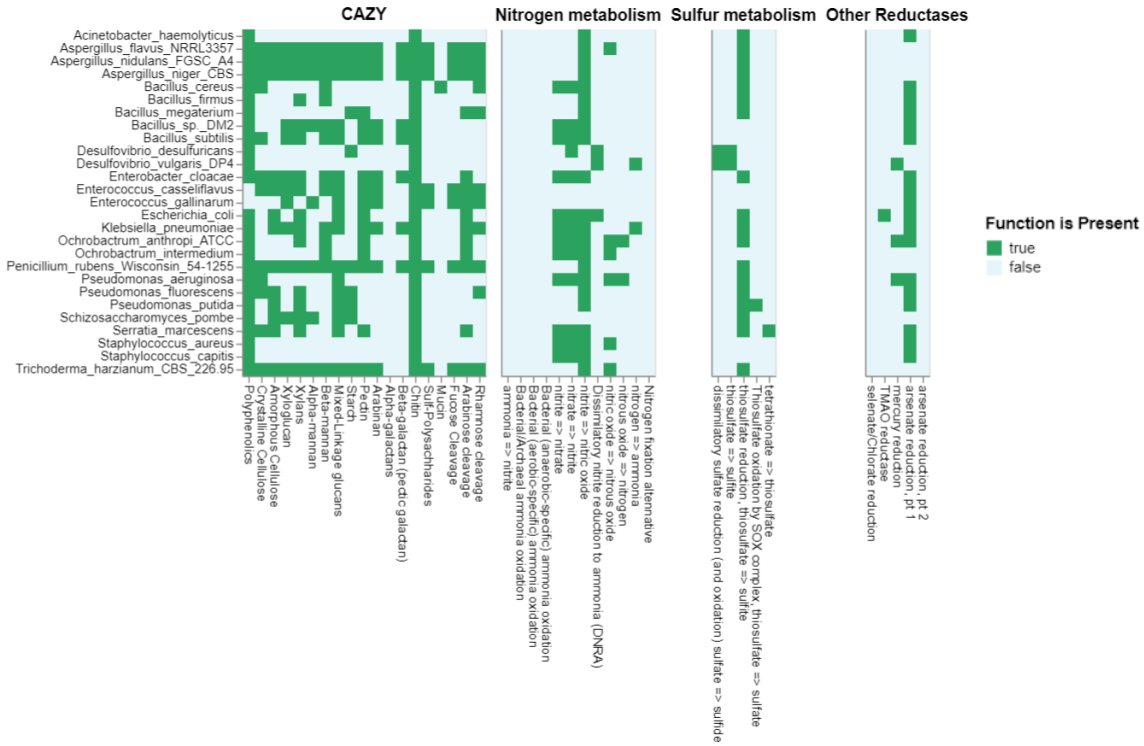

Figura 24*. Perfiles funcionales relacionados con la presencia de enzimas en diferentes procesos metabólicos en los microorganismos conocidos por remediar cromo (VI). Autoría propia (2021)*

Se puede observar que como en el caso anterior las enzimas que se encuentran presentes en la mayoría de las muestras son las responsables de la degradación de quitina y polifenoles. Lo que podría indicar que sí son factores determinantes para la remediación de cromo (VI).

Por ejemplo, el género *Cellulomonas* ha sido asociado por su potencial para remediar cromo (VI) (Field et al., 2013), además se ha encontrado en estudios que la eficiencia en la remoción o reducción de cromo (VI) puede variar de acuerdo a la fuente de carbono. Las *Cellulomonas* tienen la capacidad de degradar quitina como fuente de energía, este proceso ocurre de manera predominante en sedimentos lo que sugiere que toman un papel importante en la degradación en suelos y en los ciclos del carbono y nitrógeno (Reguera & Leschine, 2001).

Las especies *Pseudomonas putida* y *Bacillus sp*. han sido estudiadas por su capacidad para degradar fenoles y adicional por su capacidad para reducir cromo (VI)(Q. Gu et al., 2016). Esto nos permite confirmar la idea planteada anteriormente donde se indica que los productos de la degradación pueden ser usados como donantes de electrones para la remediación de cromo (VI).

Se puede observar que *Escherichia coli* es la única especie que en el análisis presenta la acción del óxido de trietilamina (TMAO). La variedad de vías para el flujo de carbono y electrones en este microorganismo le permiten usar TMAO cuando se encuentra en condiciones de anaerobiosis, con el fin de generar energía (Ansaldi et al., 2007). Por ello, tiene resistencia a condiciones de estrés y tiene la capacidad de equilibrar el crecimiento de sus células de manera eficiente (Gunsalus & Park, 1994) en ambientes como los efluentes de curtiembre.

8.2.4.2 Análisis funcional relacionado al ciclo del nitrógeno

Se realizó el análisis funcional de los microorganismos remediadores relacionados al ciclo del nitrógeno. En este caso se obtuvo lo que se puede apreciar en la figura 25.

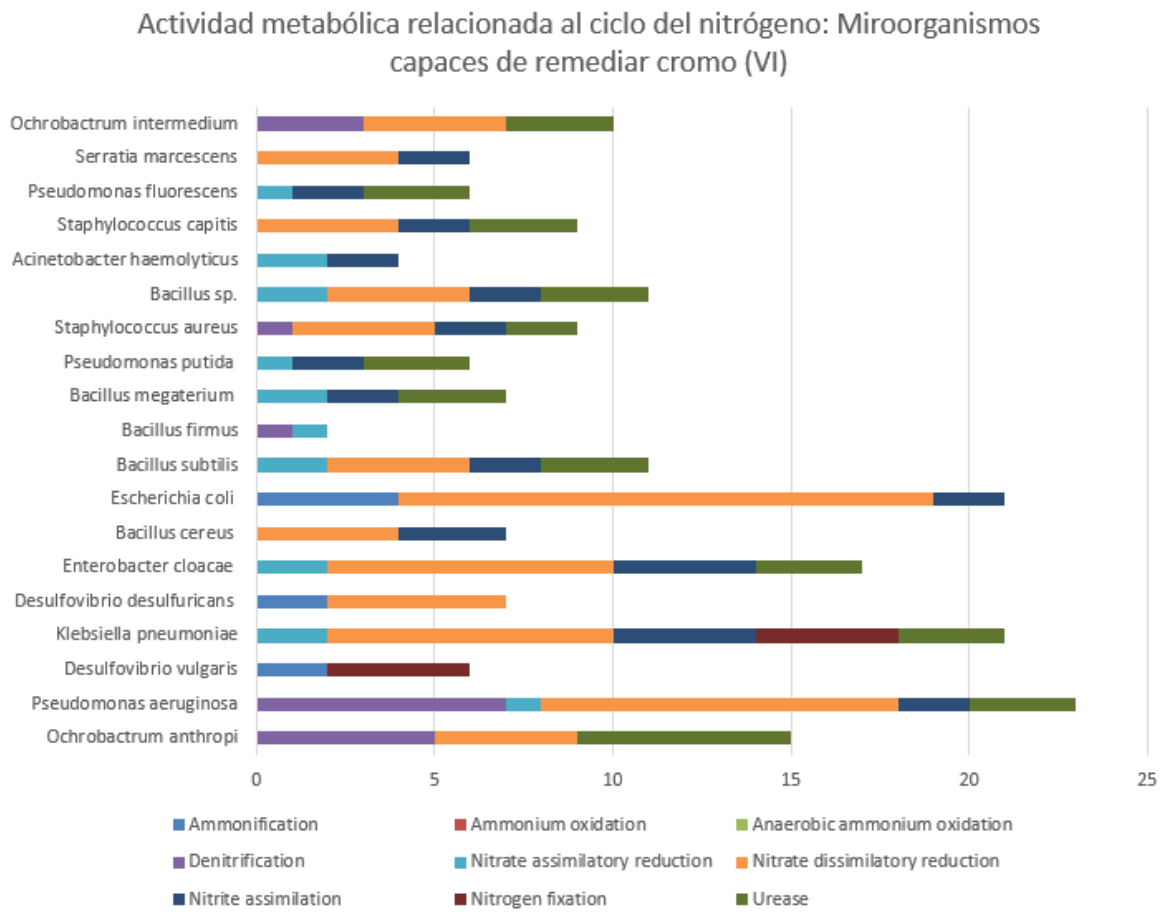

Figura 25.*Perfiles funcionales relacionados al ciclo del nitrógeno en los microorganismos capaces de remediar Cr (VI). Autoría propia (2021)*

En los resultados se presentan varias muestras que pueden realizar reducción disimilatoria de nitrógeno como fue en el caso de los microorganismos que se encontraron en ambientes contaminados con cromo (VI). Se encontró que este puede ser un factor importante y podría relacionarse con la remediación de cromo (VI). Por ejemplo, especies nitrato dependientes como *Candidatus methanoperedens* tienen potencial para degradar cromo (VI) a partir de metano, ya que pueden usar este compuesto como donante de electrones (Luo et al., 2019).

La presencia de cromo (VI) puede inhibir la actividad desnitrificante de los microorganismos (González-Blanco et al., 2020; Kim et al., 2016). Sin embargo, se han reportado algunos géneros de bacterias que pueden tolerar este contaminante en diversas concentraciones como *Acinetobacter*, *Bacillus*, *Ochrobactrum*, *Escherichia* y *Pseudomonas* (Zheng et al., 2018) que a su vez se reconocen por su potencial de remediación.

8.2.5 Análisis bioinformático de los datos de amplicón mediante QIIME 2

Se tuvieron en cuenta para el análisis 37 datos de amplicón 16s rRNA que fueron posteriormente transformados para poder usarlos en la plataforma bioinformática. Se seleccionaron todas las secuencias de 16S presentes en la base de datos.

8.2.5.1 Creación de archivo de metadata

El archivo de metadata creado contiene secuencias que fueron recolectadas de Xiangcheng, China, Argentina, Whenzou e India. Los datos fueron muestreados en 2014, 2019 y 2020.

La mayoría de los datos corresponden a dos estudios realizados en Argentina, estos se enuncian a continuación: i) *Impact assessment of bioaugmented tannery effluent discharge on the microbiota of water bodies* (2020) ii) *How the bacterial community of a tannery effluent responds to bioaugmentation with the consortium SFC 500-1. Impact of environmental variables* (2019).

El primer estudio busca evaluar los cambios en la diversidad microbiana del agua de un efluente de curtiduría antes y después de realizar una descarga de efluentes bioaumentados (Fernandez et al., 2020). El segundo estudio busca de igual manera evaluar los efectos que tiene la bioaumentación sobre comunidades nativas, con el fin de encontrar un método para el tratamiento de efluentes de curtiembre (Fernandez et al., 2019).

El documento de metadata realizado se puede encontrar en el *Anexo 2*.

8.2.5.2 Importación de secuencias al entorno de QIIME 2

Posterior a importar las secuencias a la plataforma bioinformática, se obtuvo la gráfica de calidad que se muestra en la figura 26.

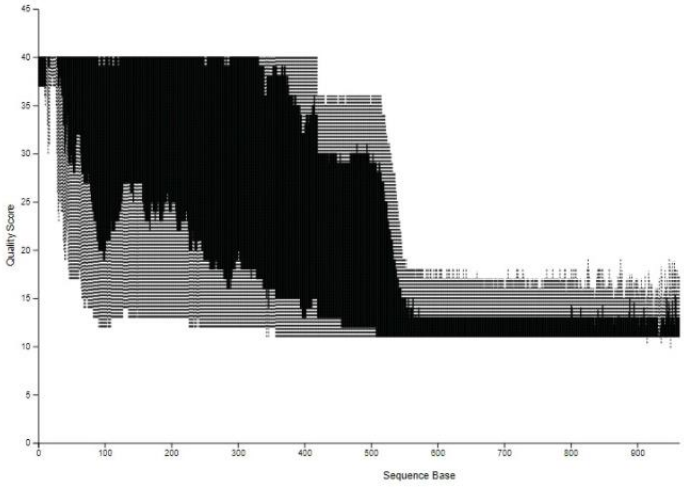

Figura 26.*Calidad de las secuencias de amplicón importada. Autoría propia (2021)*

Se puede observar que las secuencias en un rango de 50 a 500 pb presentan un coeficiente máximo de Phred de 40 y un coeficiente de Phred mínimo de 12 aproximadamente. Esto nos

permite saber que el error es menor del 10% por lo tanto, se cuenta con un buen puntaje de calidad de acuerdo al coeficiente de Phred (GATK team, 2021).

