UNIVERSIDAD AUTONOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACION EN CIENCIAS BASICAS Y APLICADAS
CENTRO DE INVESTIGACIONES DE CIENCIAS (CINC)

Desarrollo de un nuevo método computacional para discriminar taxonómicamente las secuencias de virus generadas por tecnologías de secuenciación masiva de ADN para estudios de metagenómica

T E S I S

Que para obtener el Grado de:

DOCTORADO EN CIENCIAS

Presenta

ELIZABETH CADENAS CASTREJÓN

Directoras de tesis

Dra. Lorena Díaz González

Dra. Blanca Itzelt Taboada Ramírez

CUERNAVACA, MORELOS Marzo, 2023

LISTA DEL JURADO REVISOR DE TESIS

JURADO	NOMBRE	ADSCRIPCIÓN
Presidente	Dra. Sonia Dávila Ramos	CIDC-UAEM
Secretario	Dr. Juan Manuel Rendón Mancha	CInC-UAEM
Vocal	Dr. Guillermo Santamaría Bonfil	INEEL
Vocal	Dr. Edgar Francisco Román Rangel	ITAM
Vocal	Dr. Jorge Hermosillo Valadez	CInC-UAEM
Suplente	Dr. Outmane Oubram	FCQeI-UAEM
Suplente	Dra. Lorena Díaz González	CInC-UAEM

PUBLICACIÓN RELACIONADA CON ESTA TESIS

Elizabeth Cadenas-Castrejón, Jérôme Verleyen, Celia Boukadida, Lorena Díaz-González, Blanca Taboada, Evaluation of tools for taxonomic classification of viruses, Briefings in Functional Genomics, Volume 22, Issue 1, January 2023, Pages 31–41, https://doi.org/10.1093/bfgp/elac036

DEDICATORIA

A mis familiares (padres, hermanas, hermano, sobrinos, tías, tíos, primos y primas) quienes me han animado a continuar y en especial a mi hermana Rosario quien me ha ayudado tanto cuando más la he necesitado.

A mis directoras de tesis, la Dra. Lorena Díaz González y la Dra. Blanca Itzelt Taboada Ramírez, quienes me han ayudado, guiado y aconsejado durante todo el doctorado.

A mi esposo Francisco Christian y a mi hijo Christian Alexander que ha estado a mi lado en los momentos difíciles, en mis alegrías, que con su amor y apoyo incondicional me han dado la fuerza para lograr mis metas, seguir adelante y no rendirme en las dificultades.

A dios que me ha dado vida, salud, muchas personas que me quieren y sobre todo una gran perseverancia y fuerza que me hace continuar todos los días.

AGRADECIMIENTO

Al Consejo Nacional de Ciencias y Tecnología (CONACYT) por el apoyo económico proporcionado para mis estudios, los cuales me permitieron realizar esta investigación de tesis.

A la Universidad Autónoma del Estado de Morelos (UAEM) por brindarme la oportunidad de efectuar mi doctorado.

Al Instituto de Biotecnología de la Universidad Nacional Autónoma de México, por darme acceso sus instalaciones para desarrollar esta investigación.

A mis directoras de tesis, la Dra. Lorena Díaz González y la Dra. Blanca Itzelt Taboada Ramírez, por el apoyo, la ayuda, consejos y el tiempo proporcionado en todo momento en el transcurso del desarrollo de esta investigación.

A mis revisores el Dr. Jorge Hermosillo Valadez, Dr. Juan Manuel Rendón Mancha, Dr. José Alberto Hernández Aguilar y Dr. Outmane Oubram, por las observaciones y correcciones que han permitido el progreso y finalización de esta investigación.

Al Dr. Carlos Arias, la Dra. Susana López y a todos los integrantes del grupo Arias-López del IBT-UNAM (Dr. Pavel Isa, Dr. Tomás López, Dr. Carlos Sandoval, técnicos y estudiantes) que con su apoyo me han permitido mejorar.

A Jerome Verleyen por el apoyo técnico brindado, por compartir sus conocimientos, porque siempre estuvo disponible y al pendiente en la resolución de mis dudas en Teopanzolco.

A Héctor Oliver por el apoyo brindado para solucionar los problemas en Xiuhcoatl.

A los recursos en el Clúster Teopanzolco, que pertenece a la infraestructura de HPC en la Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología (UNAM), que forma parte del Laboratorio Nacional de Apoyo Tecnológico a las Ciencias Genómicas (CONACyT).

Al Proyecto de P-10108 THE REGENTS OF THE UNIVERSITY OF CALIFORNIA por el apoyo económico proporcionado para la finalización del doctorado.

Al tiempo de cómputo otorgado por LANCAD y CONACYT en la supercomputadora Yoltla/Miztli/Xiuhcoatl en LSVP UAM-Iztapalapa/DGTIC UNAM/CGSTIC CINVESTAV.

A los recursos informáticos proporcionados por la supercomputadora MIZTLI de la Dirección General de Computación y Tecnologías de la Información y las Comunicaciones (DGTIC) de la Universidad Nacional Autónoma de México a través de los proyectos LANCAD-UNAM-DGTIC-350 y LANCAD-UNAM-DGTIC-396.

A los recursos informáticos proporcionados por el Clúster Híbrido de Supercómputo Xiuhcóatl de la Coordinación General de Sistemas de Tecnologías de la Información y las Comunicaciones (CGSTIC) del Centro de Investigación y de Estudios Avanzados (CINVESTAV) a través de los proyectos LANCAD 17-2021, LANCAD 10-2022 y LANCAD 6-2023.

A Juan Manuel Hurtado, Roberto Bahena y David Santiago Castañeda del Instituto de Biotecnología de la UNAM por su apoyo informático.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Resumen

Los virus son agentes microscópicos acelulares que requieren una célula hospedera para sobrevivir. Estos pueden infectar a todas las formas de vida en la Tierra, incluyendo los tres dominios de la vida, eucaria, bacteria y arquea. Los virus han causado algunas de las enfermedades más dramáticas y mortales en la historia humana. Sin embargo, la detección de virus permaneció muy limitada hasta el desarrollo de la metagenómica, la cual es el estudio de los fragmentos de secuencias del genoma de todos los diferentes microorganismos presentes en una muestra que se recupera directamente de un ambiente u hospedero. Los estudios metagenómicos han sido posibles gracias a las tecnologías de secuenciación de nueva generación (NGS; *Next-Generation Sequencing*), las cuales permiten obtener las secuencias de ADN de todos los ácidos nucleicos presentes en una muestra, generando grandes volúmenes de datos. Estos requieren ser analizados con métodos formales de computación; uno de los análisis es la clasificación taxonómica. La mayoría de los métodos existentes para este tipo de análisis se enfocan en la clasificación de secuencias bacterianas. Las herramientas encargadas de la clasificación de virus tienen una baja sensibilidad, debido a: i) La poca abundancia de las secuencias virales, ya que estas solo representan del 1% al 5% del ADN total obtenido de una muestra. ii) No existen genes marcadores universales, como en las bacterias, que permitan caracterizarlos fácilmente. iii) La mayoría (usualmente, entre el 60% y 99%) de las secuencias de virus obtenidas en cualquier ambiente no tienen similitud con otras secuencias en las bases de datos (BD) de referencia. Aunado a esto, el tiempo de procesamiento generalmente es muy costoso.

Con los problemas antes mencionados y aunado al aumento de datos metagenómicos, se han iniciado el uso de nuevas técnicas que sean capaces de trabajar con un gran conjunto de información y encontrar patrones de ellos, como lo son las redes neuronales profundas.

El objetivo de este proyecto fue desarrollar un nuevo método computacional que permite discriminar (clasificar) taxonómicamente las lecturas cortas de ADN de virus generadas por tecnologías de secuenciación masiva de ADN para estudios de metagenómica. El método desarrollado considera casi todas las familias virales definidas hasta enero del 2020 (169 para la información de NCBI nt y 127 para RefSeq) e incluye los virus que no pertenecen a ninguna familia viral, es decir, no tiene definida una asignación taxonómica a nivel familia. Dicho método realiza una clasificación a nivel nucleótido e identifica a qué familia pertenecen las secuencias mediante el uso de una red neuronal de convolución (en inglés *Convolution Neural Networks,* CNN), las cuales son un tipo de red neuronal profunda que identifica patrones en la información, comparte parámetros y reducen la dimensionalidad.

CAPÍTULO 1. INTRODUCCIÓN

1.1. Introducción

Los microorganismos que se pueden cultivar corresponden solamente al 1% del total presente en la Tierra, porque en su mayoría no se han encontrado aún estrategias metodológicas que permitan su cultivo, debido a la falta de conocimiento en las condiciones de cultivo como temperatura, nutrientes y su simbiosis (1). Los virus poseen diversas características tales como: (i) Son las entidades biológicas más abundantes en la Tierra; se sabe que existe más de 10^{30} bacteriófagos tan solo en los océanos (2); (ii) Son parásitos intracelulares que pueden infectar a los tres dominios de la vida (3); (iii) Son considerados como los principales actores de los ecosistemas de la naturaleza, por ejemplo, en los océanos impulsan los ciclos geoquímicos (2), e (iv) Influyen en el estado de salud de los humanos (4). Debido a limitaciones relacionadas principalmente con su identificación y cultivo, la detección de virus permaneció muy limitada hasta el desarrollo de la metagenómica (5).

La metagenómica es el estudio de los genomas de todos los microorganismos presentes en una muestra, que se recupera directamente de un sitio o huésped, sin necesidad de cultivarlos (5). Los estudios metagenómicos basados en secuenciación masiva han sido posibles gracias a las tecnologías de secuenciación de nueva generación (NGS; Next-Generation Sequencing), las cuales permiten obtener las secuencias de todos los ácidos nucleicos de los agentes microscópicos presentes en una muestra, que no se pueden cultivar, tales como los virus. El uso de la secuenciación masiva ha aumentado debido a la disminución de sus costos y diversidad de usos, y como resultado se han generado grandes volúmenes de datos de secuencias de ADN que requieren de herramientas específicas para su análisis. La clasificación taxonómica es un análisis primordial en estudios metagenómicos, dado que permite identificar todas las secuencias obtenidas en la muestra, para asignarlas a diferentes niveles taxonómicos, como lo son (de lo general a lo específico) dominio, reino, filo, clase, orden, familia, género o especie.

La mayoría de las herramientas existentes para la clasificación taxonómica se enfocan principalmente en la clasificación de secuencias de bacterias, mientras que las de virus no han mostrado resultados tan eficientes y por ello siguen siendo objeto de investigaciones con el fin de atacar sus limitaciones y mejorar los resultados obtenidos.

1.2. Planteamiento del problema y justificación

Las herramientas para la clasificación taxonómica de virus requieren un tiempo de procesamiento computacional generalmente costoso, presentando una alta precisión, pero baja sensibilidad, de acuerdo con el análisis realizado y reportado en el capítulo 4. La mayoría de estas herramientas, fueron diseñadas para realizar una clasificación de secuencias bacterianas. Las bacterias cuentan con el gen ribosomal 16S, el cual es común en todos sus genomas y permite clasificarlas fácilmente mediante este marcador único. Además, los genomas bacterianos son más conservados y anotados en las bases de datos (BD). En cuanto a la parte viral, su clasificación conlleva varios problemas, tales como:

Las secuencias virales, obtenidas después del proceso de limpieza de las muestras, son poco abundantes, ya que el ADN viral solo representa del 1 al 5% en comparación con el ADN total de la muestra (4).

En los virus no existen genes marcadores, como en las bacterias, que permitan caracterizarlos fácilmente.

La mayoría (60% - 99%) de las secuencias de virus obtenidas en cualquier ambiente, no tienen similitud con otras secuencias en las BD de referencia (6), ya que están poco anotados, y han sido menos caracterizados que las bacterias.

Algunas herramientas (7–13) que incluyen la clasificación de virus agrupan las secuencias con características similares, pero no realizan un proceso de clasificación como tal; existen otras que realizan la clasificación, pero únicamente a niveles de taxonomía muy generales. Además, todas estas herramientas presentan bajas sensibilidades y altas precisión, incluso a niveles altos de taxonomía.

Debido a los problemas mencionados anteriormente, se requiere de un nuevo método computacional que mejore la clasificación taxonómica de virus en datos de muestras metagenómicas. Este método está basado en redes neuronales de convolución, un tipo de redes neuronales profundas, debido a que tiene la capacidad de aprender, identificar y extraer patrones de grandes cantidades de datos y reducir la complejidad de la red.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un nuevo método computacional basado en redes de convolución eficiente que permita discriminar taxonómicamente, a nivel familia, las secuencias de virus generadas por tecnologías de secuenciación masiva de ADN para estudios de metagenómica viral.

1.3.2. Objetivos específicos

I. Evaluar diferentes arquitecturas de las redes neuronales de convolución (en inglés *Convolution Neural Networks*, CNN) e hiperparámetros para la clasificación de lecturas virales a nivel taxonómico de familia.

II. Entrenar la CNN para la clasificación utilizando la arquitectura final seleccionada en el objetivo I. La CNN se entrenó utilizando dos bases de datos (BD) de referencia de manera independiente. Por lo tanto, se generaron dos modelos, con la misma arquitectura, pero diferentes conjuntos de entrenamiento. El primer modelo se entrenó utilizando la BD viral de referencia curada y no redundante de RefSeq (14). Esta BD tiene una sola secuencia por cada especie viral y es una de las más utilizada por diversas herramientas de clasificación taxonómica. El segundo modelo se entrenó empleando la BD viral de nucleótidos reducida y no redundante del Centro Nacional de Información Biotecnológica (en inglés *National Center for Biotechnology Information*, NCBI) (15). Esta BD es la que contiene mayor información, ya que tiene muchas secuencias por cada especie viral y por consiguiente es más grande.

1.4. Organización de la tesis

El resto de la presente tesis se presenta en los siguientes capítulos:

- El capítulo dos presenta en forma breve los fundamentos teóricos de las redes neuronales de convolución.
- El capítulo tres presenta los antecedentes relacionados con las herramientas de clasificación taxonómica viral.
- El capítulo cuatro presenta la evaluación de once herramientas reportadas en la literatura para la clasificación taxonómica. Así también, en este capítulo se describen los conjuntos de datos usados para la validación externa tanto de las herramientas existentes como de los modelos de clasificación desarrollados en este trabajo.
- El capítulo cinco describe la metodología para la clasificación de lecturas virales a nivel familia.
- El capítulo seis presenta los resultados obtenidos de la clasificación a nivel familia implementando la metodología de solución propuesta.
- El capítulo siete expone las conclusiones y trabajos futuros de este trabajo de investigación.

CAPÍTULO 2. MARCO TEÓRICO DE LAS REDES NEURONALES DE CONVOLUCIÓN

En este capítulo se explica en forma breve la técnica de aprendizaje automático basada en redes neuronales de convolución, que ha sido empleada en este proyecto de tesis para desarrollar los modelos de clasificación taxonómica viral a nivel familia.

2.1. Redes Neuronales de Convolución

Las redes neuronales artificiales son una técnica de aprendizaje automático, cuya unidad básica es la neurona o nodo. Las neuronas reciben los datos de entrada, cada entrada tiene asociado un peso y a cada neurona se le asocia un sesgo (o bias); las entradas, los pesos y el sesgo hacen una combinación lineal. Luego, el resultado de la suma se pasa a través de una función no lineal, llamada función de activación y la salida de la neurona es enviada a otra neurona (16) (ver Figura 2.1).

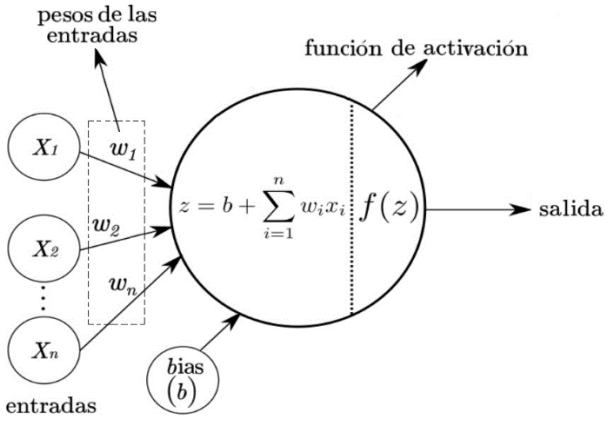


Figura 2. 1. Neurona simple.

La función de activación no lineal que utiliza las neuronas se usa para determinar las salidas de los nodos que se propaga hacia la siguiente capa. Esta se aplica en las capas de convolución o totalmente conectada. Además, realiza la transformación no lineal de los datos, lo que permite que el modelo aprenda y se adapte a una variedad de datos complejos. Existen diferentes tipos de funciones de activación, tales como sigmoide, tangente hiperbólica (tanh), softmax, ReLU (*Rectified Linear Unit*) y las variantes de ReLU (18). La función de activación más utilizada en las redes neuronales es la función ReLU (ec. 2.1), la cual es una función no lineal que permite que los valores positivos se mantengan para activar los nodos y que los valores negativos, menores o igual a cero sean cero.

$$\text{ReLU} = f(x) = max\ (0, x)\ , \tag{2.1}$$

donde *x* es el valor del resultado de la sumatoria del producto de los datos de entrada con sus pesos asociados y el sesgo calculado en un nodo (18).

En general, las redes neuronales artificiales estándares se constituyen básicamente por capas, las cuales están formadas por un conjunto de neuronas o nodos. La capa de una red se une con otra capa mediante los nodos, es decir, la salida de los nodos de una capa se utiliza como entrada de los nodos de la siguiente capa para transmitir señales (16). En la Figura 2.2 se esquematiza una red neuronal totalmente conectada, la cual es una red neuronal artificial estándar. En esta figura los círculos rellenos en azul son los datos de entrada y los círculos sin relleno representan las neuronas (ver Figura 2.1). Las flechas indican como la salida de cada neurona se transmiten a las neuronas de la siguiente capa. La última capa de la red totalmente conectada es la capa de salida, en donde cada neurona representa una clase (salida) de la red, si tiene una neurona es una red binaria (dos clases) y si tiene varía neuronas es multiclase.

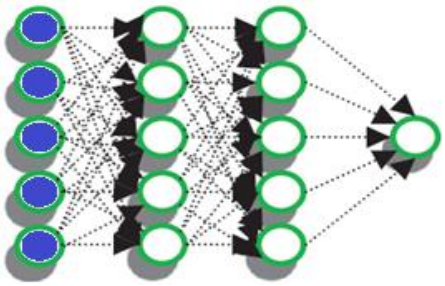


Figura 2. 2.Red totalmente conectada. Imagen obtenida y modificada de Rawat & Wang, 2017 (19).

Las redes neuronales que cuentan con gran cantidad de capas se denominan redes neuronales profundas, y pertenecen al campo de aprendizaje profundo (en inglés *Deep Learning*, DL). Estas redes permiten el uso de grandes conjuntos de datos, incluso los no estructurados. Además, son capaces de extraer, seleccionar y aprender características y patrones complejos de los datos (17). Las redes neuronales de convolución (en inglés *Convolution Neural Networks,* CNN) son un tipo de redes neuronales profundas. Particularmente, las CNN utilizan el uso compartido de parámetros y reducen la dimensionalidad (17).

Las CNN utilizan capas de convolución para aprender patrones o características específicas de los datos de entrada (Figura 2.4). Cada capa construye un mapa de características que se usa como entrada para la siguiente capa, a la cual se le aplican nuevas capas de convolución para crear un nuevo mapa de características y así sucesivamente. Después, el mapa de características final se ingresa en una red neuronal artificial totalmente

conectada, encargada de realizar la clasificación final de los datos. Las CNN cuentan con tres tipos de capas principales: capas de convolución, capas de reducción y capas totalmente conectadas. En la Figura 2.3 se muestra un esquema representativo de la CNN y sus tres tipos de capas.

La capa de convolución se conforma de un conjunto de filtros convolucionales llamados *kernels,* los cuales contienen a los pesos que se van ajustando durante el entrenamiento (Figura 2.4). Un filtro convolucional es una matriz pequeña de un tamaño específico, el cual recorre de izquierda-derecha y de arriba-abajo los datos de entrada (una matriz de dimensión fija) desplazándose una posición, y realizan una operación de convolución (una multiplicación de los elementos de las matrices) para extraer diferentes características (17, 18); cada vez que el filtro se mueve una posición se realiza una convolución. Esta operación finaliza hasta que el filtro recorre todos los datos de entrada. Finalmente, la salida de la convolución es un mapa (o matriz) de características resultante con los valores obtenidos. El proceso descrito anteriormente se repite de acuerdo con el número de filtros establecidos en cada una de las capas de convolución.

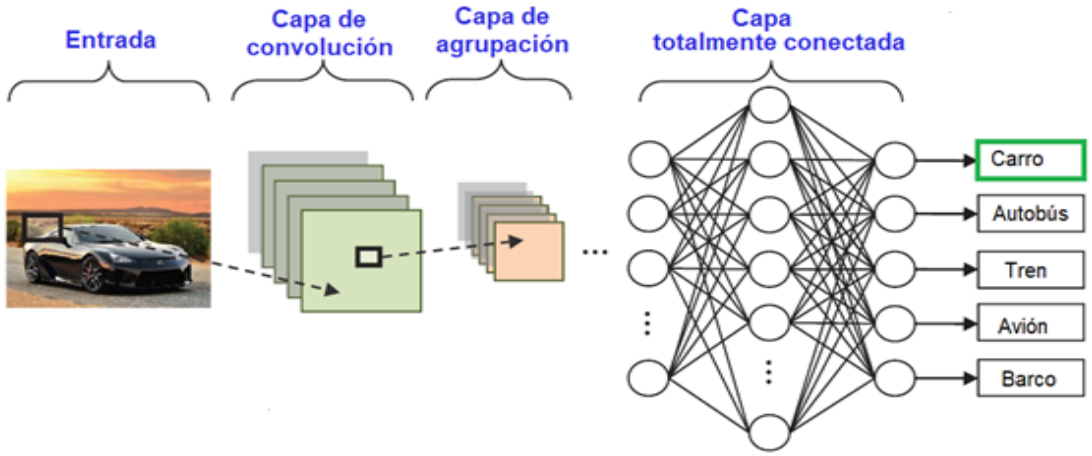


Figura 2. 3. Arquitectura básica de una CNN. Imagen tomada y modificada de Rawat & Wang, 2017 (19).La capa totalmente conectada viene siendo una red neuronal estándar (Figura 2.1).

En la Figura 2.4 se ilustra el funcionamiento de un filtro de convolución. El filtro realiza una convolución a los datos de entrada obteniendo el mapa de características resultante de la operación.

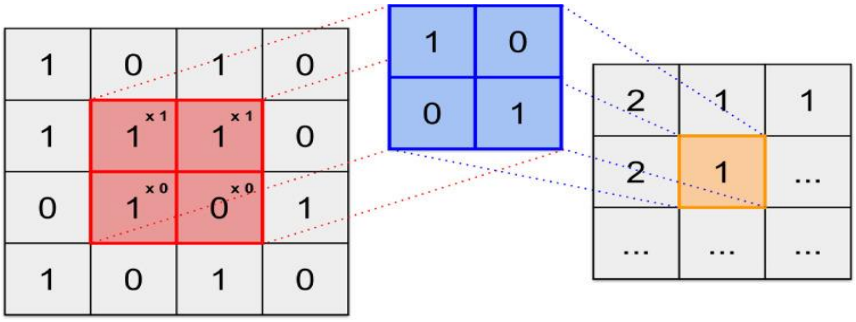


Figura 2.4. Esquema de una operación de convolución en una red CNN. El primer, segundo y tercer recuadro representan a los datos de entrada, el filtro y el resultado de la convolución, respectivamente. Imagen obtenida de Choi et al., 2020 (20).

Las capas de convolución cuentan con hiperparámetros que al modificarse cambian la dimensión de los resultados obtenidos. Por ejemplo, el *stride,* que es el número de posiciones que se mueve el filtro para recorrer los datos de entrada (18), puede cambiarse para reducir el mapa de características resultantes. También se puede aplicar el *padding,* que es el rellenado de ceros alrededor de los datos de entrada (18), el cual sirve para que el mapa de características resultante conserve el mismo tamaño que los datos de entrada y para no perder la información que se encuentra en los extremos y que puede ser relevante.

La capa de agrupación o reducción extrae la combinación de las características obtenidas por la capa de convolución en una región local (Figura 2.5). Además, regulariza la complejidad de la red al disminuir las dimensiones espaciales, reduciendo así los cálculos para las siguientes capas. Existen diferentes tipos de agrupación, las más comunes son: agrupación máxima (*max pooling*) y agrupación promedio (*average pooling*) (18). En la Figura 2.5, se muestra una capa de agrupación 2x2 dando como resultado cuatro regiones locales. La agrupación máxima (recuadro de la derecha) toma el número de mayor valor de cada región; mientras que la agrupación promedio (recuadro de la parte inferior), calcula el promedio de los números dentro de la región.

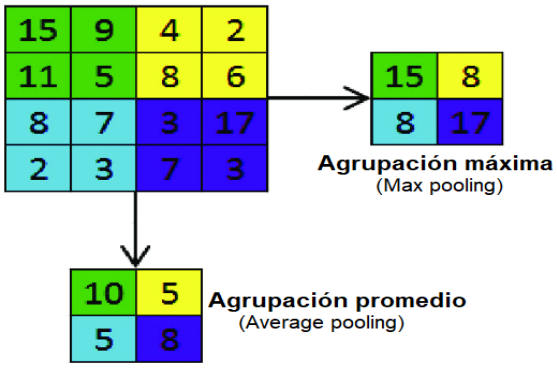


Figura 2.5. Esquema de una capa de agrupación de una red CNN de tamaño 2x2. Se muestra las dos operaciones más usuales: máxima o promedio. Figura obtenida y modificada de Rawat & Wang, 2017 (19).

Las características extraídas de la última capa de agrupación se unen para formar un vector unidimensional (Figura 2.3), el cual sirve como entrada a la última capa de CNN, la cual es una red totalmente conectada (*Fully connected*). La red totalmente conectada es una red neuronal estándar como la que se muestra en la Figura 2.2.

Para realizar el entrenamiento de una CNN, cada ejemplo del conjunto de datos de entrenamiento tiene una etiqueta que indica el objeto o clase a la que pertenece. Por ejemplo: si estamos clasificando imágenes de gatos y perros, la clase 1 sería gato y la clase 2 sería perro. En el caso de un modelo de clasificación multiclase, como los creados en este proyecto, en donde se cuenta con un conjunto de datos con varias clases, la capa de salida cuenta con múltiples clases o salidas, y para ello se aplica una función de activación *softmax.* En la función *softmax* cada nodo representa una clase (21). La función *softmax* (ec. 2.2) asigna una probabilidad a cada clase i.

$$\text{softmax } (z)_i = \frac{e^{(z_i)}}{\sum_{j=1}^{k} e^{(z_j)}}, \tag{2.2}$$

donde z_i contiene todos los valores de las entradas y k es el número de clases.

Otro punto importante en el entrenamiento de una CNN es el concepto de lote y época. El lote es un fragmento de los datos, cuyo tamaño es el número de ejemplos que la red entrena (hacia delante y hacia atrás). Una época es cuando la red ha entrenado todos los lotes del conjunto de datos.

Asimismo, existen técnicas que se aplican en las redes neuronales, como las CNN, para evitar la memorización o sobreajuste (en inglés *overfitting*) de los datos, tales como: la normalización por lotes (en inglés *Batch normalization)* y la deserción (en inglés *Dropout*). La normalización por lotes (*Batch normalization*) (ec. 2.5) sirve para estandarizar o normalizar los datos de cada lote; se puede realizar antes de una capa de convolución o de una capa totalmente conectada para ello, a cada dato del lote, se le resta la media calculada del lote (ec. 2.3) y la diferencia se divide entre la desviación estándar (ec. 2.4), obteniendo así su normalización (ec. 2.5).

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i , \tag{2.3}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i — \mu_B)^2 , \tag{2.4}$$

$$\overline{x_i} = \frac{x_i — \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \tag{2.5}$$

donde *m* es el tamaño de lote (*batch*) y x_i son los datos del lote.

Teniendo todas las características en una escala, ninguna tendrá un sesgo y, por lo tanto, la normalización de lotes (*Batch normalization*) actúa como un regularizador, que ayuda a superar el sobreajuste (*overfitting*) y mejora el aprendizaje. Además, al estandarizar los datos favorece a que las clases minoritarias no se pierdan cuando hay clases desbalanceadas.

La deserción (*Dropout*) es una otra técnica para regularizar las redes neuronales, que consiste en desactivar de manera aleatoria un porcentaje de los nodos de la red. Durante el entrenamiento, en cada lote se desactivan de manera aleatoria algunos nodos de la red, lo que lleva a que se cree una subred, que es una red más pequeña. En todo el entrenamiento se crean diversas subredes, de tal manera que cada una ven algunas características, por lo tanto, reduce el sobreajuste. Esto ofrece una regularización eficaz para reducir el sobreajuste (*overfitting)* y mejorar la capacidad de generalización en redes profundas. Por ejemplo, en la Figura 2.6-a se tiene una red totalmente conectada (los círculos en color gris representan los nodos). En la Figura 2.6-b, se muestra la misma red con los nodos desactivados (círculos de color gris claro) después de que se le aplicó la técnica de deserción. El resultado es una subred más sencilla.

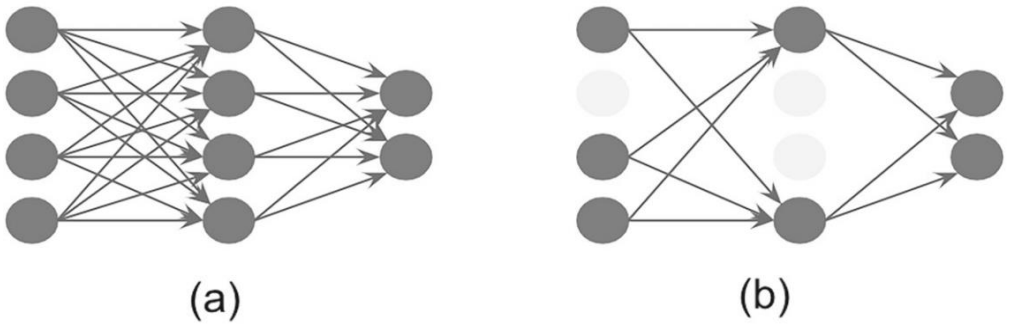


Figura 2. 6. Deserción. Esta figura obtenida de Garbin et al., 2020 (22).

Finalmente, para el proceso de entrenamiento, la CNN implementa una función de pérdida y una función de optimización. La función de pérdida sirve para evaluar el desempeño o error del modelo, por lo tanto, entre más se acerque a cero el valor de pérdida mejor será el modelo. La entropía cruzada categórica (*Categorical cross-entropy*) (ec. 2.6) es una función de pérdida que utiliza las distribuciones de las probabilidades para reducir la distancia entre el valor real y el predicho. La entropía cruzada categórica compara las probabilidades predichas por la función *softmax* con los valores de las etiquetas verdaderas, las cuales están en una codificación *one-hot.* La codificación *one-hot* crea una representación binaria (ceros y unos) única para cada clase, donde el uno indica la clase verdadera y los ceros las clases falsas. Por ejemplo: si clasificamos imágenes de gatos, perros y peces, la clase 1 sería gato y su codificación *one-hot* sería [1,0,0], mientras la clase 2 sería perro y su codificación *one-hot* sería [0,1,0] y finalmente, la clase 3 sería pez y su codificación *one-hot* sería [0,0,1].

Durante el entrenamiento, la entropía cruzada categórica pérdida calcula la divergencia de las probabilidades entre los valores verdaderos y el predicho y penaliza las probabilidades de las predicciones negativas, es decir, la probabilidades qué están lejos del valor verdadero y manteniendo las probabilidades de las predicciones positivas, por lo tanto, cuando más lejos este la probabilidad, mayor será el valor de la pérdida (18).

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)\,, \tag{2.6}$$

donde *N* es el número de ejemplos, *y* es el valor real o verdadero, y $\hat{y}$ es el valor predicho.

La función de optimización permite realizar el ajuste o modificaciones de los pesos y la tasa de aprendizaje durante el entrenamiento para reducir las pérdidas. Adam es un optimizador ampliamente utilizado en las CNN. Adam calcula las tasas de aprendizaje adaptativo para distintos parámetros utilizando el primer y segundo momento de los gradientes pasados. Para estimar el primer y segundo momento, Adam calcula los promedios móviles exponenciales de los gradientes (ec. 2.7) y de los gradientes cuadrados (ec. 2.8) (23). Luego, se corrigen los promedios móviles (ec. 2.9 y 2.10) y después se aplican para escalar la tasa de aprendizaje y actualizar los parámetros (ec. 2.11).

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)(\nabla\theta_{t-1})\,, \tag{2.7}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla\theta_{t-1})^2\,, \tag{2.8}$$

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}\,, \tag{2.9}$$

$$\hat{v}_t = \frac{m_t}{1-\beta_2^t}\,, \tag{2.10}$$

$$\theta_t = \theta_{t-1} - \alpha * \frac{\widehat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}\,, \tag{2.11}$$

donde, m_t son los promedios móviles exponenciales del gradiente (primer momento), v_t son los promedios móviles exponenciales del gradiente al cuadrado (segundo momento), θ son los parámetros (pesos), β_1 = 0.9, β_2 = 0.999, α = 0.001 y $\epsilon = 10^{-8}$. β_1 y β_2 son hiperparámetro que controlan las tasas de caída exponencial de los promedios móviles y α es la tasa de aprendizaje.

CAPÍTULO 3. ANTECEDENTES DE LAS HERRAMIENTAS DE CLASIFICACIÓN TAXONÓMICA VIRAL

En este capítulo se describe la metagenómica y las diferentes etapas que se realizan en los estudios metagenómicos. Posteriormente, se describe la clasificación general de las herramientas de asignación taxonómica, las herramientas de clasificación viral más relevantes que han sido reportadas en la literatura, y finalmente se presentan en forma breve aquellas que están basadas en redes neuronales artificiales profundas.

3.1. Metagenómica y el proceso para su estudio

En general, la metagenómica se define como el estudio del material genético (genomas) que se encarga de obtener secuencias cortas de los genomas de todos los organismos presentes en una muestra ambiental o de un huésped (5). Un genoma contiene “toda la información genética que un organismo posee” (24). La metagenómica ha cobrado gran relevancia dado que permite la identificación y el estudio de los genomas de diferentes organismos, simultáneamente, sin necesidad de cultivarlos o aislarlos experimentalmente y entender la forma en cómo se relacionan entre ellos y con sus ecosistemas.

