



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

**Reposicionamiento de fármacos como antimicrobianos
mediante aprendizaje de máquina heterólogo**

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias (Bioquímicas)

PRESENTA:

Rodrigo Andrés Nava Lara

TUTOR PRINCIPAL

Dr. Gabriel Del Río Guerra, Instituto de Fisiología Celular,
UNAM

MIEMBROS DEL COMITÉ TUTORAL

Dr. Marcelino Arciniega Castro, Instituto de Fisiología Celular,
UNAM

Dra. Romina Ma. De La Paz Rodríguez Sanoja, Instituto de
Investigaciones Biomédicas, UNAM

Ciudad De México, Septiembre, 2019



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

RECONOCIMIENTOS

Esta Tesis de Maestría se realizó bajo la dirección del Dr. Gabriel Del Río Guerra en el Laboratorio 205 Oriente, del Departamento de Bioquímica y Biología Estructural perteneciente al Instituto de Fisiología Celular, en la Universidad Nacional Autónoma de México.

El comité tutorial que asesoró el desarrollo de esta tesis estuvo formado por los siguientes profesores:

- | | | |
|---|--|---|
| ♣ | Dr. Gabriel Del Río Guerra | Instituto de Fisiología Celular, UNAM |
| ♣ | Dr. Marcelino Arciniega Castro | Instituto de Fisiología Celular, UNAM |
| ♣ | Dra. Romina Ma. De La Paz Rodríguez Sanoja | Instituto de Investigaciones Biomédicas, UNAM |

Se reconoce:

- ♣ A la Dra. María Teresa Lara Ortiz por el apoyo técnico y académico brindado.
- ♣ Al Programa de Maestría y Doctorado en Ciencias Bioquímicas de la UNAM.
- ♣ Al Programa CONACyT (números FOIN-219 y CB-252316).

ÍNDICE GENERAL	Página
<i>Índice de figuras</i> -----	1
<i>Índice de tablas</i> -----	1
<i>Índice de apéndices</i> -----	1
<i>Índice de anexos</i> -----	1
<i>Abreviaturas</i> -----	2
1. Resumen -----	3
2. Introducción -----	4
3. Antecedentes -----	5
3.1. <u>Marco teórico</u> -----	5
3.2. <u>Antecedentes inmediatos</u> -----	13
4. Justificación -----	20
5. Hipótesis -----	20
6. Objetivos -----	21
6.1. <u>Objetivo general</u> -----	21
6.2. <u>Objetivos particulares</u> -----	21
7. Metodología -----	22
8. Resultados -----	35
9. Análisis y discusión de resultados -----	54
10. Conclusiones -----	60
11. Perspectivas -----	60
12. Referencias bibliográficas -----	61
13. Apéndices -----	68
14. Anexos -----	81

ÍNDICE DE FIGURAS

Figura 1 (Los 4 tipos principales de estructuras secundarias de los PAs) -----	8
Figura 2 (Esquematación del principal experimento realizado por Maier <i>et al.</i>) -----	13
Figura 3 (Fórmula utilizada para obtener el CScore) -----	30
Figura 4 (Diagrama de flujo de la METODOLOGÍA) -----	34
Figura 5 (Evaluación de %CC sobre conjuntos de entrenamiento mayor a 90%) -----	37
Figura 6 (Evaluación de %CC para 10FCV y Split de conjuntos originales y transformados) -----	38
Figura 7 (Evaluación de %CC para 10FCV y Split de conjuntos reducidos) -----	39
Figura 8 (Promedios de %CC en clases y subclases de conjuntos de prueba) -----	40
Figura 9 (Barras de comportamiento de los mejores modelos por clase y subclase) -----	41
Figura 10 (Valores de los 5 parámetros usados para encontrar el mejor modelo) -----	43
Figura 11 (Valores de CScore para encontrar el mejor modelo) -----	44
Figura 12 (CDM por masa molecular en péptidos y CQNPs) -----	46
Figura 13 (CDM por grupos funcionales en péptidos y CQNPs) -----	47
Figura 14 (Visualización 3D del espacio de péptidos, CQNPs y heterólogos) -----	49
Figura 15 (Ecuación para obtener descriptores de tipo SP) -----	54