8.2.5.3 Eliminación de ruido de secuencia

Este proceso se realizó teniendo en cuenta los primers usados para amplificar el gen 16S rRNA, se encontró que el tamaño de fragmento esperado se encuentra entre 174 pb y 200 pb (Frank et al., 2008). Algunos de los primers usados en los estudios fueron 533R, 27F, 341F y 515R(Lu & Lu, 2014; Wang et al., 2014). De acuerdo a esto se determinó hacer un recorte de las secuencias en 190 pb. En promedio el 95.82% de las secuencias pasaron por el proceso de filtrado.

8.2.5.4 Clasificación taxonómica

En esta sección se realizó la clasificación taxonómica de las muestras, el proceso se realizó para 16s y 18s. En el *Anexo 3* se pueden encontrar los archivos de visualización resultantes.

8.2.5.4.1 Clasificación taxonómica: SILVA 16S rRNA

La asignación taxonómica a través de SILVA se puede observar a través de la figura 27.

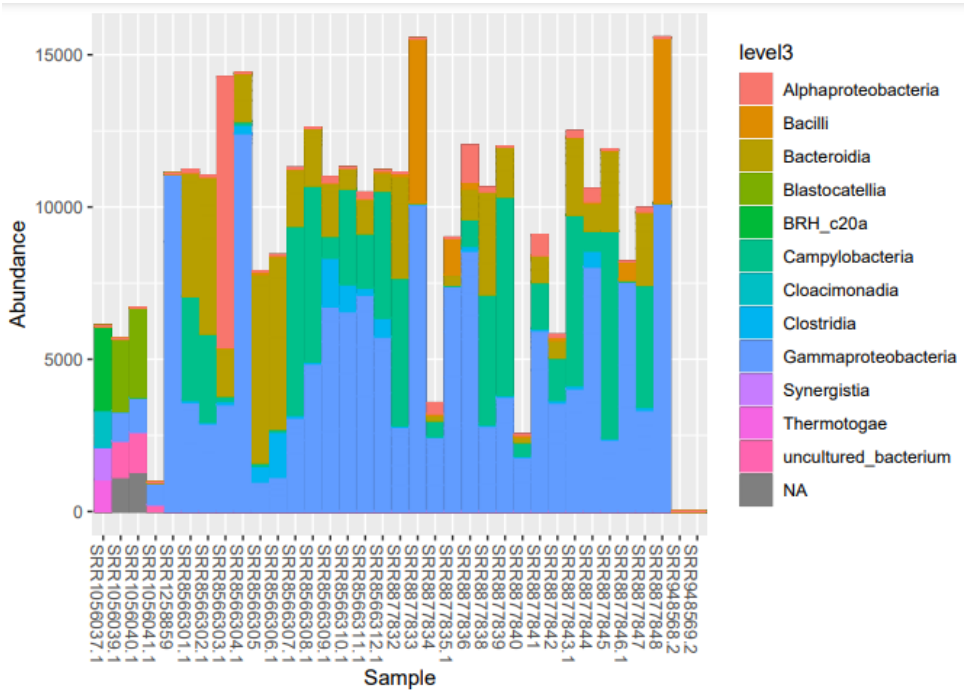

Figura 27.*Clasificación taxonómica de amplicón 16s rRNA. Autoría propia (2021)*

El resultado que se muestra hace referencia a la clasificación en la categoría de clase. Allí podemos observar abundancia de Proteobacteria, Bacteroidetes y Firmicutes como se observó en el resultado de la figura 18. Se encontraron especies relacionadas entre los dos análisis como Acinetobacter, Bacillus y Marinobacter.

En este caso se encontraron especies como *Arcobacter*, *Azoarcus*, *Myroides* y *Alicycliphilus* que no habían sido obtenidas previamente. *Azoarcus* y *Alicycliphilus* son especies que se han estudiado por su capacidad de desnitrificación y de remediación de cromo (VI), respectivamente(B. Liu et al., 2006; Solís-González & Loza-Tavera, 2019), mientras que *Myroides* y *Arcobacter* han sido estudiadas por su patogenicidad sobre animales y humanos (Lorenzin et al., 2018; Ramees et al., 2017).

8.2.5.4.2 Clasificación taxonómica: SILVA 18S rRNA

En la figura 28 se puede observar el resultado obtenido en la asignación taxonómica teniendo como base de referencia SILVA para microorganismos eucariotas.

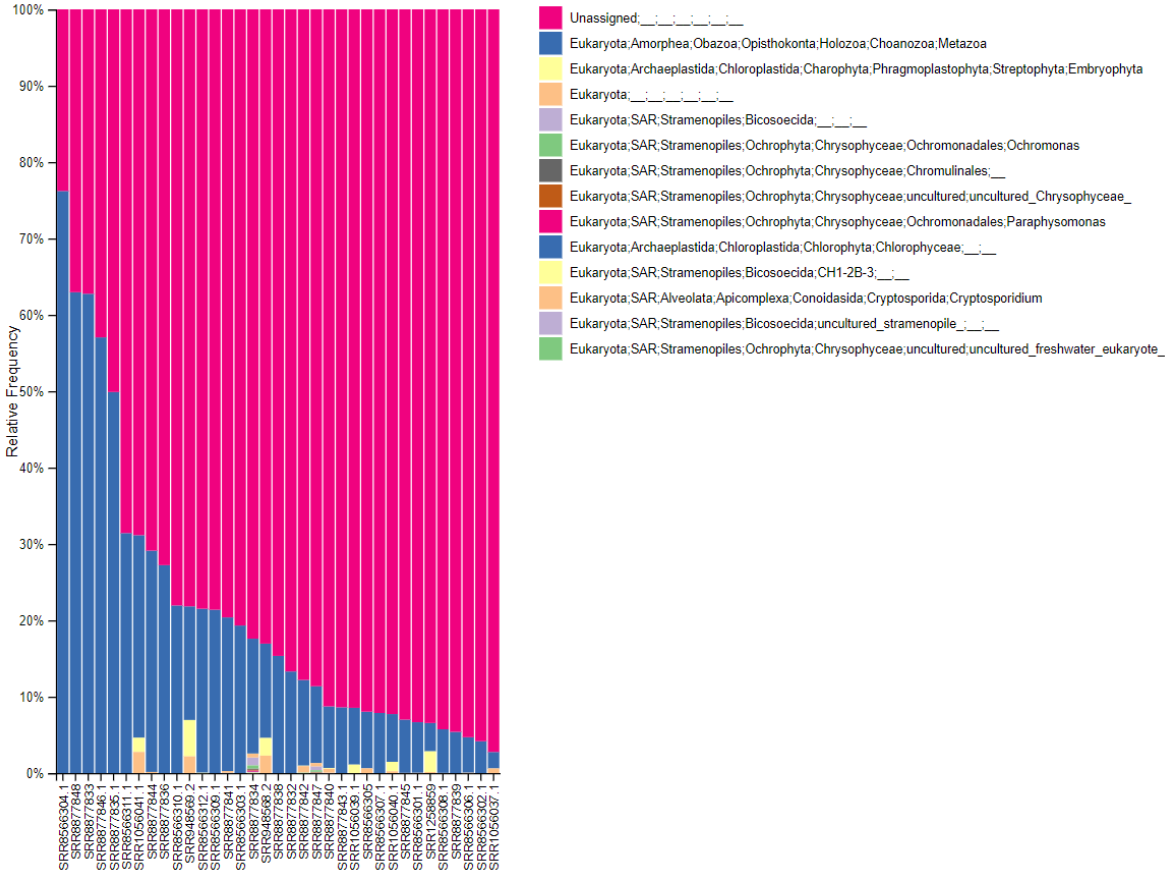

Figura 28. *Clasificación taxonómica de amplicón 18s rRNA. Autoría propia (2021)*

El 97.235% de los datos no pudieron ser clasificados, esto se pudo generar porque las muestras de amplicón eran del gen 16s rRNA. A pesar de esto, se encontró abundancia del superfilo *Stramenopiles*, esta categoría taxonómica incluye desde organismos marinos, de agua dulce, de suelo o parásitos (Medlin & Cembella, 2013). Adicional, se encontró abundancia del supergrupo *Amorphea* y este es un taxón que agrupa animales, hongos y sus respectivos parientes unicelulares(Burki et al., 2020). Se clasificaron estos microorganismos

porque es posible que en las muestras se encontraran cloroplastos de estas especies, ya que estos contienen el gen 16s rRNA (Sanschagrin & Yergeau, 2014).

8.2.6 Modelado metabólico: Análisis de la remediación de cromo (VI)

Se realizó una representación de algunas de las vías que toman los microorganismos para remediar cromo (VI) de acuerdo a la bibliografía y su relación con los resultados anteriormente obtenidos.

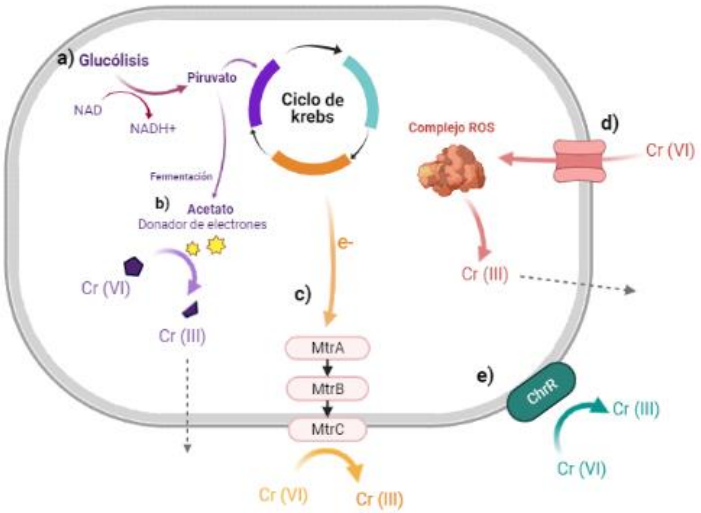

Figura 29. *Mecanismos intracelulares y extracelulares para la remediación de cromo (VI). Autoría propia (2021). Creado a través de BioRender.com*

a) *Fuente de carbono para la remediación* ***b)*** *Acetato como donador de electrones para la remediación* ***c)*** *Biorreducción de cromo (VI) a través de la vía respiratoria Mtr* ***d)*** *Influencia de las especies reactivas del oxígeno (ROS) en la remediación* ***e)*** *Cromato reductasa ChrR en la membrana celular*

En la figura 29 se puede observar la representación realizada. Se encontró que la eficiencia de la reducción de cromo (VI) puede variar de acuerdo a la fuente de carbono que esté disponible para las bacterias (Arishi & Mashhour, 2021). Algunas de las fuentes de carbono que usan los microorganismos son la glucosa, fructosa, lactosa, piruvato, lactato, citrato y acetato (Z. Rahman & Thomas, 2021). Por ello en la parte **a** de la representación gráfica se indica el proceso de glucólisis que genera el piruvato. En este caso el piruvato puede continuar y entrar al ciclo de Krebs (**a**) o fermentarse para convertirse en acetato (**b**). Cuando el piruvato es fermentado en acetato se puede hacer uso de este producto como donador de electrones para la transformación de cromo (VI) a cromo (III) (Z. Rahman & Thomas, 2021); mientras que, cuando el piruvato ingresa al ciclo de Krebs se generan electrones que pueden ser usados en la vía respiratoria Mtr (**c**). Microorganismos como *Shewanella oneidensis* tienen la capacidad de realizar esta vía que consta de una serie de componentes proteicos (MtrA, MtrB y MtrC) que desempeñan funciones importantes para la reducción de cromo (VI) (Huang et al., 2019).

Cuando los microorganismos se encuentran sometidos a situaciones de estrés se generan ROS (**d**) que pueden inducir daño al ADN. Como respuesta a estos radicales generados, la célula comienza a generar enzimas antioxidantes que permiten neutralizar los efectos de la producción de ROS y a su vez pueden ayudar a la reducción de cromo (VI) (Tang et al., 2021). Los microorganismos también pueden hacer uso de la cromato reductasa ChrR (**e**). Esta es una enzima asociada a la membrana que puede catalizar la reducción de cromo (VI) y puede encontrarse en especies como *Stenotrophomonas maltophilia* (Li et al., 2021).