El proceso general que se lleva a cabo en estudios de metagenómica se muestra en la Figura 3.1. El primer paso es la recolección de las muestras. El segundo es la extracción de todos los ácidos nucleicos de los genomas presentes en las muestras mediante un kit comercial para este fin. Los nucleótidos son pequeñas moléculas constituidas por un azúcar, un grupo fosfato y una de las cuatro bases nitrogenada: Adenina (A), Citosina (C), Guanina(G) y Timina (T) (25). El tercer paso es la preparación de las muestras, realizando una purificación del DNA y finalmente, se preparan las librerías para su secuenciación (1). Durante este último paso, el DNA es fragmentado y se les ligan los adaptadores. Posteriormente, los fragmentos de ADN son secuenciados utilizando secuenciadores de nueva generación (NGS). La secuenciación da como resultado millones de lecturas por cada muestra. Las lecturas son pequeños fragmentos de ADN que van de 75 a 1,000 nucleótidos (nt) o pares de bases (pb) de longitud, dependiendo del secuenciador y su tecnología. Estas lecturas reciben un análisis bioinformático dependiendo del estudio a realizar, por ejemplo, la clasificación taxonómica. El primer paso consiste en el preprocesamiento de las lecturas, es decir la eliminación de lecturas con baja calidad, redundantes y pertenecientes al huésped, que es el organismo de donde se extrajo la muestra. Posteriormente, se pueden crear *contigs*, que son secuencias

largas de ADN creadas de la unión de varias lecturas mediante un proceso llamado ensamble (26). Finalmente, las lecturas y/o *contigs* se alinean con las bases de datos (BD) de genomas de referencia para la clasificación taxonómica.

Específicamente, en la metagenómica viral existen diversas técnicas de extracción de DNA/RNA para obtener los genomas virales dentro de una muestra (27, 28). En el laboratorio donde se realizó esta investigación, el proceso aplicado en la preparación de las muestras permite obtener todos los virus de RNA y DNA en una misma secuenciación. El proceso consiste, en centrifugar las muestras para obtener el sobrenadante (parte viral). Después, se pasan por un filtro de poro de 0.45 µm para eliminar células del huésped y bacteriales. Posteriormente, se aplica un tratamiento con DNAse y RNAse para degradar el material genético de lo no viral. El DNA/RNA viral de la muestra, se obtiene mediante un kit de extracción de DNA/RNA viral. Posteriormente, se aplica un proceso de transcriptasa reversa para obtener cADN del RNA y la segunda cadena se genera mediante dos rondas de síntesis con Sequenase 2.0; el resultado de este proceso es lo que se utiliza en la preparación de las librerías.

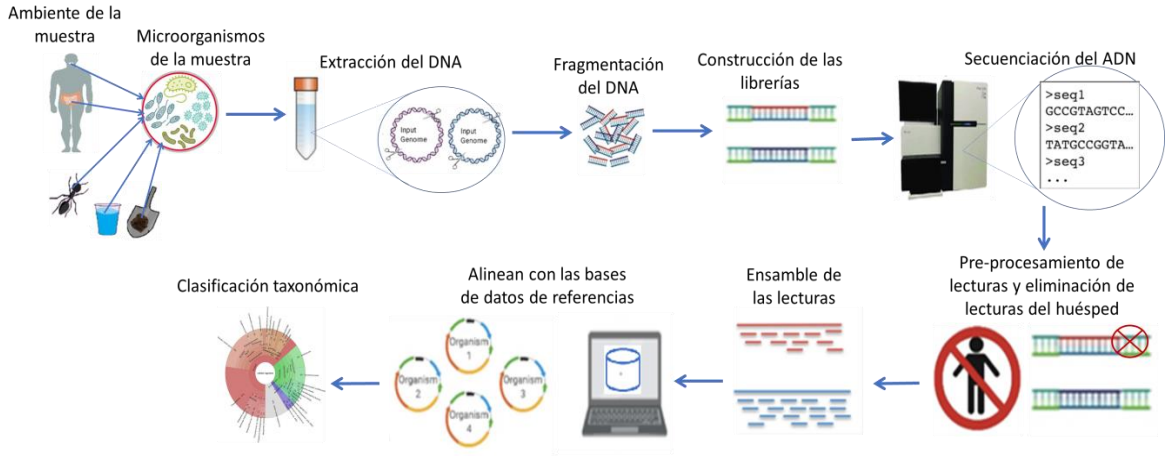


Figura 3.1. Proceso de los estudios metagenómicos.

3.2. Tipos de herramientas de clasificación taxonómica viral

Las herramientas actuales de clasificación de datos metagenómicos virales pueden dividirse en las siguientes tres categorías, basadas en similitud, composición e híbridas, a partir del proceso que utiliza para la discriminación taxonómica.

3.2.1. Basadas en similitud

Las herramientas basadas en similitud emplean algoritmos de alineamiento o mapeo, tales como, Bowtie2 (29), BLAST(siglas en inglés de *Basic Local Alignment Search Tool*) (30), Megablast (31) o programas derivados de BLAST, para comparar las lecturas contra una base de datos (BD) de genomas de referencia. Estas herramientas cuentan con una precisión alta; sin embargo, las que emplean alineadores como BLAST y sus derivados son computacionalmente costosa en tiempo de procesamiento (32), aunque las que usan Bowtie2 presenta una mejora en el tiempo.

3.2.2. Basadas en composición

Estas herramientas emplean las características de composición de las lecturas, generalmente denominados *k-mers* (subsecuencia de caracteres sucesivos de longitud *k* dentro de una secuencia (33, 34)) y/o las frecuencias de oligonucleótidos (secuencia corta de ADN o ARN con 50 pb o menos), junto con alguna técnica de agrupamiento o aprendizaje automático para formar grupos de lecturas y después realizar la clasificación taxonómica. Estas herramientas son rápidas, pero un gran porcentaje de lecturas no se pueden clasificar porque no coinciden con las secuencias de referencia que existen en la base de datos (BD) o coinciden con puntuaciones extremadamente bajas. Además, la mayoría no asignan identidad taxonómica y aquellas que si lo hacen no pueden llegar a niveles taxonómicos bajos, como los niveles de género y especie.

3.2.3. Híbridas

Las herramientas híbridas generalmente utilizan las características de composición, *k-mers* y/o las frecuencias de oligonucleótidos como las mencionadas arriba, agrupan las lecturas mediante una técnica de agrupamiento y posteriormente, los centroides de los grupos se comparan con una base de datos (BD) de referencia para realizar la clasificación taxonómica. Otras herramientas comparan las lecturas contra la BD de referencia y cuando las lecturas son clasificadas en diferentes especies, se aplica una técnica de aprendizaje automático para asignarlas a solo una de ellas. Estas herramientas son más rápidas que las basadas en similitud, pero las lecturas que no coinciden o coinciden con puntuaciones extremadamente bajas con las secuencias de referencia que existen en la BD no las pueden clasificar (32).

3.3. Herramientas de clasificación taxonómica viral

Se realizó una revisión de los trabajos reportados en la literatura que están relacionados con herramientas de clasificación taxonómica en metagenómica viral o que incluyen la clasificación viral. Se identificaron 33 herramientas que realizan la tarea de clasificación viral; a continuación se enlista el nombre de la herramienta o en su caso, el nombre de los autores del trabajo: AKE (35), Bileschi et al., 2019 (36), BLAST (30), CASTOR-KRFE (37), Cenote-Taker (38), Centrifuge (39), CHEER (40), ClassiPhages 2.0 (41), DeepVirFinder (42), DIAMOND (43), drVM (44), Fabijanska & Grabowski, 2019 (45), FastViromeExplorer (46), iVirus (47), Kraken2 (48), MaxBin 2.0 (7), MePIC (49), MetaPalette (50), MetaPORE (51), Metavir 2 (52), One Codex (53), RIEMS (54), SLIMM (55), Surpi (56), Taxonomer (57), VIGA (58), Vipie (59), ViraMiner (60), VirNet (61), VIROME (62), VirusFinder (63), VirusSeeker (64) y Zou et al., 2019 (65). De estas herramientas se seleccionaron once para ser evaluadas y comparadas en forma objetiva; dado que son unas de las más utilizadas y/o citadas.

Las herramientas de clasificación seleccionadas son aplicaciones de escritorio como BLAST, Centrifuge, DIAMOND, drVM, FastViromeExplorer, Kraken2, SLIMM y VirusFinder o, aplicaciones Web como One Codex, Taxonomer y Vipie (ver Tabla 3.1). Cada herramienta se evaluó utilizando la base de datos (BD) de referencia proporcionada por cada una de ellas, excepto DIAMOND. Esta herramienta no trae una BD propia, sino que se utiliza una propia del usuario. Por lo tanto, se evaluó con dos BD de referencia diferentes; una que es nt (15), la cual contiene múltiples secuencias por cada especie viral anotada (definida como DIAMOND1) y otra que es RefSeq (14), la cual incluye una sola secuencia por cada especie viral anotada (definida como DIAMOND2). Se utilizaron dos BD con el objetivo de analizar cómo las BD de referencia usadas pueden impactarlos resultados obtenidos. Cabe señalar, que esta parte es importante, ya que en este trabajo también se usaron ambas BD para entrenar a la CNN de manera independiente (Objetivo II, definido en la sección 1.3). Por lo tanto, los resultados de DIAMOND se compararon con los resultados obtenidos por los dos modelos generados.

Estas herramientas se pueden dividir por el método de clasificación que utilizaron, ya sea composición o similitud. Las que se basan en similitud, BLAST, DIAMOND, drVM, Vipie y VirusFinder, comparan las lecturas con una BD de referencia, mientras que las herramientas basadas en composición, Centrifuge, FastViromeExplorer, Kraken2, One Codex y Taxonomer, usan una frecuencia *k-mers* (sub-cadenas de *k* longitud) de las lecturas. Finalmente, SLIMM se basa en una metodología híbrida. En la Tabla 3.1 se presenta un análisis detallado de las once herramientas elegidas, en donde se presenta una descripción general de las principales características de las herramientas, incluido el año de desarrollo, el número de citas, el tipo de algoritmo (basado en similitudes o basado en *k-mers*), tipo de datos

de entrada, salidas y proceso llevado a cabo en la clasificación taxonómica, así como las BD de referencia utilizadas por cada uno. En general, los datos de entrada pueden ser lecturas *single-end* (secuencia única obtenida al tener el extremo de un solo fragmento de ADN), o lecturas *paired-end* (par de secuencias obtenidas de dos extremos de un fragmento de ADN), ambos tipos de lecturas, y/o *contigs* (ensamblaje de dos o más lecturas una secuencia más grande).

En lo que se refiere al proceso de clasificación taxonómica, lo hemos generalizado en tres etapas principales. i) Preprocesamiento. Depuración de lectura mediante la eliminación de lecturas de baja calidad y lecturas con una estrecha homología con el huésped, por ejemplo, genes humanos o de bacterias ribosómicas. ii) Ensamble. Se obtienen secuencias contiguas más largas (*contigs*) a partir de lecturas individuales relacionadas unidas. iii) Búsqueda. Las lecturas y/o *contigs* se comparan con un DB de referencia de nucleótidos o proteínas (viral o completo) y luego se clasifican en un nivel taxonómico específico. Finalmente, los resultados de las herramientas pueden ser la clasificación de cada lectura a un nivel taxonómico específico o tablas de perfiles de abundancia.

Tabla 3. 1. Características principales de las herramientas de clasificación taxonómica evaluadas en este trabajo. El número de citas se consultó el 20 de junio de 2022.

Herramientas	Año	# de Citas[a]	Tipo de entrada	Salida	Nivel taxonómico de salida	Pre-procesamiento	Ensamble	Método de Búsqueda	Estrategia de búsqueda	Bases de datos usadas
Herramientas de escritorio (basado en Linux)										
BLAST (30)	2009	11,984	Lecturas *single-end* y *paired-end* (en formato fasta)	Tablas de lecturas clasificadas	Especies	No	No	Smith - Waterman	Similitud	RefSeq viral
Centrifuge (39)	2016	783	Lecturas *paired-end* (en formato fastq) o *contigs*	Tabla de lecturas clasificadas con todos los rangos taxonómicos juntos	Especie, género y familia	No	No	No	Composición de *k-mers*	nt de virus, bacteria y arqueas
DIAMOND (43)	2021	5573	Lecturas *single-end* y *paired-end* (en formato fasta)	Tablas de lecturas clasificadas	Especies	No	No	BLASTX	Similitud	1) nt viral o 2) RefSeq viral
drVM (44)	2017	29	Lecturas *single-end* y *paired-end* (en formato fasta/fastq)	Tabla de abundancias con todos los rangos taxonómicos juntos	Especie, género y familia	No	SPAdes	BLASTn y SNAP	Similitud	Viral propia proporcionada por los autores
F.V.E. (46)	2018	53	Lecturas *single-end* y *paired-end* (en formato fasta/fastq)	Tabla de abundancias con todos los rangos taxonómicos juntos	Especie, género y familia	No	No	Pseudo alineamiento Kallisto[b]	Composición de *k-mers*	RefSeq viral
Kraken2 (66)	2019	1455	Lecturas *single-end* y *paired-end* (en formato fasta/fastq)	Tablas de lecturas clasificadas	Especies	Lecturas de baja complejidad es opcional	No	No	Composición de *k-mers*	Maxikraken2[g] proporcionado por los autores
SLIMM (55)	2017	36	Lecturas *paired-end* o un archivo de alineamiento (en formato BAM/SAM)	Tabla de abundancias en un rango taxonómico específico	Todos desde cepa hasta phylum	No	No	Bowtie2	Híbrida	nt viral
VirusFinder (63)	2013	153	Lecturas *single-end* y *paired-end* (en formato fasta/fastq)	Tabla de abundancias con todos los rangos	Especie, género y familia	Filtra lecturas humanas	Trinity	BLASTn	Similitud	Propia proporcionada por los autores[c]

Herramientas	Año	# de Citas[a]	Tipo de entrada	Salida	Nivel taxonómico de salida	Pre-procesamiento	Ensamble	Método de Búsqueda	Estrategia de búsqueda	Bases de datos usadas
			o un archivo de alineamiento (en formato BAM)	taxonómicos juntos						
Herramientas Web										
One Codex (53)	2015	121	Lecturas *single-end* y *paired-end* (en formato fasta/fastq)	Tabla de abundancias en un rango taxonómico específico	Especie, género y familia	No	No	No	Composición de *k-mers*	Propia proporcionada por los autores [d]
Taxonomer (57)	2016	151	Lecturas *paired-end* (en formato fastq)	Tablas de lecturas clasificadas	Especies	No	No	No	Composición de *k-mers*	Propia proporcionada por los autores [e]
Vipie (59)	2017	24	Lecturas *paired-end* (en formato fastq)	Tabla de abundancias con todos los rangos taxonómicos juntos	Especie, género, familia y orden	Galaxy para extremos y filtrar las lecturas humanas y ribosomales	Opciones: Velvet, MetaVelvet, IDBA-UD, MEGAHIT o ABySS	BLASTn	Similitud	Propia proporcionada por los autores [f]

La herramienta FastViromeExplorer es anotada como F.V.E.

[a] *Citas consultadas en Google Scholar el 20 de julio del 2022.*

[b] *En FastViromeExplorer, Kallisto busca coincidencias exactas para k-mer cortos (el tamaño predeterminado es de 31 pb) entre las lecturas metagenómicas y las secuencias de la base de datos de referencia.*

[c] *La base de datos de VirusFinder contiene secuencias virales de 32,102 clases conocidas y una base de datos humana.*

[d] *La base de datos One Codex es una combinación de genomas públicos y privados que incluyen 61,988 genomas bacterianos, 48,399 virales, 1,822 de hongos, 1,988 de arqueas y 203 de protozoarios (114,401 incluyen el huésped), y la mayoría de los genomas provienen de NCBI.*

[e] *La base de datos de Taxonomer fue construida a partir de cuatro fuentes: Greengenes para la clasificación bacteriana, UNITE para la clasificación de hongos, UniRef es la clasificación y descubrimiento viral y ENSMBL para las secuencias de referencia humana.*

[f] *La base de datos de Vipie fue construida por tres fuentes: RefSeq viral, secuencias de virus etiquetadas como "complete" en Genbank y el repositorio de fagos del European Bioinformatics Institute (EBI).*

[g] *La base de datos de Kraken2 fue construida por genomas que no son completas o representativas ("complete" o "representative"), de las bases de datos de: arqueas, bacterias, hongos, protozoarios, viral y humanos.*

3.4. Herramientas de clasificación viral que usan redes neuronales artificiales profundas

Recientemente, en estudios metagenómicos virales, se ha incrementado el uso de herramientas basadas en composición. Principalmente, porque utilizan una gran cantidad de información, tanto de secuencias a nivel nucleótido como de proteínas, para entrenar modelos de redes neuronales. A la fecha, se han reportado diversos modelos, los cuales son capaces de discriminar lecturas y *contigs* entre viral y no viral, identificar motivos y nuevos virus, así como para clasificar los virus conocidos a nivel familia, género o especie.

Entre los trabajos se encuentra la herramienta ClassiPhages 2.0 (41), que clasifica los genomas de fagos en 12 familias mediante una red neuronal artificial totalmente conectada. La red usa 5,920 perfiles obtenidos de 7,342 genomas de fagos y generados mediante modelos ocultos de Markov. Dichos perfiles fueron empleados para generar una matriz de características que fue utilizada para entrenar la red. En este modelo se empleó una función *softmax* en la capa de salida para clasificar las familias de fagos. También se aplicó una validación cruzada de 100-*folds* en el entrenamiento y se probó con un conjunto independiente. Esta herramienta reportó una exactitud del 84% y las 12 familias predichas obtuvieron una sensibilidad de hasta de 79%.

Otra herramienta es *Viral Genome Deep Classifier* (45), la cual contiene 8 modelos ensamblados de CNN para la clasificación de los subtipos de virus de dengue, hepatitis B, hepatitis C, VIH-1 e influenza A. Estos cuentan con 4, 21, 9, 49 y 169 subtipos, respectivamente. Debido a que los genomas tienen diferentes longitudes, los más pequeños, se rellenaron con ceros a la derecha hasta tener la misma longitud que el genoma más largo. Los modelos ensamblados de la CNN se pueden ver Figura 3.2. Esta arquitectura obtuvo exactitudes, precisiones y sensibilidades mayores de 84%, su rendimiento varía de acuerdo con el tipo de virus y el número de subtipos.

Finalmente, una de las mejores herramientas reportadas hasta el momento es CHEER (40), la cual tiene una arquitectura de árbol (ver Figura 3.3), que ensambla varios modelos de redes de convolución (CNN) (Capítulo 2) para clasificar taxonómicamente las lecturas a 6 órdenes, 23 familias y 55 géneros. El primer modelo es un clasificador binario que se encarga de discriminar lecturas virales (de ARN) y no virales. El siguiente modelo recibe las lecturas virales para su posterior clasificación a nivel orden. Después, cada orden específico tiene un modelo que discrimina las lecturas para a sus respectivas familias. Finalmente, para cada familia existe un modelo para la clasificación a nivel género. Además, esta herramienta implementó una función de parada anticipada, la cual utiliza un umbral para decidir si se

detiene la clasificación de lecturas en niveles taxonómicos altos (como orden o familia) o si continúa a los clasificadores más específicos (como género). Esta herramienta reporta la exactitud y sensibilidades de los tres niveles taxonómicos (orden, familia y género); a nivel orden la exactitud va del 83% al 96%, a nivel familia va del 77% al 86% y a nivel de género se encuentra entre el 66% al 79%.

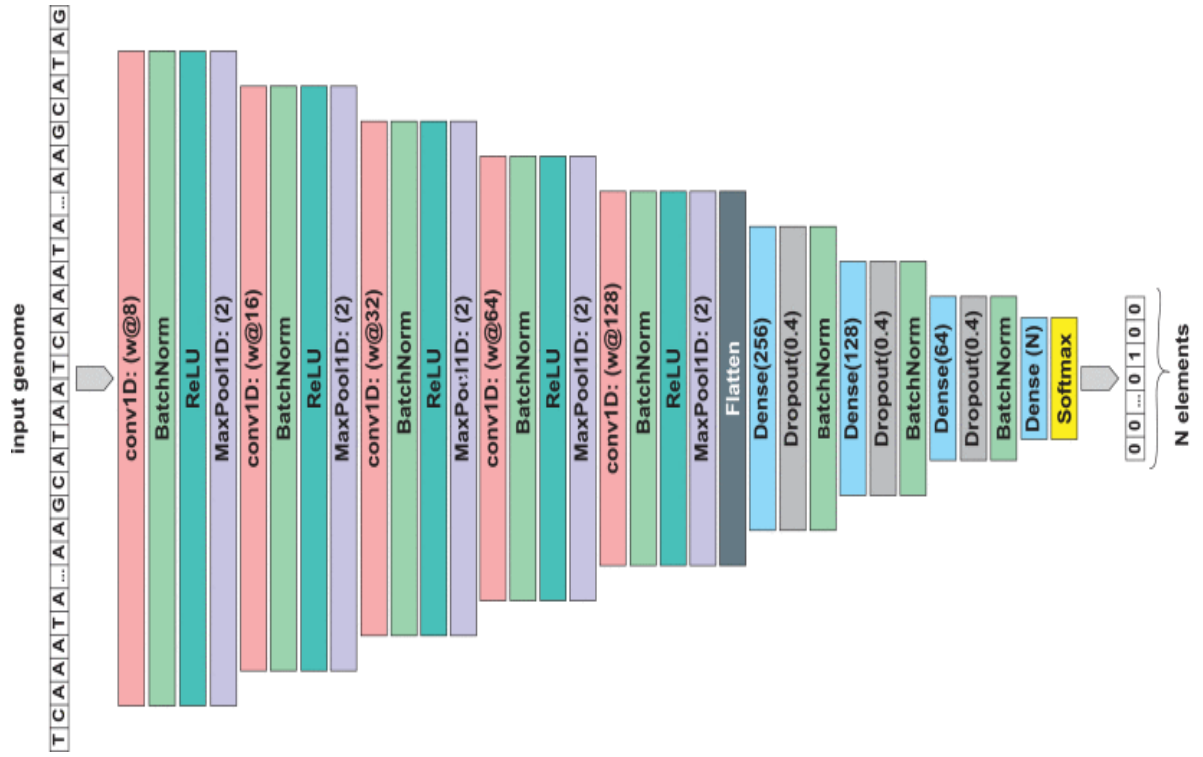


Figura 3.2. Arquitectura de Viral Genome Deep Classifier. Imagen tomada de Fabijanska & Grabowski, 2019 (45). La CNN tienen 5 capas de convolución con 8, 16, 32, 64 y 128 filtros, respectivamente, y un tamaño de filtro 7. Después de cada capa de convolución, se integra una de normalización (Batch Normalization) y después una de agrupamiento máximo (max pooling). Después, tiene tres capas totalmente conectadas con 256, 128 y 64 nodos, seguidas de una capa de deserción y una de normalización. Finalmente, el modelo tiene una capa de salida softmax con un número de neuronas que depende del número de subtipos a clasificar.

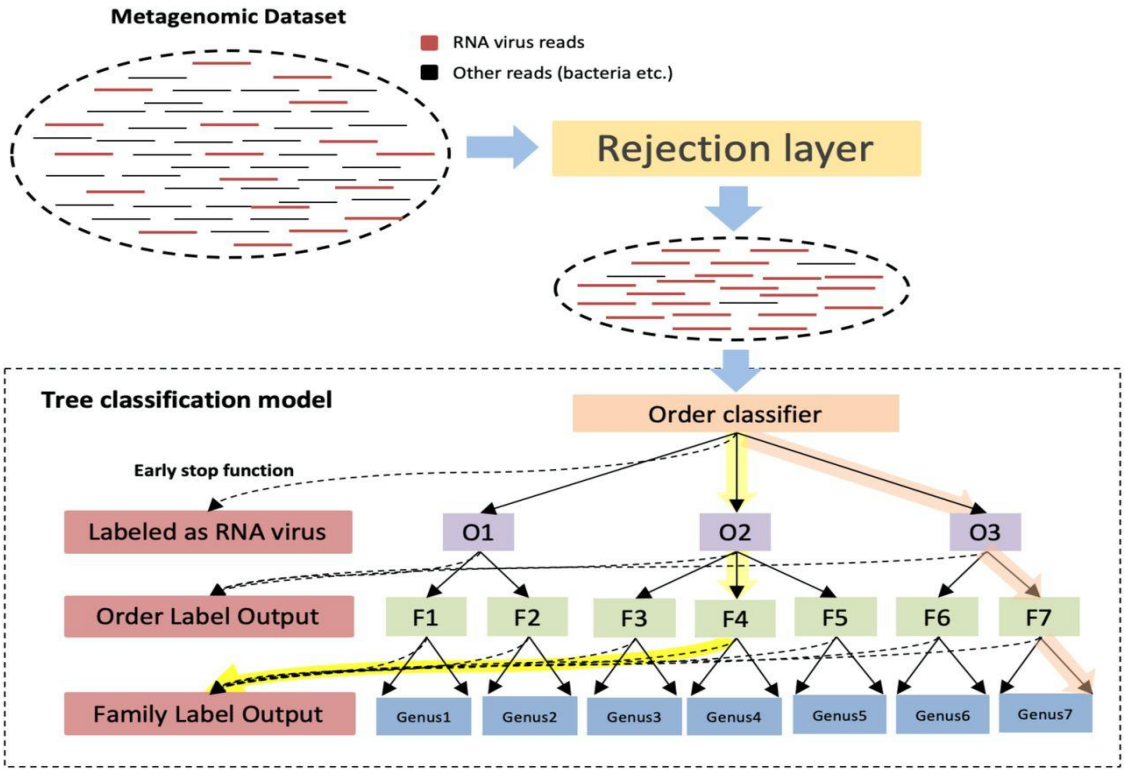


Figura 3.3. Estructura de la herramienta de clasificación CHEER. Cada CNN tiene cuatro capas de convolución con 256 filtros de tamaños de 3, 7, 11 y 15, respectivamente. Después de cada capa de convolución, se tiene una de agrupación máxima (max pooling). Posteriormente, se tienen dos capas totalmente conectadas, con 1024 y 512 nodos cada una; y una última capa que realiza la asignación de las clases finales, mediante la función de activación softmax. La imagen fue tomada de Shang & Sun, 2020 (40).

CAPÍTULO 4. EVALUACIÓN DE LAS HERRAMIENTAS DE CLASIFICACIÓN TAXONÓMICA MÁS UTILIZADAS

En este capítulo se presenta una evaluación objetiva, dado que se utilizaron los mismos conjuntos de datos y las mismas métricas, de once herramientas de clasificación taxonómica. También se describen los doce conjuntos de datos de metagenómica que se usaron para evaluar las herramientas, de los cuales ocho son simulados y cuatro son reales. Cabe señalar, que la evaluación de las herramientas existentes dio pie al primer artículo generado en este proyecto de investigación.

4.1. Conjuntos de datos evaluados

4.1.1. Simulados virales

Los conjuntos simulados virales se dividieron en dos bloques, el primero simula lecturas virales largas de la tecnología 454, el cual ya había sido reportado por Tangherlini et al., 2016 (67) y el segundo simula lecturas virales cortas de la tecnología Illumina creados para este trabajo.

4.1.1.1. Conjuntos simulados de lecturas largas

Tangherlini et al., 2016 (67) simuló tres conjuntos (*50G*, *500G* y *1000G*) de lecturas de la tecnología 454, usando 50, 500 y 1000 especies de genomas virales de referencia seleccionados al azar de la base de datos (BD) de RefSeq (14). Cada uno de los conjuntos contiene 100,000 lecturas virales únicas de 300 a 500 pb de longitud, con una tasa de error del 1.07% (68). El error es el generado por un secuenciador cuando realiza una sustitución, una inserción o una eliminación de un nucleótido (cambio de una base por otra).

Los conjuntos *50G*, *500G* y *1000G* cuentan con 30, 85 y 102 familias virales, y los dos últimos conjuntos (*500G* y *1000G*) cuentan con un grupo de genomas sin asignación taxonómica, es decir genomas que no pertenecen a ninguna familia viral. Estos conjuntos se utilizaron para evaluar el desempeño de las herramientas al clasificar lecturas grandes, como *contigs* o lecturas de Illumina de 2x251 pb o 2x301 pb y para evaluar las herramientas con diferentes niveles de complejidad viral, es decir, aumentaron el número de genomas utilizados en cada conjunto simulado.

4.1.1.2. Conjuntos simulados de lecturas cortas

Así también, para simular lecturas cortas correspondientes a la tecnología Illumina, se crearon tres conjuntos (*Eukaryotic*, *Prokaryotic* y *Unclassified*) con el simulador Grinder (69). Para ello, se seleccionó al azar un genoma viral por cada especie de todas las depositadas en la BD de GenBank (a la fecha del 30 de abril de 2020). Estos conjuntos fueron generados con cinco millones de lecturas pareadas de 150 pb de longitud, con la tasa de error de la tecnología de 0.1% (68).

El conjunto *Eukaryotic* tiene lecturas de 4,652 especies que pertenecen a 134 familias que infectan organismos eucariontes; mientras que el conjunto *Prokaryotic* cuenta con lecturas de 5,817 especies que pertenecen a 27 familias virales que infectan a bacterias y arqueas. Finalmente, el conjunto *Unclassified* contiene 558 especies diferentes sin asignación taxonómica a nivel familia. Los conjuntos *Eukaryotic* y *Prokaryotic* tienen la finalidad de evaluar el desempeño de las herramientas en todas las familias virales anotas e identificar las familias más fáciles y difíciles de clasificar, mientras que el conjunto *Unclassified* se creó para probar si las herramientas tienen la capacidad de identificar los genomas virales que no tiene una asignación taxonómica a nivel familia, debido a que son virus nuevos o menos caracterizados.

4.1.2. Simulados no virales

Se simularon dos conjuntos de datos de lecturas no virales, *Bacterial* y *Human*, mediante el programa Grinder (69). Cada conjunto cuenta con 100,000 lecturas *pair-end* de 150pb de longitud. El propósito de los dos conjuntos de datos fue evaluar el desempeño de las herramientas cuando clasifica lecturas provenientes de genomas no virales.

El conjunto de datos *Human* fue simulado utilizando un genoma humano de referencia (GCA_000001405.28) y para el conjunto de datos *Bacterial* se utilizaron cuatro genomas pertenecientes a cuatro especies: *Bacillus subtilis* (CP098491.1) de la familia *Bacillaceae*, *Streptomyces lydicamycinicuss (*CP098437.1*) de la familia Streptomycetaceae*, *Salmonella entérica* (CP003278.1) de la familia *Enterobacteriaceae*, y *Prochlorococcus marinus* (AE017126.1) de la familia *Prochlorococcaceae.*

Para la simulación del conjunto de datos *Bacterial,* se seleccionó aleatoriamente genomas de bacterias de diferentes filos (Firmicutes, Actinomycetota, Pseudomonadota y Cyanobacteriota) y que a parte dicha bacterias fueran aisladas de diversos ambientes; (terrestre (CP098491.1), humano (CP003278.1), y marino (CP098437.1 y AE017126.1). Interesantemente, la bacteria *Salmonella entérica* es un patógeno común para los humanos y *Streptomyces lydicamycinicuss* produce un antibiótico.

4.1.3. Conjuntos de datos reales

Los cuatro conjuntos reales usados se obtuvieron de distintos estudios previamente reportados, que fueron secuenciados con la tecnología Illumina. Dos conjuntos, *FISH-I* (SRA id SRX3861422) y *PB3* (SRA id SRS3103717), provienen de dos muestras recolectadas en Cuatro Ciénegas en el 2014 (70). El conjunto *FISH-I* fue obtenido de una muestra intestinal de un pez de la especie *Hemichromis guttatus,* ubicado en la poza de Churince y el conjunto *PB3* fue conseguida de una muestra de agua proveniente de una poza de Pozas Rojas. Estos dos conjuntos contienen 7,531,002 y 5,203,288 lecturas únicas de 150 pb de longitud, respectivamente.

Además, se utilizaron los conjuntos *I5-8* e *I21-1* (SRA id SRS5809795), que pertenecen a dos muestras extraídas de niños de la comunidad semi-rural de Xoxocotla, en el estado de Morelos, México (71). El conjunto *I5-8* se obtuvo a partir de una muestra de heces de un infante de cinco meses y el *I21-1* se originó de una muestra respiratoria (orofaringe) de un infante de 21 meses. Estos conjuntos cuentan con 14,296,952 y 12,126,776 lecturas pareadas únicas de 75 pb de longitud de buena calidad, respectivamente. Las especificaciones sobre la preparación, aislamiento y secuenciación de los ácidos nucleicos de las muestras se pueden consultar en las referencias originales (70, 71).

4.2. Herramientas de clasificación taxonómica evaluadas

Para este análisis comparativo, se seleccionaron once herramientas de clasificación metagenómica, dado que son las más usadas y/o recientes: BLAST, Centrifuge, DIAMOND1 (utilizó la base de datos (BD) de virus de RefSeq), DIAMOND2 (empleó la DB viral nt completa), drVM, FastViromeExplorer, Kraken2, One Codex, SLIMM, Taxonomer, Vipie y VirusFinder; fueron evaluadas usando los conjuntos de datos antes descritos y usando las mismas métricas de evaluación empleadas con frecuencia en la literatura.

4.3. Métricas de evaluación

Las métricas de evaluación son medidas cuantitativas que permite evaluar el desempeño de las herramientas. Existen numerosas métricas de evaluación, pero en este proyecto de investigación se seleccionaron las siguientes cuatro métricas: sensibilidad o *recall* (ec. 4.1) precisión (ec. 4.2), precisión equilibrada (ec. 4.3), puntuación F1 (ec. 4.4), MCC (ec. 4.5) y especificidad (ec. 4.6).

(i) **Sensibilidad (*Recall*):** Es el porcentaje de valores verdaderos positivos que se identifican correctamente del número total de valores, su fórmula matemática es:

$$Recall = \left(\frac{TP}{(TP+FN)}\right) * 100\,, \tag{4.1}$$

donde TP es el número de verdaderos positivos y FN es el número de falsos negativos.

(ii) **Precisión:** Es el porcentaje de valores verdaderos positivos que se identificaron correctamente de los valores clasificados, su fórmula es:

$$Precisión = \left(\frac{TP}{(TP+FP)}\right) * 100\,, \tag{4.2}$$

donde TP es el número de verdaderos positivos y FP es el número de falsos positivos.

(iii) Precisión equilibrada (*Balance Accuracy*): Es la media aritmética de la sensibilidad y la especificidad. La precisión equilibrada (ec. 4.3) y se define como el promedio de recuperación que obtiene cada clase. Esta medida se utiliza para clases desbalanceadas, cuando alguna de las clases aparece con mucha más frecuencia que otras.

$$Precisión\ equilibrada = \frac{Sensibilidad + Especificidad}{2}\,, \tag{4.3}$$

(iv) Puntuación F1 (*F1 score*): Es la media armónica de precisión y sensibilidad. La puntuación F1 (ec. 4.4) sirve para combinar en un solo valor la precisión y sensibilidades.

$$Puntiación\ F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad}\,, \tag{4.4}$$

(v) **Coeficiente de correlación Matthews (MCC):** Es una medida equilibrada entre clases de diferentes tamaños, que sirve para medir la calidad de las clasificaciones. Básicamente, el MCC (ec. 4.5) es el coeficiente de correlación entre las clasificaciones reales y predichas.