ÍNDICE DE TABLAS

Tabla 1 (Lista de superbugs) -----	6
Tabla 2 (Predicciones correctas en un modelo previamente generado) -----	19
Tabla 3 (Conjuntos de entrenamiento) -----	27
Tabla 4 (Conjuntos de prueba) -----	27
Tabla 5 (135 modelos generados por Auto-WEKA) -----	35
Tabla 6 (5 mejores modelos por su CScore) -----	45
Tabla 7 (20 mejores descriptores para péptidos, CQNPs y heterólogos) -----	48
Tabla 8 (Matriz de confusión para datos de descubrimiento) -----	50
Tabla 9 (Clases ATCC mayoritarias para no antibióticos con actividad antimicrobiana) -----	51
Tabla 10 (Predicciones del DiS en antibióticos de amplio espectro) -----	52

ÍNDICE DE APÉNDICES

Apéndice 1 (Lista de los 861 compuestos trabajados por Maier <i>et al.</i>) -----	68
Apéndice 2 (Abreviaturas de las 55 combinaciones algoritmo-parámetros usadas) -----	71
Apéndice 3 (Datos numéricos de la Figura 5) -----	72
Apéndice 4 (Barras de comportamiento de todos los modelos) -----	73
Apéndice 5 (Datos numéricos de las Figuras 6, 7 y 10) -----	77
Apéndice 6 (Datos numéricos del CScore de la Figura 11) -----	79
Apéndice 7 (Abreviaturas de las 11 combinaciones algoritmo-parámetros nuevas) -----	79
Apéndice 8 (Compuestos no antibióticos del DiS predichos como con actividad) -----	80
Apéndice 9 (Lista de antibióticos de amplio espectro predichos) -----	81

ÍNDICE DE ANEXOS

Anexo 1 (Artículo publicado) -----	81
---	----

ABREVIATURAS

%CC	Porcentaje de instancias clasificadas correctamente
10FCV	Validación cruzada usando 10 seccionamientos (10-Fold Cross-Validation)
aa	Aminoácido
ACP	Análisis de componentes principales
AP	Actividad primaria
ARFF	Formato de archivo con relación de atributos (Attribute-Relation File Format)
ATCC	Clasificación Anatómica, Terapéutica y Química (Anatomical, Therapeutical & Chemical)
CDM	Caracterización del dominio del modelo
CEv	1197 Compuestos Evaluados por actividad antimicrobiana o ausencia de ella
CHB	Compuesto con células humanas como blanco
CQNP	Compuesto químico no peptídico
CScore	Valor combinado obtenido mediante distancia vectorial de varios parámetros
CSV	Formato con valores separados por comas (Comma Separated Values)
DiS	Conjunto de descubrimiento (Discovery Set)
DSAMP	Conjunto de datos de péptidos antimicrobianos
DSNAMP	Conjunto de datos de péptidos no antimicrobianos
EER	Ratio de error estimado
FDA	Administración de Alimentos y Medicamentos de E.U. (Food and Drug Administration)
FN	Falso negativo
FP	Falso positivo
IGAE	Evaluador de atributos InfoGainAttributeEval
MAE	Error absoluto promedio
ML	Machine Learning
OMS	Organización Mundial de la Salud
PAs	Péptidos antimicrobianos
PNAs	Péptidos no antimicrobianos
RAE	Error absoluto relativo
ROC	Característica operativa de receptor (Receiver Operating Characteristic)
SMILES	Especificación de entrada de texto lineal molecular simplificada (Simplified Molecular Input Line Entry Specification)
TeS	Conjunto de prueba (Testing Set)
TrS	Conjunto de entrenamiento (Training Set)
VN	Verdadero negativo
VP	Verdadero positivo