8.3 Resultados del objetivo n°3

8.3.1 Búsqueda bibliográfica sobre algoritmos de machine learning empleados para el análisis de comunidades microbianas

De la búsqueda bibliográfica se obtuvo información de 34 artículos donde se presentó información de diferentes aplicaciones de modelos de machine learning en diferentes campos como lo son: microbiota intestinal, microbiota del suelo y microbiota vaginal (D. & J.A., 2015; S. F. Rahman et al., 2017; Thompson et al., 2019). Se han generado aplicaciones con algoritmos de bosques aleatorios, regresión logística, máquinas de vectores de soporte y redes neuronales (Topçuoğlu et al., 2020). Esto permitió observar lo versátil que pueden ser estos modelos y su aplicabilidad en el área de la biotecnología. La información recolectada puede ser consultada en el *Anexo 4*.

8.3.2 Selección del modelo de machine learning a emplear

En la Tabla 2. Matriz de decisión del algoritmo de machine learning a implementar se puede observar la matriz de decisión que se desarrolló para la selección del algoritmo de machine learning a emplear teniendo en cuenta la aplicación y los datos disponibles para el análisis.

Tabla 2. *Matriz de decisión del algoritmo de machine learning a implementar*

	Capacidad de clasificar datos binarios	Capacidad para trabajar con datos de alta dimensión	Complejidad de implementación	Obtención de resultados en términos probabilísticos	Total
Peso	5	2	4	4	
Árboles de decisión	5	5	5	5	75
Redes neuronales	4	5	4	5	66
Bosques aleatorios	5	5	5	5	75
Máquinas de soporte vectorial	4	5	3	0	42

En la selección se tuvieron en cuenta los criterios que se aprecian en las columnas de la tabla, mientras que las opciones de selección y puntajes se pueden encontrar en las filas.

La ponderación se realizó otorgando un peso de uno (1) a cinco (5) a cada uno de los criterios de acuerdo a la importancia de cada uno. El puntaje otorgado a las opciones de algoritmo también se tuvo en cuenta con una escala de uno (1) a cinco (5).

Para obtener el resultado se realizó el cálculo que se muestra a continuación, tomando como ejemplo la primera fila:

$$(5 * 5) + (5 * 2) + (5 * 4) + (5 * 4) = 75 \qquad \textit{Ecuación 11}$$

En el proceso se multiplicó el peso de cada criterio por la puntuación asignada y se sumó cada columna para obtener la ponderación total. De acuerdo a los resultados obtenidos se seleccionó la implementación de un algoritmo basado en **árboles de decisión y bosques aleatorios**. Uno de los criterios que más le otorgó valor a la decisión fue la capacidad de clasificar datos binarios. Este factor es de gran importancia ya que los datos que se obtuvieron dan valores de unos (1) y ceros (0) en su mayoría, siendo valores que podrían tener inconvenientes al usar algunos clasificadores. La ventaja que ofrecen los árboles de decisión es la clasificación a través de condicionales lo que permite generar una respuesta con entradas de si (1) o no (0), al igual que los bosques aleatorios (Szczerbicki, 2001).

8.3.3 Etiquetado de datos de entrada

De esta fase, se obtuvieron 27 datos de microorganismos remediadores de cromo (VI) que fueron etiquetados de acuerdo a su potencial de remediación. En la figura 30 se puede observar el potencial reportado de cada microorganismo a través de las barras verdes y la presencia de la barra azul muestra a los microorganismos etiquetados con "alto potencial para remediar". El documento resultante del proceso de etiquetado puede ser consultado en el *Anexo 5*.

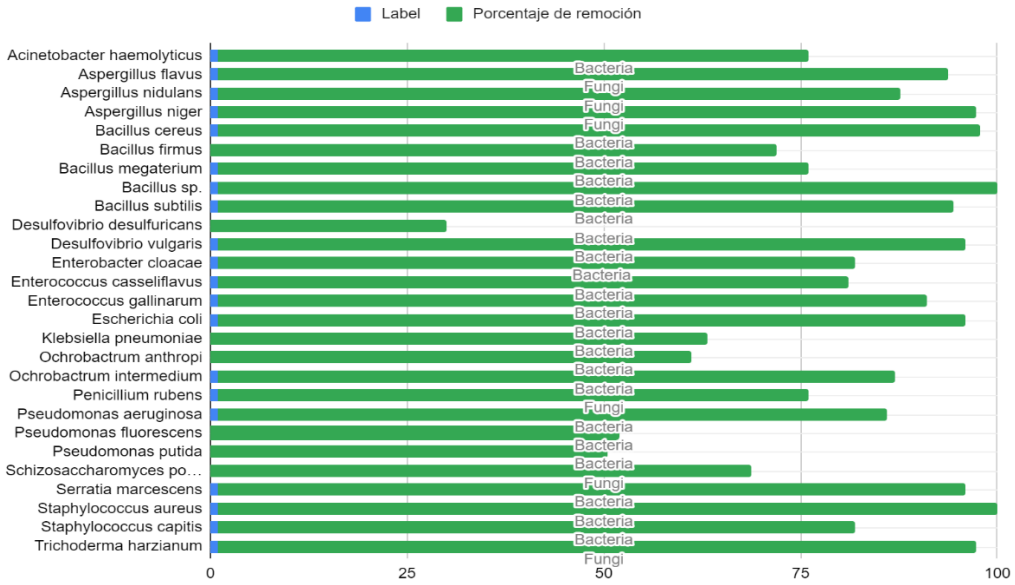

Figura 30. *Potencial de remediación de microorganismos reportado en literatura. Autoría propia (2021)*

De los 27 microorganismos para el análisis, 20 recibieron la etiqueta de "Alto potencial de remediación" y los 7 restantes recibieron la etiqueta de "bajo potencial de remediación". Este set de datos se tuvo en cuenta para el entrenamiento del modelo de machine learning. Se tuvo en cuenta que el set de datos presente no está balanceado, lo que podría tener errores al momento de realizar la clasificación. Sin embargo, se encontraron estudios que han sido realizados con datos desequilibrados, ya que en problemas cotidianos se puede presentar esta situación (Haixiang et al., 2017).

Adicional no se logró robustecer el set de datos ya que tienen que pasar por un proceso de búsqueda y análisis bioinformático, por lo tanto, es necesario enriquecer el set de datos conforme vayan aumentando las investigaciones en este campo.

8.4 Resultados del objetivo n°4

8.4.1 Implementación del algoritmo de machine learning

A continuación, se describen los resultados obtenidos en la implementación del algoritmo y en el *Anexo 6* se puede consultar el código de Python que se usó.

8.4.2 Preparación y análisis de datos

Se obtuvieron 3 archivos de Excel, cada uno de ellos con la información que se menciona a continuación: *a)* Tabla que indica las muestras con la presencia o ausencia de genes funcionales. *b)* Tabla con los microorganismos etiquetados. *c)* Archivo de Excel donde en cada hoja contiene la clasificación taxonómica de cada organismo. Cada hoja tiene el nombre

del bin al que pertenecen. El set de datos empleado para el análisis se encuentra en el *Anexo 7*.

En la figura 31 **se** puede observar un ejemplo de la información contenida en cada una de las tablas y de donde se obtuvieron.

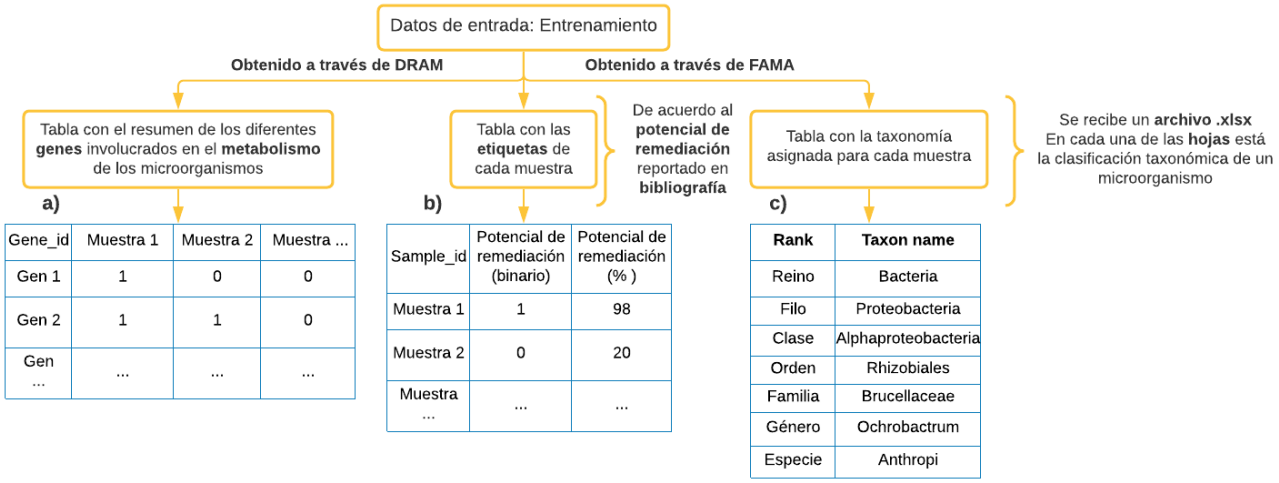

Figura 31*.Datos obtenidos del proceso bioinformático en Kbase. Autoría propia (2021)*

La tabla **a** contiene la información funcional que se usó para determinar si un microorganismo tiene un alto potencial o un bajo potencial de remediación. Cada uno de los genes hace referencia a las características que podrían ser usadas en el modelo, en total se obtuvieron **527** características.

En esta sección se realizó una descripción estadística de los datos pertenecientes en la tabla **a**, se obtuvo un resultado como el que se puede observar en la tabla 3.

Tabla 3*. Descripción estadística de los datos de entrada*

Gene ID	K00798	K01698	K01749
Count	27	27	27
Unique	3	3	4
Top	0	1	1
Freq	14	24	21

La descripción estadística permitió identificar qué características están presentes en todas las muestras, por ejemplo, los genes relacionados al ID *K01698* **y** *K01749* se encuentran en todas las muestras, ya que el valor mayor que tiene cada una de las filas es un uno (1), que representa la presencia de ese gen específico.

8.4.3 Extracción de características

8.4.3.1 Selección de características usando Chi2

Teniendo en cuenta que se obtuvieron 527 características referentes a la actividad metabólica

de los microorganismos, se decidió hacer un proceso de selección y extracción de 5 características. Esto se realizó con el fin evitar que el modelo se "aprenda" los datos (Amazon AWS, 2021) haciendo que se sobreentrene. Lo que puede hacer que clasifique correctamente únicamente los datos de entrenamiento.

A partir del modelo de selección basado en *Chi2*, se obtuvieron 5 características asociadas a diferentes procesos metabólicos en los microorganismos. Se encontró que los códigos de los genes *K02227*, *K02232, K02233*, *K00560* y *K10617* tuvieron relación con la salida esperada del modelo de machine learning, con un coeficiente de *Chi2* de 6.04, 4.11, 4.69, 5.46 y 10.39, respectivamente. Estos fueron los valores más altos obtenidos en relación a valores como el que se obtuvo con el gen K00589 que fue 0.603. En la tabla 4 se muestran algunos ejemplos de los resultados obtenidos a partir del análisis de esta estadística.

Tabla 4. *Coeficiente de Chi2 obtenido para las características*

	Característica (Gene ID)	Coeficiente de Chi^2	Posición
1	K00798	3.36	0
2	K02225	3.02	1
3	K02227	6.04	2
4	K02231	3.36	3
5	K02232	4.11	4
6	K02233	4.69	5
7	K01835	3.95	159
8	K00560	5.46	165
9	K10617	10.39	520
10	K00589	0.603	9

Los códigos *K02227, K02232 y K02233* están asociados a la biosíntesis de adenosilcobalamina (Vitamina B) de acuerdo a la base de datos UniProt. La síntesis de vitamina B se encuentra en algunas bacterias y arqueas; la producción de este compuesto depende de la fermentación bacteriana (Fang et al., 2017). Este comportamiento se ha encontrado en microorganismos con la capacidad de remediar cromo (VI) como *Pseudomonas aeruginosa*, *Escherichia coli y Bacillus megaterium* (Crespo et al., 2018).