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\,, \tag{4.5}$$

(vi) Especificidad: Es el porcentaje de valores negativos reales que se identifican correctamente como negativos, su fórmula es:

$$Especificidad = \left(\frac{TN}{TN + FP}\right) * 100\,, \quad (4.6)$$

donde TN es el número de verdaderos negativos y FP es el número de falsos positivos.

Finalmente, para evaluar los conjuntos de datos reales que se utilizaron en este proyecto, se empleó el porcentaje de lecturas asignadas como virales (ec. 4.7), debido a que no se conoce exactamente el número de lecturas virales correctas por cada familia.

$$Porcentaje\ de\ lecturas = \frac{Número\ total\ de\ lecturas\ clasificadas\ como\ virales}{Número\ total\ de\ lecturas\ en\ el\ conjunto\ de\ datos} * 100\,. \quad (4.7)$$

4.4. Resultados de la evaluación

4.4.1. Evaluación de los conjuntos simulados virales

4.4.1.1. Lecturas largas

En la Figura 4.1 se representan las sensibilidades y precisiones que fueron obtenidas de la evaluación de los tres conjuntos de lecturas largas (*50G*, *500G* y *1000G*). Con respecto a la sensibilidad, en los tres conjuntos, las herramientas BLAST, FastViromeExplorer, Kraken2 y VirusFinder mostraron las más altas sensibilidades (ec. 4.1), a nivel de familia (81%-100%) y especie (59%-99%). Las herramientas Centrifuge, DIAMOND1, DIAMOND2, SLIMM, Taxonomer y Vipie obtuvieron sensibilidades de moderadas a bajas, en ambos niveles taxonómicos (19%-71% para familia y 14%-45% para especie), en los tres conjuntos de datos. Finalmente, drVM y One Codex obtuvieron la sensibilidad más baja (3% -29%) en los tres conjuntos en ambos niveles explorados.

Como se muestra en la Figura 4.1, casi todas las herramientas obtuvieron una menor sensibilidad en el nivel de especie en comparación con el nivel de familia, a excepción de la herramienta Vipie en el conjunto *50G*. Esto puede deberse a la falta de datos taxonómicos actualizados en su base de datos (BD) de referencia, ya que algunas especies virales que tenían una familia fueron anotadas como no clasificadas. Por ejemplo, las lecturas del virus *Adoxophyes orana nucleopolyhedrovirus*, que representan casi el 6% en abundancia en este conjunto, se clasificaron como familia no clasificada (*Unclassified*) en lugar de *Baculoviridae*. Curiosamente, la sensibilidad de Centrifuge, DIAMOND1, DIAMOND2, drVM, One Codex y SLIMM fue mayor en los conjuntos 500G y 1000G que en el de 50G en contraste con el resto de las herramientas que mostró un rendimiento más bajo en muestras de mayor complejidad, es decir con más especies.

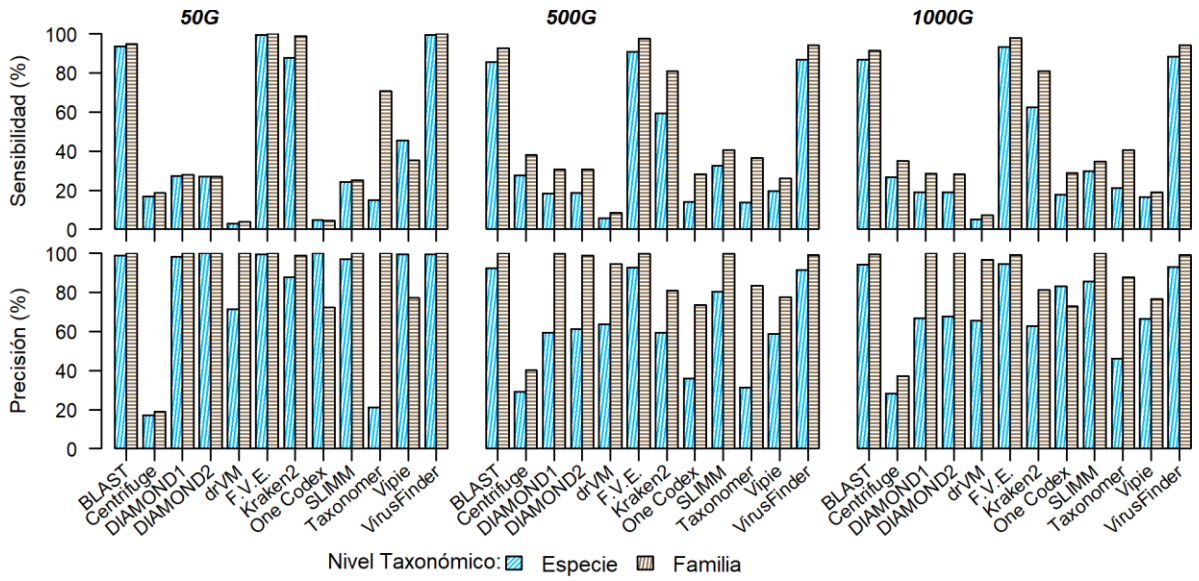


Figura 4.1. Sensibilidades y precisiones de las herramientas obtenidas, en los niveles taxonómicos de especies y familias, en los conjuntos de datos simulados 454. Cada columna representa uno de los 3 conjuntos. La herramienta FastViromeExplorer está anotada como F.E.V., DIAMOND1 es la herramienta DIAMOND evaluada con todos los genomas virales de la base de datos nt y DIAMOND2 con genomas virales de RefSeq.

En cuanto a la precisión (ec. 4.2, Figura 4.1), BLAST, DIAMOND1, DIAMOND2, drVM, FastViromeExplorer, SLIMM y VirusFinder obtuvieron las puntuaciones más altas (95%-100%) en todos los conjuntos de datos a nivel de familia, mientras que, a nivel de especie, solo BLAST, FastViromeExplorer y VirusFinder lograron precisiones mayores al 91%, mientras que las otras herramientas solo en el conjunto 50G y en los otros dos conjuntos fueron menor (59% - 85%). Por otro lado, Centrifuge mostró las precisiones más bajas (17%-40%) en ambos niveles taxonómicos, y en el caso de Taxonomer obtuvo precisiones bajas a nivel de especie (21%-46%) y mejoraron a nivel de familia (83%-100%) en todos los conjuntos.

Un análisis más detallado revela que algunas herramientas tuvieron un sesgo importante en la clasificación de ciertos tipos de familias en los tres conjuntos de datos. Por ejemplo, drVM tiende a fallar en la identificación de familias de genomas dsDNA y ssRNA (+). Mientras tanto, One Codex, SLIMM y Taxonomer tuvieron problemas en la identificación de virus de genoma dsDNA y BLAST en familias de genoma ssRNA (+). Curiosamente, *Luteoviridae* (en *500G* y *1000G*) fue la familia menos identificada por la mayoría de las herramientas. De lo contrario, las familias *Arenaviridae*, *Baculoviridae*, *Bromoviridae*, *Geminiviridae*, *Myoviridae*, *Papillomaviridae*, *Peribunyaviridae*, *Reoviridae*, *Partitiviridae* y *Parvoviridae* fueron bien identificadas por la mayoría de las herramientas.

4.4.1.2. Lecturas cortas

Los resultados de sensibilidad y precisión obtenidos en los conjuntos de datos se muestran en la Figura 4.2. En relación con la sensibilidad, todas las herramientas en los tres conjuntos y en ambos niveles no obtuvieron altas sensibilidades a excepción de BLAST a nivel familia. En el conjunto *Eukaryotic* las herramientas BLAST, Centrifuge, DIAMOND2, drVM, Kraken2, One Codex y Taxonomer obtuvieron a nivel familia sensibilidades altas (68%-81%), mientras que a nivel especie la sensibilidad fue ligeramente inferior (46%-68%).

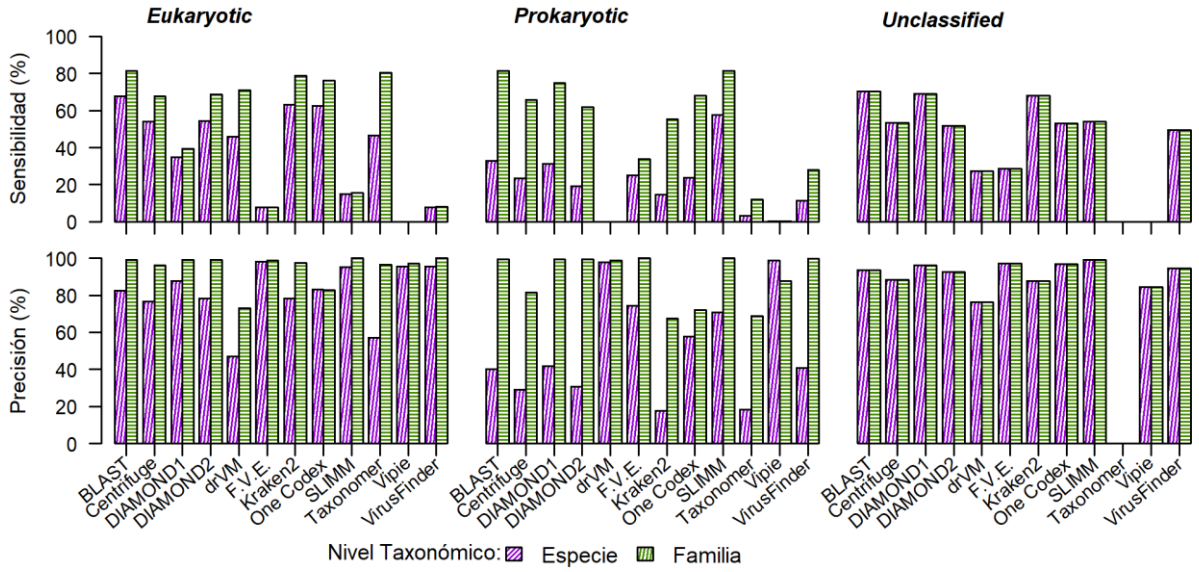


Figura 4.2. Sensibilidades y precisiones de las herramientas obtenidas, en el nivel taxonómico de familia y especie, en los conjuntos de datos simulados de Illumina. Cada panel de columnas representa un conjunto de datos. El conjunto de datos de Unclassified, el nivel familia, representa todas las especies que no tienen una familia asignada. La herramienta FastViromeExplorer está anotada como F.V.E., DIAMOND1 es la herramienta DIAMOND evaluada con todos los genomas virales en la base de datos nt y DIAMOND2 con genomas virales de RefSeq.

En el conjunto *Prokaryotic* las herramientas BLAST, Centrifuge, DIAMOND2 y One Codex obtuvieron sensibilidades moderadas a nivel familia (62%-81%) y en el conjunto *Unclassified* fueron BLAST y Kraken2 en ambos niveles taxonómicos (68%-70%), por otro lado, Centrifuge, DIAMOND2 y One Codex tuvieron una reducción (52%-68%) en estos dos últimos conjuntos de datos. Además, las herramientas DIAMOND1 y SLIMM lograron una alta sensibilidad a nivel familia (75%-81%) y especie (31%-58%) en el conjunto *Prokaryotic* y en el conjunto *Unclassified* (54%-69% en ambos niveles taxonómicos), pero presentó una gran caída en el conjunto de *Eukaryotic* (SLIMM <16% y DIAMOND1 <39%). Por otro lado, Vipie obtuvo la sensibilidad más baja en todos los casos (<1%).

En relación con la precisión, en los conjuntos *Eukaryotic* y *Unclassified* (Figura 4.2), todas las herramientas mostraron precisiones sobresalientes (>83% a nivel de familia y 77% a nivel de especie), excepto drVM y Taxonomer que mostraron una disminución en sus precisiones (73% y 76% a nivel familia, y 47%-76% a nivel especie, respectivamente) y en el conjunto *Unclassified* Taxonomer obtuvo 0%. Con respecto al conjunto *Prokaryotic,* Centrifuge, Kraken2, One Codex y Taxonomer mostraron bajas precisiones, entre 67%-81% a nivel de familia, y a nivel de especie de 18%-58%, y el resto de las herramientas obtuvieron mejores precisiones (88%-100% para familia y 71%-99% para especies), excepto por BLAST, DIAMOND1, DIAMOND2 y VirusFinder, que presentó una drástica disminución de su precisión a nivel de especie (31%-42%).

Un análisis detallado reveló que en el conjunto *Eukaryotic,* FastViromeExplorer, SLIMM y VirusFinder no clasificaron las familias con genomas dsDNA, mientras que drVM, SLIMM y Taxonomer fallaron con los genomas de ssRNA. En particular, las familias virales *Metaxyviridae*, *Leishbuviridae* y *Xinmoviridae,* que tiene pocos genomas de referencia e infectan plantas o insectos, no fueron identificadas por ninguna herramienta. Por otro lado, las familias *Arenaviridae*, *Benyviridae*, *Birnaviridae*, *Bromoviridae*, *Chrysoviridae*, *Closteroviridae*, *Geminiviridae, Nanoviridae*, *Orthomyxoviridae*, *Partitiviridae, Parvoviridae, Phenuiviridae*, *Reoviridae*, *Tospoviridae* y *Secoviridae* fueron identificadas por todas las herramientas, sin considerar Vipie. En el conjunto *Prokaryotic*, la familia *Finnlakeviridae* (con un genoma de referencia) fue la más difícil de identificar con casi todas las herramientas, ya que solo DIAMOND1 y SLIMM clasificaron sus lecturas.

4.4.2. Evaluación de los conjuntos simulados no virales

Los conjuntos *Human* y *Bacterial* fueron evaluados con la métrica de Especificidad. Los resultados arrojan que todas las herramientas tienen altas sensibilidades del 99% al 100%. Esto mostró que las herramientas tienen la habilidad de clasificar correctamente las lecturas no virales (humano y bacterial). Por otro lado, la exactitud fue evaluado en seis herramientas debió a que su base de datos (BD) de referencia incluye secuencias de humano y bacterias, presentando que Kraken2 y Centrifuge obtuvieron altas exactitudes (81% y 71%, respectivamente), con exactitudes moderadas se encuentran Once Codex y Taxonomer (62% y 71%, respectivamente). Finalmente, Vipie con el 2% y VirusFinder con el 31% obtuvieron las más bajas exactitudes.

4.4.3. Evaluación de los conjuntos reales

En los conjuntos de datos reales, los resultados obtenidos por las herramientas se presentan únicamente a nivel familia, dado que en los artículos originales es uno de los niveles

taxonómicos utilizados para reportar los resultados (70, 71) (ver Figura 4.3) y porque los modelos desarrollados en este proyecto de tesis clasifican a nivel taxonómico de familia. En las publicaciones originales (70, 71), los conjuntos reportaron entre el 4% y el 6.5% de las lecturas como virales y se analizaron con el mismo proceso, utilizando BLASTn, BLASTx, MEGAN (72) y el algoritmo LCA (*Lowest Common Ancestor*) (73).

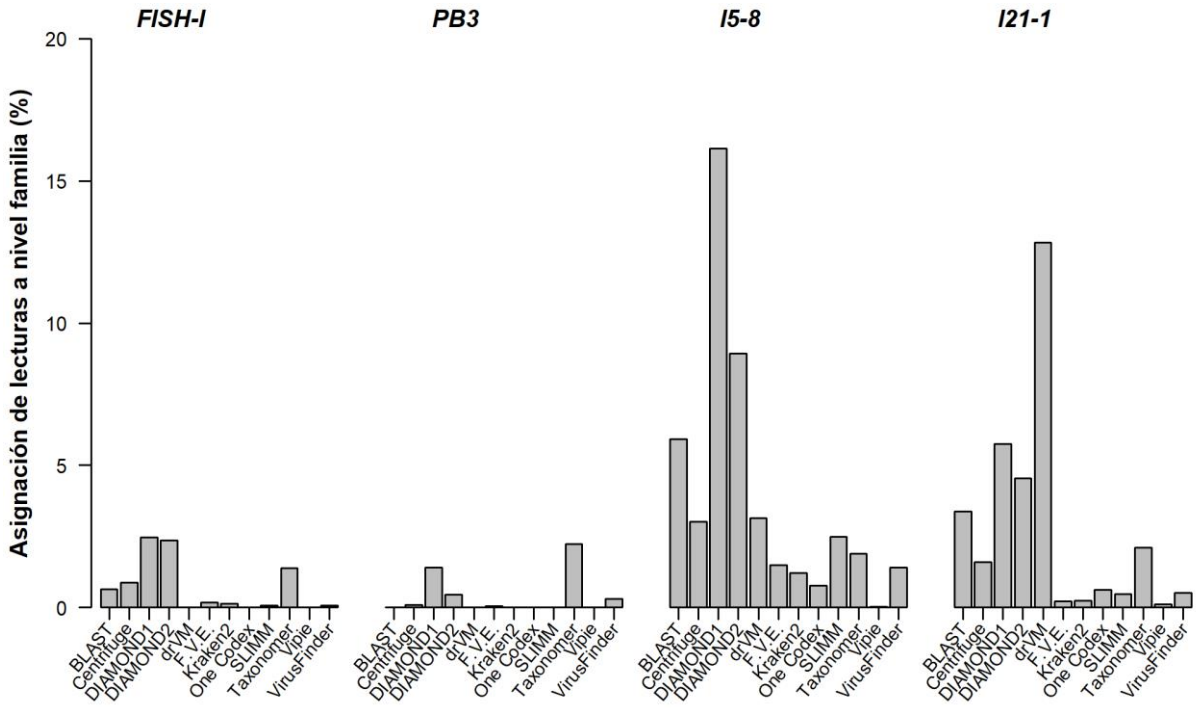


Figura 4.3. Proporción de lecturas asignadas por cada herramienta a una familia viral. La herramienta FastViromeExplorer se representa como F.V.E., DIAMOND1 es la herramienta DIAMOND evaluada con todos los genomas virales en la base de datos nt y DIAMOND2 con genomas virales de RefSeq.

Centrifuge (0.1%-3.2%) y Taxonomer (1.4%-2.2%) obtuvieron en promedio el porcentaje más alto de lecturas virales considerando los cuatro conjuntos de datos. Además, todas las herramientas asignaron menos del 3% de las lecturas virales en los conjuntos de datos *FISH-I* y *PB3*, mientras que en *I5-8* todas las herramientas obtuvieron más del doble de lecturas (<7%), excepto DIAMOND1 y DIAMOND2 que clasificaron entre 8% y 16% de las lecturas virales. En el conjunto *I21-1* se identificaron menos del 5%, a excepción de drVM que identificó más del 12% de las lecturas virales. Curiosamente, los dos primeros conjuntos de datos (*FISH-I* y *PB3*), no son de origen humano, que tuvieron la clasificación de lectura viral más baja en todas las herramientas (en promedio 0.7% y 0.4%, respectivamente), seguidos por el conjunto de datos *I21-1* (con 2.7%), que es de origen humano nasofaríngeo, y finalmente los porcentajes más altos se obtuvieron en *I5-8* (con 3.9%), que es de heces humanas. Estos resultados sugieren una clasificación más alta de

lecturas virales en muestras que provienen de entornos más estudiados, por ejemplo, el humano.

4.4.4. Requisitos de memoria y tiempo utilizados

La memoria y el tiempo de ejecución de todas las herramientas de clasificación en cada conjunto se presenta en la Tabla 4.1. Las herramientas BLAST, DIAMOND1 y VirusFinder fueron las herramientas que consumieron más tiempo, por otro lado, Centrifuge y FastViromeExplorer fueron las que menos tiempo consumieron. Por ejemplo, BLAST empleó 27 horas CPU para analizar 10 millones de lecturas y Centrifuge utilizó 6 minutos. En relación con la memoria, la herramienta Kraken2 fue la que mayor cantidad de memoria consumió, debido a que requirió de por lo menos 140 GB para su funcionamiento. Curiosamente, BLAST (con la base de datos (BD) viral de RefSeq) fue la que menos memoria consumió con menos del 500 MB.

Además, para las herramientas Web (One Codex, Taxonomer y Vipie) no fue posible medir la memoria que consumieron y en el caso de Taxonomer no fue posible medir el tiempo, ya que actualmente se encuentra en mantenimiento.

Tabla 4. 1. Comparación de la memoria y tiempo de ejecución requeridos por cada herramienta en el proceso de clasificación de cada conjunto de datos.

Conjunto de datos	BLAST*	Centrifuge*	DIAMOND1*	DIAMOND2*	drVM*	F.V.E.*	Kraken2*	One Codex	SLIMM*	Taxonomer	Vipie	VirusFinder*
	Tiempo de ejecución (Minutos) calculados por core.											
50G	83	3	112	27	9	2	64	6	10	ns	8	43
500G	89	5	119	25	19	2	64	9	21	ns	9	70
1000G	102	4	111	22	18	2	64	8	20	ns	9	75
Eukaryotic	1729	5	1644	502	519	6	69	44	49	ns	51	1296
Prokaryotic	1768	7	1536	547	208	8	68	47	51	ns	38	1264
Unclassified	1426	5	857	389	208	5	68	48	47	ns	29	1253
FISH-I	927	5	321	135	64	6	72	33	46	ns	56	476
PB3	664	5	241	104	47	5	70	31	48	ns	37	546
I5-8	1265	7	354	139	61	5	73	38	49	ns	42	500
I21-1	928	3	149	35	55	5	71	19	49	ns	29	472
Bacterial	66	2	54	14	10	2	65	6	3	ns	7	19
Human	64	2	51	11	12	2	65	5	1	ns	6	57
Total (Minutos)	9111	54	5549	1950	1230	**50**	813	294	394	ns	321	7801
	Memoria (MB)											
50G	282	16,998	3246	922	1556	6835	143,524	ns	26,788	ns	ns	4618
500G	393	17,009	13,343	1044	1690	6836	143,667	ns	26,788	ns	ns	4618
1000G	433	17,009	3584	2591	1536	7022	143,811	ns	26,788	ns	ns	4618
Eukaryotic	475	17,213	11,991	5366	7967	7655	144,241	ns	26,757	ns	ns	4516
Prokaryotic	401	17,162	12,216	9339	1812	7662	144,179	ns	26,757	ns	ns	4516
Unclassified	235	17,029	12,052	10,179	4045	7644	144,230	ns	26,757	ns	ns	4516
FISH-I	189	17,009	7322	4608	1823	6837	143,964	ns	26,757	ns	ns	4096
PB3	189	17,009	6308	6318	1823	6835	143,974	ns	26,757	ns	ns	4516
I5-8	187	17,060	16,005	8376	1853	6919	143,985	ns	26,767	ns	ns	4096
I21-1	183	17,142	15,933	2273	1802	6838	144,067	ns	26,767	ns	ns	4096
Bacterial	147	16,998	2775	768	1833	6834	143,606	ns	26,757	ns	ns	4618
Human	152	16,998	2785	748	1833	6836	143,585	ns	26,757	ns	ns	4618
Promedio (MB)	**272**	17053	8963	4378	2464	7063	143903		26766			4454
Desviación Estándar (MB)	*120*	*76*	*5174*	*3515*	*1857*	*361*	*259*		*14*			*220*

*La herramienta FastViromeExplorer está representado por F.V.E.; ns significa no especificada. * Todas las herramientas fueron ejecutadas en el mismo clúster de computadoras con las siguientes características: Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz que tiene 6 nodos (AMD Opteron(tm) Processor 6376) con 64 cores y 512 GB de memoria.*

4.5. Conclusiones de la evaluación de las herramientas

Se presentó una comparación objetiva, dado que se utilizaron los mismos conjuntos de datos y las mismas métricas, en once herramientas de clasificación taxonómica de conjuntos de datos metagenómicos, utilizando los mismos conjuntos de datos y métricas de evaluación. De manera global, todas las herramientas mostraron altas precisiones en la clasificación taxonómica a nivel familia, pero las sensibilidades que obtuvieron fueron menores.

BLAST (con la base de datos (BD) de RefSeq viral) y Kraken2 fueron las herramientas que mostraron un alto desempeño, tanto en sensibilidad como precisión, a excepción de Kraken2 en el conjunto de datos *Prokaryotic* donde presentó un bajo desempeño. Además, la herramienta Kraken2 mostró la más alta especificidad y precisión, pero BLAST no porque su BD únicamente cuenta con secuencias virales. Es importante mencionar que Kraken2 fue la herramienta que mayor cantidad de memoria utilizó, con 140 GB, mientras que BLAST fue la herramienta que más tiempo utilizó para realizar la clasificación, con 27 horas por cada 10 millones de lecturas.

Al observar los resultados, se identificaron varios factores que impactan en el desempeño de las herramientas. Un factor es la longitud de las lecturas, en el caso de las lecturas largas (conjuntos *50G*, *500G* y *1000G*) se clasifican mejor, con sensibilidades altas, con las herramientas BLAST, FastViromeExplorer, Kraken2, Vipie y VirusFinder. Estas herramientas presentaron una sensibilidad de al menos 10%, 64%, 2%,19% y 44% mayor respectivamente, en las lecturas largas en comparación con las cortas. Por su parte, en las lecturas cortas las herramientas Centrifuge, DIAMOND y One Codex obtuvieron altas sensibilidades en comparación con las largas; por ejemplo, para Centrifuge su sensibilidad fue de al menos 15% mayor en las lecturas cortas que en las largas, para DIAMOND fue del 9% y para One Codex fue del 24%.

En nivel taxonómico de clasificación es otro factor que influye en el desempeño de las herramientas. Todas las herramientas mostraron una mayor sensibilidad y precisión a niveles taxonómicos generales, como familia, en contraste con los niveles específicos como especie (32). Esta tendencia es más clara en las lecturas cortas, a nivel familia la sensibilidad media fue del 16% y la precisión fue del 23%, y son superiores que a nivel especie; por otro lado, en las lecturas largas las mismas métricas fueron del 7% y 11% respectivamente. Los resultados obtenidos sugieren que las lecturas cortas carece de información que pueden compartir con otras especies filogenéticamente cercanas, por lo tanto, las lecturas cortas se clasifican mejor en los niveles taxonómicos más altos, como se informó anteriormente (74).

Otro factor que afecta el desempeño de las herramientas es la riqueza del viroma (número de especies virales). En las lecturas largas, las herramientas Centrifuge, drVM, One Codex y SLIMM presentaron un incremento en sus sensibilidades, entre 4% al 25%, al aumentar la riqueza del viroma. Por el contrario, las herramientas BLAST, FastViromeExplorer, Kraken2, Taxonomer, Vipie y VirusFinder mostraron una disminución es sus sensibilidades, entre un 2% al 34%, como fue descrito previamente en algunas herramientas (67).

Considerando el host, en las lecturas cortas, las herramientas BLAST, Centrifuge, DIAMOND2 y One Codex alcanzaron las más altas sensibilidades y precisiones (>62% y >72% respectivamente) a nivel familia en los virus eucariontes y procariontes. Para los virus procariontes, DIAMOND1 y SLIMM obtuvieron buenas sensibilidades y altas precisiones (>75% y del 99% respectivamente), y para los virus eucariontes, las herramientas drVM, Kraken2 y Taxonomer presentaron buenas sensibilidades (>71%) y precisiones (>73%).

Un análisis más detallado detectó que algunas familias como *Bromoviridae*, *Closteroviridae*, *Geminiviridae*, *Orthomyxoviridae*, *Reoviridae*, *Partitiviridae* y *Parvoviridae* fueron mejor identificadas por la mayoría de las herramientas; debido a que las familias antes mencionadas han sido relativamente bien estudiadas, por lo que en las BD de referencia disponen de mayor información genética sobre las mismas. Otro punto a examinar es la mala clasificación de las familias virales, los resultados sugieren que las familias con pocos genomas anotados en las BD de referencia fueron mal clasificados por casi todas las herramientas; problema descrito previamente (9, 74), por ejemplo, las familias virales *Leishbuviridae*, *Xinmoviridae* o *Barnaviridae*.

En relación con las muestras reales, las herramientas DIAMOND y Taxonomer asignaron como virales más del 1% de las lecturas de los conjuntos *FISH-I* y *PB3*, mientras que BLAST, DIAMOND y drVM identificaron más del 5% de las lecturas de los conjuntos *I5-8* e *I21-1*. Los resultados sugieren que en muestras de origen humano las herramientas clasifican más lecturas virales, sobre todos en muestras de heces, ya que son las muestras humanas más estudiadas (75, 76). Por otra parte, en los ambientes poco investigados, se clasifican pocas lecturas como virales, lo que sugiere que las muestras contienen virus con baja identidad con los genomas de referencia.

En el caso de los conjuntos de datos no virales, todas las herramientas presentaron alta especificidad, es decir, pueden clasificar correctamente las lecturas no virales (de origen humano y bacteriano). Además, las herramientas Centrifuge, Kraken2, One Codex, Taxonomer, Vipie y VirusFinder cuentan con una BD que contiene genomas no virales, por lo que alcanzaron especificidades ligeramente mayores (99.9%) al resto de las herramientas (99.6%).

En conclusión, para elegir una herramienta metagenómica se debe considerar varios factores, como: la longitud de la lectura, la riqueza de la muestra, el tipo de muestra analizada (enfocada en fagos y virus eucariontes), nivel taxonómico, requisitos de memoria, tiempo de procesamiento, entre otros. Esto representa el éxito o fracaso de un análisis metagenómico. Además, la identificación de las lecturas cortas a nivel especie continúan siendo un reto debido a que todas las herramientas presentaron menos del 70% de sensibilidad. Los resultados revelan que las lecturas virales que fueron mal clasificadas guían a conclusiones erróneas. Finalmente, si la baja sensibilidad se produce en los conjuntos simulados que no cuentan con una variación genética, se incrementará en los estudios reales.

CAPÍTULO 5. METODOLOGÍA DE SOLUCIÓN

El método computacional de clasificación taxonómica viral desarrollado durante este trabajo de investigación realiza una anotación a nivel taxonómico de familia y está basado en Redes Neuronales Convolucionales. Se decidió usar esta técnica debido a que las redes neuronales de convolución permiten el uso de grandes conjuntos de datos; extraer, seleccionar y aprender características y patrones complejos; comparten el uso de parámetros, reducen la dimensión de los datos y pueden contender con clases desbalanceadas, es decir, donde hay una diferencia muy grande entre el número de elementos de cada clase. Estas particularidades se tienen presentes en los conjuntos de datos metagenómicos que se emplean en este proyecto.

La presente metodología para discriminar taxonómicamente las secuencias de virus en alguna de las familias virales existentes fue realizada a nivel nucleótido, lo cual permite identificar lecturas virales con alta identidad a las secuencias de referencia de la base de datos (BD). Esta metodología cuenta con los siguientes tres módulos principales: (i) Preprocesamiento (análisis y reducción) de las BD de referencia; (ii) Preparación de los conjuntos de datos, y (iii) Clasificación a nivel familia. A continuación, se describen cada uno de los módulos.

5.1. Definición de clases usando las bases de datos de referencia RefSeq y nt

Como se vio en el Capítulo 4, varias herramientas utilizan la base de datos (BD) RefSeq, la cual es una BD reducida, no redundante (no tiene genomas repetidos) y contiene un genoma de cada especie viral. Debido a que los virus tienen una alta tasa de mutaciones, se requiere una BD más completa que contenga varias secuencias por cada especie viral. Por lo tanto, en este trabajo se planea evaluar la eficiencia del modelo en función a la BD a utilizar en el entrenamiento (Objetivo II, definido en la sección 1.3). Por un lado, la de RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) y por otro es la de nucleótidos (nt) de NCBI (https://www.ncbi.nlm.nih.gov/nucleotide/), que contiene todo el universo viral.

Cada una de las familias virales presente en la BD, representa una de las clases que predice cada uno de los modelos CNN, por lo tanto, el número de clases (familias) que tiene cada modelo depende de la BD. El proceso para la creación de las clases se detalla en las siguientes subsecciones.

5.1.1. Descarga de las bases de datos

Para realizar el proceso de entrenamiento de los dos modelos de CNN, se utilizaron dos bases de datos (BD) de referencia, la primera fue RefSeq viral y la segunda fue toda la BD viral de nucleótidos (nt, abreviatura de nucleótido) que pertenecen al NCBI. La información de ambas BD fue descargada en el 1 de enero del 2022. De la BD de RefSeq, se descargaron 14,748 genomas virales que pertenecen a 183 familias definidas y un grupo, el cual fue definido como "*Unclassified*", dado que contiene los genomas que no están asignados en una familia. Por otro lado, de la BD de referencia nt se descargaron 6,451,040 secuencias virales, entre los que se encuentran genomas completos y parciales, y que pertenecen a 190 familias definidas y su respectivo grupo "*Unclassified*" con las secuencias que carecen de asignación taxonómica a nivel familia.

5.1.2. Análisis y reducción de la base de datos

Hay algunos virus que han sido más estudiados que otros, porque causan enfermedades o porque son de interés. Esto ha ocasionado un inmenso desbalance de secuencias de genomas virales entre las familias. En otras palabras, hay familias virales que tienen solo un genoma anotado, mientras que hay otras que tiene un gran número de genomas anotados. Por ejemplo, en la base de datos (BD) de RefSeq la familia *Alvernaviridae* tiene un genoma, mientras que la familia *Siphoviridae* tiene 2,267 genomas.

Para reducir este sesgo del desbalance de genomas que existe entre las familias de virus, se realizó un proceso de reducción. Primero, se identificaron las familias con mayor y menor número de secuencias. Después, se eliminan las familias con menos de cinco genomas anotados debido a que no proporcionan información suficiente para que los modelos de aprendizaje profundo aprendan durante el entrenamiento. También, se eliminaron las secuencias menores a 240 pb y de aquellas que fueran menores al 90% de la secuencia de referencia principal. Esto se debe a que al eliminar las secuencias cortas o fragmentos de genomas se mantienen únicamente las secuencias casi completas. Posteriormente, se realizó la reducción en las familias con el mayor número de secuencias, mediante la herramienta CD-Hit (77). CD-Hit permitió agrupar las secuencias duplicadas o con alta similitud entre ellas. A continuación, se describe el procesamiento realizado con CD-Hit v.4.6.8 (58, 59):

(i) Se ordenaron las secuencias de la longitud mayor a la menor.

(ii) La primera secuencia fue clasificada como la secuencia representativa del primer clúster.

(iii) La segunda secuencia se comparó con la secuencia representativa del clúster existente; si la similitud fue mayor o igual al porcentaje establecido, la secuencia fue

agregada al primer clúster, de lo contrario, fue asignada como representante de un nuevo clúster.

Para el resto de las secuencias, se sigue este mismo procedimiento. Finalmente, las secuencias representativas de cada clúster formado constituyeron la BD depurada. Para los porcentajes de similitud y cobertura utilizados en el CD-HIT fueron establecidos de forma empírica consideraron los rangos intercuartílicos de la distribución previa a la reducción, los cuales se describen a continuación:

(i) A las familias del segundo cuartil (50%) se les aplicó CD-HIT con una similitud del 99% y una cobertura al 95%.

(ii) A las familias del tercer cuartil (75%) se les aplicó una similitud del 98% y una cobertura al 90%.

Las familias que fueron obtenidas después del proceso de reducción en una BD, son las clases del modelo CNN para dicho BD.