1. RESUMEN

En este trabajo se identificó actividad antimicrobiana contra la microbiota intestinal en compuestos aprobados por la FDA usando modelos generados por Machine Learning (ML) de tipo heterólogo, es decir, utilizando conjuntos de entrenamiento que combinaran péptidos de distintas bases de datos y compuestos químicos no peptídicos (CQNP) obtenidos desde un estudio (Maier *et al.*, 2018), esto para combinar la abundancia de los péptidos y el uso clínico ya establecido de los CQNP. Para ello, utilizando 1444 descriptores fisicoquímicos del programa PaDEL, a partir de datos de ambos tipos de sustancias se generaron 9 tipos de conjuntos de entrenamiento (1 **SóloPéptidos**, 4 **SóloCQNP**, y 4 **Heterólogos** [estos 8 últimos con datos subdivididos de 4 maneras]) con sus correspondientes 9 conjuntos de prueba, todos con sólo CQNP. A partir de estos conjuntos se generaron 135 modelos, que al elegir los que aprendieran de sí mismos con eficacia mayor a 90%, quedaron 77, que al evaluarse con su conjunto de prueba, se vio que la manera en que se subdividen los conjuntos **SóloCQNP** y **Heterólogos** influye en el aprendizaje. De estos modelos, el mejor fue elegido con una medida combinatoria que usó 5 parámetros estadísticos (% clasif. correcta, curva ROC, error EER, %10FCV y %split). Este modelo fue después usado para hallar posible actividad antimicrobiana en un conjunto diferente de CQNP (que incluye antifungales/colorantes) aprobados por la FDA (Conjunto de descubrimiento), hallando 118 CQNP originalmente no usados como antimicrobianos, pero que podrían tener actividad contra la microbiota y en consecuencia podrían promover resistencia a antibióticos. También se usó para buscar antibióticos de amplio espectro, que serían aquellos predichos como activos contra microbiota. Se encontró que 54 antibióticos no serían de amplio espectro de acuerdo a este criterio, mientras que 59 sí podrían serlo, incluyendo 3 (amoxicilina, cefalexina, fenoximetilpenicilina) que tradicionalmente se consideraban de espectro reducido.

2. INTRODUCCIÓN

Hay reportes de compuestos químicos no peptídicos (CQNPs) aprobados por la FDA no usados como antimicrobianos, pero que podrían tener esta actividad y en consecuencia promoverían resistencia a los antibióticos. Este trabajo tiene como objetivo identificar computacionalmente actividad antimicrobiana en CQNPs ya aprobados por la FDA para una actividad primaria (AP) que no es antimicrobiana, sino de otro tipo (antipsicótica, antitumoral, etc.). Para ello, se generaron modelos de clasificación de actividad antimicrobiana de estos CQNPs aprobados, a partir de datos que reportan actividad antimicrobiana (o su ausencia) en péptidos y en otros CQNPs; estos modelos heterólogos se obtuvieron por medio de la evaluación de múltiples algoritmos de Machine-Learning (ML) sobre datos de péptidos y de CQNPs, para hallar los modelos más adecuados para este fin. Para lograrlo, se usaron datos de péptidos antimicrobianos (PAs, 8000) y no antimicrobianos (PNAs, 3546) reportados en 20 bases de datos públicas para entrenar los modelos; y también datos de CQNPs (861 CQNPs, de los que al menos 428 están aprobados por la FDA) reportados en un trabajo previo (Maier *et al.*, 2018) donde se encontró que fármacos como antipsicóticos, antineoplásicos y antidiabéticos afectan a al menos 1 de 40 cepas de la microbiota intestinal *in vitro*. Estos datos se usaron para entrenar y evaluar los modelos, identificando los más confiables. Estos modelos se utilizaron para obtener resultados de predicciones en un conjunto de descubrimiento, consistente en CQNPs aprobados por la FDA de los que no existe aún reporte de actividad contra la microbiota (829 compuestos, de los que la mayoría provino de la base de datos pública ZINC15, y algunos corresponden a ciertos antifungales y colorantes ya previamente reportados a nivel experimental (Peña-Díaz *et al.*, 2009)). También se obtuvo información sobre si ciertos antibióticos son de amplio o reducido espectro, y si existen CQNPs clasificados originalmente como no antimicrobianos que sin embargo, pudieran tener esa actividad.