En los resultados se presentó una enzima denominada timidilato sintasa (*K00560*), estas enzimas actúan catalizando la producción de timidilato que es importante para que la célula pueda desempeñar funciones clave como el control del ciclo celular, la biosíntesis de ADN y la apoptosis (Chu & Allegra, 1996; Myllykallio et al., 2018). Se podría relacionar con la recuperación que tienen los microorganismos cuando se encuentran en condiciones de estrés como la presencia de cromo (VI) (Tang et al., 2021).

La última característica que se obtuvo corresponde al código *K10617*, que obtuvo el mayor valor en el coeficiente de *Chi2*. Esta es una alcohol deshidrogenasa (*cymB*), que hace parte de procesos de biodegradación de xenobióticos (Eaton, 1997), lo que hace que sea un hallazgo importante, ya que esta es una de las maneras en las que los microorganismos se adaptan para sobrevivir a condiciones de estrés. Algunas cepas bacterianas como

Acinetobacter, Trichoderma, Pseudomonas y Bacillus han sido estudiadas por su capacidad xenobiótica donde usan los contaminantes orgánicos como fuente de carbono o nitrógeno lo que les permite continuar con su desarrollo (Mishra et al., 2021).

Se obtuvo una matriz de covarianza de un tamaño de (527,527), ya que este fue el número de características. Sin embargo, se evaluaron las 5 que fueron seleccionadas previamente. En la tabla 5 se puede observar la matriz de las características que se tuvieron en cuenta.

Tabla 5. *Matriz de covarianza de las características seleccionadas*

	K02227	K02232	K02233	K00560	K10617
K02227	0.72	0.3	0.26	0.01	-0.12
K02232	0.3	0.25	0.18	-0.07	0.01
K02233	0.26	0.18	0.26	-0.11	0.08
K00560	0.01	-0.07	-0.11	0.57	-0.27
K10617	-0.12	0.01	0.08	-0.27	1.28

A partir de la tabla obtenida se puede encontrar la relación que tienen las características seleccionadas entre sí. Por ejemplo, la característica K00560 tiene una relación negativa con las características K02232, K02233 y K10617. A partir del análisis de covarianza se puede conocer la dispersión que tienen las muestras entre sí y cómo pueden estar situadas espacialmente, esto de acuerdo al signo del valor de covarianza.

8.4.3.2 Selección de características usando bosques aleatorios (Scikit-learn)

A partir del modelo de bosques aleatorios fueron seleccionadas automáticamente **136** características de las **527** que se tenían inicialmente. Comparando este proceso con el anteriormente realizado se encontró que las características K02227, K02232, K02233 y K10617 se relacionan en ambos procesos de selección, lo que indica que pueden permitir la clasificación entre microorganismos con alto potencial y bajo potencial de remediación.

8.4.4 Creación del set de datos (Entrenamiento y prueba)

Teniendo en cuenta el proceso de validación cruzada, se obtuvieron 27 divisiones, en las que se obtuvo un set de datos de 1 dato de validación y 26 datos de entrenamiento.

8.4.5 Validación y comparación del modelo de árbol de decisión con otros modelos de machine learning

8.4.5.1 Validación con las características seleccionadas a partir del coeficiente Chi2

En la tabla 6, se muestran los resultados obtenidos de la validación mediante la matriz de confusión realizada para el modelo de árbol de decisión y bosque aleatorio. Este análisis se realizó teniendo en cuenta el modelo con la extracción de características a través del método de Chi2.

Tabla 6. *Resultado de las métricas (árbol de decisión y bosque aleatorio): Chi2*

Validación mediante LOOCV	**Árbol de decisión**	**Bosques aleatorios**

División	Exactitud	
1	1.0	1.0
…	…	…
7	0.87	0.85
8	0.88	0.87
9	0.9	0.88
10	0.81	0.9
11	0.83	0.81
..	…	…
27	0.77	0.81

A partir de cada una de las divisiones obtenidas, se determinó la media aritmética de la exactitud obtenida para cada iteración en la validación cruzada, de lo que se obtuvo resultado del 78% (entrenamiento) y 78% (validación) para el algoritmo de árbol de decisión y de un 88% (entrenamiento) y 88% (validación) para el algoritmo de bosques aleatorios. De acuerdo a esto se observa que los bosques aleatorios tuvieron un mejor desempeño teniendo en cuenta las características seleccionadas.

Para poder confirmar los resultados obtenidos y conocer el comportamiento que tienen los modelos de machine learning para diferenciar microorganismos con alto potencial entre los que tienen un bajo potencial se obtuvieron las métricas de validación y entrenamiento que se observan en la tabla 7.

Tabla 7. *Métricas del set de validación y entrenamiento: Chi2*

Clase		0			1		
		Precisión	**Sensibilidad**	**F-Score**	**Precisión**	**Sensibilidad**	**F-Score**
Árbol de decisión	Validación y entrenamiento	0.56	0.71	0.63	0.89	0.80	0.84
Bosques aleatorios	Validación	0.62	0.85	0.67	0.89	0.85	0.87
	Entrenamiento	0.70	0.1	0.82	1	0.84	0.91

Clase 0: Bajo potencial de remediación; Clase 1: Alto potencial de remediación

En la tabla 7 se muestran las métricas de validación que se tuvieron en cuenta, donde se realiza la comparación del modelo de árbol de decisión con bosques aleatorios. Para el caso del árbol de decisión se obtuvieron los mismos resultados tanto para el set de validación como para el set de entrenamiento con una precisión del 89% para clasificar la clase 1 y del 56% para clasificar la clase 0. Al comparar los datos es posible identificar que en ambos modelos se tiene una mayor precisión para clasificar la clase positiva (1), esto se debe a que únicamente se contó con 7 muestras con etiquetas de cero (0) y con 20 muestras con etiquetas de uno (1).

En la tabla 8 se puede observar la exactitud obtenida del set de validación y entrenamiento en el algoritmo de bosques aleatorios. Teniendo en cuenta que para el set de validación se obtuvo un 81% de exactitud y para el set de entrenamiento un 88% es posible decir que el

modelo tiene la capacidad de generalizar nuevos datos que ingresan al modelo sin generar sobreentrenamiento, ya que ambos sets tienen una exactitud en un rango de 80-90%.

***Tabla 8**. Exactitud del set de entrenamiento y validación: Bosques aleatorios (Chi2)*

Modelo de bosques aleatorios	Set de validación	Set de entrenamiento
Exactitud	81%	88%

Teniendo en cuenta los resultados obtenidos se decidió usar el algoritmo basado en bosques aleatorios para la prueba en los datos de comunidades microbianas presentes en ríos contaminados con cromo (VI). Sin embargo, el modelo diseñado queda abierto a modificaciones y a ser alimentado con más datos con el fin de poder obtener un resultado más robusto. Esto teniendo en cuenta que únicamente se usaron 27 datos para entrenar el modelo.

8.4.5.1 Validación con las características seleccionadas a partir del modelo de bosques aleatorios

Se realizó la validación del modelo teniendo en cuenta las **136** características que se determinaron a partir del modelo de selección de los bosques aleatorios. En la tabla 9 se puede observar la comparación de los valores obtenidos en cada una de las iteraciones de la validación cruzada para este caso. Las características que fueron seleccionadas por este modelo se pueden encontrar en la carpeta de Drive compartida con el proyecto.

***Tabla 9.** Métricas del set de validación y entrenamiento: Características relacionadas a bosques aleatorios*

Validación mediante LOOCV	**Árbol de decisión**	**Bosques aleatorios**
División	**Exactitud**	
1	0.5	1.0
…	…	…
7	0.62	0.75
8	0.66	0.77
9	0.6	0.7
10	0.54	0.63
11	0.58	0.66
..	…	…
27	0.59	0.62

Se obtuvo que la exactitud del set de entrenamiento y validación fue de un 100% y 59%, respectivamente. En el caso del modelo de bosques aleatorios se obtuvieron unos valores de 100% (set de entrenamiento) y 63% (set de validación). La clasificación para el set de entrenamiento en ambos casos fue del 100%, mientras que el set de validación tuvo un porcentaje de exactitud más bajo, lo que puede indicar que el modelo está sufriendo de sobreentrenamiento. Esto se puede identificar porque el modelo estaría teniendo un menor error en los datos de entrenamiento, pero en los datos de validación tiene un error alto (IBM Cloud Education, 2021).

Tabla 10. Métricas del set de validación y entrenamiento: Características relacionadas a bosques aleatorios

Clase		0			1		
		Precisión	**Sensibilidad**	**F-Score**	**Precisión**	**Sensibilidad**	**F-Score**
Árbol de decisión	Entrenamiento	0.1	0.1	0.1	0.1	0.1	0.1
	Validación	0.25	0.29	0.27	0.74	0.70	0.72
Bosques aleatorios	Entrenamiento	0.1	0.1	0.1	0.1	0.1	0.1
	Validación	0.20	0.14	0.17	0.73	0.80	0.76

En la figura 10 se puede observar con mayor detalle las métricas de precisión, sensibilidad y F-Score que se obtuvieron para las clases uno (1) y cero (0). Así como en el caso de las características seleccionadas por Chi2, se puede observar una diferencia significativa entre las métricas de la clasificación de la clase negativa y positiva, lo que denota la importancia de contar con un set de datos balanceado.

8.4.6 Implementación del modelo entrenado con la información de comunidades microbianas presentes en ríos contaminados con cromo (VI)

La implementación del modelo se realizó con los datos provenientes de artículos con información asociada a microorganismos presentes en ambientes contaminados con cromo (VI). Posterior al análisis bioinformático se obtuvieron las tablas correspondientes para poder analizarlos a través del algoritmo de bosques aleatorios. De acuerdo a las métricas identificadas se decidió realizar la implementación del modelo que tuvo en cuenta la selección de características basadas en el coeficiente de Chi2, ya que se obtuvo un mejor porcentaje de exactitud a comparación de los otros modelos implementados.

Luego de organizar los datos e implementar el modelo se obtuvo una tabla con las siguientes columnas: *Super Kingdom, phylum, Class, Order, Family, Genus, Specie, Prediction* a través de un archivo de Excel con el nombre "*ResultadoPredicción_DT.xlsx*" que se puede observar en los archivos que hacen parte del notebook de colab. En este archivo se encuentra la predicción y los microorganismos a los que pertenecen. En el *Anexo* **s**e encuentra el documento obtenido.

En la figura 32 se puede observar una representación de los resultados obtenidos en esta sección. Allí encuentran los microorganismos que fueron identificados en las muestras por tener alto o bajo potencial para remediar cromo (VI).

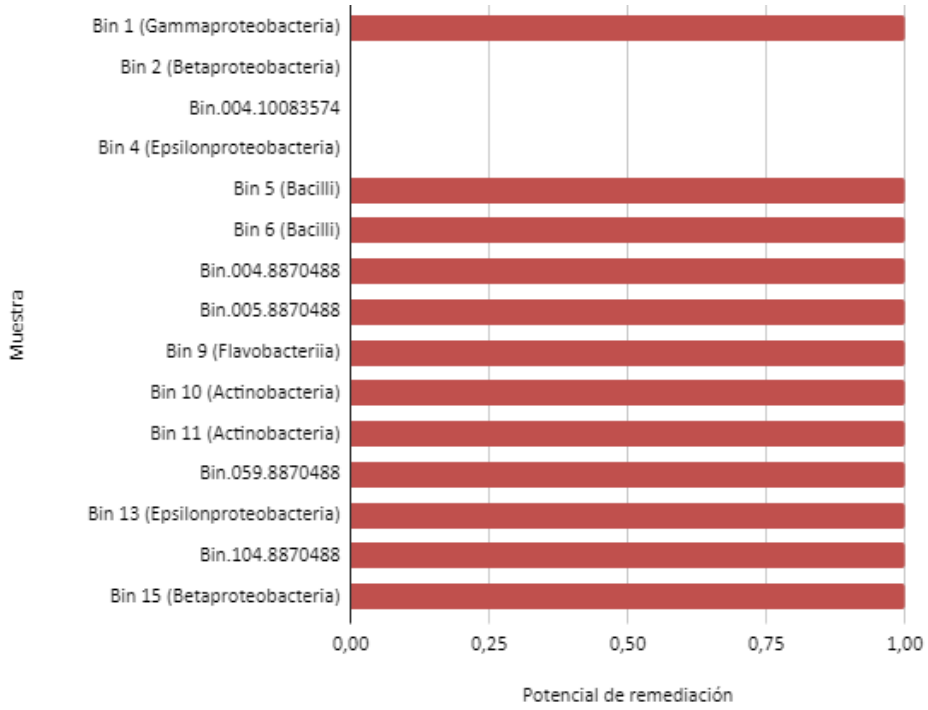

Figura 32. *Predicción realizada por el algoritmo de bosques aleatorios. Autoría propia (2021)*

Las barras rojas de la gráfica representan la predicción realizada a partir del modelo de bosques aleatorios. Las muestras que están nombradas como "bin", no pudieron ser clasificadas por lo que es necesario hacer la búsqueda directamente desde el proceso bioinformático para conocer a qué microorganismo pertenecen.