5.1.2.1. Base de datos viral de RefSeq

Se efectuó un análisis exploratorio del contenido la base de datos (BD) RefSeq. Como se mencionó en la sección 5.1.1, la BD contenía 14,748 secuencias, las cuales pertenecen a 184 familias virales (59.2% de genoma de DNA y 40.8% de RNA). De estas, 128 familias tuvieron más de 5 genomas (Figura 5.1-a). El rango intercuartílico indica que el 25% de las familias tuvieron 3 genomas, el 50% tuvieron 14 genomas y 75% tuvieron 76 genomas. Además, dentro del análisis realizado se identificó que la familia con mayor número de genomas fue la familia *Siphoviridae* (que se encuentra en el tercer cuartil) con 2,267 genomas.

Después, de que se aplicaron los filtros de reducción mencionados en la sección 5.1.2, el conjunto final fue constituido por 14,628 genomas virales. Estos pertenecen a 127 familias de virus (58.8% de DNA y 40.3% de RNA) y un grupo extra definido como *Unclassified,* el cual representa las especies virales que no tienen asignada una familia. Por lo tanto, el modelo CNN que usa la BD RefSeq tiene 128 clases, una por cada clase. En este conjunto final, el 25% de las familias tuvieron 13 genomas, el 50% tuvieron 46 genomas y el 75% tuvieron 108 genomas (Figura 5.1-b).

Cabe destacar que, a pesar de haber eliminado las familias con menos genomas, la BD final aún presenta un desbalance, el cual va desde 5 genomas por familia a 2,267 genomas, siendo *Siphoviridae* y *Unclassified* las familias mayoritarias (ver Figura 5.2).

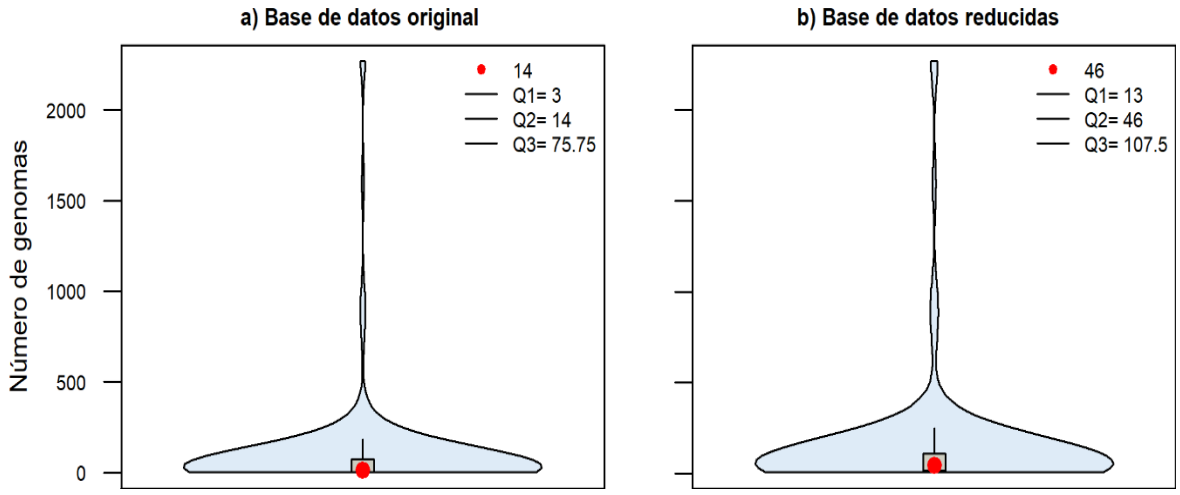


Figura 5.1. Análisis del número de genomas en todas las familias de la base de datos RefSeq: (a) Distribución de los genomas antes de la reducción (b) Distribución después de la reducción. El punto rojo índica la mediana del conjunto.

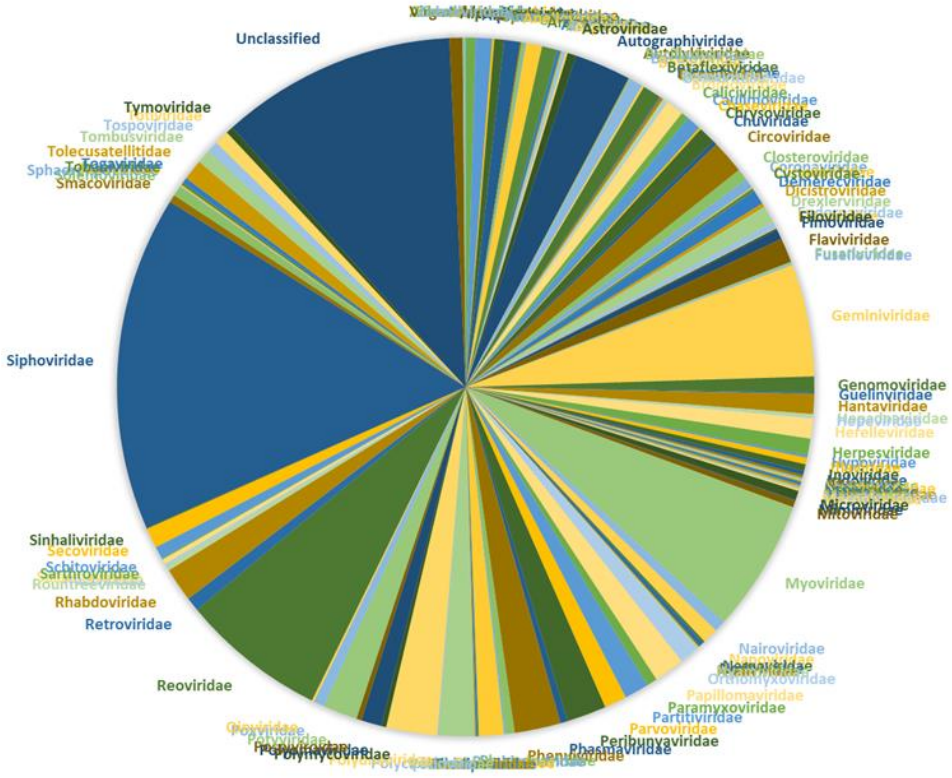


Figura 5. 2. Desbalance de las clases en la base de datos RefSeq. Entre mayor es la clase (más genomas tiene) mayor es la sección del gráfico.

5.1.2.2. Base de datos nt

En la base de datos (BD) nt, el análisis inicia con 6,451,040 números de secuencias, las cuales pertenecen a 191 familias (7% de los genomas son de DNA y 97% de RNA). De estas, 170 familias tuvieron más de cinco secuencias. El rango intercuartílico indicó que el 25% de las familias tuvieron 26 secuencias, el 50% tuvieron 344 secuencias y 75% tuvieron 4,466 secuencias (Figura 5.3-a). Esté análisis exhibió la existencia de un enorme desbalance de secuencias que contiene cada familia. Particularmente, se identificaron que solo tres familias (*Coronaviridae*, *Orthomyxoviridae* y *Reoviridae*) tuvieron más de 500,000 secuencias.

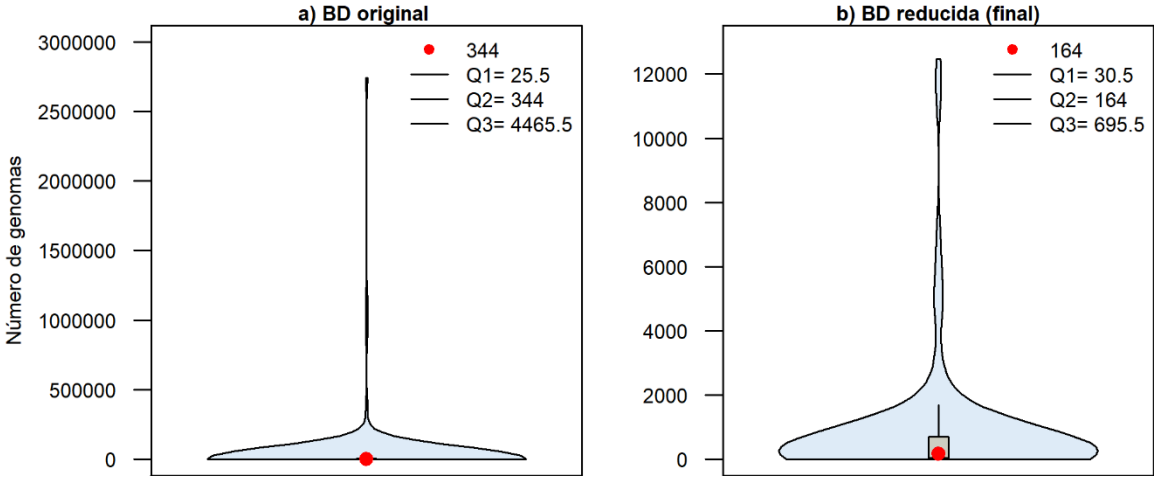


Figura 5.3. Análisis del número de genomas en todas las familias de la base de datos nt: (a) Distribución antes de la reducción, (b) Distribución después de la segunda reducción. El punto rojo índica la mediana del conjunto.

Posteriormente, se realizó una reducción de la BD, al igual que RefSeq aplicando los filtros antes mencionados en la sección 5.1.2. Después, se realizó el proceso de eliminar la redundancia en la BD mediante la herramienta CD-HIT. Este proceso es necesario en la BD, ya que contiene múltiples secuencias por cada especie viral, a diferencia de RefSeq que contiene una sola secuencia por cada especie.

En la Figura 5.3-b se presenta la distribución final de la BD nt reducida. El 25%, 50% y 75% de las familias tuvieron 30, 164 y 696 secuencias, respectivamente. la BD terminó constituida por 132,089 secuencias, lo que representa solo el 2% de la información original. Se obtuvieron169 familias definidas (1% de los genomas son de DNA y 1% de RNA) y un grupo *Unclassified*, por lo tanto, la CNN que utiliza la BD nt tiene 170 clases. Interesantemente, en ambos tipos de genomas (DNA y RNA) existe una reducción, siendo sobresaliente en los genomas de RNA. Es importante mencionar que aun con el proceso de reducción que fue aplicado, se presenta aún un alto desbalance entre las familias, como se muestra en la Figura 5.4.

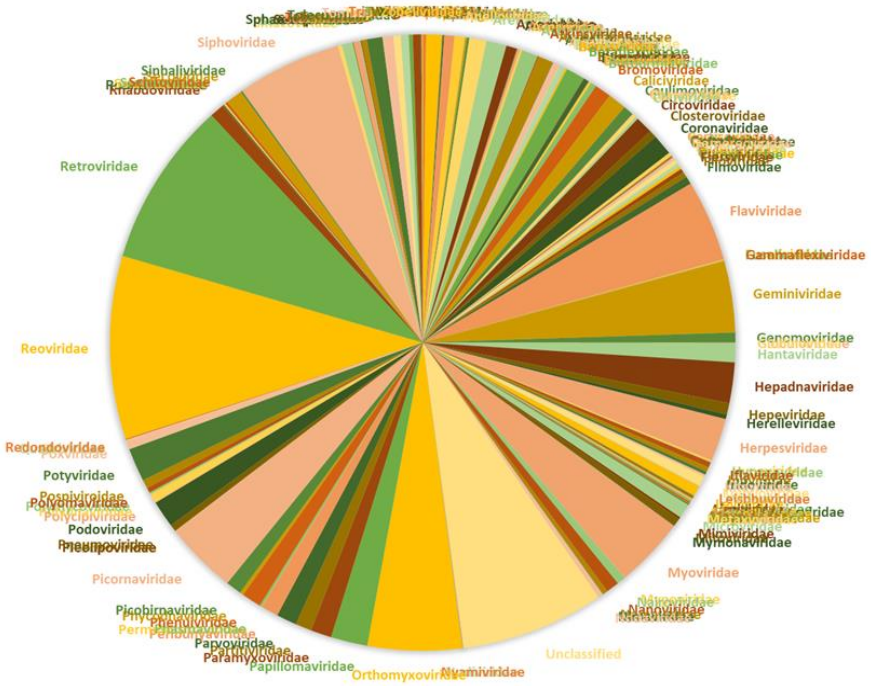


Figura 5. 4. Desbalance de las clases en la base de datos nt. Entre mayor es la clase (más genomas tiene) mayor es la sección del gráfico.

De este punto en adelante, la información de la BD nt se establece como genomas, debido a que las secuencias tienen una cobertura mayor o igual al 90% del genoma completo de la especie.

5.1.2.3. Resumen comparativo de la reducción de RefSeq y nt.

En la Tabla 5.1 se presenta el número de genomas y familias, así como los resultados obtenidos en cada uno de los pasos del proceso de reducción de las bases de datos (BD) RefSeq (sección 5.1.2.1) y nt (sección 5.1.2.2). Estos resultados indican que nt contuvo 9 veces más genomas que RefSeq, en otras palabras, RefSeq representa el 14% de la información contenida en nt. Las BD de RefSeq inicialmente tenía 14,748 genomas y termina con 14,628 genomas, por lo tanto, se hizo una reducción del 0.90%. En la BD nt inició con 6,451,040 genomas y finalizó 132,628 genomas por consiguiente se redujeron el 98%.

Tabla 5. 1. Resultados de la reducción de la base de datos RefSeq y nt.

Número de genomas (o secuencias en el caso de la BD nt) y familias obtenidas en las diferentes etapas de la reducción.	**RefSeq**	**nt**
Número de genomas antes del proceso de reducción:	14,748	6,451,040

Número de genomas (o secuencias en el caso de la BD nt) y familias obtenidas en las diferentes etapas de la reducción.	RefSeq	nt
Número de familias que existían originalmente en las BD (antes del proceso de reducción):	184	191
Número de familias poco anotadas en las BD (con menos de 5 genomas):	56	21
Número de familias o clases finales en las BD (con más de 5 genomas):	128	170
Número de genomas después del proceso de reducción (eliminar las familias con menos de 5 genomas, con longitudes menores de 240 pb y con cobertura menor de 90%):	14,628	3,854,755
Número de genomas después de aplicar CD-HIT:	NA	132,089

NA significa No Aplica.

En la Figura 5.5 se puede observar, las familias divididas en virus de eucariontes, procariontes o virus sin asignación taxonómica (*Unclassified*). La BD de RefSeq está formada por un 57% de las familias de virus eucariontes, 32% de las familias de virus procariontes y el 11% de virus sin asignación, mientras que la BD nt tiene el 76% de las familias virus eucariontes y un 16% para las familias virus procariontes y un 9% para virus sin asignación taxonómica.

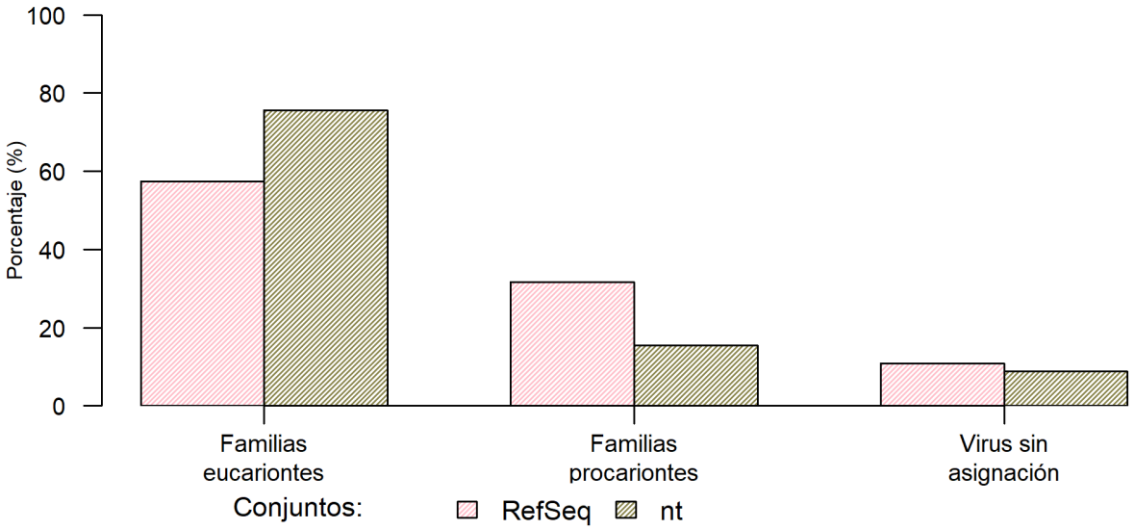


Figura 5. 5. Porcentajes de genomas por el tipo de familias viral (eucariontes, procariontes y virus sin asignación).

5.2. Preparación de los conjuntos de datos de la CNN

Para simular las lecturas generadas por NGS, las secuencias de los genomas de cada una de las bases de datos (BD) reducidas RefSeq y nt, se dividieron en *k-mers,* donde *k* = 150 pb de longitud, con saltos de 10 pb; en donde un salto es un número de caracteres sin considerar entre cada *k-mer.* Cabe mencionar, que el valor *k*=150 se definió considerando que es el tamaño de las lecturas generado más frecuentemente por los secuenciadores Illumina, los cuales son los más utilizados (78); y el tamaño del salto de 10 pb fue seleccionado después de diversas pruebas con varios tamaños, en donde se identificó que este salto no afecta el rendimiento de las CNN. En forma esquemática, en la Figura 5.6-a, se muestra como un genoma fue dividido en *k-mers* de 150 pb, y los *k-mers* resultantes fueron identificados como v_1, v_2, v_3, …, v_n. A cada *k-mer* se le agregó un *k-mer* complementario, que representa el complemento inverso (ver Figura 5.6-b) de la lectura, para simular las dos hebras del ADN. El complemento inverso se conformó invirtiendo los nucleótidos de una secuencia e intercambiándolos por su correspondiente nucleótido complementario; es decir, la Adenina (A) por la Timina (T), la Guanina (G) por la Citosina (C) y viceversa.

Posteriormente, para asegurar que todos los *k-mers* tengan la misma longitud, se les realizó un rellenado (*padding*) de ceros hacia la derecha, con la finalidad de que la información no relevante o que no pertenece al *k-mer* se ubique al final de la secuencia. Esto garantizó que los *k-mers* con menor longitud tengan el mismo tamaño de los *k-mers* establecidos (150 pb), lo cual también se aplicará cuando se haga la predicción de las lecturas reales. Después, cada uno de los nucleótidos base (“A”, “C”, “G”, “T”) son transformados a su respectiva codificación *one-hot,* formándose así un vector binario (con 0 y 1) (Figura 5.6-c). También se agregó el carácter “N” a la codificación, dado que este carácter es colocado por el secuenciador cuando no reconoce alguno de los cuatro nucleótidos base. Cada *k-mer* representa un ejemplo en el conjunto de datos y está formado por una matriz de tamaño 150 x 6.

Finalmente, cada *k-mer* fue etiquetado con la familia a la que corresponde, ya que cada familia representa una clase o salida del modelo. Cada familia fue codificada con un *one-hot* para tener una representación única (ver Figura 5.7). Para el conjunto de datos pertenecientes a la BD de RefSeq, los *k-mers* fueron etiquetados con alguna de las 128 clases y para la BD nt los *k-mers* son etiquetados a alguna de las 170 clases, en ambos casos el conjunto *Unclassified* es considerado una clase. Esta metodología fue aplicada a las dos BD RefSeq y nt.

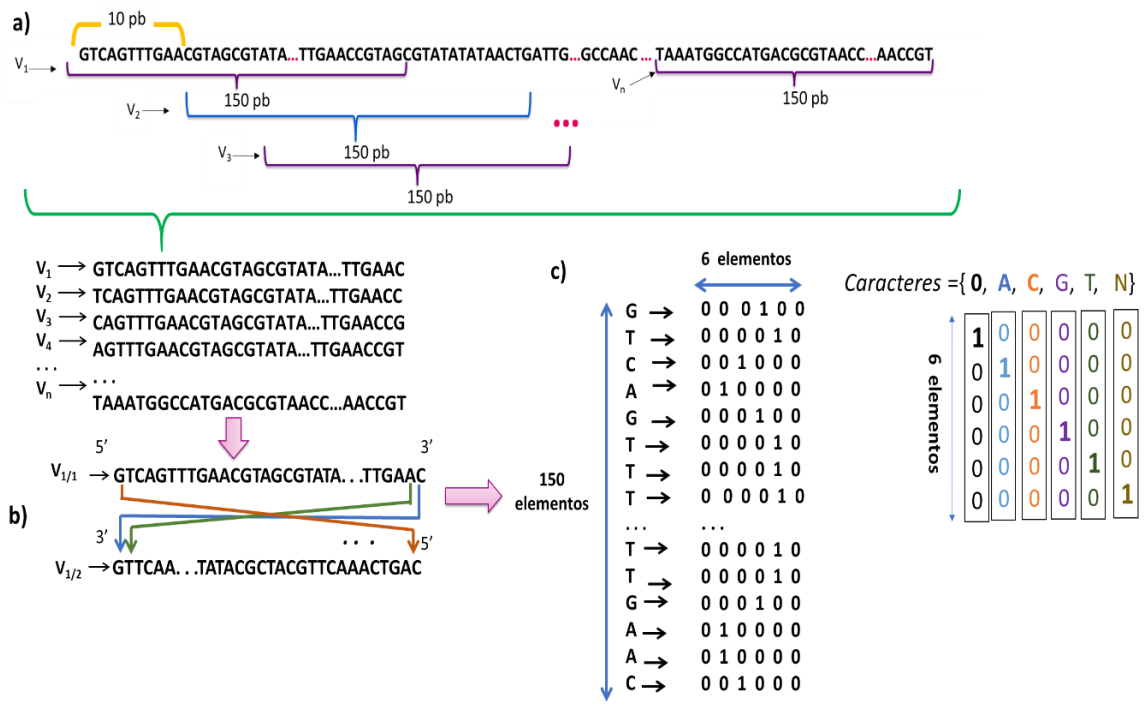


Figura 5.6. Preparación del conjunto de datos de entrada a la CNN: (a) Esquema de fragmentación de un genoma en k-mers de 150 pb con saltos de 10 pb; (b) Esquema de la adición de un complemento inverso reverso a cada k-mer; (c) Esquema de la transformación de los nucleótidos a su codificación binaria one-hot.

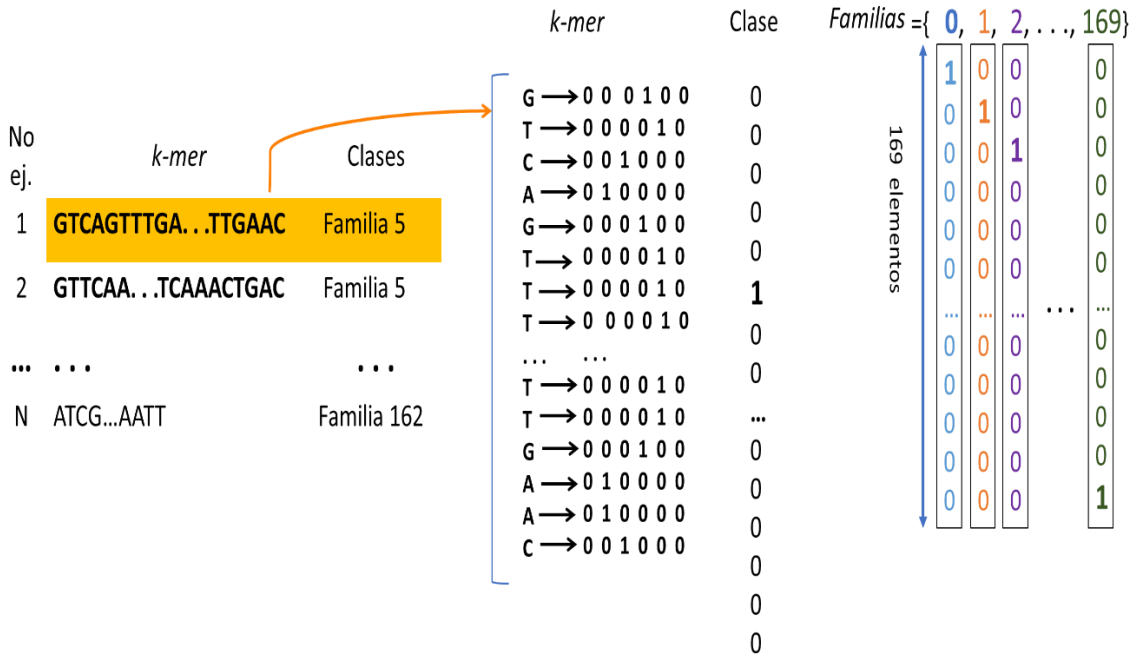


Figura 5.7. Etiquetado de los datos de entrada a su correspondiente familia mediante una codificación binaria one-hot.

5.2.1. Resumen de la preparación de los conjuntos de datos

En la Tabla 5.2 se presenta la cantidad de *k-mers* obtenidos de la preparación de los conjuntos de las bases de datos (BD) RefSeq y nt. Estos resultados indican que nt contuvo 7 veces más *k-mers* que RefSeq, en otras palabras, RefSeq representa el 14% de la información contenida en nt. Para los conjuntos de entrenamiento, la validación y prueba fue separado el 70%, 20% y 10% de la información total, respectivamente.

Tabla 5. 2. Resumen de la preparación de los conjuntos de ambas bases de datos.

Conjuntos de datos	**Número de k-mers**	
	RefSeq	**nt**
Global	43,523,298	309,624,849
Global incluyendo el complemento inverso	87,046,596	619,249,698
C*onjunto de entrenamiento*	60,932,506	433,474,626
C*onjunto de validación*	17,417,986	123,911,800
C*onjunto de prueba*	8,696,104	61,863,272

En la Figura 5.8-a se muestra el número de *k-mers* que tienen las familias de la BD RefSeq. Existen 118 familias con menos de un millón de *k-mers*, 7 familias que tienen 1 millón de *k-mers*, 1 que tiene 2 millones de *k-mers* y 2 con más de 11 millones de *k-mers*, las cuales son *Myoviridae* y *Siphoviridae* con un total de 11,001,560 y 11,864,581 *k-mers,* respectivamente. Estos resultados no son sorpresa porque ambas familias cuentas con los genomas virales más grandes en la BD RefSeq, de hasta 497,513 pb y 322,272 pb de longitud respectivamente. Mientras que en la BD nt (ver Figura 5.8-b) se detectaron 133 familias con menos de un millón de *k-mers*, 13 con un millón de *k-mers*, 6 con 2 millones de *k-mers,* 12 de 3 a 9 millones de *k-mers* y 6 con más de 11 millones de *k-mers.* Al igual que en RefSeq, las seis familias (*Mimiviridae, Poxviridae, Unclassified, Siphoviridae, Herpesviridae* y *Myoviridae*) con más de 11,000,000 de *k-mers* fueron las que contienen genomas de gran longitud. Por ejemplo, un genoma de la familia *Mimiviridae* cuenta con una longitud de 1,259,197 pb y generó 125,905 *k-mers*, mientras que un genoma de la familia *Alvernaviridae* cuenta con 4,375 pb de longitud y generó 425 *k-mers*.

En la Figura 5.9 se puede observar, las familias divididas en virus de eucariontes, procariontes o virus sin asignación taxonómica (*Unclassified*), proporcionando sus porcentajes de *k-mers*. La DB de RefSeq generó el 73% de los *k-mers* de las familias de virus procariontes, mientras que las familias de virus eucariontes generaron el 21% de *k-mers*,

permitiendo que los virus sin asignación aportan solo el 6% de *k-mers*. En el caso de la BD nt, las de familias de virus procariontes generaron un 40% de las lecturas, mientras que las familias de virus eucariontes proporcionaron el 53% de los *k-mers* totales y el resto de los *k-mers* (7%) provienen de los virus sin asignación.

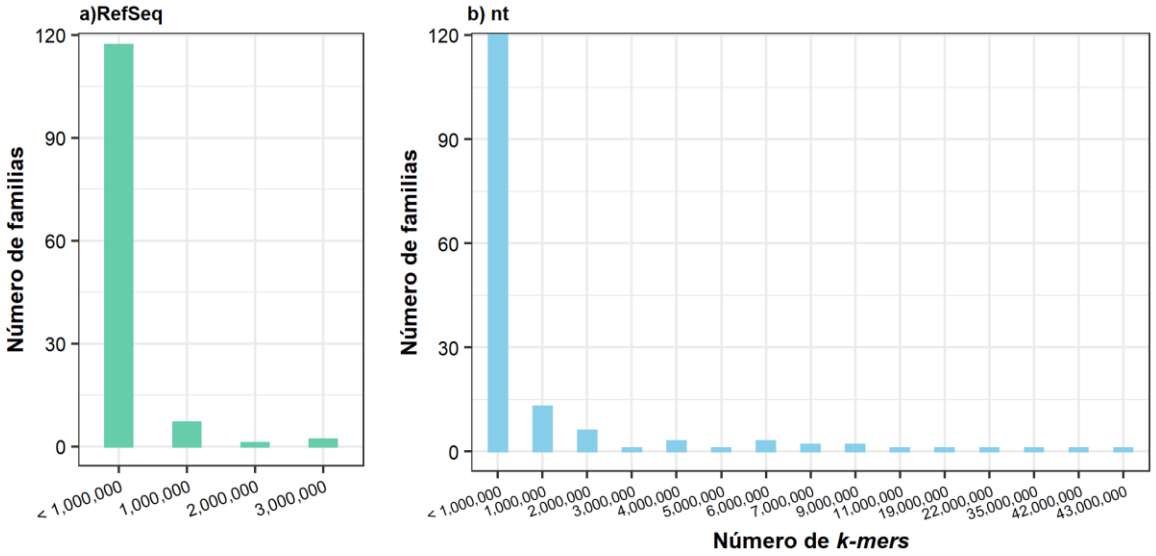


Figura 5.8. Número de k-mers que tienen las familias en las bases de datos: (a) RefSeq; (b) nt

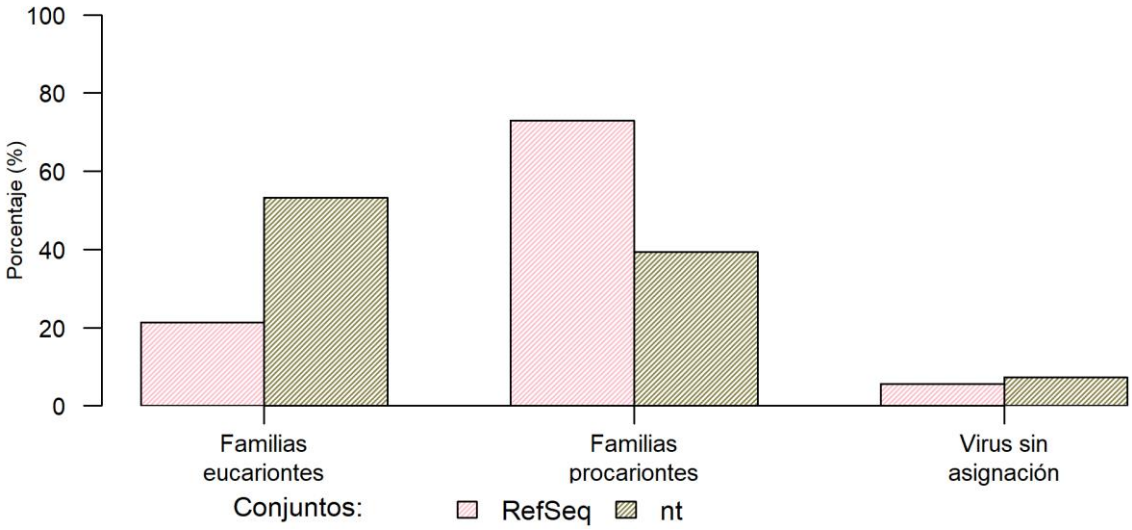


Figura 5.9. Porcentajes de k-mers por el tipo de familias viral (eucariontes, procariontes y virus sin asignación).

Una vez creados los dos conjuntos de datos, cada uno fue dividido en tres subconjuntos: entrenamiento (70%), validación (20%) y prueba (10%) (Tabla 5.2). El conjunto de entrenamiento fue usado para establecer los pesos o parámetros configurables en el proceso

de aprendizaje del modelo, mientras que el conjunto de validación afinó los pesos establecidos, y finalmente, el conjunto de prueba evaluó el rendimiento del modelo.

5.3. Modelo para la clasificación a nivel familia

Para la clasificación a nivel familia, se desarrolló un modelo global capaz de asignar cada entrada del conjunto de datos a una de las clases (incluido el grupo *Unclassified,* que representa a las lecturas virus sin asignación taxonómica), las cuales se definieron en la sección 5.1.1.

Para realizar los modelos CNN propuestos en este proyecto, que clasifican a nivel taxonómico familia, se tomó como base el modelo CNN propuesto por Fabijanska & Grabowski, 2019 (45), que cuenta con cinco capas de convolución y tres totalmente conectadas (sección 3.4). Este modelo fue considerado como antecedente de este proyecto debido a que los autores lo usaron para clasificar subtipos de virus, quienes usaron genomas completos y obtuvieron sensibilidades y precisiones entre 84% y el 100%, dependiendo del tipo de virus y el número de subtipos.

Para la arquitectura y ajuste de los hiperparámetros de los modelos CNN propuestos en este proyecto se definió el siguiente rango de valores a evaluar.

- Número de filtros: 16, 32, 64, 128, 256, 512, 768, 896, 1024, 1152,1164, 1408, 1920, 2048, 2304, 2560 y 4096.
- Tamaño de filtro: 3, 5, 7, 9,11, 13 y 15.
- Tamaño de lote (*batch)* de 1024 y 2048.

La evaluación de estas combinaciones de arquitecturas e hiperparámetros generaron como resultado diversos modelos de CNN, los cuales presentaron un desempeño que va del 40% hasta 92% de exactitud y de 38% hasta 90% de sensibilidad. Después de hacer un análisis comparativo de los resultados obtenidos, se estableció la cantidad y tamaño de filtros que proporcionaron los mejores desempeños con la menor cantidad de recursos. La arquitectura del modelo CNN propuesto se implementó con las dos bases de datos (BD) (RefSeq y nt), dado que el primer modelo es una CNN que fue entrenado con *k-mers* de 150 pb de la BD RefSeq se define como “CNN_150_RefSeq” y el segundo modelo que es una CNN que fue entrenado con *k-mers* de 150 pb de la BD nt se define como “CNN_150_NT”; se utiliza “NT” en mayúscula para hacer referencia a la BD de nucleótidos, pero para distinguirla de la abreviatura de nucleótidos.

En la Tabla 5.3 se resume la arquitectura del modelo CNN_150 para cada BD; ambos modelos contienen cinco capas de convolución con 768, 1024, 1408, 1920 y 2560 filtros

respectivamente, en el modelo “CNN_150_RefSeq” se utilizó un tamaño de filtro de 9 en todas sus capas de convolución y en el modelo “CNN_150_NT” se utilizó un tamaño de filtro de 9, 9, 11, 13 y 15 para cada una de las capas de convolución, respectivamente. En cada una de las capas de convolución se aplicó una función de activación ReLU, un *stride* igual a 1 y un *padding*; el *padding* fue agregado (mediante *same)* para conservar la dimensión de los datos de entrada, lo cual no estaba considerado en el modelo base. Después de cada capa de convolución, se agregó una capa de agrupamiento máxima (*MaxPooling) para* extraer los valores máximos y reducir a la mitad la dimensión de los datos procedentes de las capas de convolución al aplicar un *stride* igual a 2. Posteriormente, los valores máximos obtenidos fueron normalizados (en las capas de *Batch normalization)*.