3. ANTECEDENTES

3.1. MARCO TEÓRICO

EL PROBLEMA DE LA MULTIRRESISTENCIA

En los últimos años, ha surgido un problema serio dentro de los círculos de importancia pública en el sector salud: el aumento en la incidencia de enfermedades infecciosas a nivel mundial. Esta situación se ha atribuido a la aparición de microorganismos (bacterias, parásitos, hongos y virus) denominados "superbichos" o "*superbugs*", esto es, que son resistentes a una amplia gama de antibióticos de uso común tales como cefalosporinas, fluoroquinolonas, fluconazol y artemisinina, entre otros. Se han definido 2 tipos de resistencia: La múltiple, que implica resistencia a antibióticos que pertenezcan a 3 o más categorías diferentes por su mecanismo de acción; y la extendida, que implica resistencia a todos los antibióticos del antibiograma común (Chavolla-Canal *et al.*, 2016; Tanwar *et al.*, 2014). Estos *superbugs* ejercen la resistencia ya sea mediante enzimas que inhabilitan a los antibióticos, modificación del blanco del antibiótico el cual ya no lo reconoce, reducción de la permeabilidad, o expulsión mediante bombas de eflujo (Rossolini, 2016).

Esta problemática parece haber sido causada por un uso inadecuado de estas sustancias por parte de la población, y se ve agravada porque existen pacientes en situaciones inmunocomprometidas (enfermos de VIH, personas bajo tratamiento de enfermedades autoinmunes, pacientes receptores de transplantes) que son un blanco potencial de estos *superbugs*, que dadas sus características de resistencia, pueden esparcir enfermedades a un mayor número de individuos, con el consiguiente riesgo de epidemias (Tanwar *et al.*, 2014; Tonarelli y Simonetta, 2013), las cuales inciden en un aumento en las tasas de morbilidad y mortalidad a nivel mundial, así como en los costos de salud. Además, esta situación ha influido en el hecho de que cada vez se descubren menos antibióticos eficientes, lo cual conlleva el riesgo inherente de que en algún momento,

ya no existan compuestos antimicrobianos eficaces contra ciertas infecciones (Nagarajan *et al.*, 2018; Rossolini, 2016).

El problema de la aparición de organismos multirresistentes a antibióticos tiene un nivel de seriedad tal que la Organización Mundial de la Salud (OMS), expidió en 2017 una alerta en la cual enlista 12 *superbugs*, separados en 3 grupos correspondientes al nivel de urgencia que requiere tratar con el problema que representan. Estos datos se observan en la

Tabla 1.

TABLA 1. Lista de *superbugs* para los que se requieren urgentemente nuevos tratamientos antibióticos.

Superbug	Resistencia	Prioridad
<i>Acinetobacter baumannii</i>	Carbapenem	Crítica
<i>Pseudomonas aeruginosa</i>	Carbapenem	Crítica
<i>Enterobacteriaceae</i>	Carbapenem	Crítica
<i>Enterococcus faecium</i>	Vancomicina	Alta
<i>Staphylococcus aureus</i>	Meticilina	Alta
<i>Helicobacter pylori</i>	Claritromicina	Alta
<i>Campylobacter spp.</i>	Fluoroquinolona	Alta
<i>Salmonellae</i>	Fluoroquinolona	Alta
<i>Neisseria gonorrhoeae</i>	Cefalosporina	Alta
<i>Streptococcus pneumoniae</i>	Penicilina	Media
<i>Haemophilus influenzae</i>	Ampicilina	Media
<i>Shigella spp.</i>	Fluoroquinolona	Media

Datos obtenidos a partir de la página de la OMS, publicados el 12 de febrero de 2017

(<http://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed>).

Este reporte habla sobre la importancia y urgencia que representa el riesgo de contraer infecciones por alguno de estos u otros *superbugs*, dado lo cual se ha promovido la búsqueda y el desarrollo de otro tipo de sustancias con capacidad antibiótica.

CARACTERÍSTICAS DE LOS PÉPTIDOS ANTIMICROBIANOS

Algunas de las sustancias probadas para hacer frente a esta situación han sido los péptidos antimicrobianos (PAs), que se sabe que se encuentran en varias especies correspondientes a todos los niveles de la vida, y que

constituyen la primera línea de defensa de estos organismos contra diversos patógenos o competidores, pudiendo tener además otras funciones como actividad anticancerígena, liberación de prostaglandinas, quimiotaxis, promoción de angiogénesis e inducción de reparación de lesiones, entre otras (Tonarelli y Simonetta, 2013; Wang, 2015). Generalmente están codificados en el genoma y se sintetizan en los ribosomas, aunque se conocen PAs de origen no ribosomal, que pueden poseer aminoácidos (aa) no proteicos (Beisswenger y Bals, 2005; Tonarelli y Simonetta, 2013).