Entre los microorganismos identificados por tener un alto potencial para remediar cromo (VI) según los resultados obtenidos está la clase *Bacilli*, en la que se encontraron especies como *Bacillus paralicheniformis* (Bin 5) y *Marinobacter hydronocarbonoclasticus* (Bin 1), que son especies que han sido asociadas por su capacidad para remediar cromo (VI) como se discutió en los resultados de la sección 8.2.

Entre los microorganismos que no fueron identificados debido a su "bajo potencial", se encontró el orden *Campylobacterales* que ha sido asociado por tener microorganismos con actividad patogénica como *Campylobacter sp*. que es una bacteria causante de enfermedades transmitidas por alimentos y que puede encontrarse en agua y suelos (Lapierre A., 2013), por este motivo pudo encontrarse en las muestras de efluentes de curtiduría.

A pesar de los resultados obtenidos no se descarta la posibilidad de que los microorganismos que no hayan sido clasificados en la clase de bajo potencial puedan llegar a remediar fuentes hídricas contaminadas con cromo (VI), se hace la aclaración porque los microorganismos pueden adaptarse a diferentes situaciones y así mismo modificar sus vías metabólicas.

CONCLUSIONES

De acuerdo a la información recolectada (50 artículos), se pudo evidenciar que la mayoría de ellos contienen datos relacionados con de amplicón 16S rRNA (71.2%). Se encontró que el 17.3% de los datos eran metagenómicos y no había estudios que se centraran en las comunidades microbianas de ambientes contaminados en Colombia; la mayor parte de los estudios se centraba en ríos de la India.

A partir del análisis taxonómico realizado en Kbase y QIIME 2 se encontró la abundancia de los phylum *Proteobacteria, Firmicutes, Bacteroidetes* y *Acidobacteria* y especies como *Bacillus sp., Halomonas sp.* y *Comamonas* en las muestras provenientes de efluentes de curtiduría que han sido estudiadas por su potencial para remediar cromo (VI), lo que nos permite confirmar que los microorganismos que se encuentran en condiciones de estrés desarrollan mecanismos para la eliminación o reducción de compuestos contaminantes.

El análisis funcional permitió identificar diversos mecanismos que tienen los microorganismos para la reducción de cromo (VI), donde los productos metabólicos que se usan como donadores de electrones cumplen un papel importante en la remediación. Se encontró que de acuerdo a la fuente de carbono el potencial de remediación de los microorganismos puede aumentar o disminuir. Adicional, la desnitrificación está relacionada con las vías catabólicas del carbono como la degradación de carbohidratos, ya que esto puede generar donadores de electrones para la reducción de nitratos y cromo (VI).

Se encontró que el modelo de Scikit-learn basado en Chi2 para la selección de características permitió reducir la dimensionalidad de las variables de entrada. Las características que se obtuvieron (*K02227, K02232, K02233, K00560* y *K10617*) se pueden relacionar con la capacidad que tienen los microorganismos para remediar cromo (VI), ya que se relacionaron con la degradación de xenobióticos y la regulación celular. Siendo estos mecanismos naturales de los microorganismos, pero que a su vez pueden influir en la reducción de cromo (VI) a cromo (III).

Los modelos de aprendizaje automático implementados tuvieron una mayor precisión para clasificar la clase 1 (alto potencial de remediación), lo que hace necesario que se implemente el modelo con una mayor cantidad de datos para que sea posible generar un resultado más robusto.

Se obtuvieron **136** características a partir del modelo de selección basado en bosques aleatorios, entre ellas fue posible encontrar las características *K02227, K02232, K02233* y *K10617* que fueron identificadas por el análisis mediante Chi2, lo que confirma la relación que tienen con los procesos de remediación de cromo (VI).

Se encontró que el modelo de bosques aleatorios basado en la selección de características a partir del coeficiente de Chi2 tuvo una exactitud del 81% para el set de validación y de un 88% para el set de entrenamiento, lo que indica que no existe sobreentrenamiento en el modelo y que tiene la capacidad de clasificar datos con alto y bajo potencial de remediación. A comparación del modelo realizado teniendo en cuenta 136 características, donde se obtuvo una exactitud del 100% y 63% (set de entrenamiento; set de validación), lo que indica que en

este modelo se presentó un sobreentrenamiento probablemente por la cantidad de características y la naturaleza de las mismas.

Con la implementación del modelo de bosques aleatorios se encontraron microorganismos conocidos por remediar cromo (VI) como *Marinobacter hidrocarbonoclasticus* y *Bacillus paralicheniformis*. Adicional, se clasificó como microorganismo con bajo potencial para remediar cromo (VI) el orden *Campylobacterales* que ha sido asociado en bibliografía contener especies patógenas.

RECOMENDACIONES

Teniendo en cuenta el desarrollo del proyecto y los resultados conseguidos a continuación se realiza una serie de recomendaciones que es importante tener en cuenta al momento de decidir trabajar con datos correspondientes a comunidades microbianas:

- Es importante a la hora de iniciar a implementar el proyecto conocer la naturaleza de los datos que se van a analizar. Esto con el fin de saber de qué manera se puede generar un análisis que descarte la menor cantidad de información posible.

- Se recomienda que al hacer un análisis para determinar el potencial de remediación de diferentes microorganismos se use un modelo de bosques aleatorios que pueda ajustarse mejor a los datos.

- Incluir y diferenciar la clasificación de acuerdo al microorganismo presente en la muestra. Teniendo en cuenta bacterias, hongos y microalgas con el fin de poder generar un análisis más riguroso.

- Sería importante en proyectos futuros incluir en la programación la posibilidad de analizar cómo se puede aumentar el potencial de remediación a través de consorcios bacterianos que trabajen de manera simbiótica.

BIBLIOGRAFÍA

Abdel-Razik, M. A., Azmy, A. F., Khairalla, A. S., & AbdelGhani, S. (2020). Metal bioremediation potential of the halophilic bacterium, Halomonas sp. strain WQL9 isolated from Lake Qarun, Egypt. *Egyptian Journal of Aquatic Research*, *46*(1), 19–25. https://doi.org/10.1016/j.ejar.2019.11.009

Adewoyin, M. A., & Okoh, A. I. (2018). The natural environment as a reservoir of pathogenic and non-pathogenic Acinetobacter species. *Reviews on Environmental Health*, *33*(3), 265–272. https://doi.org/10.1515/reveh-2017-0034

Amazon AWS. (2021). *Model Fit: Underfitting vs. Overfitting*. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Andrews, S. (2020). *Fastqc Tutorial & Faq*. Research Technology Support Facility.

Ansaldi, M., Théraulaz, L., Baraquet, C., Panis, G., & Méjean, V. (2007). Aerobic TMAO respiration in Escherichia coli. *Molecular Microbiology*, *66*(2), 484–494. https://doi.org/10.1111/j.1365-2958.2007.05936.x

Arishi, A., & Mashhour, I. (2021). Microbial mechanisms for remediation of hexavalent chromium and their large-scale applications; Current research and future directions. *Journal of Pure and Applied Microbiology*, *15*(1), 53–67. https://doi.org/10.22207/JPAM.15.1.32

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and their applications. *Journal of Chromatographic Science*, *55*(2), 182–196. https://doi.org/10.1093/chromsci/bmw167

Aslam, F., Yasmin, A., & Sohail, S. (2020). Bioaccumulation of lead, chromium, and nickel by bacteria from three different genera isolated from industrial effluent. *International Microbiology*, *23*(2), 253–261. https://doi.org/10.1007/s10123-019-00098-w

Ayele, A., & Godeto, Y. G. (2009). Chromium (VI) resistance and removal by Acinetobacter haemolyticus. *World Journal of Microbiology and Biotechnology*.

Ayele, A., & Godeto, Y. G. (2021). Bioremediation of Chromium by Microorganisms and Its Mechanisms Related to Functional Groups. *Journal of Chemistry*, *2021*. https://doi.org/10.1155/2021/7694157

Baldiris, R., Acosta-Tapia, N., Montes, A., Hernández, J., & Vivas-Reyes, R. (2018). *molecules Reduction of Hexavalent Chromium and Detection of Chromate Reductase (ChrR) in Stenotrophomonas maltophilia*. https://doi.org/10.3390/molecules23020406

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). The Glyoxylate Cycle Enables Plants and Bacteria to Grow on Acetate. In *Biochemistry* (p. section 17.4). http://www.ncbi.nlm.nih.gov/books/NBK22383/

Bhattacharya, A., Gupta, A., Kaur, A., & Malik, D. (2019). Alleviation of hexavalent chromium by using microorganisms: Insight into the strategies and complications. *Water Science and Technology*, *79*(3), 411–424. https://doi.org/10.2166/wst.2019.060

Bisong, E. (2019). Building Machine Learning and Deep Learning Models on Google Cloud Platform. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. https://doi.org/10.1007/978-1-4842-4470-8

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borisov, V. B., Gennis, R. B., Hemp, J., & Verkhovsky, M. I. (2011). The cytochrome bd respiratory oxygen reductases. *Biochimica et Biophysica Acta - Bioenergetics*, *1807*(11), 1398–1413. https://doi.org/10.1016/j.bbabio.2011.06.016

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology and Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Cai, L. J., Lv, S., & Shi, K. B. (2021). Application of an Improved CHI Feature Selection Algorithm. *Discrete Dynamics in Nature and Society*, *2021*. https://doi.org/10.1155/2021/9963382

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Campos, V. L., Moraga, R., Yánez, J., Zaror, C. A., & Mondaca, M. A. (2005). Chromate reduction by Serratia marcescens isolated from tannery effluent. *Bulletin of Environmental Contamination and Toxicology*, *75*(2), 400–406. https://doi.org/10.1007/s00128-005-0767-z

Can T. (2014). Introducción a la bioinformática. En: Yousef M., Allmer J. (eds) miRNomics: MicroRNA Biology and Computational Analysis. *Métodos En Biología Molecular (Métodos y Protocolos)*, *1107*.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pẽa, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Cardona, L. A. S., Vargas-Cardona, H. D., González, P. N., Peña, D. A. C., & Gutiérrez, Á. Á. O. (2020). Classification of categorical data based on the chi-square dissimilarity and t-sne. *Computation*, *8*(4), 1–15. https://doi.org/10.3390/computation8040104

Caviedes Rubio, D. I., Muñoz Calderón, R. A., Perdomo Gualtero, A., Rodríguez Acosta, D., & Sandoval Rojas, I. J. (2015). Tratamientos para la Remoción de Metales Pesados Comúnmente Presentes en Aguas Residuales Industriales. Una Revisión. *Ingeniería y Región*, *13*(1), 73. https://doi.org/10.25054/22161325.710

Chávez Porras, Á. (2010). Descripción de la nocividad del cromo proveniente de la industria curtiembre y de las posibles formas de removerlo. *Revista Ingenierías Universidad de Medellín*, *9*(17), 41–49.

Chen B., Ye X., Zhang B., Jing L., L. K. (2018). No Title. *World Seas*, *3*, 407–426.

Chen, J., & Tian, Y. (2021). Hexavalent chromium reducing bacteria: mechanism of reduction and characteristics. *Environmental Science and Pollution Research*, *28*(17), 20981–20997. https://doi.org/10.1007/s11356-021-13325-7

Chernikova, T. N., Bargiela, R., Toshchakov, S. V., Shivaraman, V., Lunev, E. A., Yakimov, M. M., Thomas, D. N., & Golyshin, P. N. (2020). Hydrocarbon-Degrading Bacteria Alcanivorax and Marinobacter Associated With Microalgae Pavlova lutheri and Nannochloropsis oculata. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.572931

Chivian, D., Clark, M., & Jungbluth, S. P. (2020). *Genome Extraction from Shotgun Metagenome Sequence Data*.