Tabla 5. 3. Detalles de las arquitecturas de los modelos CNN_150_RefSeq y CNN_150_NT.

No. de la capa	**Capa**	**Configuración de capa**		**Función de activación**	***padding***	***stride***
		RefSeq	**NT**			
1	*Convolution1D*	768 x (9)	768 x (9)	ReLU	*Same*	1
2	*MaxPooling1D*	-	-	-	-	2
3	*Batch normalization*	-	-	-	-	-
4	*Convolution1D*	1024 x (9)	1024 x (9)	ReLU	*Same*	1
5	*MaxPooling1D*	-	-	-	-	2
6	*Batch normalization*	-	-	-	-	-
7	*Convolution1D*	1408 x (9)	1408 x (11)	ReLU	*Same*	1
8	*MaxPooling1D*	-	-	-	-	2
9	*Batch normalization*	-	-	-	-	-
10	*Convolution1D*	1920 x (9)	1920 x (13)	ReLU	*Same*	1
11	*MaxPooling1D*	-	-	-	-	2
12	*Batch normalization*	-	-	-	-	-
13	*Convolution1D*	2560 x (9)	2560 x (15)	ReLU	*Same*	1
14	*MaxPooling1D*	-	-	-	-	2
15	*Batch normalization*	-	-	-	-	-
16	*Fully connected*	256 neuronas		ReLU	-	-
17	*Dropout*	*r* = 0.4		-	-	-
18	*Batch normalization*	-		-	-	-
19	*Fully connected*	128 neuronas		ReLU	-	-
20	*Dropout*	*r* = 0.4		-	-	-
21	*Batch normalization*	-		-	-	-

No. de la capa	Capa	Configuración de capa		Función de activación	*padding*	*stride*
		RefSeq	**NT**			
22	*Fully connected*	64 neuronas		ReLU	-	-
23	*Dropout*	*r* = 0.4		-	-	-
24	*Batch normalization*	-		-	-	-
25	*Output layer*	número de clases		*Softmax*	-	-

La columna "No. de la capa" indica el número de capas en el modelo. La columna "Capa" indica el nombre de la capa del modelo. La columna "Configuración de capa" contiene la información de las capas, en las capas Convolution1D (capas de convolución) el primer número es el número de filtros y el segundo número (que se encuentra entre paréntesis) es el tamaño del filtro, en las capas Fully connected (capas totalmente conectadas) se tiene el número de neuronas y en las capas Dropout (capas de deserción) el valor r es la tasa de abandono o deserción. La columna "Función de activación" indica la función de activación aplicada en la capa. En la columna "padding" la palabra "same" indica que se está aplicando un rellenado de ceros alrededor de conjunto de entrada. La columna "stride" indica el número de posiciones que se desplazan los filtros sobre los conjuntos de datos de entrada.

Al final de la última capa de convolución, los datos de salida fueron aplanados para formar un vector unidimensional, el cual entra a la primera capa totalmente conectada. Las tres capas totalmente conectadas se constituyeron con 256, 128 y 64 neuronas respectivamente, a las cuales se les aplicó una función de activación ReLU. Después de cada una de las capas totalmente conectadas se incluyó una capa de deserción (*dropout*) con una tasa de 0.4 y una capa de normalización de lote (*Batch normalization*). Cabe recordar que la capa de deserción (*dropout*) apaga aleatoriamente el 40% de las neuronas para evitar el sobre aprendizaje u sobreajuste (*overfitting)*. Por último, los datos pasaron por una capa de salida con el mismo número de neuronas o nodos que el número de clases o familias a clasificar. Dado que es una salida multiclase, se aplicó una función *softmax*. El aprendizaje se realizó usando la función de optimización Adam y una función *categorical cross-entropy* para la pérdida.

En la Figura 5.9, se presenta de manera visual el modelo de CNN_150. Esta CNN cuenta con un nodo por cada clase, donde cada clase es una familia viral, para el modelo CNN_150_RefSeq la nC es igual a 128 clases y para el modelo CNN_150_NT la nC es igual a 170 clases.

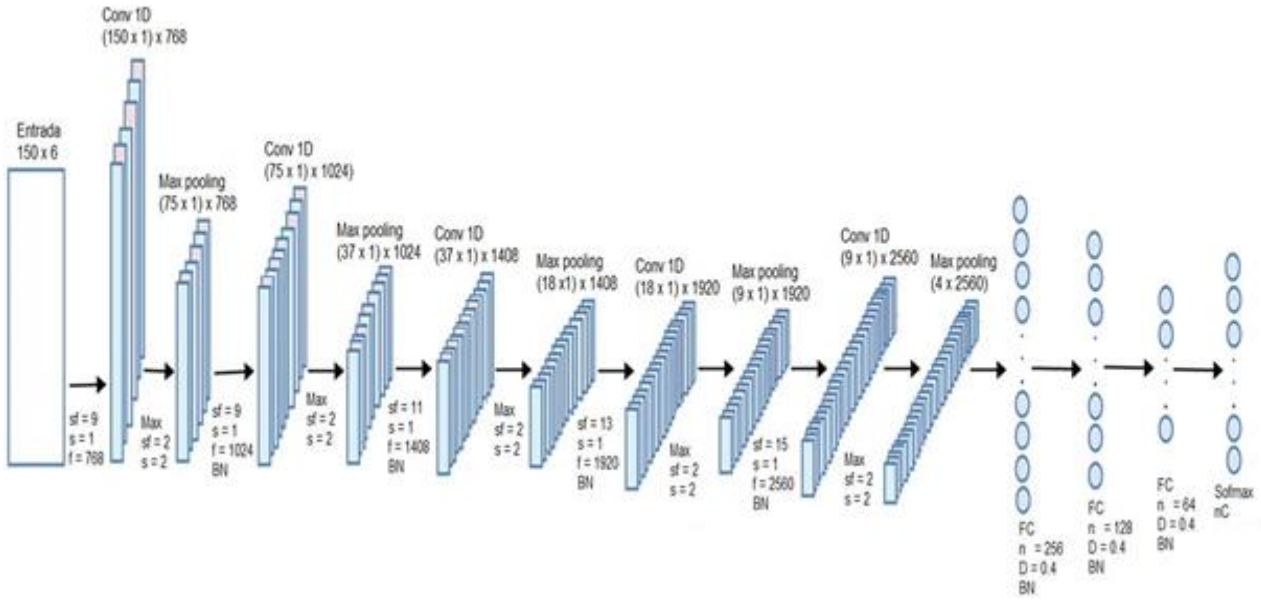


Figura 5.10. Modelo CNN_150_NT para la clasificación a nivel familia (modelo aplicado a las bases de datos nt). Las capas de convolución están definidas con Conv1D, número y tamaño de los filtros como f y sf respectivamente. La capa de reducción como Max pooling, la de normalización como BN, las totalmente conectadas como FC, la capa de deserción como D con la tasa de abandono que utilizó. Finalmente, la capa de salida está representada como nC.

5.4. Evaluación del modelo

Para evaluar los modelos CNN_150_RefSeq y CNN_150_NT, se realizó un post-procesamiento con las lecturas predichas. Primero, con las predicciones correctas e incorrectas del conjunto de datos se establecieron umbrales de probabilidad para cada modelo. Posteriormente, ambos modelos con sus umbrales fueron aplicados a los ocho conjuntos de datos simulados y a los cuatro reales. Las lecturas que pasaron el umbral son las lecturas que fueron consideradas como virales y las cuales fueron utilizadas para calcular las métricas de sensibilidad o *recall* (ec. 4.1) precisión (ec. 4.2), precisión equilibrada (ec. 4.3), puntuación F1 (ec. 4.4), MCC (ec. 4.5) y especificidad (ec. 4.6), descritas en forma detallada en la sección 4.3 y que fueron utilizadas para medir el desempeño de los modelos.

5.5. Aplicación de los modelos

A continuación, se describe el proceso de la predicción de las lecturas con los modelos de CNN (ver Figura 5.10).

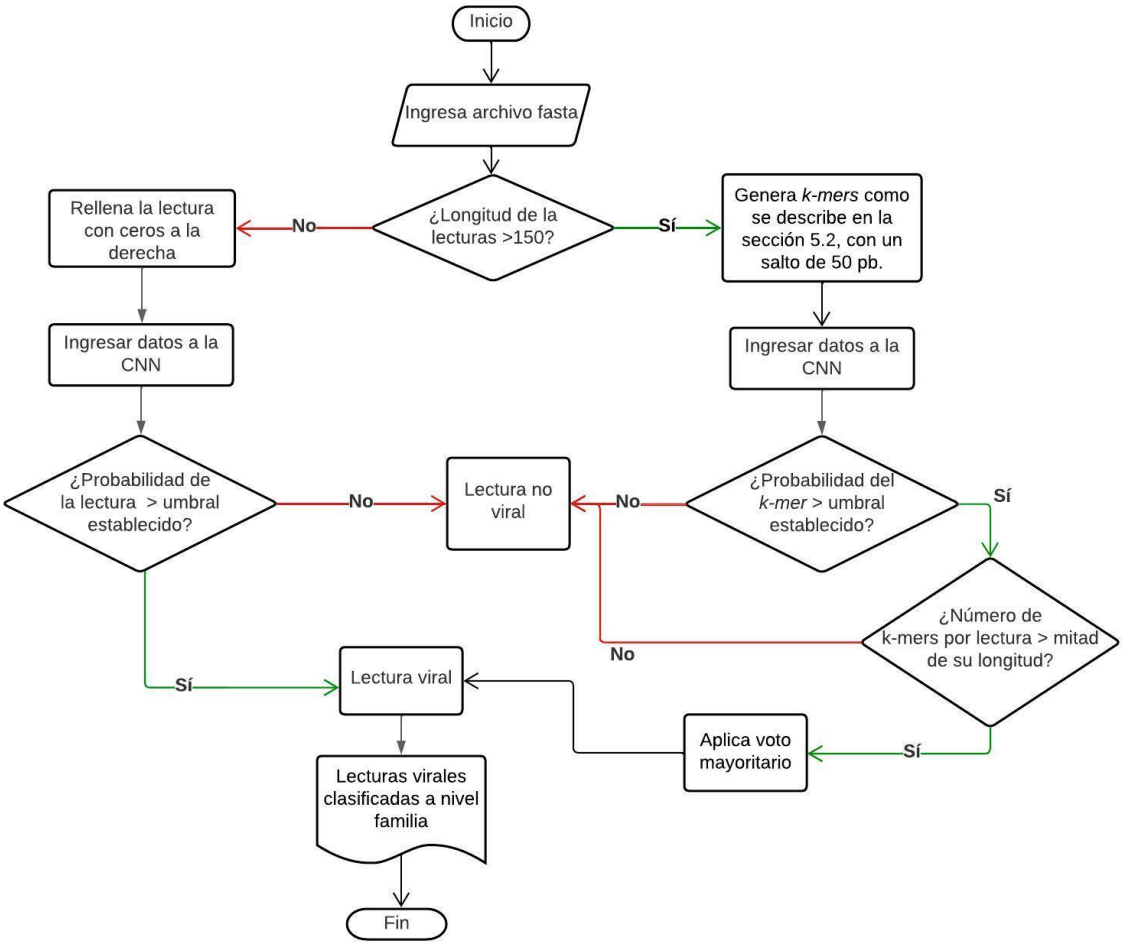


Figura 5.11. Proceso de la aplicación de los modelos para la predicción de lecturas o contigs.

CAPÍTULO 6. RESULTADOS Y DISCUSIONES

En este capítulo se presentan los resultados obtenidos de la reducción, preparación de los datos y de la evaluación de los modelos desarrollados en este trabajo de investigación para la clasificación a nivel familia de lecturas virales, en diferentes conjuntos de datos.

6.1. Evaluación de los modelos para la clasificación de lecturas metagenómicas a nivel familia

Los modelos de redes neuronales de convolución (CNN) desarrollados para predecir a qué familia taxonómica viral se clasifican las lecturas de DNA procedentes de estudio metagenómicos virales emplearon como referencia la información de dos bases de datos (BD) (RefSeq y nt); cada modelo utilizó una BD distinto en su entrenamiento. El modelo que empleó la BD RefSeq viral fue definido como CNN_150_RefSeq, mientras que el modelo que usó la BD viral de nucleótidos (nt) de Genbank fue definido como CNN_150_NT y están explicados en detalle en la sección 5.3 del capítulo 5.

Los dos modelos CNN (CNN_150_RefSeq y CNN_150_NT) fueron entrenados usando sus respectivos conjuntos de entrenamiento y validación. Los resultados obtenidos por el modelo CNN_150_RefSeq se muestran en la Figura 6.1, mientras los del modelo CNN_150_NT en la Figura 6.2. Ambos gráficos muestras las métricas de pérdida y exactitud obtenidas por los modelos en cada época en el conjunto de entrenamiento y validación. Como se puede ver en la Figura 6.1, en el modelo CNN_150_RefSeq, la pérdida del entrenamiento (*Training Loss)* inicio en 1.42 y conforme avanzó el entrenamiento, terminó en 0.39, mientras que la pérdida de la validación (*Validation Loss)* inició en 1.04 y finalizó en 0.42. Estos resultados mostraron que durante la época 5 del entrenamiento se presentó un pequeño sobreajuste (*overfitting)* en el modelo, existiendo una diferencia pequeña entre la pérdida del conjunto de entrenamiento y el de validación. Este modelo se entrenó durante 14 épocas, pero se presentó un sobreajuste, por lo tanto, para el modelo final de la CNN_150_RefSeq se utilizó el modelo entrenado hasta la época 11. Por otra parte, la exactitud del entrenamiento (*Training Accuracy)* comenzó en el 0.66 y la exactitud de la validación (*Validation Accuraccy)* en 0.75, pero ambos finalizaron en 0.90 aproximadamente. Asimismo, en el modelo CNN_150_NT (Figura 6.2), la pérdida del entrenamiento (*Training Loss)* empezó 1.81 y terminó en 0.61 y la pérdida de la validación (*Validation Loss)* comenzó en 1.39 y finalizó en 0.60, no presentándose un sobreajuste (*overfitting)*. Por otra parte, la exactitud del entrenamiento (*Training Accuracy)* empezó en 0.55, la exactitud de la validación (*Validation Accuraccy)* en 0.64 y ambas finalizaron en 0.83.

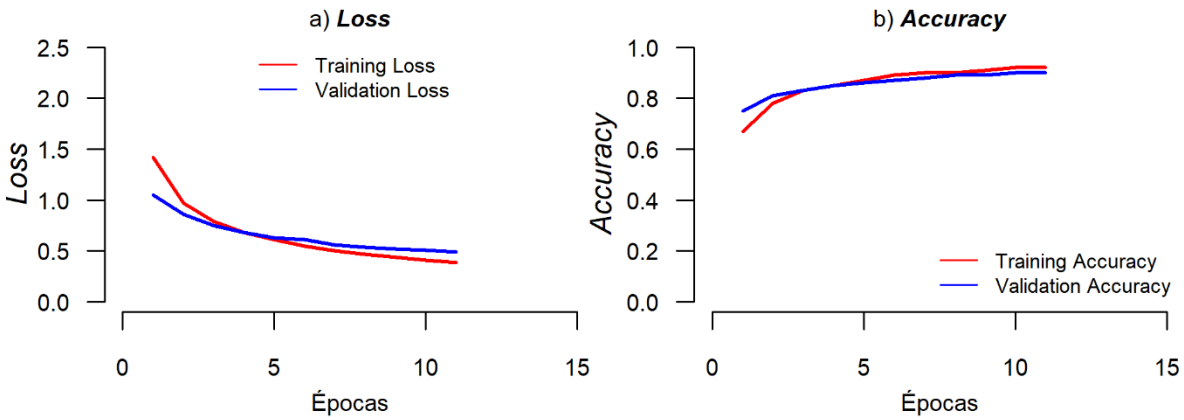


Figura 6.1. Resultados del entrenamiento del modelo CNN_150_RefSeq. a) Pérdida (Loss). La pérdida del conjunto de entrenamiento es señalada como "Training Loss" (rojo), mientras que la pérdida del conjunto de validación se encuentra como "Validation Loss" (azul); b) Exactitud (Accuracy). La exactitud que alcanza el conjunto de entrenamiento y validación están marcados como "Trainning Accuracy" (rojo) y "Validation Accuraccy" (azul), respectivamente.

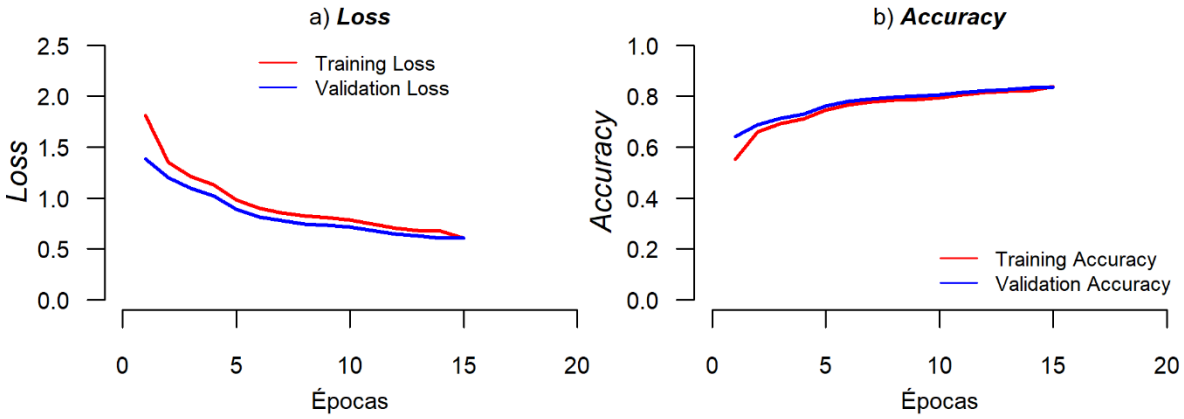


Figura 6.2. Resultados del entrenamiento del modelo NT. Gráfico a) Pérdida (Loss). La pérdida del conjunto de entrenamiento es señalada como "Training Loss" (rojo), mientras que la pérdida del conjunto de validación se encuentra como "Validation Loss" (azul). Gráfico b) Exactitud (Accuracy). La exactitud que alcanza el conjunto de entrenamiento y validación están marcados como "Trainning Accuracy" (rojo) y "Validation Accuraccy" (azul), respectivamente.

Adicionalmente, para evaluar el desempeño de ambos modelos CNN (CNN_150_RefSeq y CNN_150_NT) se utilizaron sus respectivos conjuntos de prueba, porque no fueron utilizados para el entrenamiento. Para ello, se calcularon las métricas mencionadas en el capítulo 4 en la sección 4.3. En la Figura 6.3 se presentan los gráficos de sensibilidad (Figura 6.3-a), precisión (Figura 6.3-b), precisión equilibrada (Figura 6.3-c), puntuación F1 (Figura 6.3-d) y MCC (Figura 6.3-e). Cada gráfico representa los resultados de la evaluación de las diferentes métricas usadas, indican si los modelos son capaces de clasificar una gran cantidad de lecturas virales y si las clasifican correctamente a la familia a la que pertenece. En este

sentido, los modelos obtuvieron valores por arriba del 81% para la red CNN_150_NT y 90% para la red CNN_150_RefSeq. Asimismo, el MCC (Figura 6.3-e), la cual es una métrica para evaluar el desempeño con clases desbalanceadas, muestra que la red CNN_150_RefSeq obtuvo un valor de 0.9 y la CNN_150_NT uno de 0.8.

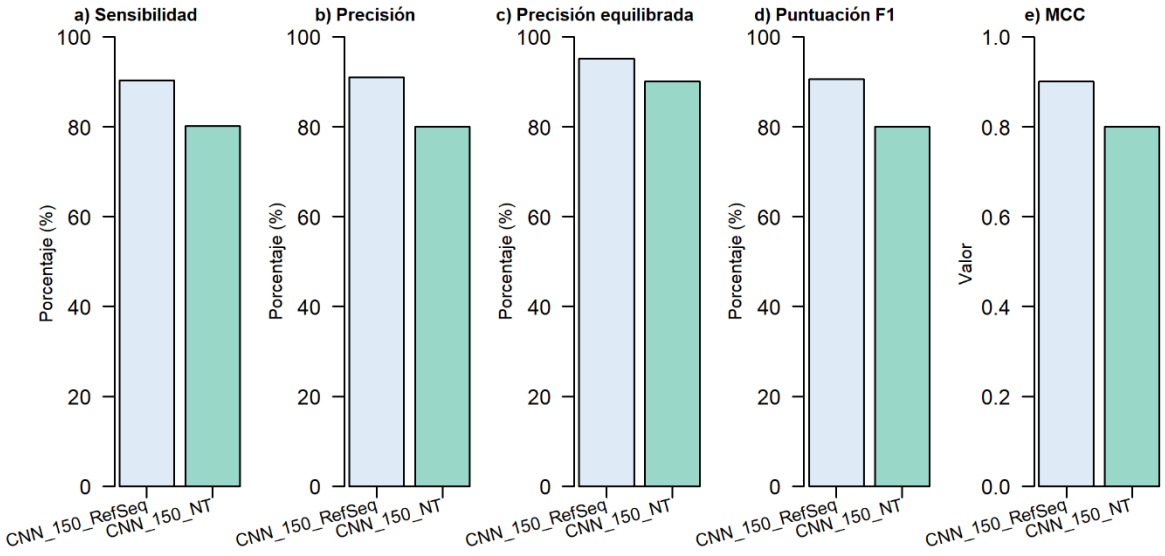


Figura 6.3. Resultados de los modelos CNN_150_RefSeq y CNN_150_NT en el conjunto de prueba. a) Sensibilidad, b) Precisión, c) Precisión equilibrada, d) Puntuación F1 y e) MCC.

6.2. Análisis de los resultados obtenidos

Se realizó un análisis de los resultados obtenidos en los conjuntos de prueba de cada modelo. En la Figura 6.4 se presentan las sensibilidades obtenidas en cada una de las 127 clases (familias) usando el modelo CNN_150_RefSeq. En este modelo, en 51 clases (familias) no se logró clasificar ninguna lectura correctamente. Estas clases representan el 1.4% de las lecturas totales del conjunto de prueba. Por otro lado, se encontró que las clases *Myoviridae* y *Siphoviridae* fueron las familias mejor identificadas por el modelo (Figura 6.4-b), las cuales; representan el 24% y 27% del total del conjunto de prueba. Las clases que fueron moderadamente identificadas por el modelo (ver Figura 6.4-a) son: i) *Autographiviridae*, *Herelleviridae, Podoviridae y Unclassified* recuperaron 3% - 5%; ii) *Ackermannviridae*, *Baculoviridae*, *Demerecviridae y Poxviridae* recuperaron el 2%; iii) *Mimiviridae*, *Schitoviridae*, *Drexlerviridae* y *Phycodnaviridae* recuperaron el 1%. Mientras 62 clases tuvieron una baja identificación (entre 1% - 0%); la suma de las 62 clases representa el 9% del total del conjunto de prueba.

En la Figura 6.5 se presentan las sensibilidades del modelo CNN_150_NT en cada una de las 170 clases (familias). Para este modelo, se identificó que en 39 clases el modelo no logró clasificar ninguna lectura correctamente. Estas clases representan el 0.5% de las lecturas

totales del conjunto de prueba. Por otro lado, se encontró que las clases *Myoviridae* y *Herpesviridae* fueron las familias mejor identificadas por el modelo (ver Figura 6.4-b y 6.4-a respectivamente). Estas familias representan el 12% y 13% del total del conjunto de prueba. Las clases que fueron identificadas por el modelo en porcentajes moderados (ver Figura 6.4-a) fueron las siguientes: i) *Poxviridae, Siphoviridae y Unclassified* recuperaron el 5% - 9%; ii) *Herelleviridae*, *Mimiviridae* y *Retroviridae* recuperaron el 3%; iii) *Ackermannviridae, Asfarviridae, Baculoviridae, Flaviviridae y Phycodnaviridae* recuperaron el 2%; iv) *Demerecviridae* e *Iridoviridae* recuperaron el 1%. Finalmente, 114 clases presentaron una baja identificación (entre 1% - 0%) por parte del modelo CNN_150_NT. Estas 114 clases representan el 15% del total del conjunto de prueba.

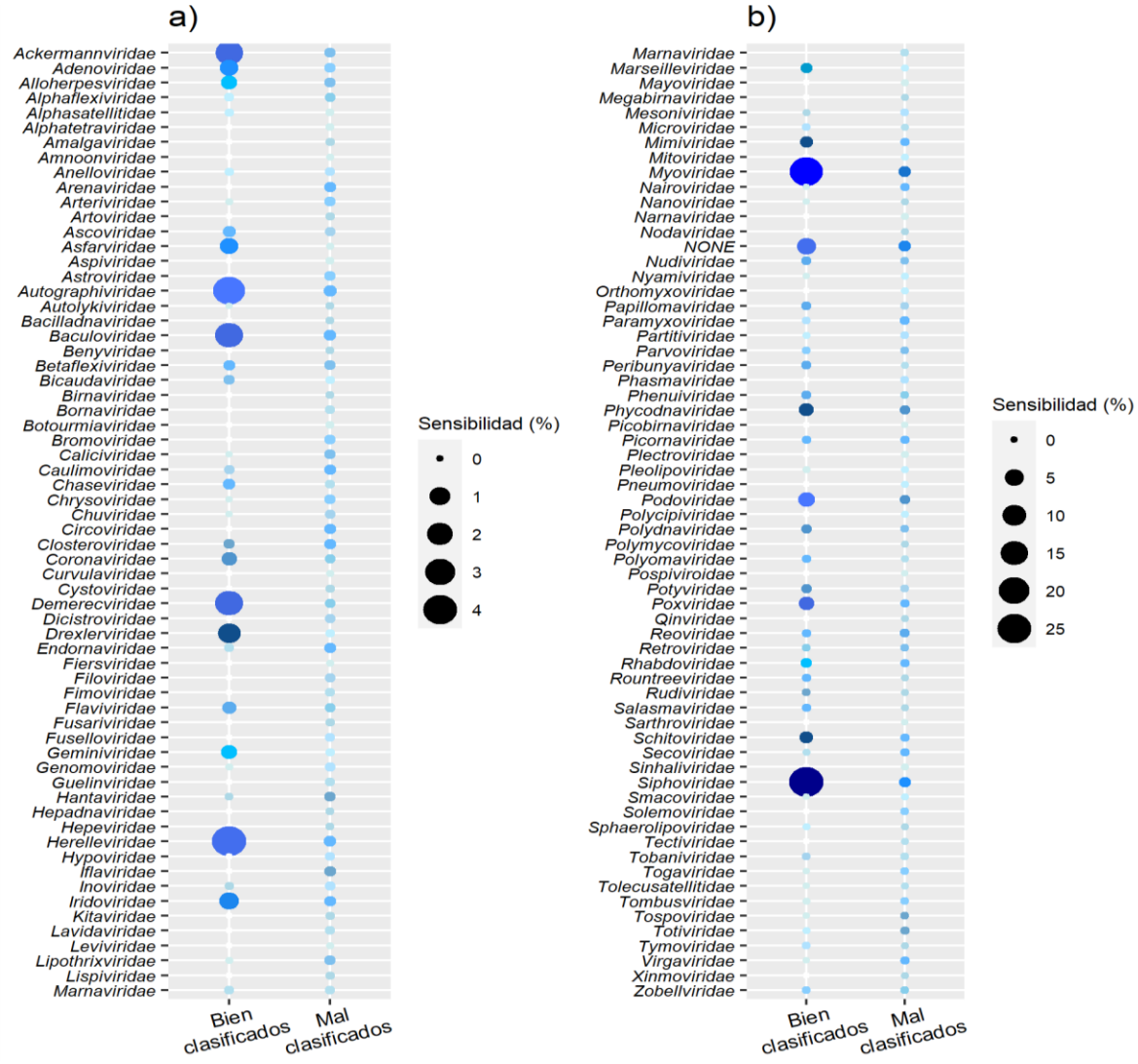


Figura 6.4. Las clases del modelo CNN_150_RefSeq bien y mal clasificadas. En el panel izquierdo se tiene el nombre de cada clase (familia) y en el panel inferior se tiene sí estuvieron bien o mal clasificadas. En el gráfico a) se tiene las primeras 64 clases del conjunto y en el gráfico b) las otras 63 clases. En ambos gráficos el tamaño de la burbuja indica el porcentaje de sensibilidad.

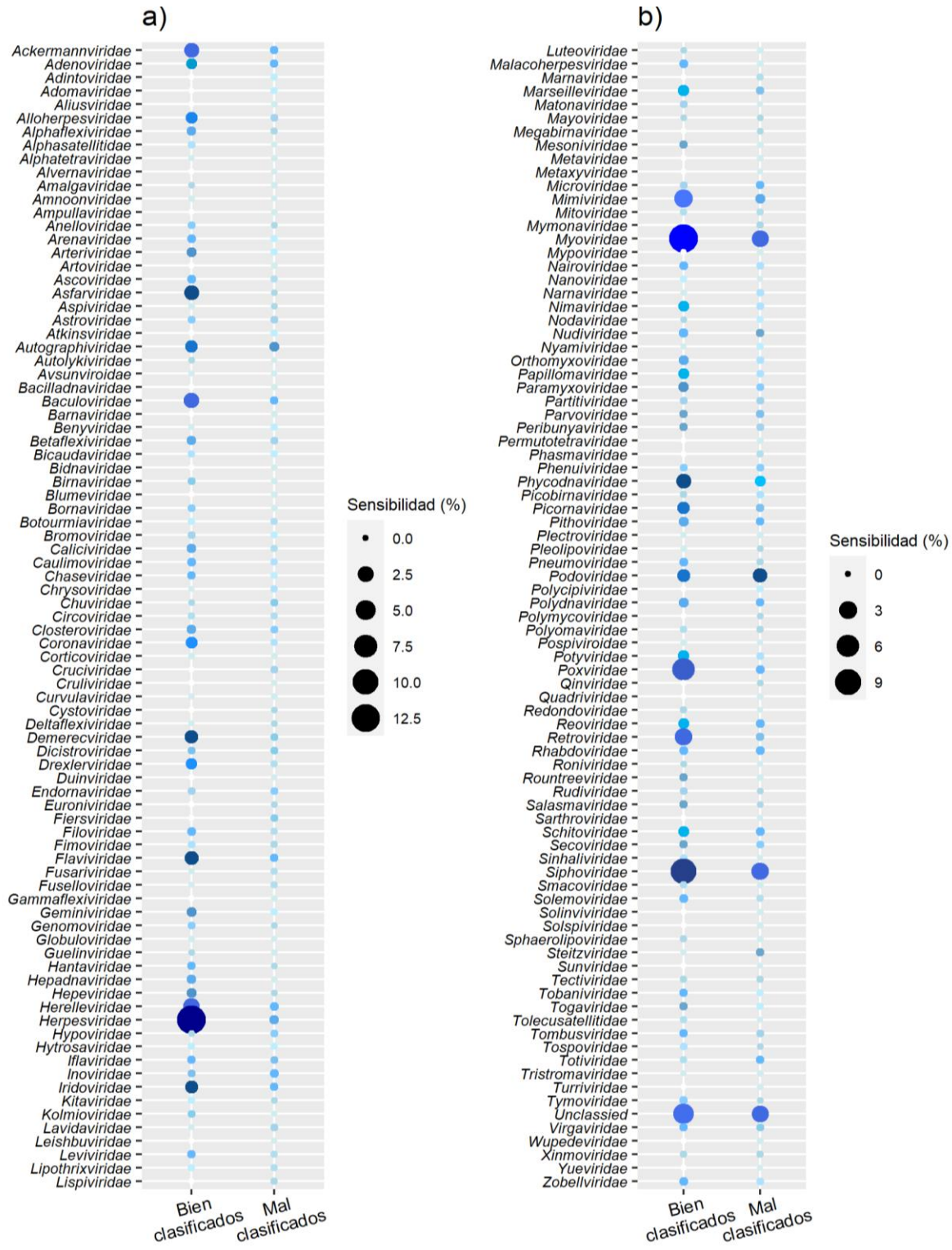


Figura 6.5. Las clases del modelo CNN_150_NT bien y mal clasificado. En el panel izquierdo se tiene el nombre de cada clase (familia) y en el panel inferior sí estuvieron bien o mal clasificadas. En el gráfico a) se tiene las primeras 64 clases del conjunto y en el gráfico b) están las otras 63 clases. En ambos gráficos el tamaño de la burbuja indica el porcentaje de sensibilidad.

A pesar de que ambas BD se sometieron a una limpieza para quitar las familias menos anotadas y que en el caso de la BD de nt viral las clases mayoritarias se redujeron para quitar la mayor cantidad de información redundante, los resultados sugieren que en ambos modelos la mayoría de las clases minoritarias son las clases que no recuperaron correctamente ninguna lectura. En otras palabras, las clases mayoritarias al ser tan grandes han provocado que los modelos aprendan más sobre ellas que del resto de las clases. Además, en ambos modelos las clases minoritarias tienden a ser clasificadas por los modelos a la clase *Unclassifed* (ver Anexo 1 y Anexo 2). Esto se debe a que en esta clase existe mucha información tanto de virus eucariontes como de bacteriófagos. En la BD de nt viral, la clase *Unclassifed* cuenta con 45,076,444 lecturas, mientras que las clases minoritarias tiene entre 3,188 a 411,040 lecturas.

6.3. Post-procesamiento

Se realizó un post-procesamiento de las predicciones porque cualquier lectura predicha por los modelos CNN_150_RefSeq y CNN_150_NT es clasificada a alguna de las familias virales, debido a que los modelos fueron creados a partir de la información de base de datos (BD) de referencias virales y no cuentan con una clase negativa. En otras palabras, no se cuenta con información no viral como humano, bacterias, hongos, etc. Dicho post-procesamiento se basa en establecer un umbral considerando las probabilidades obtenidas de la función de activación softmax que se aplicó en las capas de salida de los modelos. Este umbral permite establecer si las lecturas son virales o no virales y reducir la cantidad de falsos positivos.

Para establecer el umbral, primeramente, se predijeron los *k-mer* de toda la BD. Después, se dividieron los *k-mer* correcta e incorrectamente clasificados por los modelos. Posteriormente, se calculó y analizó la frecuencia acumulada de las predicciones obtenida de los rangos de probabilidades de los *k-mers*, los cuales van en intervalos de 0.05, para establecer el valor del umbral. La Figura 6.6 muestra que en el modelo CNN_150_RefSeq los *k-mers* clasificados incorrectamente corresponden al 7% del conjunto total y generalmente obtuvieron probabilidades bajas; mientras que el 84% de los *k-mers* correctamente clasificados alcanzaron probabilidades altas (>0.90). Esto permitió establecer y determinar que el umbral de corte para este modelo fuera del 0.95. Al aplicar este umbral en cualquier conjunto de datos se logró descartar lecturas no virales. Por ejemplo, lecturas de bacterias y humano que están en las muestras, y se disminuyó la predicción de falsos positivos en las familias virales. Por lo tanto, si una lectura predicha obtuvo una probabilidad igual o mayor al umbral, dicha lectura fue considera como una viral y asignada a la familia predicha; de lo

contrario se consideró que la lectura no pertenece a la clase o es una lectura no viral, en consecuencia, debe ser descartada.