Se ha reportado que estos PAs presentan secuencias de 6 a 100 aa, aunque en la mayoría de ellos este rango se reduce a 10/15-40/50 aa, con un promedio de aproximadamente 23 aa (Kang *et al.*, 2014; Mahlapuu *et al.*, 2016; Polanco-González, 2009). Estos PAs generalmente poseen las siguientes características:

Los PAs son *catiónicos*, es decir, poseen predominantemente residuos con carga positiva. Esto los lleva a tener cargas netas positivas, que van en un rango de 2⁺-11⁺, aunque comúnmente su rango es de 4⁺-6⁺ (Kang *et al.*, 2014; Mahlapuu *et al.*, 2016; Tonarelli y Simonetta, 2013). Estas cargas positivas son las responsables de facilitar la interacción con las membranas bacterianas, que tienen carga negativa gracias a los lipopolisacáridos (en bacterias Gram⁻) o a los ácidos teicoicos (en bacterias Gram⁺), y a los fosfolípidos en ambos tipos de bacterias (Kang *et al.*, 2014; Mahlapuu *et al.*, 2016; Nagarajan *et al.*, 2018); siendo éste un mecanismo de especificidad sobre membranas bacterianas, puesto que la toxicidad en membranas eucariontes (por ejemplo, las de las células del paciente tratado con PAs) es menor dado que éstas tienen una incidencia mucho más baja de cargas negativas (Beisswenger y Bals, 2005; Kang *et al.*, 2014).

Los PAs son *anfipáticos* (o *anfifílicos*), lo que significa que poseen a la vez una parte hidrofílica y una parte hidrofóbica, lo cual se debe a que suelen

poseer aa polares, no polares y con cargas positivas (Mojsoska y Jenssen, 2015; Rodríguez-Plaza, Rivas-Santiago, *et al.*, 2014). Esta característica cobra importancia al momento de interactuar con las membranas del microorganismo blanco, pues los aa pueden agruparse tanto en regiones hidrofílicas como hidrofóbicas. Las regiones hidrofílicas y catiónicas tendrían interacción con las cargas negativas en la membrana, mientras que las regiones hidrofóbicas interactuarían con las cadenas lipídicas de los ácidos grasos, lo que llevaría a desestabilización y ruptura de la membrana (Kang *et al.*, 2014; Mojsoska y Jenssen, 2015).

Los PAs *presentan diversidad en su estructura secundaria*, lo que da lugar a un gran espectro de posibilidades en cuanto a mecanismos de acción y a características fisicoquímicas. De entre todos los tipos de estructura secundaria existentes en PAs, se reconocen 4 clases principales, tal como se muestra en la **Figura 1**:



Figura 1. Representación de los 4 tipos principales de estructuras secundarias de los PAs: A) PAs α -helicoidales (mostrando en azul la región catiónica, y en violeta la región hidrofóbica). B) PAs con β -plegadas (mostradas en verde). C) PAs con estructuras mixtas (puentes disulfuro mostrados en amarillo). D) PAs sin estructura determinada (mostrada en rojo). (Modificado de Mojsoska y Jensen, 2015).

PAs α -helicoidales - Son los más comunes y de los que hay más reportes. Están configurados para que de un lado de la hélice aparezcan los residuos hidrofóbicos, y del otro lado los residuos catiónicos. Esta conformación la adoptan al interactuar con las membranas, en solución acuosa su

estructura es intrínsecamente desordenada (Ageitos *et al.*, 2017; Mahlapuu *et al.*, 2016; Mojsoska y Jenssen, 2015).

PAs con β -plegadas - Son estabilizados por puentes disulfuro, lo que hace que en solución acuosa conserven su estructura. También poseen residuos catiónicos de un lado de sus láminas (Mahlapuu *et al.*, 2016; Nagarajan *et al.*, 2018).