Christensen Editor, H. (2018). Introduction to Bioinformatics in Microbiology. *Learning Materials in Biosciences*, 219. https://doi.org/10.1007/978-3-319-99280-8%0Ahttp://link.springer.com/10.1007/978-3-319-99280-8

Chu, E., & Allegra, C. J. (1996). The role of thymidylate synthase in cellular regulation. *Advances in Enzyme Regulation*, *36*, 143–163. https://doi.org/10.1016/0065-2571(95)00004-6

Coelho, L. M., Rezende, H. C., Coelho, L. M., de Sousa, P. A. R., Melo, D. F. O., & Coelho, N. M. M. (2015). Bioremediation of Polluted Waters Using Microorganisms. *Advances in Bioremediation of Wastewater and Polluted Soil*, *10*(60770). https://doi.org/10.5772/60770

Committee, N. R. C. (US). (2007). Data Management and Bioinformatics Challenges of Metagenomics Metagenomics. In *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (pp. 1–158). https://www.ncbi.nlm.nih.gov/books/NBK53999/

Crespo, A., Blanco-Cabra, N., & Torrents, E. (2018). Aerobic vitamin B12 biosynthesis is essential for Pseudomonas aeruginosa Class II ribonucleotide reductase activity during planktonic and biofilm growth. *Frontiers in Microbiology*, *9*(MAY). https://doi.org/10.3389/fmicb.2018.00986

D., B., & J.A., F. (2015). Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. *BioData Mining*, *8*(1), 1–9. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L605597958%0Ahttp://dx.doi.org/10.1186/s13040-015-0055-3

DDBJ Release Statistics. (2020). https://www.ddbj.nig.ac.jp

Dijkhuizen, L., & Harder, W. (1984). Current views on the regulation of autotrophic carbon dioxide fixation via the Calvin cycle in bacteria. *Antonie van Leeuwenhoek*, *50*(5–6), 473–487. https://doi.org/10.1007/BF02386221

Domingo-Pueyoa, A., Sanz-Valeroa, J., & Wanden-Bergheb, C. (2014). Efectos sobre la salud de la exposición laboral al cromo y sus compuestos: revisión sistemática. *Arch Prev Riesgos Labor*, *17(3)*, 142–153.

Dotaniya, M. L., Rajendiran, S., Meena, V. D., Saha, J. K., Coumar, M. V., Kundu, S., & Patra, A. K. (2017). Influence of Chromium Contamination on Carbon Mineralization and Enzymatic Activities in Vertisol. *Agricultural Research*, *6*(1), 91–96. https://doi.org/10.1007/s40003-016-0242-6

Doucet, J.-D. H. and A. (2008). Renal Ion-Translocating ATPases: The P-Type Family. In *Seldin and Giebisch's The Kidney (Fourth Edition)*.

Eaton, R. W. (1997). p-cymene catabolic pathway in Pseudomonas putida F1: Cloning and characterization of DNA encoding conversion of p-cymene to p-cumate. *Journal of Bacteriology*, *179*(10), 3171–3180. https://doi.org/10.1128/jb.179.10.3171-3180.1997

Emmanuel, S. A., Sul, W. J., Seong, H. J., Rhee, C., Ekpheghere, K. I., Kim, I. S., Kim, H. G., & Koh, S. C. (2019). Metagenomic analysis of relationships between the denitrification process and carbon metabolism in a bioaugmented full-scale tannery wastewater treatment plant. *World Journal of Microbiology and Biotechnology*, *35*(10). https://doi.org/10.1007/s11274-019-2716-8

Ertani, A., Mietto, A., Borin, M., & Nardi, S. (2017). Chromium in Agricultural Soils and Crops: A Review. *Water, Air, and Soil Pollution*, *228*(5), 190. https://doi.org/10.1007/s11270-017-3356-y

Espinosa-Victoria, D., López-Reyes, L., Carcaño-Montiel, M. G., & Serret-López, M. (2020). The Burkholderia genus: between mutualism and pathogenicity. *Revista Mexicana de Fitopatología, Mexican Journal of Phytopathology*, *38*(3), 337–359. https://doi.org/10.18781/r.mex.fit.2004-5

Fakhar, A., Gul, B., Gurmani, A. R., Khan, S. M., Ali, S., Sultan, T., Chaudhary, H. J., Rafique, M., & Rizwan, M. (2020). Heavy metal remediation and resistance mechanism of Aeromonas, Bacillus, and Pseudomonas: A review. *Critical Reviews in Environmental Science and Technology*, 1–48. https://doi.org/10.1080/10643389.2020.1863112

Fang, H., Kang, J., & Zhang, D. (2017). Microbial production of vitamin B12: A review and future perspectives. *Microbial Cell Factories*, *16*(1). https://doi.org/10.1186/s12934-017-0631-y

Fernandez, M., Pereira, P. P., Agostini, E., & González, P. S. (2019). How the bacterial community of a tannery effluent responds to bioaugmentation with the consortium SFC 500-1. Impact of environmental variables. *Journal of Environmental Management*, *247*, 46–56. https://doi.org/10.1016/j.jenvman.2019.06.055

Fernandez, M., Pereira, P. P., Agostini, E., & González, P. S. (2020). Impact assessment of bioaugmented tannery effluent discharge on the microbiota of water bodies. *Ecotoxicology*, *29*(7), 973–986. https://doi.org/10.1007/s10646-020-02237-w

Ferreira, H. (2018). *Confusion matrix and other metrics in machine learning*. https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a

Field, E. K., Gerlach, R., Viamajala, S., Jennings, L. K., Peyton, B. M., & Apel, W. A. (2013). Hexavalent chromium reduction by Cellulomonas sp. strain

ES6: The influence of carbon source, iron minerals, and electron shuttling compounds. *Biodegradation*, *24*(3), 437–450. https://doi.org/10.1007/s10532-012-9600-7

Focardi, S., Pepi, M., & E., S. (2013). Microbial Reduction of Hexavalent Chromium as a Mechanism of Detoxification and Possible Bioremediation Applications. In *Biodegradation - Life of Science*. https://doi.org/10.5772/56365

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Applied and Environmental Microbiology*, *74*(8), 2461–2470. https://doi.org/10.1128/AEM.02272-07

Gamba, K. T., & Pedraza, A. M. (2017). Evaluación de estrategias de biorremediación para el tratamiento de aguas residuales industriales contaminadas con aceites usados. *Ingeciencia*, *2*(2), 18–30. http://editorial.ucentral.edu.co/ojs_uc/index.php/Ingeciencia/article/view/2679

GATK team. (2021). *Phred-scaled quality scores*. GATK Technical Documentation - Glossary. https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores

Gil Garzón, M., Soto, A., Usma Gutierrez, J., & Gutiérrez Florez, O. (2012). Contaminantes emergentes en aguas, efectos y posibles tratamientos. *Producción + Limpia*, *7*(2), 52–73.

Giuffrè, A., Borisov, V. B., Arese, M., Sarti, P., & Forte, E. (2014). Cytochrome bd oxidase and bacterial tolerance to oxidative and nitrosative stress. *Biochimica et Biophysica Acta - Bioenergetics*, *1837*(7), 1178–1187. https://doi.org/10.1016/j.bbabio.2014.01.016

González-Blanco, G., Pérez-Pérez, V., Aguirre-Garrido, J. F., Beristain-Cardoso, R., & Buendía-González, L. (2020). Kinetics and microbial structure of nitrogen cycle bacteria contained in the rhizosphere of natural wetland polluted with chromium. *Revista Mexicana de Ingeniera Quimica*, *19*(2), 543–553. https://doi.org/10.24275/rmiq/IA660

Gu, Q., Wu, Q., Zhang, J., Guo, W., Wu, H., & Sun, M. (2016). Community analysis and recovery of phenol-degrading bacteria from drinking water biofilters. *Frontiers in Microbiology*, *7*(APR). https://doi.org/10.3389/fmicb.2016.00495

Gu, Y., Xu, W., Liu, Y., Zeng, G., Huang, J., Tan, X., Jian, H., Hu, X., Li, F., & Wang, D. (2015). Mechanism of Cr(VI) reduction by Aspergillus niger: enzymatic characteristic, oxidative stress response, and reduction product. *Environmental Science and Pollution Research*, *22*(8), 6271–6279. https://doi.org/10.1007/s11356-014-3856-x

Gualdrón Durán, L. E. (2018). Evaluación de la calidad de agua de ríos de colombia usando parámetros físicoquímicos y biológicos. *Dinamica Ambiental*, *1*(1), 83–102. https://doi.org/10.18041/2590-6704/ambiental.1.2016.4593

Guevara, D. (2010). Biorremoción De Cromo (Cromo Total Y Cromo Vi) En Agua Sintética Por Dos Inóculos Bacterianos Nativos Compuestos , a Escala De Laboratorio. In *Escuela Politécnica Del Ejército* (Issue Previa a la obtención del Título de Ingenieraen Biotecnología).

Gunsalus, R. P., & Park, S. J. (1994). Aerobic-anaerobic gene regulation in Escherichia coli: control by the ArcAB and Fnr regulons. *Research in Microbiology*, *145*(5–6), 437–450. https://doi.org/10.1016/0923-2508(94)90092-2

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Handelsman, J. (2005). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *69*(1), 195–195. https://doi.org/10.1128/mmbr.69.1.195.2005

Hashmi, I., Bindschedler, S. and Junier, P. (2020). Beneficial Microbes in Agro-Ecology. *Beneficial Microbes in Agro-Ecology*, 363–396. https://doi.org/10.1016/c2020-0-00594-3

Huang, X. N., Min, D., Liu, D. F., Cheng, L., Qian, C., Li, W. W., & Yu, H. Q. (2019). Formation mechanism of organo-chromium (III) complexes from bioreduction of chromium (VI) by Aeromonas hydrophila. *Environment International*, *129*, 86–94. https://doi.org/10.1016/j.envint.2019.05.016

IBM Cloud Education. (2021). *Overfitting*. https://www.ibm.com/cloud/learn/overfitting

Initiative, A. G. (2000). *Chromium in freshwater and marine water*. Australian Goverment Initiative. https://www.waterquality.gov.au/anz-guidelines/guideline-values/default/water-quality-toxicants/toxicants/chromium-2000#:~:text=The current analytical practical quantitation,analysed to 0.5 µg%2FL.

Jaiswal, S., & Shukla, P. (2020). Alternative Strategies for Microbial Remediation of Pollutants via Synthetic Biology. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.00808

Jamil, M., Zeb, S., Anees, M., Roohi, A., Ahmed, I., ur Rehman, S., & Rha, E. shik. (2014). Role of Bacillus licheniformis in Phytoremediation of Nickel Contaminated Soil Cultivated with Rice. *International Journal of Phytoremediation*, *16*(6), 554–571. https://doi.org/10.1080/15226514.2013.798621

Jiménez, R. R., & Ph, D. (2021). *Practical Metagenomics : Microbiome tutorial with QIIME 2*. https://doi.org/https://doi.org/10.7490/f1000research.1118734.1

Jin, R., Liu, Y., Liu, G., Tian, T., Qiao, S., & Zhou, J. (2017). Characterization of Product and Potential Mechanism of Cr(VI) Reduction by Anaerobic Activated Sludge in a Sequencing Batch Reactor. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-01885-z

Jobby, R., Jha, P., Yadav, A. K., & Desai, N. (2018). Biosorption and biotransformation of hexavalent chromium [Cr(VI)]: A comprehensive review. *Chemosphere*, *207*, 255–266. https://doi.org/10.1016/j.chemosphere.2018.05.050

Joo, J. ock, Choi, J. H., Kim, I. H., Kim, Y. K., & Oh, B. K. (2015). Effective bioremediation of Cadmium (II), nickel (II), and chromium (VI) in a marine environment by using Desulfovibrio desulfuricans. *Biotechnology and Bioprocess Engineering*, *20*(5), 937–941. https://doi.org/10.1007/s12257-015-0287-6

Jung, I. (2020). *Bioinformatics Databases*. Bioinformatics. https://www.longdom.org/scholarly/bioinformatics-databases-journals-articles-ppts-list-2876.html

Kalam, S., Basu, A., Ahmad, I., Sayyed, R. Z., El-Enshasy, H. A., Dailin, D. J., & Suriani, N. L. (2020). Recent Understanding of Soil Acidobacteria and Their Ecological Significance: A Critical Review. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.580024

Kalsoom, A., Batool, R., & Jamil, N. (2021). Highly Cr(vi)-tolerant Staphylococcus simulans assisting chromate evacuation from tannery effluent. *Green Processing and Synthesis*, *10*(1), 295–308. https://doi.org/10.1515/gps-2021-0027

Katyal, P., & Kaur, G. (2018). Reduction of Cr (VI) by Micrococcus luteus isolate from Common Effluent Treatment Plants (CETPs). *International Journal of Current Microbiology and Applied Sciences*, *7*(07), 693–710. https://doi.org/10.20546/ijcmas.2018.707.084

Kazakov A, N. P. (2019). *Fama: a computational tool for comparative analysis of shotgun metagenomic data*.