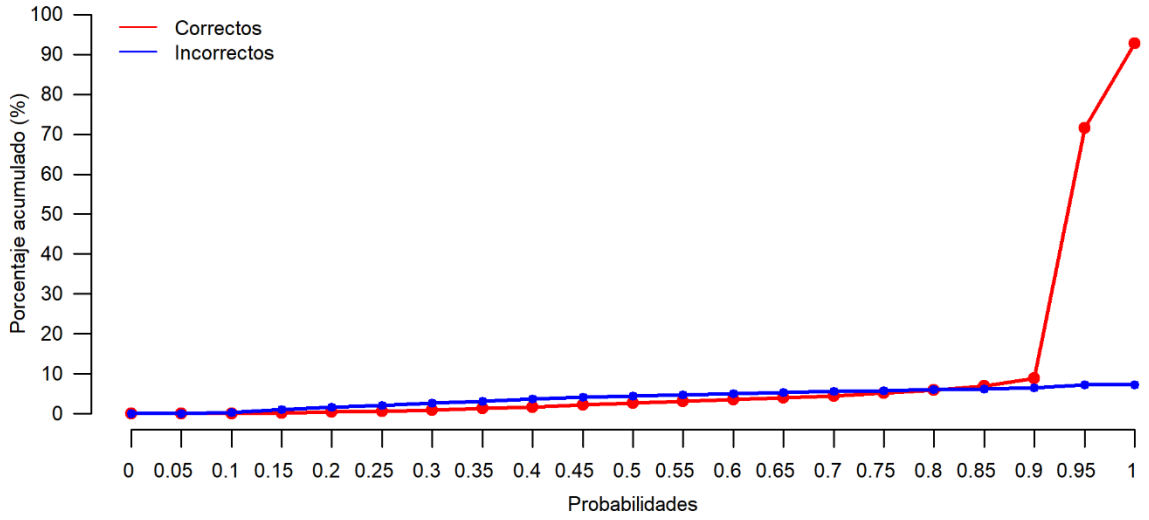


Figura 6.6. Probabilidades de predicciones correctas e incorrectas del conjunto de datos creado con la base de datos de RefSeq.

En el caso del modelo entrenado con la BD nt, obtuvo un 85% y 15% de los *k-mers* bien y mal clasificados, respectivamente. El 69% de los *k-mers* correctamente clasificados tuvieron una probabilidad mayor al 0.90 y el 1% fueron clasificados incorrectamente. Para tratar de recuperar la mayor cantidad de lecturas correctas con el mínimo porcentaje de falsos positivos, se estableció un umbral de 0.6 para las familias que infectan virus eucariontes y un umbral de 0.8 para las familias de bacteriófagos y para la clase de *Unclassified*. En los conjuntos reales se estableció los mismos umbrales que en los conjuntos simulados, con el objetivo de minimizar las predicciones de lecturas falsas positivas sin perder el porcentaje de lecturas correctamente clasificadas (ver Figura 6.7); pero las lecturas predichas a la clase *Unclassified* se descartaron (se consideran falsas positivas).

Finalmente, las lecturas clasificadas como virales fueron utilizadas para evaluar el desempeño de los modelos usando las métricas de sensibilidad o recall (ec. 4.1), precisión (ec. 4.2) y MCC (ec. 4.3). Los resultados obtenidos por los modelos fueron comparados con los resultados de las herramientas de clasificación taxonómica descritos en el Capítulo 4 y los resultados se presentan en la sección 6.4 en este capítulo.

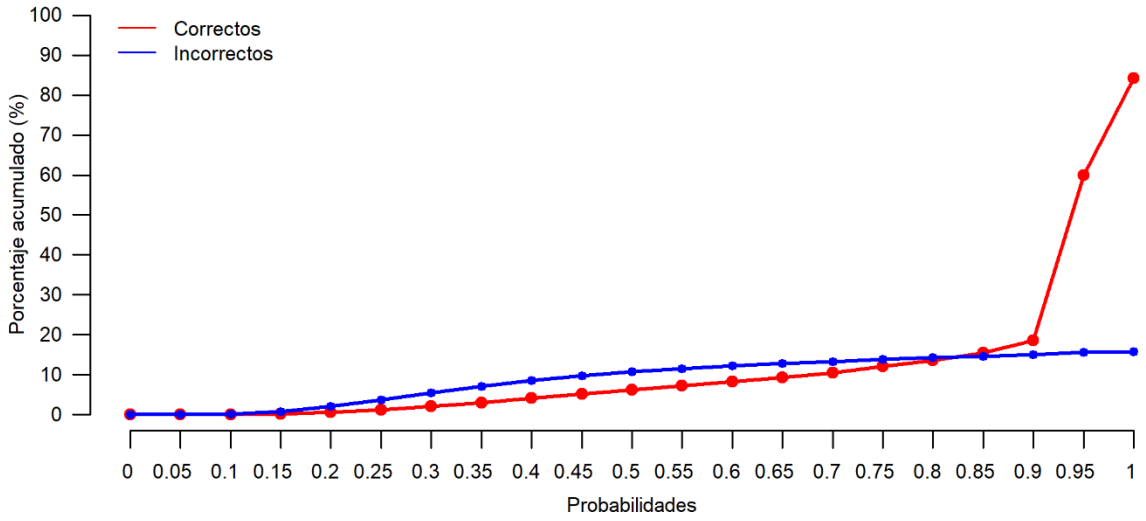


Figura 6.7. Probabilidades de las predicciones correctas e incorrectas del conjunto de datos creado con la base de datos de nt.

6.4. Comparación de los modelos con otras herramientas

Los modelos CNN_150_RefSeq y CNN_150_NT fueron evaluados utilizando ocho conjuntos de datos simulados y 4 cuatro conjuntos reales, los cuales fueron descritos en el capítulo 4 sección 4.1. Además, se aplicó el umbral establecido de 0.95 para el modelo 150_RefSeq y para el CNN_150_NT se aplicaron dos umbrales, uno de 0.6 y otro de 0.8. Solo las lecturas que pasaron los umbrales fueron empleadas para el cálculo de las métricas de evaluación.

6.4.1. Conjuntos simulados virales

6.4.1.1. Lecturas largas

Los modelos CNN_150_RefSeq y CNN_150_NT fueron diseñados para clasificar lecturas de 150 pb de longitud, mientras los conjuntos de datos que simularon lecturas largas son de una longitud promedio de 450 pb (ver sección 4.1.1.1 para mayores detalles). Por lo tanto, estas lecturas fueron preprocesadas como se indica en el capítulo 5 sección 5.2, pero con un salto de 50 pb, medida elegida después de diversas pruebas con varios tamaños porque no afecta los resultados de predicción. Este preprocesamiento generó como resultado aproximadamente 6 sub-lecturas o *k-mers* por cada lectura simulada de los conjuntos. Posteriormente, cada sub-lectura fue clasificada por la CNN. Para definir la familia o clase de

cada una, se aplicó el umbral establecido de acuerdo con cada modelo. A las sub-lecturas que pasaron el umbral y que representaron más de la mitad de la lectura original, se les aplicó la técnica del voto mayoritario. Esta técnica consiste principalmente en que si la mayoría de las sub-lecturas de una lectura original son clasificadas a una familia o clase, la lectura original es asignada a esa clase; en otras palabras, si la mayoría de los *k-mers* de una lectura fueron asignadas a una familia específica toda la lectura es asignada a esa familia. Por ejemplo, si la lectura 1 está fragmentada en 3 *k-mers* o sub-lecturas y el modelo CNN predice que 2 *k-mers* pertenecen a la familia Circoviridae y un *k-mer* pertenece a la familia Adenoviridae, entonces la lectura 1 es clasificada a la familia Circoviridae. Después de utilizar del voto mayoritario para establecer a que familia pertenece cada lectura simulada, se calculó la sensibilidad, precisión, precisión equilibrada, puntuación F1 y el MCC de cada conjunto de datos.

En la Figura 6.8 se presentan las sensibilidades, precisiones, precisiones equilibradas, puntuación F1 y MCC obtenidas en la clasificación de las lecturas virales simuladas largas. Asimismo, los modelos CNN son comparados con los resultados obtenidos por otras herramientas de clasificación taxonómica viral descritos en detalle en el Capítulo 4 y sección 4.4.1.1.

Los modelos CNN_150_RefSeq y CNN_150_NT alcanzaron sensibilidades del 55% y 72% respectivamente en el conjunto *50G*, en el conjunto *500G* las sensibilidades fueron de del 64% y 67% respectivamente, y en el de 1000G fueron del 58% y 63% respetivamente. Ambos modelos superaron en sensibilidad a ocho herramientas, excepto a BLAST, FastViromeExplorer, Kraken2 y VirusFinder que alcanzaron más del 95% de sensibilidad en los tres conjuntos. Además, de manera general se observó que ambos modelos disminuyeron su sensibilidad cuando aumentó la riqueza del viroma en los conjuntos, al igual que las herramientas BLAST, FastViromeExplorer, Kraken2, Taxonomer, Vipie y VirusFinder.

En relación con la precisión, ambos modelos obtuvieron una alta precisión, mayor al 96%, al igual que las otras herramientas. Las precisiones del modelo CNN_150_RefSeq en los conjuntos *500G* y *1000G* fueron ligeramente más bajas en comparación con las del modelo CNN_150_NT; ambos modelos superaron a las herramientas Centrifuge, Kraken2, One Codex y Vipie.

En los tres conjuntos de datos (*50G*, *500G* y *100G*), el modelo CNN_150_RefSeq obtuvo una precisión equilibrada del 78% al 82%, mientras el modelo CNN_150_NT alcanzó el 82% al 86%; ambos modelos fueron superados por las herramientas BLAST, FastViromeExplorer, Kraken2 y VirusFinder con valor mayor al 91%. El resto de las herramientas obtuvieron una precisión equilibrada que va del 52% al 71.

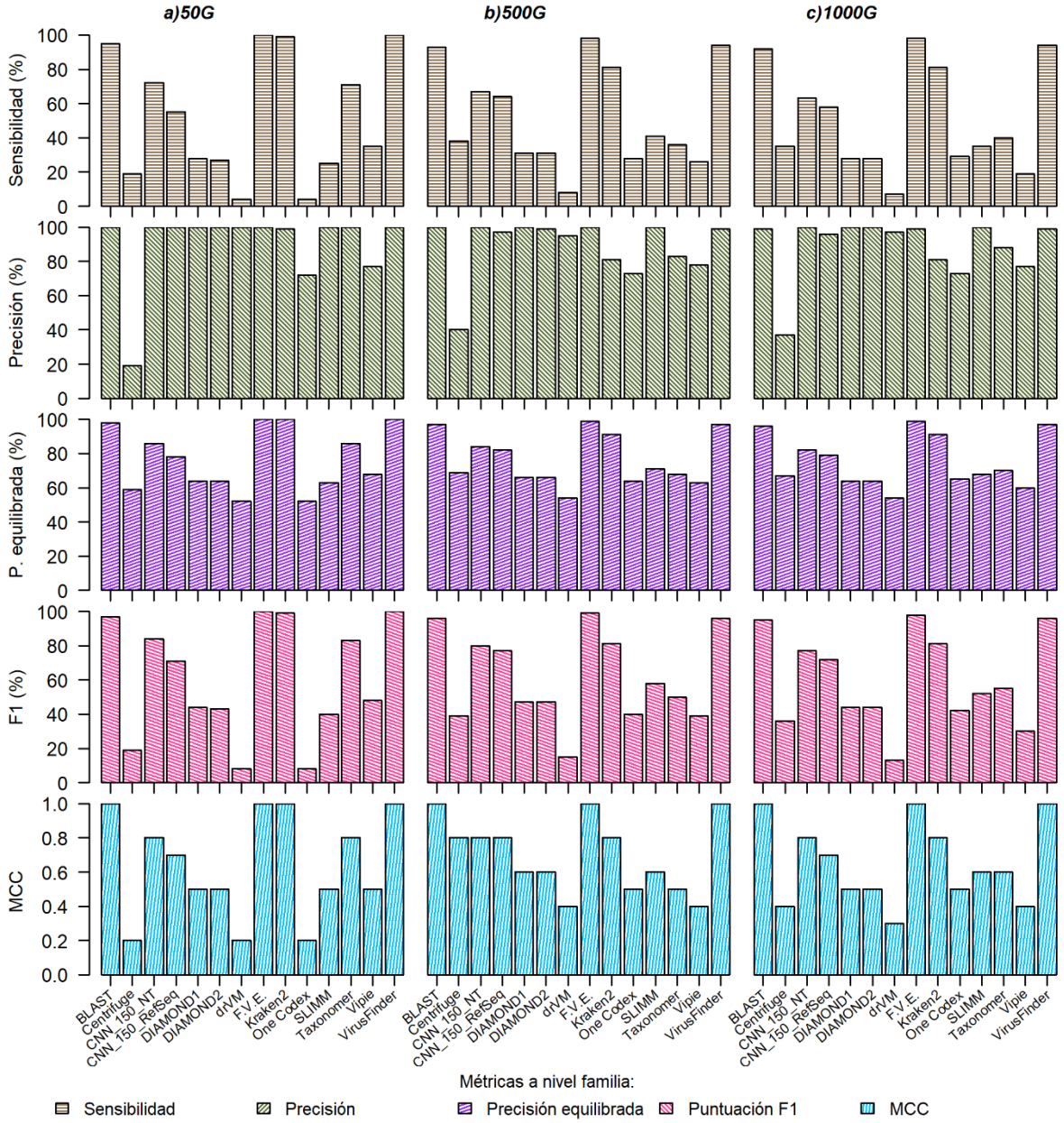


Figura 6.8. Resultados obtenidos de la clasificación de las redes CNN_150_NT y CNN_150_RefSeq en los conjuntos simulados de lecturas largas (454). Se realiza una comparación de los resultados con otras herramientas de clasificación metagenómica. a) 50G; b) 500G; c)1000G. La herramienta FastViromeExplorer está representada como F. V. E., DIAMOND1 es la herramienta DIAMOND empleando la base de datos completa y DIAMOND2 la de RefSeq.

Los modelos CNN_150_RefSeq y CNN_150_NT presentaron una puntuación F1 respectiva de 71% y 84% en el conjunto 50G, de 77% y 80% en el conjunto *500G*, y de 72% y 77% en el de *1000G*. BLAST, FastViromeExplorer y VirusFinder (en los tres conjuntos de datos) y Kraken2 (en el conjunto *50G*) obtuvieron una alta puntuación F1, la cual es mayor al 95%. Por otro lado, Kraken2 (en los conjuntos *500G* y *1000G*) y Taxonomer (en el conjunto

50G) presentaron un valor F1 de 81% y 83%, respectivamente. El resto de las herramientas alcanzaron en los tres conjuntos una puntuación F1 de 8% al 58%.

Finalmente, ambos modelos en los tres conjuntos de datos tuvieron un valor de MCC entre el 0.7 y 0.8, siendo superados únicamente por BLAST, FastViromeExplorer y VirusFinder.

Es importante señalar que los *k-mers* que fueron descartados por estar debajo del umbral establecido, para los modelos CNN_150_RefSeq y CNN_150_NT, corresponden al 17% y al 3% respectivamente de las lecturas totales del conjunto de datos *50G*, para *500G* fueron el 19% y 16% respectivamente y para el conjunto *1000G* fueron del 18% y 14%, respectivamente. Las lecturas que fueron excluidas porque representaron a los *k-mers* que pasaron el umbral, pero que fueron menos de la mitad de la lectura original, en el conjunto *50G* corresponden al 28% en la CNN_150_RefSeq y el 25% en la CNN_150_NT, en el conjunto *500G* fueron del 15% y 17% respectivamente, y en el conjunto *1000G* corresponden al 21% (CNN_150_RefSeq) y 22% (CNN_150_NT). Finalmente, en ambos modelos las lecturas mal clasificadas fueron predichas a las clases *Myoviridae*, *Siphoviridae* y *Unclassified*, las cuales son las clases mayoritarias.

6.4.1.2. Lecturas cortas

Estos conjuntos simulan lecturas cortas de la tecnología Illumina con longitudes de 150 pb (para más detalles ver sección 4.1.1.2). En estos conjuntos, los modelos CNN_150_NT y CNN_150_RefSeq realizaron la clasificación de cada lectura simulada a una familia viral.

Las sensibilidades, precisiones, precisiones equilibradas, puntuación F1 y MCC de los conjuntos simulados de lecturas cortas se presentan en la Figura 6.9. En el conjunto *Eukaryotic,* los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron sensibilidades del 21% y 61%, respectivamente. Estos modelos (CNN_150_RefSeq y CNN_150_NT) fueron superados por las herramientas BLAST, Centrifuge, DIAMOND2, drVM, Kraken2, One Codex y Taxonomer, cuyas sensibilidades van del 68% al 81%, pero fueron más sensibles que las herramientas FastViromeExplorer, SLIMM, Vipie y VirusFinder. En el conjunto *Prokaryotic,* los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron sensibilidades del 76% y 69%, respectivamente; en este conjunto ambos modelos fueron superados por BLAST, SLIMM y DIAMNOND (81%-75%), mientras que el resto de las herramientas obtuvieron menos del 75% de sensibilidad. En el conjunto *Unclassified,* los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron sensibilidades del 6% y 5%, respectivamente, siendo superados por todas las herramientas, excepto por Taxonomer y Vipie (que obtuvieron valores nulos).

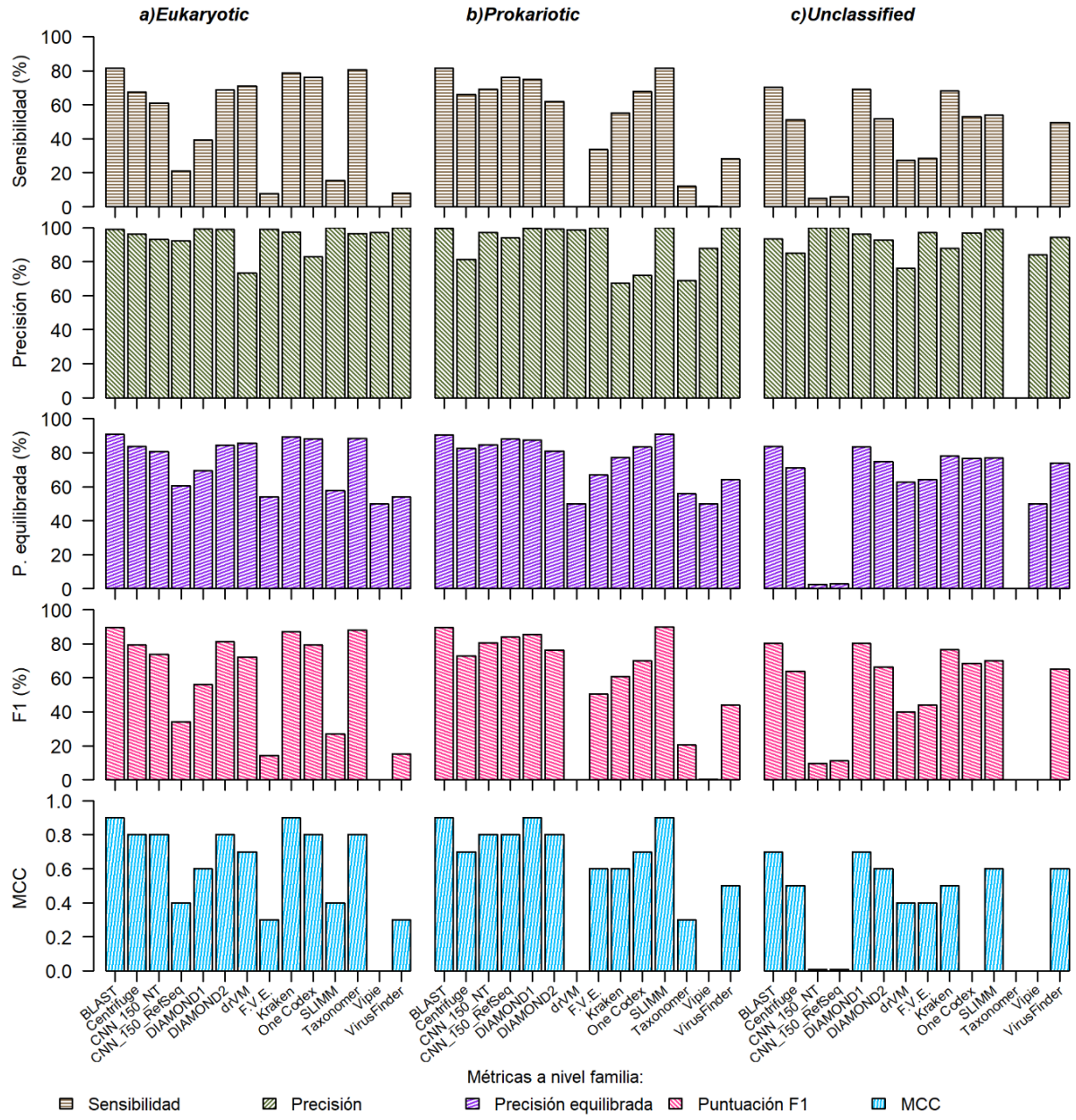


Figura 6.9. Resultados obtenidos de la clasificación de las redes CNN_150_NT y CNN_150_RefSeq de los conjuntos simulados de lecturas cortas (Illumina). Se realiza una comparación de los resultados con otras herramientas de clasificación. a) Eukaryotic; b) Prokaryotic; c) Unclassified. La herramienta FastViromeExplorer está representada como F. V. E., DIAMOND1 es la herramienta DIAMOND empleando la base de datos completa y DIAMOND2 la de RefSeq.

En el conjunto *Eukaryotic*, los modelos CNN_150_RefSeq y CNN_150_NT consiguieron precisiones de 92% y 93%, respectivamente, aunque fueron superados por el resto de las herramientas con precisiones superiores al 96%. Por otra parte, en el conjunto *Prokaryotic*, ambos modelos CNN_150_RefSeq y CNN_150_NT consiguieron precisiones de 94% y 97%, respectivamente, y fueron superados por BLAST, DIAMOND1, DIAMOND2, drVM,

FastViromeExplorer y VirusFinder que alcanzaron más del 99%. En el conjunto *Unclassified*, ambos modelos alcanzaron el 100%, mientras que el resto de las herramientas obtuvieron valores del 99%-76%, a excepción de Taxonomer que nuevamente obtuvo una precisión nula.

Con lo que respecta a la precisión equilibrada, en el conjunto *Eukaryotic* los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron un 61% y 81%, siendo superados por BLAST, Centrifuge, DIAMOND2, drVM, Kraken2, One Codex y Taxonomer que obtuvieron de 84% a 91%. En tanto que SLIMM, FastViromeExplorer, Vipie y VirusFinder obtuvieron una precisión equilibrada de 50% a 58%. En el conjunto *Prokaryotic*, los modelos CNN_150_RefSeq y CNN_150_NT consiguieron el 88% y el 85% de precisión equilibrada. BLAST y SLIMM obtuvieron un 91% superando a todas las herramientas, por otro lado, DIAMON1 alcanzó un 87% superando a la CNN_150_NT, y el resto de las herramientas obtuvieron de 50% al 83%. En el conjunto *Unclassified,* ambos modelos alcanzaron el 3%, siendo superados por todas las herramientas (con 50% a 84% de precisión balanceada), excepto por Taxonomer que tiene un valor nulo.

En la puntuación F1, los modelos CNN_150_RefSeq y CNN_150_NT alcanzaron el 34% y 74%, respectivamente en el conjunto *Eukaryotic,* mientras que las herramientas BLAST, Centrifuge, DIAMOND2, Kraken2, One Codex y Taxonomer obtuvieron valores superiores al 79% que superaron a los dos modelos CNN. Por otra parte, DIAMOND1 y drVM (56% y 72%) superaron solo al modelo CNN_150_RefSeq. Finalmente, las herramientas SLIMM, FastViromeExplorer y VirusFinder obtuvieron puntuaciones F1 de 14% al 27% y Vipie presentó el valor más bajo (0.1%). En el conjunto *Prokaryotic* las herramientas BLAST, DIAMOND1 y SLIMM obtuvieron las puntuaciones F1 (85% al 90%) más altas que los modelos CNN_150_RefSeq y CNN_150_NT, los cuales lograron el 81% y 84%, respectivamente. Además, Centrifuge, DIAMOND2, FastViromeExplorer, Kraken2, One Codex, Taxonomer y VirusFinder alcanzaron una puntuación F1 de 20% al 76%, mientras que drVM y Vipie obtuvieron los valores más bajos con menos del 1%. Para el conjunto *Unclassified* los modelos CNN_150_RefSeq y CNN_150_NT lograron una puntuación F1 de 11% y 10% respectivamente, siendo superadas por todas las herramientas con valores que oscilan entre 40% a 80%, excepto por las herramientas Vipie y Taxonomer que tiene valores menores del 1%.

Finalmente, en el conjunto *Eukaryotic*, los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron un valor del MCC de 0.4 y 0.8, respectivamente. Este último fue superado solo por las herramientas BLAST y Kraken2 con un valor de MCC 0.9. En el conjunto *Prokaryotic* ambos modelos de CNN obtuvieron un MCC de 0.8, siendo superados con un MCC del 0.9 por BLAST, DIAMOND1 y SLIMM. Por último, en el conjunto *Unclassified,* los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron un MCC de 0.1 y fueron superadas por todas las herramientas, excepto por One Codex, Taxonomer y Vipie que obtuvieron un valor de 0.

Los resultados del MCC muestran que los modelos son buenos para clasificar las lecturas de fagos y eucariontes a pesar del desbalance que existe entre las clases.

Cabe resaltar que para los modelos CNN_150_RefSeq y CNN_150_NT, las lecturas abajo del umbral establecido fueron del 73% y 33% respectivamente en el conjunto de datos de *Eukaryotic*, del 19% y 28% respectivamente para el conjunto *Prokaryotic* y para el conjunto *Unclassified* del 82% y 85% respectivamente. En ambos modelos las lecturas que fueron mal clasificadas fueron predichas a las clases *Myoviridae*, *Siphoviridae* y *Unclassified.*

6.4.2. Conjuntos simulados no virales

Para obtener la especificidad global que los modelos CNN_150_Refseq y CNN_150_NT, se utilizaron los conjuntos de datos *Bacterial* y *Human*. Los resultados mostraron que los modelos obtuvieron alta especificidad, del 98% y 73%, respectivamente. Sin embargo, los resultados están por debajo del resto de las herramientas que tuvieron una especificidad del 99%. Esto probablemente se debe a que los modelos fueron entrenados solo con la información de las secuencias virales. Para que los modelos no realicen clasificaciones incorrectas cuando se introducen lecturas no virales, como las de humano o bacterias, se aplicó un umbral especificó para determinar lo que es viral y no viral; si las lecturas se encuentren por debajo del valor del umbral son clasificadas como no virales.

6.4.3. Conjuntos reales

En la Figura 6.10 se presentan los resultados de la evaluación de los conjuntos reales. Al igual que en los conjuntos no virales, se aplicó un umbral como en la sección previa, uno para el modelo CNN_150_RefSeq y otro para el modelo CNN_150_NT. Este umbral ayuda a descartar las lecturas de origen no viral, disminuyendo los falsos positivos.

Los modelos CNN_150_RefSeq y CNN_150_NT obtuvieron el mayor porcentaje de lecturas asignadas en los conjuntos *FISH-I* (31% y 17%) y *PB3* (34% y 16%); mientras que el resto de las herramientas obtuvieron menos del 2% en ambos conjuntos.

Por su parte, en el conjunto *I5-8* el modelo CNN_150_RefSeq obtuvo el 1%, mientras que la CNN_150_NT obtuvo 9% de lecturas asignadas. El modelo CNN_150_RefSeq superó a las herramientas FastViromeExplorer, Kraken2, One Codex, Vipie y VirusFinder, ya que tienen menos del 1%, y el modelo CNN_150_NT fue la segunda herramienta con mayor porcentaje de lecturas asignadas y fue superada por DIAMOND1 (16%).

Por otro lado, en el conjunto *I21-1,* la CNN_150_RefSeq obtuvo 0.6% de asignación y superó a las herramientas FastViromeExplorer, Kraken2, SLIMM y Vipie que tiene menos del

0.6% de asignación (0.1% - 5%); en cambio, la CNN_150_NT fue la primera herramienta con mayor porcentaje de lecturas asignadas con un 14% superando ligeramente a drVM (13%).

Los resultados sugieren que el modelo CNN_150_RefSeq hace una sobreestimación en las familias de bacteriófagos, en otras palabras, el modelo está asignando más lecturas a familias de bacteriófagos. Esto puede deberse a que el 32% de los genomas son de bacteriófagos y generaron un 73% de los *k-mers* utilizados para entrenar el modelo. Por lo que, los conjuntos *FISH-I* y *PB3* mostraron altos porcentajes de lecturas asignadas con el modelo CNN_150_RefSeq.

En las Figuras 6.11, 6.12 y 6.13 se muestran las familias que fueron identificadas por el modelo CNN_150_RefSeq y CNN_150_NT. En la Figura 6.11 se presentan todas las familias identificadas por el modelo CNN_150_RefSeq, es decir que superaron el umbral de 0.95. El modelo de CNN_150_RefSeq identificó 34 familias, siendo *Myoviridae* y *Siphoviridae* las familias con el mayor porcentaje de lecturas asignadas con 12% y 9%, respectivamente, en el conjunto *FISH-I,* y en el conjunto *PB3* fue de 17% y 9%, respectivamente; otras familias como *Autographiviridae*, *Herelleviridae* y *Podoviridae* en los conjuntos *FISH-I* y *PB3* obtuvieron un porcentaje de asignación de 1% - 2% y el resto de las familias alcanzó menos del 1% de las lecturas asignadas. En la Figura 6.7 se muestra que las familias con mayor número de lecturas asignadas en los conjuntos. Estas familias son de bacteriófagos, dado que el modelo asignó más del 11% tan solo en dos familias, mientras que el resto de las herramientas está por debajo de este valor.

En los conjuntos *I5-8* e *I21-1,* las familias con mayor porcentaje de lecturas fueron *Herelleviridae, Myoviridae, Poxviridae*, *Polydnaviridae* y *Siphoviridae* con 0.4% - 0.1% y el resto de las herramientas obtuvieron menos del 0.1%.

En las Figuras 6.12 y 6.13 se muestran las familias que fueron identificadas por el modelo CNN_150_NT. Este modelo identificó 118 familias, siendo *Myoviridae* y *Siphoviridae* algunas de las familias con el mayor porcentaje de lecturas asignadas, entre el 3% y 7%, en los conjuntos *FISH-I* y *PB3*; mientras que *Herelleviridae*, *Herpesviridae* y *Microviridae* obtuvieron el 2% - 1% en el conjunto *FISH-I* y el resto de las herramientas obtuvieron menos del 1%. Por su parte, en los conjuntos *I5-8* e *I21-1,* la familia con más lecturas asignadas fue *Siphoviridae,* con 4% y 8%, respectivamente. Por otro lado, las familias *Iridoviridae* y *Poxviridae (para los conjuntos I5-8* e *I21-1), Herpesviridae* y *Picornaviridae (para los conjuntos I21-1), y Mimiviridae* y *Phycodnaviridae* (para el conjunto *I5-8*) alcanzaron entre el 1% - 2% de las lecturas asignadas, y las demás familias asignaron menos del 1%.

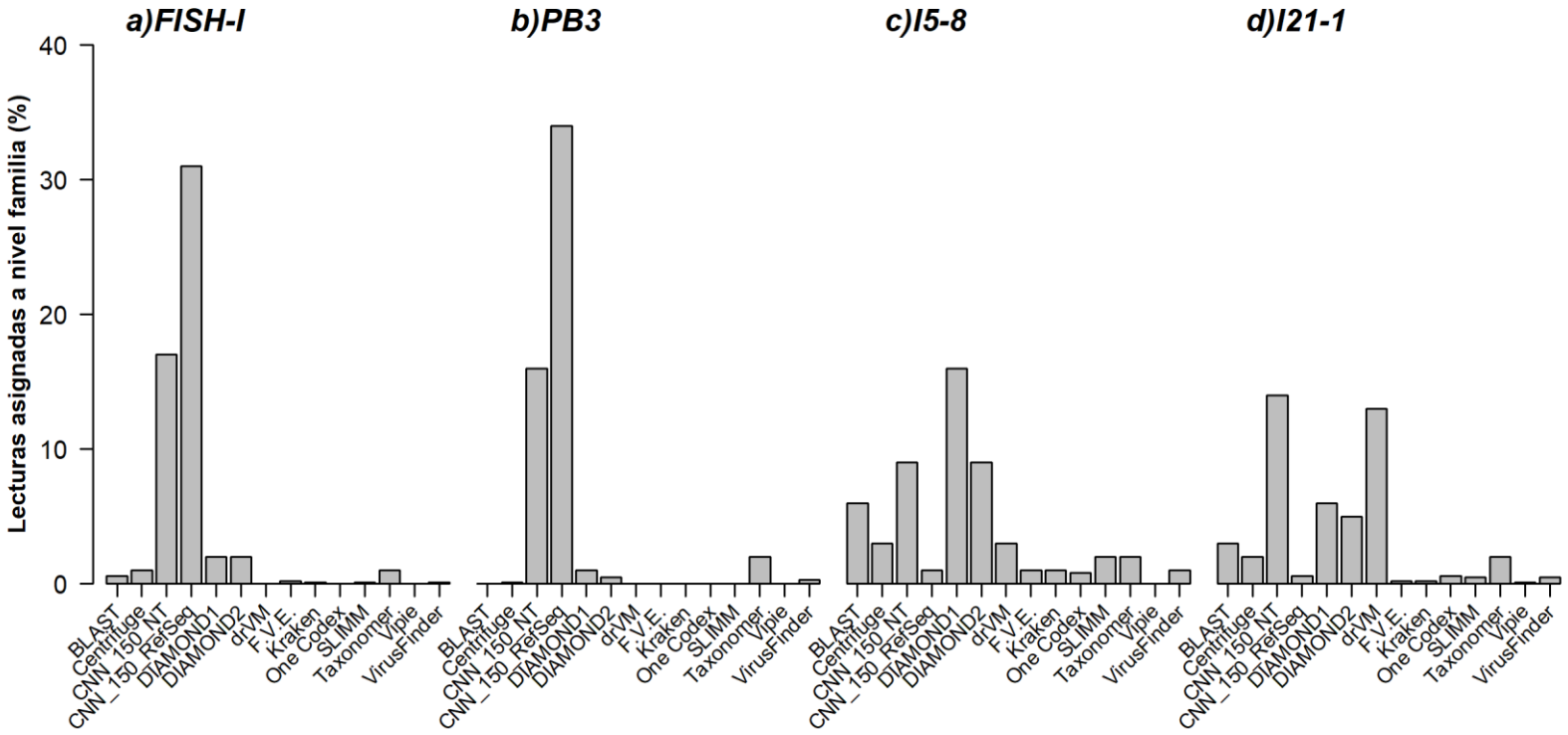


Figura 6.10. Asignación de las lecturas de los conjuntos reales. En el panel inferior se encuentra los nombres de las herramientas y en el panel vertical el porcentaje de lecturas asignadas. La herramienta FastViromeExplorer es F.V.E., DIAMON1 es la herramienta DIAMOND utilizando la base de datos de genomas completas y DIAMOND2 utilizando la base de datos RefSeq.