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods in Molecular Biology*, *1399*, 207–233. https://doi.org/10.1007/978-1-4939-3369-3_13

Kelleher, J. D., Namee, B. Mac, & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. *Igarss 2014*, *1*, 1–691.

Kevin Ahern, Indira Rajagopal, & Taralyn Tan. (2021). *Metabolism: Citric Acid Cycle & Related Pathways Citric acid cycle*. https://bio.libretexts.org/@go/page/7837

Kim, Y. M., Park, H., & Chandran, K. (2016). Nitrification inhibition by hexavalent chromium Cr(VI) - Microbial ecology, gene expression and off-gas emissions. *Water Research*, *92*, 254–261. https://doi.org/10.1016/j.watres.2016.01.042

Kleine, L. L. (2012). Estadistica genómica (orientada a la predicción funcional de proteinas). *Universidad Nacional de Colombia*, 127.

Kuang, J., Huang, L., He, Z., Chen, L., Hua, Z., Jia, P., Li, S., Liu, J., Li, J., Zhou, J., & Shu, W. (2016). Predicting taxonomic and functional structure of microbial communities in acid mine drainage. *ISME Journal*, *10*(6), 1527–1539. https://doi.org/10.1038/ismej.2015.201

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2011). Using QIIME to analyze 16S rrna gene sequences from microbial communities. *Current Protocols in Bioinformatics*, *SUPPL.36*. https://doi.org/10.1002/0471250953.bi1007s36

Lamichhane, S., Sen, P., Dickens, A. M., Hyötyläinen, T., & Orešič, M. (2018). Data Analysis for Omic Sciences: Methods and Applications. *Comprehensive Analytical Chemistry*.

Langille, M. G. I. (2018). Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *MSystems*, *3*(2), e00163-17. https://doi.org/10.1128/msystems.00163-17

Lapierre A., L. (2013). Factores de virulencia asociados a especies zoonóticas de Campylobacter spp. *Avances En Ciencias Veterinarias*, *28*(1). https://doi.org/10.5354/0716-260x.2013.27866

Lehtinen, S., Lees, J., Bähler, J., Shawe-Taylor, J., & Orengo, C. (2015). Gene function prediction from functional association networks using kernel partial least squares regression. *PLoS ONE*, *10*(8). https://doi.org/10.1371/journal.pone.0134668

Li, L., Shang, X., Sun, X., Xiao, X., Xue, J., Gao, Y., & Gao, H. (2021). Bioremediation potential of hexavalent chromium by a novel bacterium Stenotrophomonas acidaminiphila 4-1. *Environmental Technology and Innovation*, *22*. https://doi.org/10.1016/j.eti.2021.101409

Linusson, H., & Olausson, A. (2012). *Feature Selection Using Random Forest* (Vol. 6, Issue 3, pp. 1319–1324). https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

Liu, B., Zhang, F., Feng, X., Liu, Y., Yan, X., Zhang, X., Wang, L., & Zhao, L. (2006). Thauera and Azoarcus as functionally important genera in a denitrifying quinoline-removal bioreactor as revealed by microbial community structure comparison. *FEMS Microbiology Ecology*, *55*(2), 274–286. https://doi.org/10.1111/j.1574-6941.2005.00033.x

Liu, Z., Chen, D., Sheng, L., & Liu, A. Y. (2013). Class Prediction and Feature Selection with Linear Optimization for Metagenomic Count Data. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0053253

Lorenzin, G., Piccinelli, G., Carlassara, L., Scolari, F., Caccuri, F., Caruso, A., & De Francesco, M. A. (2018). Myroides odoratimimus urinary tract infection in an immunocompromised patient: An emerging multidrug-resistant micro-organism. *Antimicrobial Resistance and Infection Control*, *7*(1). https://doi.org/10.1186/s13756-018-0391-4

Louppe, G. (2014). *UNDERSTANDING RANDOM FORESTS*. https://arxiv.org/pdf/1407.7502.pdf

Lu, X. M., & Lu, P. Z. (2014). Characterization of bacterial communities in sediments receiving various wastewater effluents with high-throughput sequencing analysis. *Microbial Ecology*, *67*(3), 612–623. https://doi.org/10.1007/s00248-014-0370-0

Luo, J. H., Wu, M., Liu, J., Qian, G., Yuan, Z., & Guo, J. (2019). Microbial chromate reduction coupled with anaerobic oxidation of methane in a membrane biofilm reactor. *Environment International*, *130*. https://doi.org/10.1016/j.envint.2019.104926

Mala, J. G. S., Takeuchi, S., & Mani, U. (2020). Microbial Chromate Reductases: Novel and Potent Mediators in Chromium Bioremediation-A Review. *Applied Microbiology: Theory & Technology*, 32–44. https://doi.org/10.37256/amtt.112020222

Mancera, N., & Álvarez Ricardo. (2006). Current State of Knowledge of the Concentration of Mercury and Other Heavy Metals in Fresh Water Fish in Colombia. *Acta Biológica Colombiana*, *11*, 21.

Marino, R. (2006). *Evaluación de las tecnologías de tratamiento de aguas subterráneas contaminadas con Cromo*. Universitat Politecnica Catalunya.

Mathur, S. (2017). Role of Prokayotic P-Type ATPases. *International Journal of Cell Science & Molecular Biology*, *3*(1). https://doi.org/10.19080/ijcsmb.2017.03.555602

Mayssara A. Abo Hassanin Supervised, A. (2014a). Bacterial mechanisms for Cr(VI) resistance and reduction: an overview and recent advances. *Paper Knowledge . Toward a Media History of Documents*, *59*, 321–332.

Mayssara A. Abo Hassanin Supervised, A. (2014b). 済無No Title No Title No Title. *Paper Knowledge . Toward a Media History of Documents*, 31–55.

Medlar, A. J., Törönen, P., & Holm, L. (2018). AAI-profiler: Fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. *Nucleic Acids Research*, *46*(W1), W479–W485. https://doi.org/10.1093/nar/gky359

Medlin, L. K., & Cembella, A. D. (2013). Biodiversity of Harmful Marine Algae. In *Encyclopedia of Biodiversity: Second Edition* (pp. 470–484). https://doi.org/10.1016/B978-0-12-384719-5.00404-4

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11257

Minematsu, A., Miyazaki, T., Shimamura, S., Nishikawa, H., Nakayama, H., Takazono, T., Saijo, T., Yamamoto, K., Imamura, Y., Yanagihara, K., Kohno, S., Mukae, H., & Izumikawa, K. (2019). Vacuolar proton-translocating ATPase is required for antifungal resistance and virulence of Candida glabrata. *PLoS ONE*, *14*(1). https://doi.org/10.1371/journal.pone.0210883

Mishra, S., Lin, Z., Pang, S., Zhang, W., Bhatt, P., & Chen, S. (2021). Recent Advanced Technologies for the Characterization of Xenobiotic-Degrading Microorganisms and Microbial Communities. *Frontiers in Bioengineering and Biotechnology*, *9*. https://doi.org/10.3389/fbioe.2021.632059

Mittal, P., Prasoodanan Pk, V., Dhakan, D. B., Kumar, S., & Sharma, V. K. (2019). Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes. *Environmental Microbiomes*, *14*(1), 1–12. https://doi.org/10.1186/s40793-019-0345-3

Moon, C. D., Young, W., Maclean, P. H., Cookson, A. L., & Bermingham, E. N. (2018). Metagenomic insights into the roles of Proteobacteria in the gastrointestinal microbiomes of healthy dogs and cats. *MicrobiologyOpen*, *7*(5). https://doi.org/10.1002/mbo3.677

Muccee, F., & Ejaz, S. (2020). Whole genome shotgun sequencing of POPs degrading bacterial community dwelling tannery effluents and petrol contaminated soil. *Microbiological Research*, *238*. https://doi.org/10.1016/j.micres.2020.126504

Mukherjee, A., Chettri, B., Langpoklakpam, J. S., Basak, P., Prasad, A., Mukherjee, A. K., Bhattacharyya, M., Singh, A. K., & Chattopadhyay, D. (2017). Bioinformatic Approaches Including Predictive Metagenomic Profiling Reveal Characteristics of Bacterial Response to Petroleum Hydrocarbon Contamination in Diverse Environments. *Scientific Reports*, *7*(1), 1–22. https://doi.org/10.1038/s41598-017-01126-3

Muralikrishna V., I., M. V. (2017). No Title. *Environmental Management. Science and Engineering for Industry*, 1–4.

Myllykallio, H., Sournia, P., Heliou, A., & Liebl, U. (2018). Unique features and anti-microbial targeting of folate- and flavin-dependent methyltransferases required for accurate maintenance of genetic information. *Frontiers in Microbiology*, *9*(MAY). https://doi.org/10.3389/fmicb.2018.00918

Nikolai, J. (2018). Understanding the Covariance Matrix. In *Datascienceplus* (pp. 1–11). https://datascienceplus.com/understanding-the-covariance-matrix/

Ninla Elmawati Falabiba. (2019a). 済無No Title No Title No Title. *Green Chemistry*, 261–290.

Ninla Elmawati Falabiba. (2019b). 済無No Title No Title No Title. *International Encyclopedia of Public Health*, *2*, 240–247.

O'Neill, A. G., Beaupre, B. A., Zheng, Y., & Liu, D. (2020). NfoR: Chromate Reductase or Flavin Mononucleotide Reductase? *Applied and Environmental Microbiology*, *86*(22). https://doi.org/10.1128/AEM.01758-20

Ome Barrera, Ó., & Zafra Mejía, C. (2018). Factores clave en procesos de biorremediación para la depuración de aguas residuales. Una revisión. *Revista U.D.C.A Actualidad & Divulgación Científica*, *21*(2), 573–585. https://doi.org/10.31910/rudca.v21.n2.2018.1037

Ortiz-Estrada, Á. M., Gollas-Galván, T., Martínez-Córdova, L. R., & Martínez-Porchas, M. (2019). Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. *Reviews in Aquaculture*, *11*(1), 234–245. https://doi.org/10.1111/raq.12237

Ozawa, K., Meikari, T., Motohashi, K., Yoshida, M., & Akutsu, H. (2000). Evidence for the presence of an F-type ATP synthase involved in sulfate respiration in Desulfovibrio vulgaris. *Journal of Bacteriology*, *182*(8), 2200–2206. https://doi.org/10.1128/JB.182.8.2200-2206.2000

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Pedregosa, F. \emphet al., Pedregosa, F., Weiss, R., Brucher, M., Pedregosa, F. et al., Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490

Pei, Y., Yu, Z., Ji, J., Khan, A., & Li, X. (2018). Microbial community structure and function indicate the severity of chromium contamination of the Yellow River. *Frontiers in Microbiology*, *9*(JAN). https://doi.org/10.3389/fmicb.2018.00038

Portillo, F. (2000). Regulation of plasma membrane H+-ATPase in fungi and plants. *Biochimica et Biophysica Acta - Reviews on Biomembranes*, *1469*(1), 31–42. https://doi.org/10.1016/S0304-4157(99)00011-8

Qu, K., Guo, F., Liu, X., Lin, Y., & Zou, Q. (2019). Application of machine learning in microbiology. *Frontiers in Microbiology*, *10*(APR). https://doi.org/10.3389/fmicb.2019.00827

Qurbani, K., & Hamzah, H. (2020). Intimate communication between Comamonas aquatica and Fusarium solani in remediation of heavy metal-polluted environments. *Archives of Microbiology*, *202*(6), 1397–1406. https://doi.org/10.1007/s00203-020-01853-8

Rahman, S. F., Olm, M. R., Morowitz, M. J., & Banfield, J. F. (2017). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *BioRxiv*. https://doi.org/10.1101/185348

Rahman, Z., & Thomas, L. (2021). Chemical-Assisted Microbially Mediated Chromium (Cr) (VI) Reduction Under the Influence of Various Electron Donors, Redox Mediators, and Other Additives: An Outlook on Enhanced Cr(VI) Removal. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.619766

Ramees, T. P., Dhama, K., Karthik, K., Rathore, R. S., Kumar, A., Saminathan, M., Tiwari, R., Malik, Y. S., & Singh, R. K. (2017). Arcobacter: An emerging food-borne zoonotic pathogen, its public health concerns and advances in diagnosis and control - A comprehensive review. *Veterinary Quarterly*, *37*(1), 136–161. https://doi.org/10.1080/01652176.2017.1323355

Rathore, A. S., & Gupta, R. D. (2015). Chitinases from Bacteria to Human: Properties, Applications, and Future Perspectives. *Enzyme Research*, *2015*. https://doi.org/10.1155/2015/791907

Reguera, G., & Leschine, S. B. (2001). Chitin degradation by cellulolytic anaerobes and facultative aerobes from soils and sediments. *FEMS Microbiology Letters*, *204*(2), 367–374. https://doi.org/10.1016/S0378-1097(01)00429-3

Ren, Y., Niu, J., Huang, W., Peng, D., Xiao, Y., Zhang, X., Liang, Y., Liu, X., & Yin, H. (2016). Comparison of microbial taxonomic and functional shift pattern along contamination gradient. *BMC Microbiology*, *16*(1), 110. https://doi.org/10.1186/s12866-016-0731-6

Ren, Z., Wang, F., Qu, X., Elser, J. J., Liu, Y., & Chu, L. (2017). Taxonomic and functional differences between microbial communities in Qinghai Lake and its input streams. *Frontiers in Microbiology*, *8*(NOV). https://doi.org/10.3389/fmicb.2017.02319

Richardson, A. R., Somerville, G. A., & Sonenshein, A. L. (2015). Regulating the intersection of metabolism and pathogenesis in gram-positive bacteria. *Metabolism and Bacterial Pathogenesis*, 129–165. https://doi.org/10.1128/9781555818883.ch7

Romero López, T. de J., & Vargas Mato, D. (2017). Uso de microorganismos eficientes para tratar aguas contaminadas. *Ingeniería Hidráulica y Ambiental*, *38*(3), 88–100.

Rudakiya, D. M. (2013). Evaluation of remediation in heavy metal tolerance and removal by Comamonas acidovorans MTCC 3364. *IOSR Journal Of Environmental Science, Toxicology And Food Technology*, *5*(5), 26–32. https://doi.org/10.9790/2402-0552632

Ruiz-Lozano, J. M., & Azcón, R. (2011). Brevibacillus, Arbuscular Mycorrhizae and Remediation of Metal Toxicity in Agricultural Soils. *Soil Biology*, *27*, 235–258. https://doi.org/10.1007/978-3-642-19577-8_12

Ruiz-Moreno, H. A., López-Tamayo, A. M., Caro-Quintero, A., Husserl, J., & González Barrios, A. F. (2019). Metagenome level metabolic network reconstruction analysis reveals the microbiome in the Bogotá River is functionally close to the microbiome in produced water. *Ecological Modelling*, *399*, 1–12. https://doi.org/10.1016/j.ecolmodel.2019.02.001

Ryan, M. P., & Tony Pembroke, J. (2020). The genus ochrobactrum as major opportunistic pathogens. *Microorganisms*, *8*(11), 1–30. https://doi.org/10.3390/microorganisms8111797

Salinero, K. K., Keller, K., Feil, W. S., Feil, H., Trong, S., Di Bartolo, G., & Lapidus, A. (2009). Metabolic analysis of the soil microbe Dechloromonas aromatica str. RCB: Indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation. *BMC Genomics*, *10*. https://doi.org/10.1186/1471-2164-10-351

Sammut, C., & Webb, G. I. (Eds.). (2010). Leave-One-Out Cross-Validation. In *Encyclopedia of Machine Learning* (pp. 600–601). Springer US. https://doi.org/10.1007/978-0-387-30164-8_469

Sanschagrin, S., & Yergeau, E. (2014). Next-generation sequencing of 16S ribosomal RNA gene amplicons. *Journal of Visualized Experiments*, *90*. https://doi.org/10.3791/51709

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L. 0003 3527 8101, Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883–8900. https://doi.org/10.1093/nar/gkaa621

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Solís-González, C. J., & Loza-Tavera, H. (2019). Alicycliphilus: current knowledge and potential for bioremediation of xenobiotics. *Journal of Applied Microbiology*, *126*(6), 1643–1656. https://doi.org/10.1111/jam.14207

Solutions, E. (2016). *Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures @ blog.exsilio.com*. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Soueidan, H., & Nikolski, M. (2016). Machine learning for metagenomics: methods and tools. *Metagenomics*, *1*(1). https://doi.org/10.1515/metgen-2016-0001

Speigth G., J. (2017). No Title. *Environmental Organic Chemistry for Engineers*, . 387-432.

Speigth G., J. (2018). Reaction Mechanisms in Environmental Engineering. *Reaction Mechanisms in Environmental Engineering*, 337–384. https://doi.org/10.1016/c2013-0-16045-x

Sul, W. J., Kim, I. S., Ekpeghere, K. I., Song, B., Kim, B. S., Kim, H. G., Kim, J. T., & Koh, S. C. (2016). Metagenomic insight of nitrogen metabolism in a tannery wastewater treatment plant bioaugmented with the microbial consortium BM-S-1. *Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering*, *51*(13), 1164–1172. https://doi.org/10.1080/10934529.2016.1206387

Szczerbicki, E. (2001). Management of Complexity and Information Flow. In *Agile Manufacturing: The 21st Century Competitive Strategy* (pp. 247–263). https://doi.org/10.1016/b978-008043567-1/50013-9

Tang, X., Huang, Y., Li, Y., Wang, L., Pei, X., Zhou, D., He, P., & Hughes, S. S. (2021). Study on detoxification and removal mechanisms of hexavalent chromium by microorganisms. *Ecotoxicology and Environmental Safety*, *208*. https://doi.org/10.1016/j.ecoenv.2020.111699

Tariq, M., Waseem, M., Rasool, M. H., Zahoor, M. A., & Hussain, I. (2019). Isolation and molecular characterization of the indigenous Staphylococcus aureus strain K1 with the ability to reduce hexavalent chromium for its application in bioremediation of metal -contaminated sites. *PeerJ*, *2019*(10). https://doi.org/10.7717/peerj.7726

Thomas, F., Hehemann, J. H., Rebuffet, E., Czjzek, M., & Michel, G. (2011). Environmental and gut Bacteroidetes: The food connection. *Frontiers in Microbiology*, *2*(MAY). https://doi.org/10.3389/fmicb.2011.00093

Thompson, J., Johansen, R., Dunbar, J., & Munsky, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS ONE*, *14*(7). https://doi.org/10.1371/journal.pone.0215502

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., & Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *MBio*, *11*(3). https://doi.org/10.1128/mBio.00434-20

Ueno, S., Kaieda, N., & Koyama, N. (2000). Characterization of a P-type Na+-ATPase of a facultatively anaerobic alkaliphile, Exiguobacterium aurantiacum. *Journal of Biological Chemistry*, *275*(19), 14537–14540. https://doi.org/10.1074/jbc.275.19.14537

Upadhyay, M. K., Yadav, P., Shukla, A., & Srivastava, S. (2018). Utilizing the potential of microorganisms for managing arsenic contamination: A feasible and sustainable approach. *Frontiers in Environmental Science*, *6*(MAY). https://doi.org/10.3389/fenvs.2018.00024

Valencia, A., Suárez Castaño, R., Sánchez, A., Cardozo, E., Bonilla, M., & Buitrago, C. (2009). Gestión de la contaminación ambiental: cuestión de corresponsabilidad. *Revista de Ingeniería*, *30*(30), 90–99. https://doi.org/10.16924/revinge.30.11

Vegge, C. S., Jansen van Rensburg, M. J., Rasmussen, J. J., Maiden, M. C. J., Johnsen, L. G., Danielsen, M., MacIntyre, S., Ingmer, H., & Kelly, D. J. (2016). Glucose metabolism via the entner-doudoroff pathway in campylobacter: A rare trait that enhances survival and promotes biofilm formation in some isolates. *Frontiers in Microbiology*, *7*(NOV). https://doi.org/10.3389/fmicb.2016.01877

Vélez, J. A., Quiroz, L. F., Ruiz, O. S., Montoya, O. I., Turrión, M., & Ordúz, S. (2021). Hexavalent chromium-reducing bacteria on biosolids from the San Fernando Wastewater Treatment Plant in Medellín (Colombia) Bacterias aisladas de biosólidos de la PTAR San Fernando en Medellín-Colombia con capacidad para reducir cromo hexavalente. *Revista Colombiana de Biotecnología*, *XXIII*(1), 32–45.

Verma, S. K., & Sharma, P. C. (2020). NGS-based characterization of microbial diversity and functional profiling of solid tannery waste metagenomes. *Genomics*, *112*(4), 2903–2913. https://doi.org/10.1016/j.ygeno.2020.04.002

Wang, Z., Zhang, X. X., Lu, X., Liu, B., Li, Y., Long, C., & Li, A. (2014). Abundance and diversity of bacterial nitrifiers and denitrifiers and their functional genes in tannery wastewater treatment plants revealed by high-throughput sequencing. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0113603

Wolfe, A. J. (2015). Glycolysis for Microbiome Generation. *Microbiology Spectrum*, *3*(3). https://doi.org/10.1128/microbiolspec.mbp-0014-2014

Woloszynek, S., Mell, J. C., Zhao, Z., Simpson, G., O'Connor, M. P., & Rosen, G. L. (2019). Exploring thematic structure and predicted functionality of 16S rRNA amplicon data. *PLoS ONE*, *14*(12). https://doi.org/10.1371/journal.pone.0219235

Wu, B., Liu, F., Fang, W., Yang, T., Chen, G. H., He, Z., & Wang, S. (2021). Microbial sulfur metabolism and environmental implications. *Science of the Total Environment*, *778*. https://doi.org/10.1016/j.scitotenv.2021.146085

Zakaria, Z. A., Zakaria, Z., Surif, S., & Ahmad, W. A. (2007). Hexavalent chromium reduction by Acinetobacter haemolyticus isolated from heavy-metal contaminated wastewater. *Journal of Hazardous Materials*, *146*(1–2), 30–38. https://doi.org/10.1016/j.jhazmat.2006.11.052

Zhang, SY, Wang, QF, Wan, R. y Xie, S. (2011). Cambios en la comunidad bacteriana de la biorremediación del antraceno en el suelo de compostaje de residuos sólidos urbanos. *Revista de La Universidad de Zhejiang*, *12 (9)*, 760–768.

Zhang, N. (2012). *Chromate reduction by Desulfovibrio desulfuricans ATCC 27774*.

Zheng, X. ying, Lu, D., Wang, M. yang, Chen, W., Zhou, G., & Zhang, Y. (2018). Effect of chromium (VI) on the multiple nitrogen removal pathways and microbial community of aerobic granular sludge. *Environmental Technology (United Kingdom)*, *39*(13), 1682–1696. https://doi.org/10.1080/09593330.2017.1337230

Zúñiga Trejos, S. (2014). *Desarrollo de herramientas bioinformáticas aplicadas al Diagnóstico Genético mediante Secuenciación Masiva*.

ANEXOS

*Los archivos se encuentran adjuntos con el documento

Para dirigirte a la carpeta de drive con todos los documentos tenidos en cuenta para el desarrollo del proyecto ingresa a el siguiente link: https://drive.google.com/drive/folders/1qrVA_LznLqrSReraNyjpWHOXjQisLqX-?usp=sharing

Anexo 1. Base de datos relacionada a microorganismos en ambientes contaminados con cromo (VI).

Anexo 2. Archivo de metadata para el análisis en QIIME 2.

Anexo 3. Archivos de visualización del análisis bioinformático en QIIME 2.

Anexo 4. Base de datos relacionada a algoritmos de Machine learning aplicados en comunidades microbianas.

Anexo 5. Etiquetas asignadas a partir del potencial de remediación.

Anexo 6. Código en Google colaboratory: Algoritmo de árbol de decisión y bosques aleatorios.

Anexo 7. Set de datos para el modelo de machine learning.

Anexo 8. Salida del algoritmo: Clasificación.