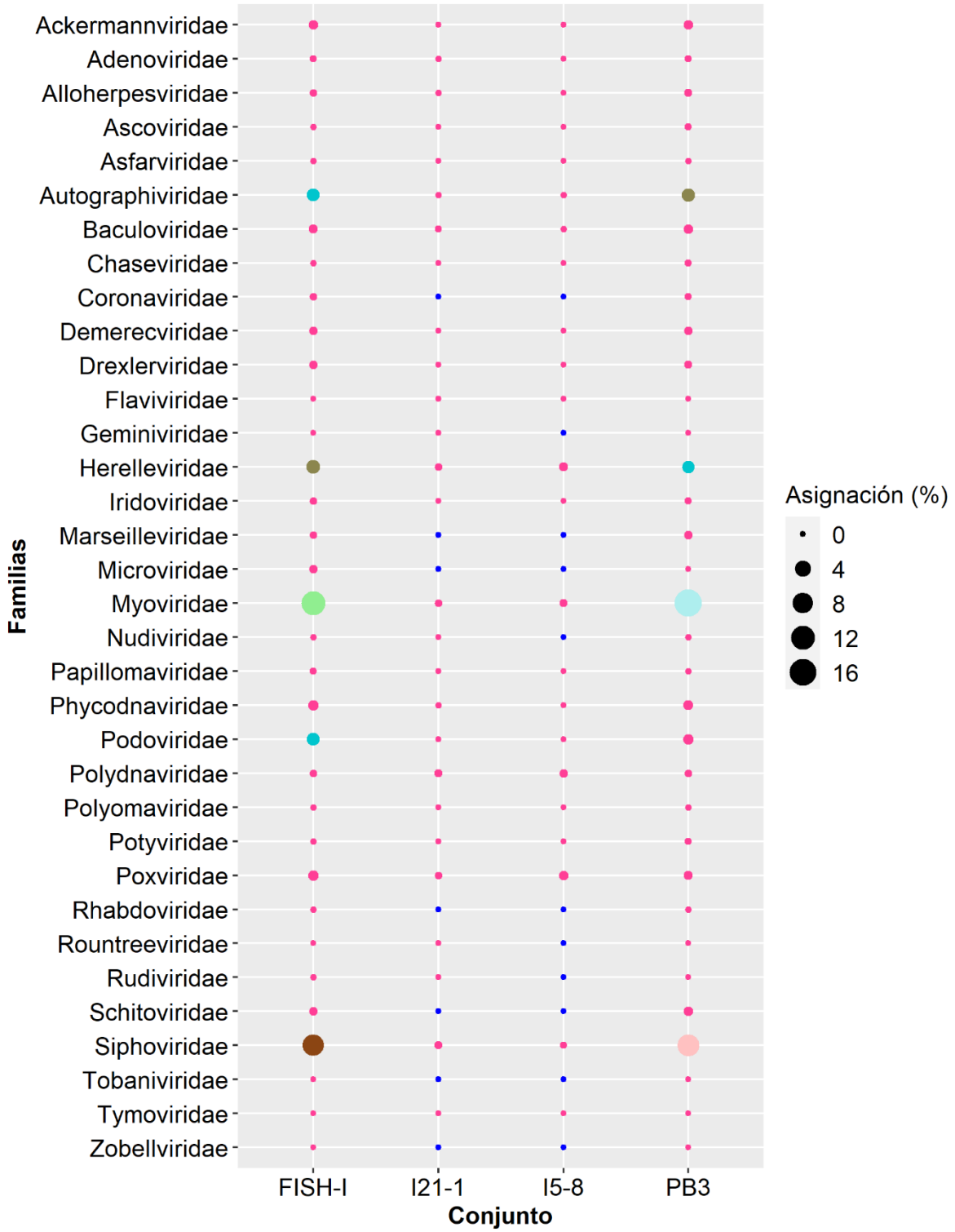


Figura 6.11. Familias identificadas en los conjuntos reales con el modelo CNN_150_RefSeq. En el panel inferior se tiene el nombre de los conjuntos, en el lateral derecho los nombres de las familias identificadas y en el izquierdo están representados en círculos (de color negro) el porcentaje de las lecturas asignadas.

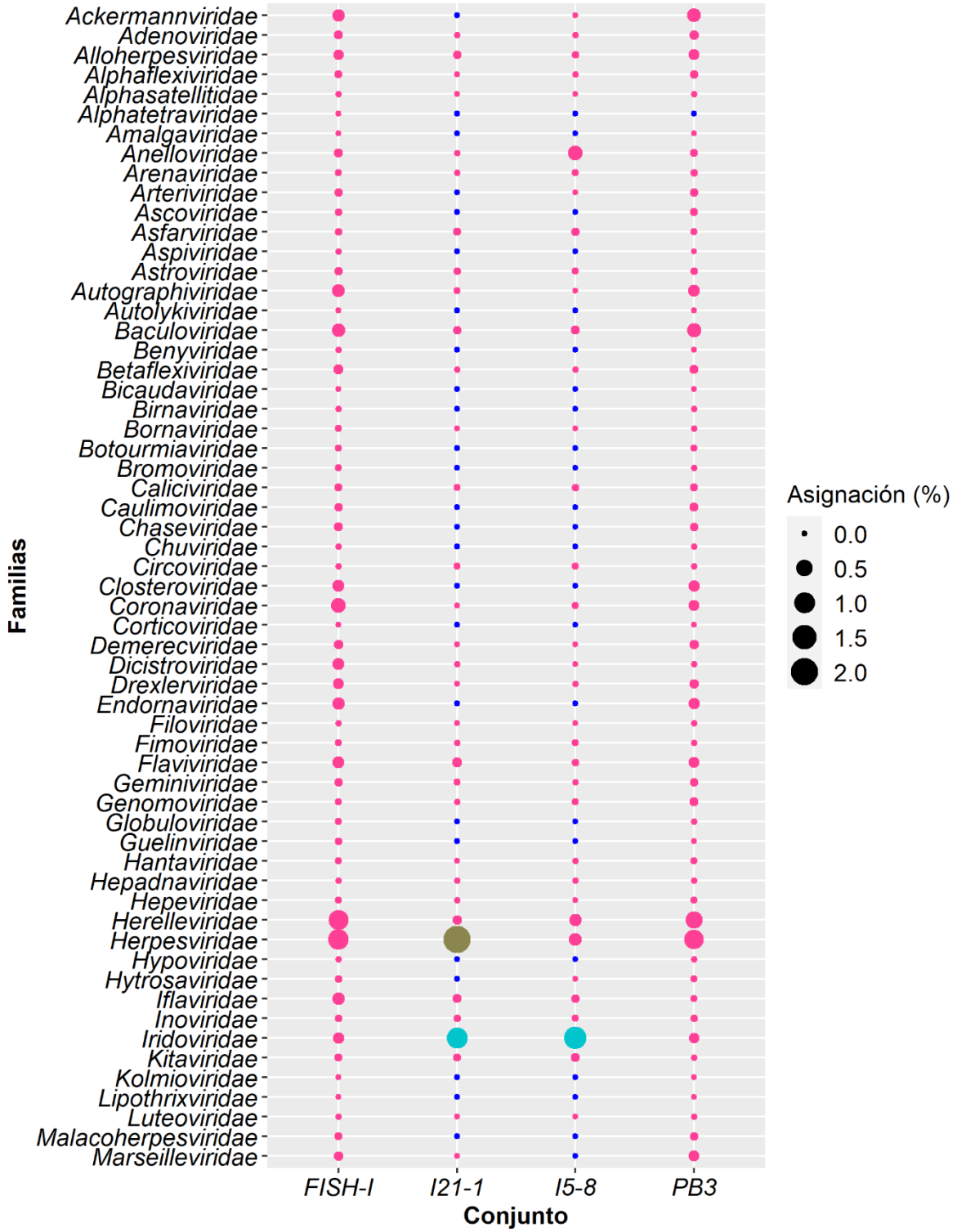


Figura 6.12. Familias identificadas (de la Ackermannviridae - Marseilleviridae) en los conjuntos reales con el modelo CNN_150_NT. En el panel inferior se tiene los conjuntos reales ordenados alfabéticamente, en el lateral derecho están los nombres de las familias identificadas, en el izquierdo están representados en círculos (de color negro) el porcentaje de las lecturas asignadas.

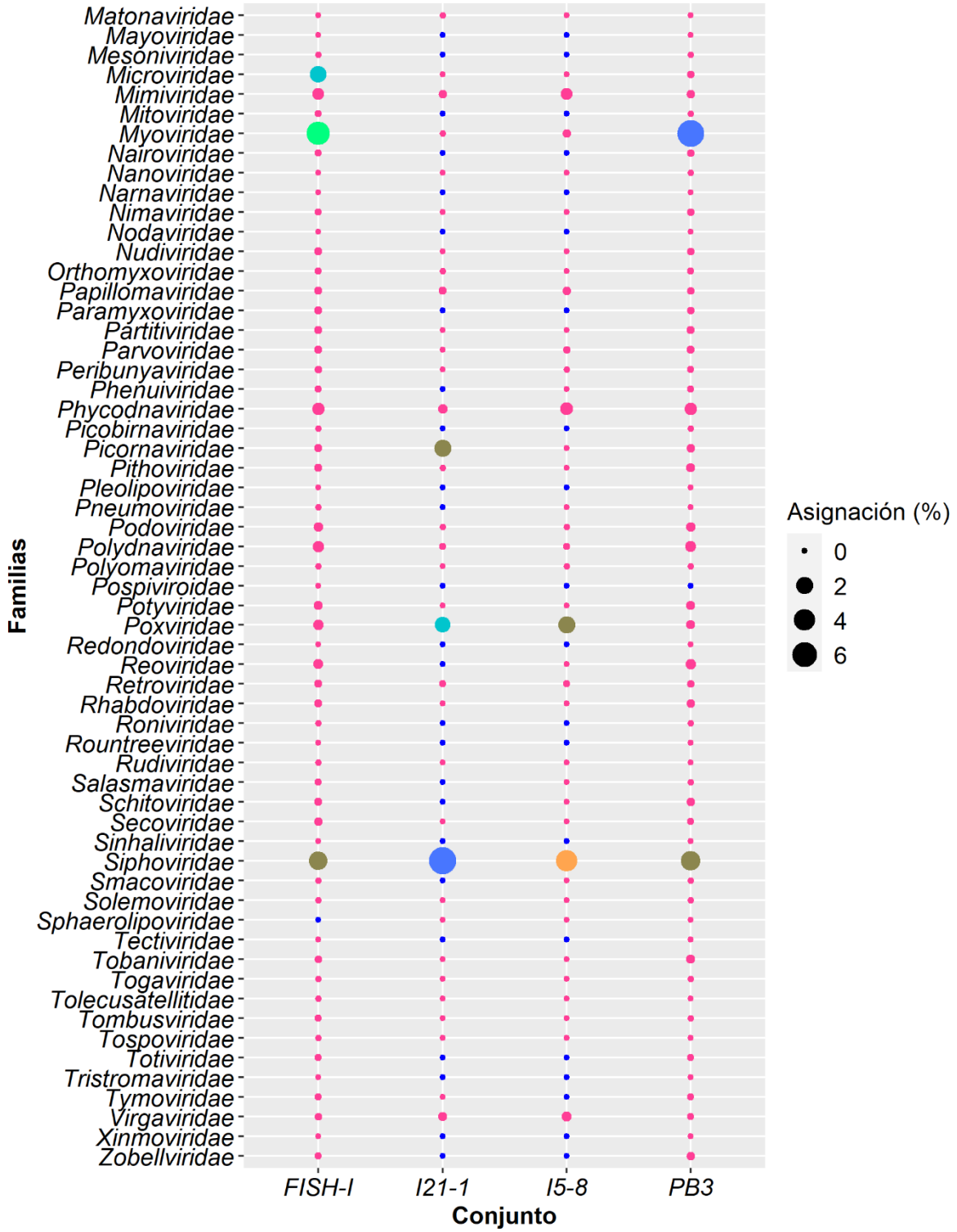


Figura 6.13. Familias identificadas (de la familia Matonaviridae - Zobellviridae) en los conjuntos reales con el modelo CNN_150_NT. En el panel inferior se tiene los nombres de los conjuntos, en el lateral derecho están los nombres de las familias identificadas y en el izquierdo están representados en círculos (de color negro) el porcentaje de las lecturas asignadas.

6.4.4. Memoria y tiempo utilizados

En la tabla 6.1 se compara la memoria y el tiempo de ejecución obtenido por los modelos, CNN_150_NT y CNN_150_RefSeq, y el resto de las herramientas. Ambos modelos utilizaron poco tiempo en el proceso de clasificación, aproximadamente 45 minutos CPU, para analizar 10 millones de lecturas. Solo quedó por arriba Centrifuge y FastVimoreExplorer (con 6 minutos), siendo más rápidos que el resto de las herramientas. Por otro lado, los modelos obtuvieron un consumo de memoria media de 7714 MB y 6322 MB, respectivamente; superando a BLAST, DIAMOND2, drVM y VirusFinder (<500 MB y 4454 MB de memoria), y en el caso de CNN_150_NT también superó a FastViromeExplorer.

Tabla 6. 1. Comparación de la memoria y tiempo de ejecución requeridos por de modelos CNN_150 con respecto a las otras herramientas en el proceso de clasificación de cada conjunto de datos.

Conjunto de datos	BLAST*	Centrifuge*	CNN_150_NT[1]	CNN_150_RefSeq[1]	DIAMOND1*	DIAMOND2*	drVM*	F.V.E.*	Kraken2*	One Codex	SLIMM*	Taxonomer	Vipie	VirusFinder*
Tiempo de ejecución (Minutos) calculados por core.														
50G	83	3	4	3	112	27	9	2	64	6	10	ns	8	43
500G	89	5	4	3	119	25	19	2	64	9	21	ns	9	70
1000G	102	4	4	3	111	22	18	2	64	8	20	ns	9	75
Eukaryotic	1729	5	44	43	1644	502	519	6	69	44	49	ns	51	1296
Prokaryotic	1768	7	44	43	1536	547	208	8	68	47	51	ns	38	1264
Unclassified	1426	5	44	43	857	389	208	5	68	48	47	ns	29	1253
FISH-I	927	5	21	20	321	135	64	6	72	33	46	ns	56	476
PB3	664	5	15	15	241	104	47	5	70	31	48	ns	37	546
I5-8	1265	7	41	40	354	139	61	5	73	38	49	ns	42	500
I21-1	928	3	14	13	149	35	55	5	71	19	49	ns	29	472
Bacterial	66	2	3	3	54	14	10	2	65	6	3	ns	7	19
Human	64	2	3	3	51	11	12	2	65	5	1	ns	6	57
Total (Minutos)	9111	54	241	232	5549	1950	1230	**50**	813	294	394	ns	321	7801
Memoria (MB)														
50G	282	16,998	2830	2870	3246	922	1556	6835	143,524	ns	26,788	ns	ns	4618
500G	393	17,009	2840	2640	13,343	1044	1690	6836	143,667	ns	26,788	ns	ns	4618
1000G	433	17,009	2840	2680	3584	2591	1536	7022	143,811	ns	26,788	ns	ns	4618
Eukaryotic	475	17,213	15360	15154	11,991	5366	7967	7655	144,241	ns	26,757	ns	ns	4516
Prokaryotic	401	17,162	15210	15060	12,216	9339	1812	7662	144,179	ns	26,757	ns	ns	4516
Unclassified	235	17,029	15610	1537	12,052	10,179	4045	7644	144,230	ns	26,757	ns	ns	4516
FISH-I	189	17,009	8610	8120	7322	4608	1823	6837	143,964	ns	26,757	ns	ns	4096
PB3	189	17,009	6499	6460	6308	6318	1823	6835	143,974	ns	26,757	ns	ns	4516
I5-8	187	17,060	10250	10060	16,005	8376	1853	6919	143,985	ns	26,767	ns	ns	4096
I21-1	183	17,142	5420	5310	15,933	2273	1802	6838	144,067	ns	26,767	ns	ns	4096
Bacterial	147	16,998	3500	2960	2775	768	1833	6834	143,606	ns	26,757	ns	ns	4618
Human	152	16,998	3600	3010	2785	748	1833	6836	143,585	ns	26,757	ns	ns	4618
Promedio (MB)	**272**	17053	7714	6322	8963	4378	2464	7063	143903		26766			4454
Desviación estándar (MB)	*120*	*76*	*5196*	*4829*	*5174*	*3515*	*1857*	*361*	*259*		*14*			*220*

*La herramienta FastViromeExplorer está representado por F.V.E.; ns significa no especificada. * Todas las herramientas fueron ejecutadas en el mismo clúster de computadoras con las siguientes características: Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz que tiene 6 nodos (AMD Opteron(tm) Processor 6376) con 64 cores y 512 GB de memoria. [1] Los modelos CNN_150_NT y CNN_150_Refseq fueron ejecutados en un equipo con las siguientes características: Intel Xeon CPU @2.20 GHz, 13 GB RAM y Tesla K80 GPU.*

6.5. Evaluación de los resultados

Para analizar los resultados obtenidos se aplicó un método de Análisis de Decisión Multicriterio (en inglés, *MultiCriteria Decision Analysis* (MCDA)). Este método permite estructurar procesos para la toma decisiones con base en un conjunto criterios. Como un primer paso se realizó la normalización de los valores obtenidos por todas las herramientas en todos los conjuntos. Esta normalización puso a los valores de 0 a 10, donde 10 es el valor mayor y cero el menor. Para los resultados de las métricas como la sensibilidad, precisión, precisión equilibrada, puntuación F1 y MCC se aplica una normalización para valores máximos, dado que el valor máximo representa a la mejor herramienta (ec. 6.1), mientras que para los valores de tiempo y memoria se realizó una normalización mínima, dado que la herramienta que utilizó menos tiempo y memoria es la mejor (ec. 6.2) (79).

$$Normalización\ de\ datos_i = \frac{SP_i - SP_{min}}{SP_{max} - SP_{min}} * 10\ , \tag{6.1}$$

$$Normalización\ de\ datos_i = \frac{SR_{max} - SR_i}{SR_{max} - SR_{min}} * 10\ , \tag{6.2}$$

donde SP_i es el registro estadístico y SR_i es residual estadístico (elementos a normalizar según el caso).

Para las lecturas largas (*50G*, *500G* y *1000G*) se promedió los valores de las métricas evaluadas de los tres conjuntos por cada herramienta y posteriormente fue normalizado (ver Tabla Anexo 3.1). En los conjuntos reales, antes de ser normalizados, se promedió los valores de asignación de los conjuntos que tiene un huésped humano (*I5-8 e I21-1*) y los que no tiene huésped humano (*FISH-I* y *PB3*), los resultados finales están en la Tabla Anexo 3.5 y Tabla Anexo 3.6, respectivamente. Para tiempo y memoria se normalizó el promedio de todos los conjuntos (ver Tabla Anexo 3.7). Después se creó un escenario (ver Tabla 6.2) para las lecturas simuladas, mismo peso indica que se le está dado la misma relevancia. Para los conjuntos reales que solo tiene una métrica (el porcentaje de asignación) se utilizó el valor normalizado; se aplicó lo mismo para el tiempo y la memoria.

El valor final (V(s)) (ec. 6.3) obtenido de cada herramienta en los diferentes conjuntos en cada escenario se calcula de la siguiente manera:

$$V(s) = \sum_{i=1}^{E} w_i u(s)\ , \tag{6.3}$$

donde w_i son los pesos usados en cada métrica, $u(s)$ es el valor de cada métrica y E es el número de métricas a evaluar.

Tabla 6. 2. Un escenario para los conjuntos simulados, con una ponderación igual para todas las métricas.

Métricas de evaluación	Escenario 1	
	Pesos ponderados	Pesos (%)
Precisión	10	20
Sensibilidad	10	20
Precisión equilibrada	10	20
Puntuación F1	10	20
MCC	10	20
Total	50	100%

Para elegir cuál fue la mejor herramienta se utiliza el valor promedio de desempeño del escenario.

Tabla 6. 3. Resultados finales (valores promedios) del MCDA para el análisis de las herramientas, considerando sus métricas.

Herramientas	Lecturas largas	Lecturas cortas			Lecturas reales	
		Eukaryotic	*Prokaryotic*	*Unclassified*	*Sin huésped humano*	*Con huésped humano*
BLAST	9.71	**9.94**	9.96	9.87	0.09	3.89
Centrifuge	2.04	8.60	7.25	7.86	0.17	2.14
CNN_150_NT	8.36	7.91	8.78	2.47	5.08	**10.00**
CNN_150_RefSeq	7.76	4.09	9.02	2.55	**10.00**	0.66
DIAMOND1	3.90	6.46	9.55	**9.88**	0.46	9.56
DIAMOND2	4.46	8.89	8.48	8.48	0.38	6.07
drVM	2.07	6.66	1.92	5.93	0.00	6.94
F.V.E.	**9.99**	3.29	6.10	6.53	0.03	0.48
Kraken2	8.46	9.63	5.37	8.89	0.02	0.48
One Codex	2.63	8.03	6.71	6.97	0.00	0.57
SLIMM	4.89	4.25	**10.00**	8.81	0.02	1.05
Taxonomer	5.78	9.35	1.79	0.00	0.46	1.70
Vipie	3.32	1.79	1.26	2.88	0.00	0.00
VirusFinder	9.85	3.40	5.45	8.40	0.06	0.39

La herramienta FastViromeExplorer está representado por F.V.E.

Tabla 6. 4. Resultados finales (valores promedios) del MCDA para el análisis de las herramientas considerando el tiempo y la memoria.

Herramientas	Tiempo	Memoria
BLAST	0.00	**10.00**
Centrifuge	**10.00**	8.83
CNN_150 _NT	9.79	9.48
CNN_150_RefSeq	9.80	9.58
DIAMOND1	3.93	9.39
DIAMOND2	7.90	9.71
drVM	8.70	9.85
F.V.E.	**10.00**	9.53
Kraken2	9.16	0.00
One Codex	9.73	ns
SLIMM	9.62	8.16
Taxonomer	ns	ns
Vipie	9.70	ns
VirusFinder	1.45	9.71

La herramienta FastViromeExplorer está representado por F.V.E.; ns significa no especificada.

Los modelos CNN_150_RefSeq y CNN_150_NT son buenos para la clasificación de virus procariontes dado que obtuvieron valores de 9.02 y 8.70 (ver Tabla 6.3). En los conjuntos reales provenientes de muestras que no tiene un huésped humano, como FISH-I y PB3, obtuvieron un alto porcentaje de asignación de 10 y 5.8, mientras que en las muestras que provienen de un huésped humano, como I5-8 e I21-1, el modelo CNN_150_NT fue el mejor con un valor de 10. Además, ambos modelos muestran que utilizan poco tiempo de procesamiento con un valor de 9.80 y 9.79 (ver Tabla 6.4) y consumen memoria de forma moderada 9.58 (CNN_150_RefSeq) y 9.42 (CNN_150_NT).

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

En este trabajo de investigación se desarrollaron y evaluaron dos modelos de Deep learning CNN_150_RefSeq y CNN_150_NT usando la técnica de redes neuronales convolucionales (CNN), para discriminar taxonómicamente las secuencias de virus generadas por tecnologías de secuenciación masiva de ADN, en datos metagenómicos. De este trabajo se concluye que:

a) En los conjuntos simulados de lecturas largas, los modelos CNN_150_NT y CNN_150_RefSeq tiene buenos desempeños en precisión, por otro lado, en sensibilidad, precisión equilibrada, puntuación F1 y MCC son superadas por BLAST, FastViromeExplorer, Kraken2 y VirusFinder, las cuales fueron las mejores herramientas en estos conjuntos.
b) En los conjuntos simulados de lecturas cortas, los modelos CNN_150_NT y CNN_150_RefSeq presentan una disminución en la sensibilidad, precisión equilibrada, puntuación F1 y MCC.
c) En el conjunto simulado *Eukaryotic*, ambos modelos obtuvieron altos desempeños en precisión (92% - 93%). En la precisión equilibrada y en el MCC, solo la CNN_150_NT presenta un alto desempeño (81% y 0.8 respectivamente).
d) En el conjunto simulado *Prokaryotic*, ambos modelos fueron de las mejores herramientas de clasificación viral. Ambos modelos presentaron altos desempeños en precisiones (>94%), en la precisión equilibrada y en la puntuación F1 (>85%), y en el MCC (0.8).
e) En el conjunto simulado *Unclassified*, los modelos CNN_150_NT y CNN_150_RefSeq tienen bajo desempeño en todas las métricas.
f) Los modelos CNN_150_NT y CNN_150_RefSeq presentaron una disminución en su desempeñó cuando aumenta la riqueza del viroma de las muestras.
g) La información que se utiliza para entrenar la CNN es de gran importancia, ya que permite tener un mejor modelo de clasificación viral. El modelo CNN_150_NT tiene una mayor sensibilidad, precisión, precisión equilibrada, puntuación F1 y MCC que el modelo CNN_150_RefSeq, debido a que contiene más información en el entrenamiento, siendo la base de datos (BD) nt 9 veces más grande que la BD RefSeq. Por lo tanto, en este trabajo se propone como herramienta final la CNN_150_NT.

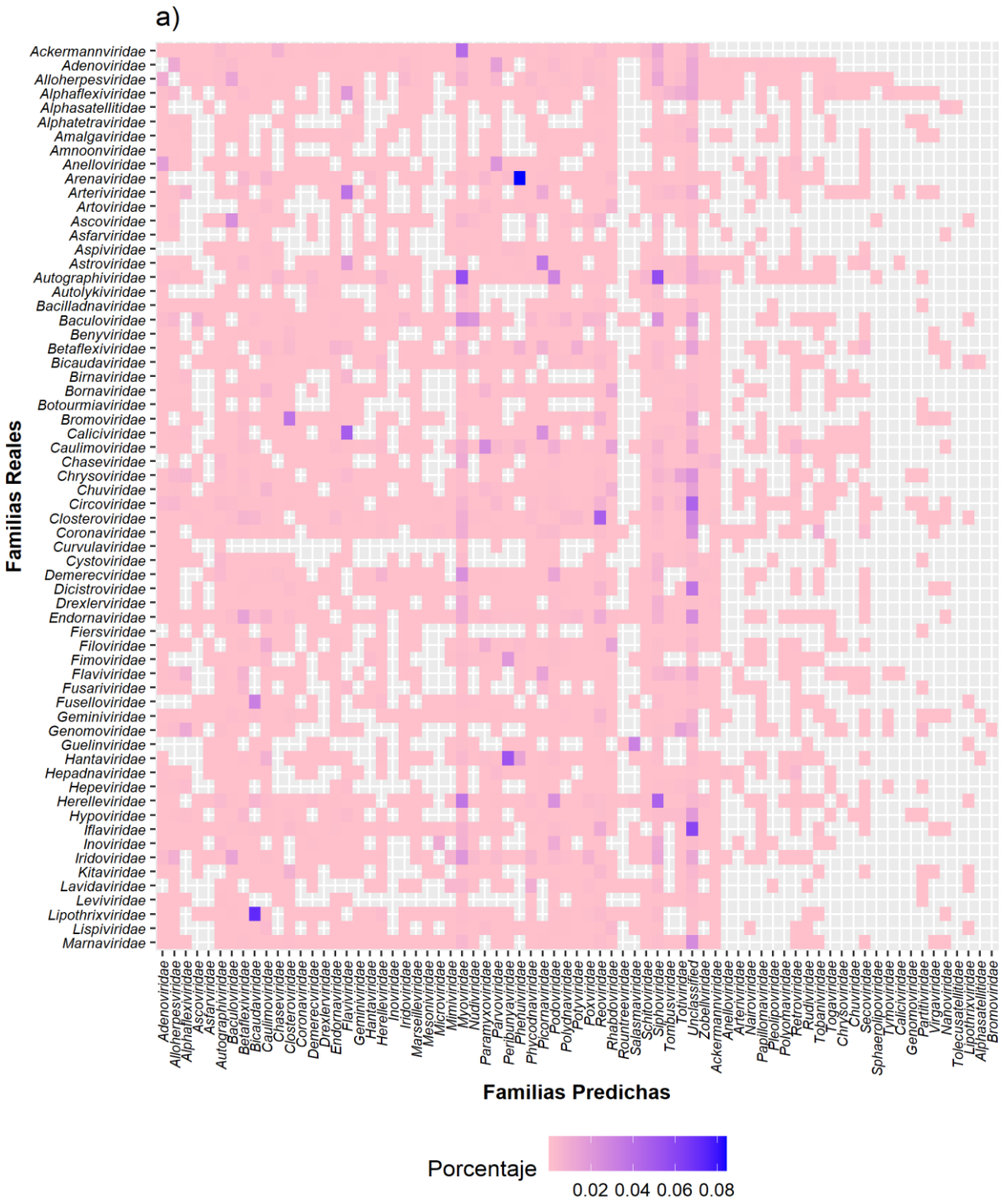
h) Dado que en los datos empleados para entrenar las CNN no cuentan con una clase negativa, por lo tanto, se determinó un umbral que permite filtrar las lecturas y disminuye los falsos positivos.
i) La CNN_150_RefSeq y CNN_150_NT tienen la clase *Unclassified* que al final representó un sesgo importante en su desempeño. Esto debido a que provocó la pérdida de 51 y 38 clases respectivamente, las cuales son las clases con la menor cantidad de lecturas (clases minoritarias), siendo asignadas en general y manera equivocada a dicha clase mayoritaria. En general las clases que se perdieron cuentan con menos de 426,294 *k-mers* en el entrenamiento, mientras la de *Unclassified* cuenta con 31,553,510 *k-mers*.

Considerando lo antes mencionado, cabe destacar que las CNN_150_NT y CNN_150_RefSeq son modelos que predicen rápidamente lecturas cortas. Además, son buenos para la predicción de lecturas de virus procariontes y las provenientes de ambientes no humanos. En el caso de las CNN_150_NT es excelente en la predicción de lecturas que tiene como huésped al humano. También se debe señalar que los modelos son capaces de predecir lecturas de mayor y menor longitud que la establecida (150 pb) en poco tiempo con una memoria moderada. Todos los resultados obtenidos se deben principalmente a que los modelos consideran casi todas las familias virales existentes, algo que ninguna otra herramienta de este tipo ha logrado realizar.

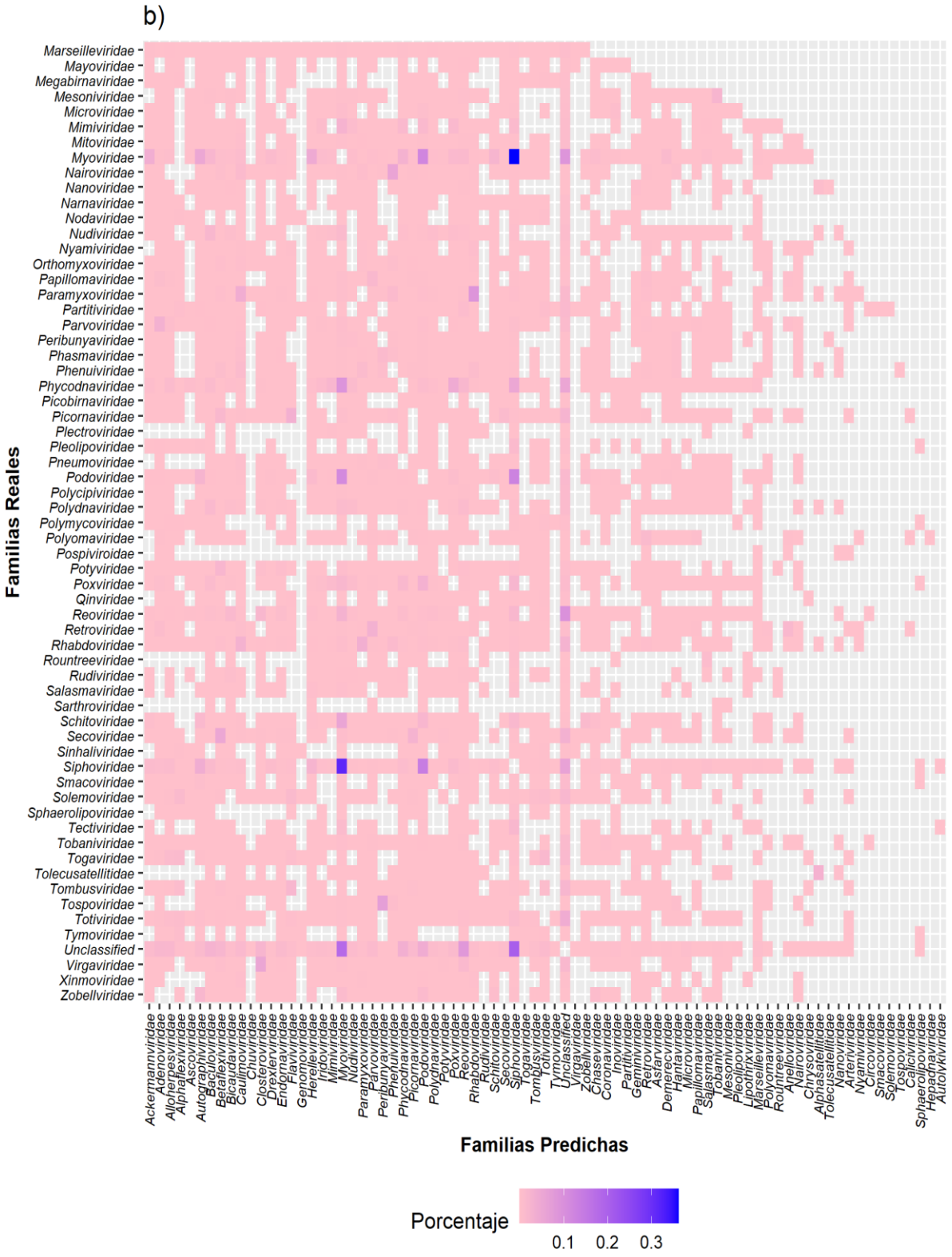
7.2. Trabajos futuros

1. Entrenar dos modelos de CNN por separado, uno para clasificar familias de bacteriófagos y otro para las familias que infectan eucariontes, y que ambos modelos no contengan la clase *Unclasified*.
2. Se requiere de una mejor base de datos para el entrenamiento de los modelos. Esta base de datos debe identificar las partes conservadas de las especies de cada familia y partes únicas, lo cual facilitaría la clasificación viral.
3. Crear un modelo para cada familia que permita clasificar las lecturas virales a niveles taxonómicos más bajos o específicos, como por ejemplo especie.
4. Explorar otras técnicas que no usen *k-mers* para no incrementar el tamaño de la base de datos y que permite reducir el tiempo de entrenamiento.

Anexo 1: Relación de las familias mal clasificadas modelo CNN_150_RefSeq



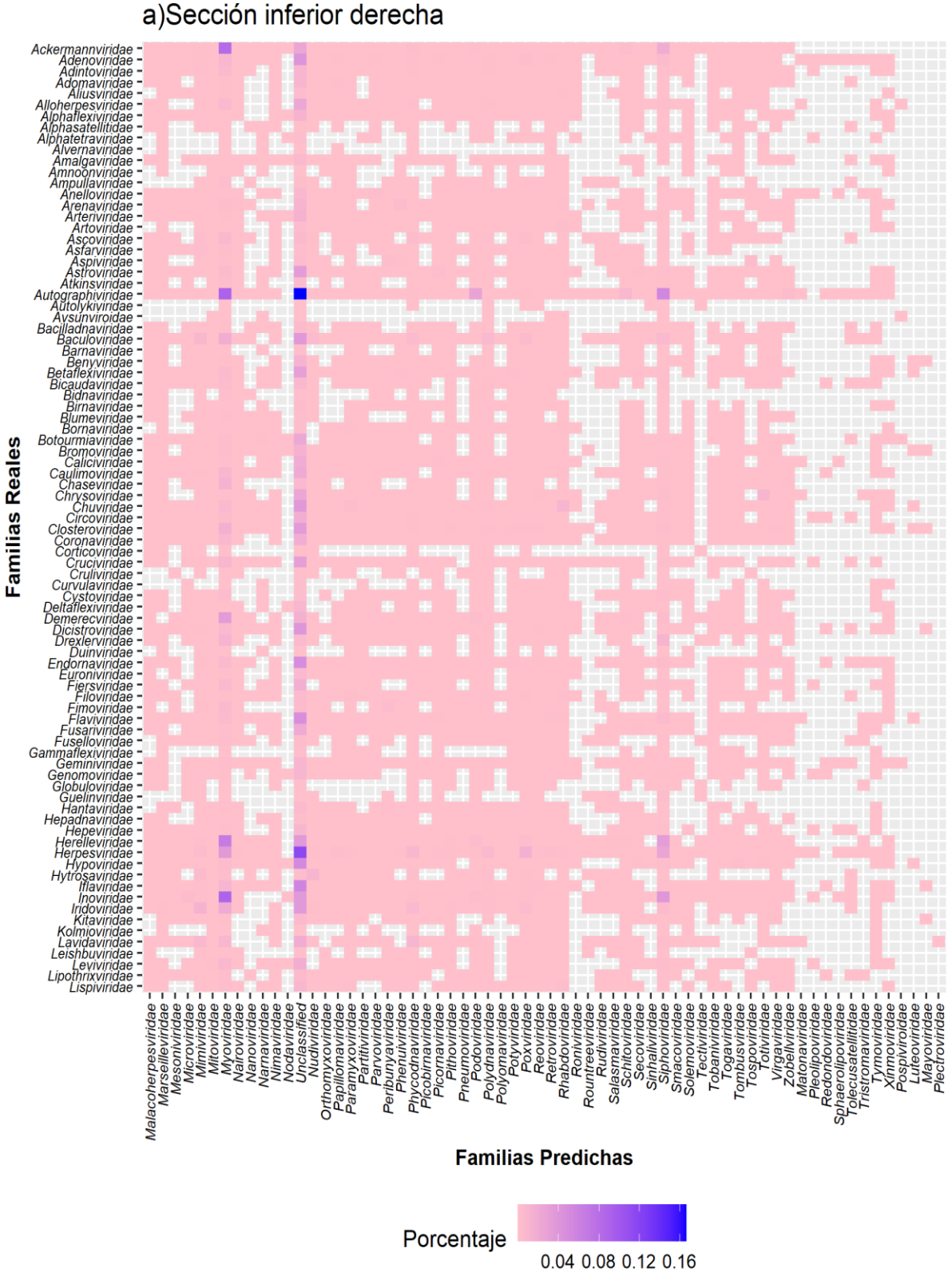
Anexo 1.1. Relación de las familias (Ackermannviridae - Marnaviridae) mal clasificadas. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichas las lecturas reales. Entre más intenso es el color del gradiente mayor porcentaje de lecturas mal clasificadas.



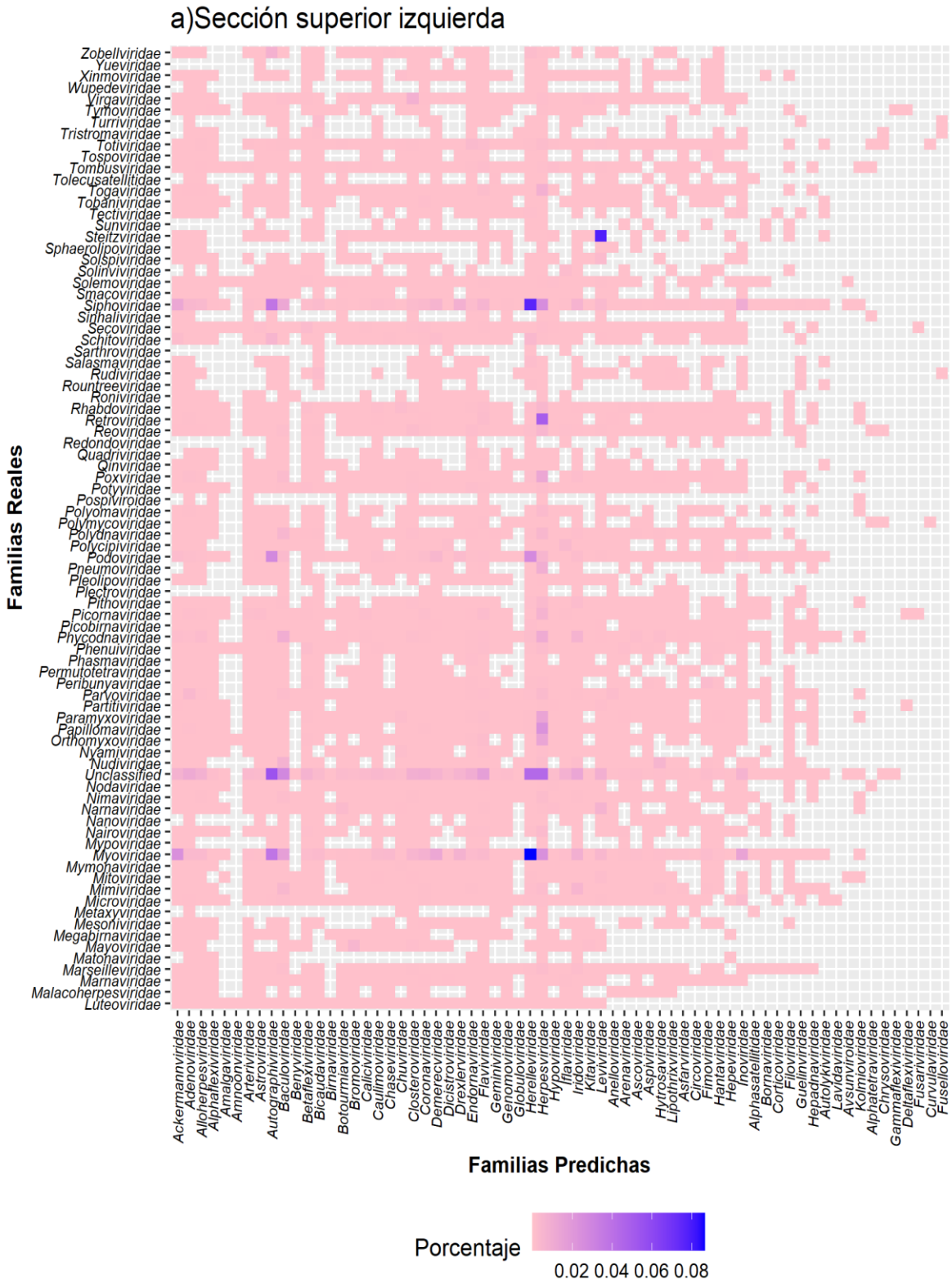
Anexo 1.2. Relación de las familias (Marseilleviridae - Zobellviridae) mal clasificadas. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichos las lecturas. Entre más intenso es el color del gradiente mayor porcentaje de lecturas mal clasificadas.

Anexo 2: Relación de las familias mal clasificadas modelo CNN_150_NT

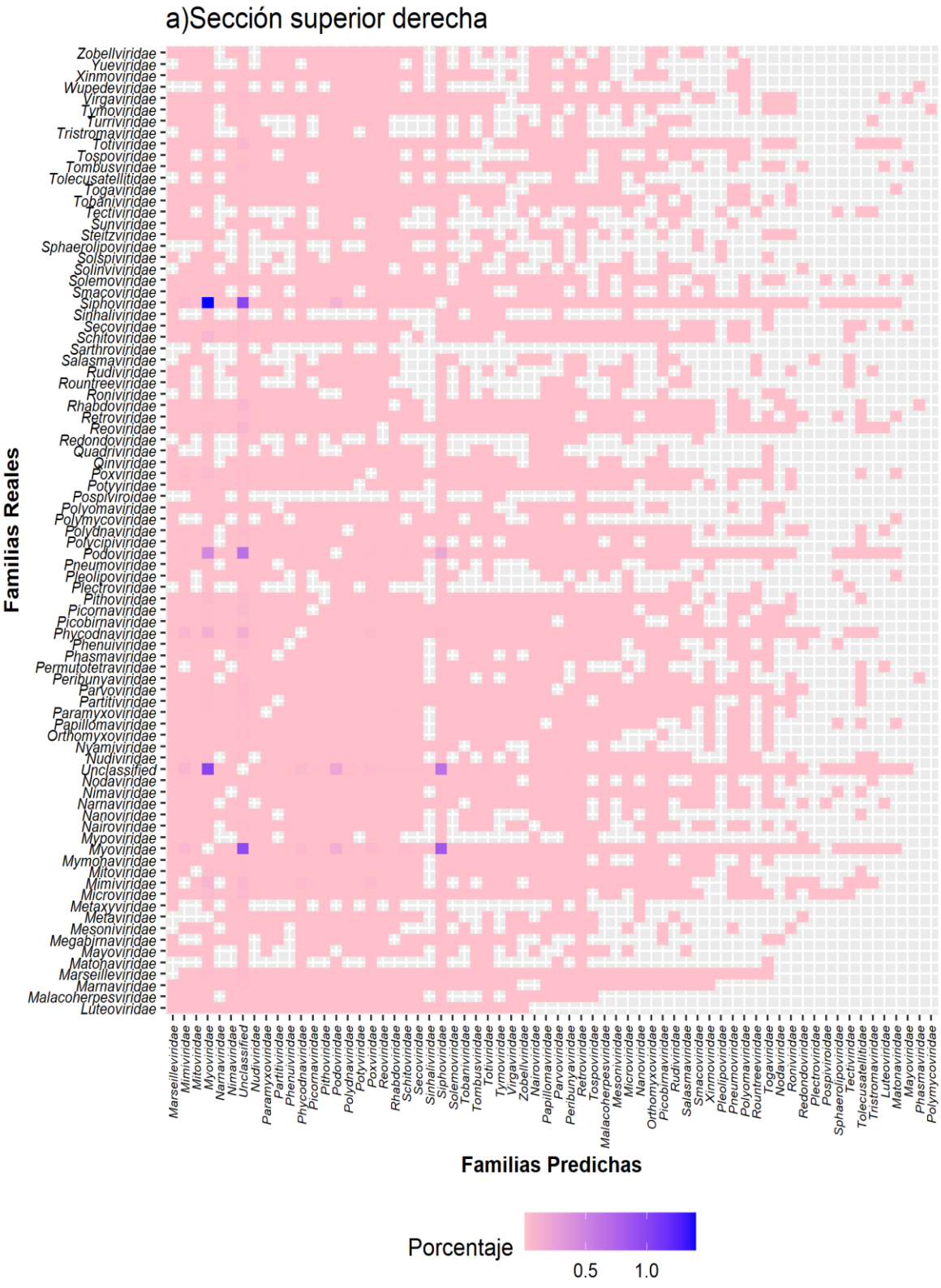
Anexo 2.1. Relación de las familias mal clasificadas del modelo CNN_150_NT parte 1. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichas las lecturas. Entre más intenso es el color del gradiente, se tiene un mayor porcentaje de lecturas mal clasificadas.



Anexo 2.2. Relación de las familias mal clasificadas del modelo CNN_150_NT parte 2. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichas las lecturas. Entre más intenso es el color del gradiente mayor porcentaje de lecturas mal clasificadas.



Anexo 2.3. Relación de las familias mal clasificadas del modelo CNN_150_NT parte 3. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichas las lecturas. Entre más intenso es el color del gradiente mayor porcentaje de lecturas mal clasificadas.



Anexo 2.4. Relación de las familias mal clasificadas del modelo CNN_150_NT. En el panel izquierdo se tiene las familias reales y en el inferior las familias a las que fueron predichas las lecturas. Entre más intenso es el color del gradiente mayor porcentaje de lecturas mal clasificadas.

Anexo 3: Tablas normalizadas del análisis de decisión multicriterio

Tabla Anexo 3.1. Resultados de las métricas normalizadas en las lecturas largas (0 – 10).

Herramientas	Precisión	Sensibilidad	Precisión equilibrada	Puntuación F1	MCC
BLAST	10	9	9	10	10
Centrifuge	0	3	3	2	2
CNN_150_NT	9	8	8	9	8
CNN_150_RefSeq	9	7	7	8	8
DIAMOND1	10	2	0	4	3
DIAMOND2	10	2	3	4	3
drVM	10	0	1	0	0
FastViromeExplorer	10	10	10	10	10
Kraken2	8	9	9	9	8
One Codex	6	2	2	2	1
SLIMM	10	3	3	4	4
Taxonomer	9	5	5	6	5
Vipie	7	2	3	3	2
VirusFinder	10	10	10	10	10

Tabla Anexo 3.2. Resultados de las métricas normalizadas en el conjunto Eukaryotic (0 – 10).

Herramientas	Precisión	Sensibilidad	Precisión equilibrada	Puntuación F1	MCC
BLAST	10	10	10	10	10
Centrifuge	9	8	8	9	9
CNN_150_NT	7	7	7	8	9
CNN_150_RefSeq	7	3	3	4	4
DIAMOND1	10	5	5	6	7
DIAMOND2	10	8	8	9	9
drVM	0	9	9	8	8
FastViromeExplorer	10	1	1	2	3
Kraken2	9	10	10	10	10
One Codex	4	9	9	9	9
SLIMM	10	2	2	3	4
Taxonomer	9	10	9	10	9
Vipie	9	0	0	0	0
VirusFinder	10	1	1	2	3

Tabla Anexo 3.3. Resultados de las métricas normalizadas en el conjunto Prokaryotic (0 – 10).

Herramientas	Precisión	Sensibilidad	Precisión equilibrada	Puntuación F1	MCC
BLAST	10	10	10	10	10
Centrifuge	4	8	8	8	8
CNN_150_NT	9	8	8	9	9
CNN_150_RefSeq	8	9	9	9	9
DIAMOND1	10	9	9	10	10
DIAMOND2	10	8	8	8	9
drVM	10	0	0	0	0
FastViromeExplorer	10	4	4	6	7
Kraken2	0	7	7	7	7
One Codex	1	8	8	8	8
SLIMM	10	10	10	10	10
Taxonomer	0	1	1	2	3
Vipie	6	0	0	0	0
VirusFinder	10	3	3	5	6

Tabla Anexo 3.4. Resultados de las métricas normalizadas en el conjunto Unclassified (0 – 10).

Herramientas	Precisión	Sensibilidad	Precisión equilibrada	Puntuación F1	MCC
BLAST	9	10	10	10	10
Centrifuge	9	7	8	8	7
CNN_150_NT	10	1	0	1	0
CNN_150_RefSeq	10	1	0	1	0
DIAMOND1	10	10	10	10	10
DIAMOND2	9	7	9	8	9
drVM	8	4	7	5	6
FastViromeExplorer	10	4	8	5	6
Kraken2	9	10	9	10	7
One Codex	10	8	9	9	0
SLIMM	10	8	9	9	9
Taxonomer	0	0	0	0	0
Vipie	8	0	6	0	0
VirusFinder	9	7	9	8	9

Tabla Anexo 3.5. Resultados de las métricas normalizadas en los conjuntos sin huésped humano (0 – 10).

Herramientas	Porcentaje de asignación
BLAST	0
Centrifuge	0
CNN_150_NT	5
CNN_150_RefSeq	10
DIAMOND1	0
DIAMOND2	0
drVM	0
FastViromeExplorer	0
Kraken2	0
One Codex	0
SLIMM	0
Taxonomer	0
Vipie	0
VirusFinder	0

Tabla Anexo 3.6. Resultados de las métricas normalizadas en los conjuntos con huésped humano (0 – 10).

Herramientas	Porcentaje de asignación
BLAST	4
Centrifuge	2
CNN_150_NT	10
CNN_150_RefSeq	1
DIAMOND1	10
DIAMOND2	6
drVM	7
FastViromeExplorer	0
Kraken2	0
One Codex	1
SLIMM	1
Taxonomer	2
Vipie	0
VirusFinder	0

Tabla Anexo 3.7. Resultados de tiempo y memoria normalizados (0 – 10).

Herramientas	Tiempo	Memoria
BLAST	0	10
Centrifuge	10	9
CNN_150 _NT	10	9
CNN_150_RefSeq	10	10
DIAMOND1	4	9
DIAMOND2	8	10
drVM	9	10
FastViromeExplorer	10	10
Kraken2	9	0
One Codex	10	ns
SLIMM	10	8
Taxonomer	ns	ns
Vipie	10	ns
VirusFinder	1	10

Referencias

1. Kumar V, Maitra SS, Shukla RN. 2015. Environmental Metagenomics: The Data Assembly and Data Analysis Perspectives. J Inst Eng India Ser A 96:71–83.
2. Suttle CA. 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol 5:801–812.
3. Lin HH, Liao YC. 2017. drVM: A new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. GigaScience 6:1–10.
4. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, De La Cruz Peña MJ, Martínez JM, Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F, Sullivan MB, Acinas SG, Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun 8.
5. Handelsman Jo. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol Mol Biol Rev 68:669–685.
6. Palermo CN, Fulthorpe RR, Saati R, Short SM. 2019. Metagenomic Analysis of Virus Diversity and Relative Abundance in a Eutrophic Freshwater Harbour. Viruses 11:792.
7. Wu Y, Simmons BA, Singer SW. 2015. MaxBin 2 . 0 : an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinforma Oxf Engl 1–2.
8. Wu Y-W, Ye Y. 2011. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples. J Comput Biol 18:523–534.
9. Le VV, Tran LV, Tran HV. 2016. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. BMC Bioinformatics 17:22.
10. Vinh LV, Lang TV, Binh LT, Hoai TV. 2015. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. Algorithms Mol Biol AMB 10:2.
11. Alneberg J, Bjarnason BS, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.
12. Chen J, Shang J, Wang J, Sun Y. 2019. A binning tool to reconstruct viral haplotypes from assembled contigs. BMC Bioinformatics 20:544.
13. Amgarten D, Braga LPP, da Silva AM, Setubal JC. 2018. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. Front Genet 9.

14. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:61–65.

15. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. Nucleic Acids Res 47:D94–D99.

16. Krenker A, Bešter J, Kos A. 2011. Artificial Neural Networks. Introd Artif Neural Netw 376.

17. Koumakis L. 2020. Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J 18:1466–1473.

18. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directionsJournal of Big Data. Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8.

19. Rawat W, Wang Z. 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 29:2352–2449.

20. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. 2020. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 9:1–12.

21. Khan A, Sohail A, Zahoora U, Qureshi AS. 2020. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53:5455–5516.

22. Garbin C, Zhu X, Marques O. 2020. Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl 79:12777–12815.

23. Kingma DP, Ba J. 2017. Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs.

24. Cobián A, Eguiarte LE. 2002. Estructura y complejidad del genoma humano. Cienc Unam 68:56–64.

25. Lawrence C. 2019. Nucleótido | NHGRI.

26. Cóntigo (Contig) | NHGRI. Genome.gov. https://www.genome.gov/es/genetics-glossary/C%C3%B3ntigo. Retrieved 23 June 2021.

27. Reinholt SJ, Baeumner AJ. 2014. Microfluidic Isolation of Nucleic Acids. Angew Chem Int Ed 53:13988–14001.

28. Pyrc K, Jebbink MF, Berkhout B, van der Hoek L. 2008. Detection of New Viruses by VIDISCA, p. 73–89. *In* Cavanagh, D (ed.), SARS- and Other Coronaviruses: Laboratory Protocols. Humana Press, Totowa, NJ.

29. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

31. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. 2008. Database indexing for production MegaBLAST searches. Bioinforma Oxf Engl, 2008/06/21 ed. 24:1757–1764.

32. Dröge J, Mchardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief Bioinform 13:646–655.

33. Zhang R, Cheng Z, Guan J, Zhou S. 2015. Exploiting Topic Modeling to Boost Metagenomic Sequences Binning. BMC Bioinformatics 16:1–12.

34. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6:25373.

35. Langenkämper D, Goesmann A, Nattkemper TW. 2014. AKE - the Accelerated k-mer Exploration web-tool for rapid taxonomic classification and visualization. BMC Bioinformatics 15:384.

36. Bileschi ML, Belanger D, Bryant D, Sanderson T, Carter B, Sculley D, DePristo MA, Colwell LJ. 2019. Using Deep Learning to Annotate the Protein Universe. bioRxiv 1–21.

37. Lebatteux D, Remita AM, Diallo AB. 2019. Toward an Alignment-Free Method for Feature Extraction and Accurate Classification of Viral Sequences. J Comput Biol 26:519–535.

38. Tisza MJ, Pastrana DV, Welch NL, Stewart B, Peretti A, Starrett GJ, Pang YYS, Krishnamurthy SR, Pesavento PA, McDermott DH, Murphy PM, Whited JL, Miller B, Brenchley J, Rosshart SP, Rehermann B, Doorbar J, Ta'ala BA, Pletnikova O, Troncoso JC, Resnick SM, Bolduc B, Sullivan MB, Varsani A, Segall AM, Buck CB. 2020. Discovery of several thousand highly diverse circular DNA viruses. eLife 9:1–26.

39. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 26:1721–1729.

40. Shang J, Sun Y. 2020. CHEER: hierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. Methods https://doi.org/10.1016/j.ymeth.2020.05.018.

41. Liesegang H. 2019. ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks. bioRxiv https://doi.org/10.1101/558171.

42. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. 2020. Identifying viruses from metagenomic data using deep learning. Quant Biol 8:64–77.

43. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368.

44. Lin HH, Liao YC. 2017. drVM: A new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. GigaScience 6:1–10.

45. Fabijanska A, Grabowski S. 2019. Viral Genome Deep Classifier. IEEE Access 7:81297–81307.

46. Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ 6:e4227.

47. Bolduc B, Youens-Clark K, Roux S, Hurwitz BL, Sullivan MB. 2017. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J 11:7–14.

48. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

49. Takeuchi F, Sekizuka T, Yamashita A, Ogasawara Y, Mizuta K, Kuroda M. 2014. MePIC, Metagenomic Pathogen Identification for Clinical Specimens. Jpn J Infect Dis 67:62–65.

50. Science E, Koslicki D, Falush D. 2016. MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. mSystems 1:1–18.

51. Loman NJ, Watson M, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM, Dodd R, Mulembakani P, Schneider BS, Muyembe-Tamfum J-J, Stramer SL, Chiu CY. 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Nat Methods 12:303–304.

52. Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15:76.

53. Minot S, Krumm N, Greenfield N. 2015. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. bioRxiv 027607.

54. Scheuch M, Höper D, Beer M. 2015. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinformatics 16:69.

55. Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. 2017. SLIMM: species level identification of microorganisms from metagenomes. PeerJ 5:1–15.

56. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk K-C, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Grard G, Leroy E, Schneider BS, Fair JN, Martínez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J, Miller S, Chiu CY, Chiu CY. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 24:1180–92.

57. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, Graf EH, Tardif KD, Kapusta A, Rynearson S, Stockmann C, Queen K, Tong S, Voelkerding KV, Blaschke A, Byington CL, Jain S, Pavia A, Ampofo K, Eilbeck K, Marth G, Yandell M, Schlaberg R. 2016. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol 17:111.

58. González-Tortuero E, Sean Sutton TD, Velayudhan V, Shkoporov AN, Draper LA, Stockdale SR, Ross RP, Hill C. 2018. VIGA: A sensitive, precise and automatic de novo VIral Genome Annotator. bioRxiv https://doi.org/10.1101/277509.

59. Lin J, Kramna L, Autio R, Hyöty H, Nykter M, Cinek O. 2017. Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. BMC Genomics 18:378.

60. Garretto A, Hatzopoulos T, Putonti C. 2019. VirMine: Automated detection of viral sequences from complex metagenomic samples. PeerJ 2019:e6695.

61. Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. 2019. VirNet: Deep attention model for viral reads identification. Proc - 2018 13th Int Conf Comput Eng Syst ICCES 2018 623–626.

62. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci 6:427–439.

63. Wang Q, Jia P, Zhao Z. 2013. VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. PLoS ONE 8:e64465.

64. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW, Wang D. 2017. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology 503:21–30.

65. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019. A primer on deep learning in genomics. Nat Genet 51:12–18.

66. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. bioRxiv 1–13.

67. Tangherlini M, Dell'Anno A, Zeigler Allen L, Riccioni G, Corinaldesi C. 2016. Assessing viral taxonomic composition in benthic marine ecosystems: Reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. Sci Rep 6:28428.

68. Chen S, Liu M, Zhou Y. 2018. Bioinformatics Analysis for Cell-Free Tumor DNA Sequencing Data. Methods Mol Biol Clifton NJ 1754:67–95.

69. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. 2012. Grinder: A versatile amplicon and shotgun sequence simulator. Nucleic Acids Res 40.

70. Taboada B, Isa P, Gutiérrez-Escolano AL, M. del ángel R, Ludert JE, Vázquez N, Tapia-Palacios MA, Chávez P, Garrido E, Espinosa AC, Eguiarte LE, López S, Souza V, Arias CF. 2018. The geographic structure of viruses in the Cuatro Ciénegas Basin, a unique oasis in northern Mexico, reveals a highly diverse population on a small geographic scale. Appl Environ Microbiol 84:1–25.

71. Taboada B, Morán P, Serrano-Vázquez A, Iša P, Rojas-Velázquez L, Pérez-Juárez H, López S, Torres J, Ximenez C, Arias CF. 2021. The gut virome of healthy children during the first year of life is diverse and dynamic. PLOS ONE 16:1–18.

72. Huson D, Mitra S, Ruscheweyh H. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560.

73. Aho AV, Hopcroft JE, Ullman JD. 1973. On Finding Lowest Common Ancestors in Trees, p. 253–265. *In* Proceedings of the Fifth Annual ACM Symposium on Theory of Computing. Association for Computing Machinery, New York, NY, USA.

74. Menzel P, Krogh A. 2015. Kaiju : Fast and sensitive taxonomic classification for metagenomics. bioRxiv 7:1–9.

75. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. 2020. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. Cell Host Microbe 28:724-740.e8.

76. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312:1355–1359.

77. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinforma Oxf Engl 26:680–682.

78. Illumina Inc. 23AD. Illumina sequencing platforms. Serv Learn. https://www.illumina.com/systems/sequencing-platforms.html. Retrieved 3 February 2023.

79. Acevedo-Anicasio A, Santoyo E, Pérez-Zárate D, Pandarinath K, Guevara M, Díaz-González L. 2021. GaS_GeoT: A computer program for an effective use of newly improved gas geothermometers in predicting reliable geothermal reservoir temperatures. Geotherm Energy 9.

DR. JEAN MICHEL GRÉVY MACQUART
COORDINADOR DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada **Desarrollo de un nuevo método computacional para discriminar taxonómicamente las secuencias de virus generadas por tecnologías de secuenciación masiva de ADN para estudios de metagenómica**, que presenta la alumna **Elizabeth Cadenas Castrejón (10010089)** para obtener el título de **Doctor en Ciencias.**

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dra. Sonia Dávila Ramos CIDC-UAEM	APROBADO	
Dr. Juan Manuel Rendón Mancha CInC-UAEM	APROBADO	
Dr. Guillermo Santamaría Bonfil INEEL	APROBADO	
Dr. Edgar Francisco Román Rangel ITAM	APROBADO	
Dr. Jorge Hermosillo Valadez CInC-UAEM	APROBADO	
Dr. Outmane Oubram FCQeI-UAEM	APROBADO	
Dra. Lorena Díaz González CInC-UAEM	APROBADO	

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LORENA DIAZ GONZALEZ | Fecha:2023-01-17 11:28:55 | Firmante
NTZcKPtURmJShD5uyL+Nm83kM91JzLVgrh/M8gnsWDBMtN4bapcp0nlhlXD5Oy8NSE2RCLrdHAMhSKBba/c8O4DnMTXDWOoE+8Qq2Xh7vLBGsaId57qtjQ6sqbeEgmTzaAO
NGBFNFFz7eyNtsZKnrhCkMnwjXciM3Sq33YaS+gfQD8tHZ2bfSK2sXp3zECbkUFmu9A13ysy15FWJ4cMqSY5nt9kgpNEclxY28xEAFlW7lux/fESyWNv6DTOqf3bDw9joZqIgCrJxI
hGkVU+0VVjW0XnnPDRFFTrGObwwggTIgbRHmwS35PZKZ0fKC2a8OCtMVJPgfYvzZHpZ9b6A5g==

GUILLERMO SANTAMARÍA BONFIL | Fecha:2023-01-17 11:38:38 | Firmante
ybPuhU7l4KWjf5sEF7Jugi6hI+PAxU0Di4U0UmuzkqsFIGPmRRyRLZ0llNJ13E1rxsTRSpwfcjfY06r7wC9TYp5JMLtq93QnoGu/xdOMRequC1iE+zSiMF6dxXtSt0d717yUULllaZCE
VS0/kcq+66204du/lGwe9AoQpzMadC6iRdhpIcIepHHgfj/IRrxmaCSw22/BTWQUegLDjCtZMesofYk9+PNCS0yS6RzNFOzeX1CDp2As8LmkhVJwozTyVCaJCHjqvyWq4i8fivLGca
KIXEaVda2JgGUfsyQWKYZmIWeS8BFlLKBtlib0IClDFmu6ek7bporfP9IlYfNLvQ==

JUAN MANUEL RENDON MANCHA | Fecha:2023-01-17 12:07:43 | Firmante
uqgbhagzIqKmuIlgAx5nD7/GQZrhlx+S2DsPTuIJVPL90NWwO36SfhlhJDQ8+ZJrUDDffr2H22iSdc2UDheClPz2MRMclruDGB7kl1otYaCGf2SSd6yOMjKE8FkwOVJjQUlg9Kx+4M8
x5nAR3YeLm8P+7xSmREgx65ifpACoe0H2qimMpRjvxhJN2T8RRIBNCqc5o41lRm3s8mm44WI1+UuOGcMFQ+mDdSKtV0NuAsVzFRvnJ/co7OgkoPWeJiAsps0Pk+cUwmVJJ3T
f/ZkozagtnRPbEYwK01808zlr9SQ0SEBoEHEViaUHJ4fUxWuXIyBgEXjFneH0eOkCoIRAcA==

JORGE HERMOSILLO VALADEZ | Fecha:2023-01-17 12:25:02 | Firmante
ScetmtUW2EkMYHUarP0JQIfuvvSskilRxAEvFGJJBtA105UTokcdyfbEKJ7Sq/+0LYeRHjx9A0tJjxYIRH8zP/YtyzMBO1tzGZ/UF5GskmuQwoAXfzvl70500PObs5uZdEVuVbj/WbZN
xR4z173CdAMhGWymsGQjS14BhDAybU5gk5ORBg+Ly3o1q78FFuLJK5rb1wwDGmFLdWgV1Gl02q41HUzOUZTERIY0WDLITaNtzHf4HofjMdw0YckoF83t54dlMFUBws0e0kpC
7yGCqAJryCpCKoENsmIXMwIetkId39t7bAnmJdUCF+A90xI0QVjmjYHS0+YlVRplDt3d5w==

EDGAR FRANCISCO ROMÁN RANGEL | Fecha:2023-01-17 14:33:02 | Firmante
bFE8cpN/tq6FPHa8hx4QHEuqXty8mXEbcKPyvnD2s4nUKwIcSLe0ymwYtZfnPHDEx7QX/IHxl/+XsAHjKFd6yOS//nhEY89mUkydrwx3pDOr003wNVq2NtkZMLGFg6zgIaHPkD+O
GLvA4oSRafQM0LOd2nyp92pJMRXrMMlk9AuPim9ynoJbZHjomDS1LhLW3kOvJgFu42N+D+n+7szPiMcrTbGODgYCwFEsccAULUtAGkTack6NWq021qroULFXzSxCijeRbEsho
8O6reHuW1mZ/5Qed6Vx5xOi5Q6HBMd5eplBjXRcCdfkPZyUlkdPsG5dgr1lsbLl5S5tMJ4hvw==

OUTMANE OUBRAM | Fecha:2023-01-18 11:07:02 | Firmante
Id3io15J3r/vTtBMT4SoTbHC0fd1zunCpxQScjRFcYJavwDP0hG244kZHQHlYNeoCxRl2fS0mISKlZUclXqPxyZ/UqOdqn39XiNtnov/3EdxwBDfgYFM8+1e/+iQij+COGM4xsErgOx3
+I4SoZsOcL6zBG9RA3j9v8ok3tRlSfFsDP1gQDEANagRNFZdA4J6deEr9wszq4l5NYTV0ENgsDfiaOpXha3BZDAb1BVTtq65gV7hK12Br9+P1LxFJPWR8qNmIAtPAR+cVdT9mH
4JDxoo4e3wnQKkYbO+HitufEQZ5jSJtIOz+De0E/Mq+3XfCQpklBkmYyg3xjTOGuoBhA==

SONIA DAVILA RAMOS | Fecha:2023-01-18 15:39:04 | Firmante
R0MM9NrMNgK6KsboqV+MX4JZ+PsGEJzOSO1zgA5cr7uAbty3kYsbuifOmrO0a/XLiMFzaIXdBtW8teAdeP0bxH5spKPCwY04HaVPiMkSIK0VtKL9pWj6pNcFNQo3tiC4gLSOFvjk
jy01lwZGAI16HZwgt9BgGOQp9M838qR4VTXQFIudnCgyYAA5YOYCnW0Vf6PCDLYmiXtB9mLQjBaZyD3t/1s/lPOcuzSqhJ+LpRPRUrXtP4qjeUqvH9I+4idk6PR6y5zWedhZv+LR
Rus/ds6wD86/P0d2T1hXG7yfjj9vF1ahgIJ+6I1oJ5ZhS2hZOMaff2omQ56nBz+vWx88Cg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

w4rVvKmSo

https://efirma.uaem.mx/noRepudio/P8bAxwf3n75xiAcMMO2Distcx2mCjezH