PAs con estructuras mixtas - Poseen estructuras tanto de tipo α -hélices como de tipo β -plegadas. Algunos de ellos se producen de manera natural en el humano (Mojsoska y Jenssen, 2015).

PAs sin estructura determinada - Son los menos comunes. Normalmente no forman ni α -hélices ni β -plegadas, lo cual se debe a que poseen un alto contenido de aa específicos, tales como prolina, arginina, histidina y triptófano (Mojsoska y Jenssen, 2015; Nagarajan *et al.*, 2018).

Algunas otras características que convierten a los PAs en alternativas potenciales a los antibióticos convencionales son su tamaño pequeño (10 a 50 residuos la mayoría), su acción rápida y, especialmente, su escasa tendencia a inducir resistencia (Wang, 2015). Esta baja tendencia se debe a que el mecanismo de acción de los PAs no es ciertamente específico de un blanco, puesto que consiste 1) en la interacción con las membranas celulares, provocando su disrupción y la muerte del microorganismo; y 2) en que se pueden unir de manera electrostática a varios blancos dentro de la célula como los ácidos nucleicos y las proteínas, lo cual priva a la bacteria de las funciones vitales básicas (Kang *et al.*, 2014; Mahlapuu *et al.*, 2016; Rodríguez-Plaza, Rivas-Santiago, *et al.*, 2014). La resistencia que las bacterias muestran ante los PAs normalmente suele ser mucho más baja que la ejercida ante otras sustancias, pues no parece haber un mecanismo general por el cual las bacterias se hagan resistentes a cada PA. Existe un estudio en el que a un cultivo de *Pseudomonas aeruginosa* se le realizaron 30 pases en 2 condiciones diferentes, una en presencia de un PA, y otra en presencia del antibiótico gentamicina. En la

primera condición, la resistencia de la bacteria al PA al final de los 30 pases aumentó de 2 a 4 veces, mientras que la resistencia a gentamicina aumentó 190 veces (Yeung *et al.*, 2011).

Otra ventaja de los PAs es que la gran mayoría de ellos se produce naturalmente, lo cual reduce la necesidad de buscar estrategias para síntesis de moléculas con propiedades antibióticas.

DESVENTAJAS DE LOS PÉPTIDOS A NIVEL CLÍNICO

A pesar de todas las ventajas ya mencionadas, a la fecha son pocos los PAs que han sido aprobados para su uso clínico, debido a algunas limitantes que se ha observado que presentan. Estas inconveniencias pueden agruparse del modo siguiente:

Toxicidad, la cual puede deberse a 2 factores principales: las altas dosis que se requieren para inducir los efectos deseados, y que la especificidad por membranas bacterianas no sea suficiente y también se afecte a las membranas de las células del paciente (Kang *et al.*, 2014; Méndez-Samperio, 2014). También existe el riesgo de que se vean afectadas las bacterias de la microbiota intestinal, lo cual puede llevar a problemas de desbalance bacteriano y en consecuencia, a enfermedades (Maier *et al.*, 2018).

Incapacidad de activación, la cual puede darse debido a 2 factores: Baja biodisponibilidad, la cual normalmente se debe a que dadas las características catiónicas de los PAs, éstos se absorben pobremente en el intestino, y los que se absorben tienen dificultades para llegar a torrente sanguíneo por acción del hígado (Mahlapuu *et al.*, 2016; Méndez-Samperio, 2014; Uhlig *et al.*, 2014); y anulación por alta concentración de sales, dado que las interacciones que tienen los PAs con las membranas bacterianas son de carácter electrostático, y siendo las sales moduladores de la fuerza iónica, una alta concentración de éstas alteraría

o incluso cancelaría las interacciones (Kang *et al.*, 2014; Tonarelli y Simonetta, 2013).

Susceptibilidad a degradación, lo cual se traduce en una escasa vida media. Normalmente sucede con PAs que alcanzan a llegar a torrente sanguíneo pese a las barreras mencionadas en el punto anterior, o a PAs que tienen que ser inyectados directamente a sangre vía intravenosa; y se debe al efecto de proteasas ya sea de la bacteria blanco o del huésped, o al efecto causado por los anticuerpos del paciente (Uhlig *et al.*, 2014).

Costos de producción poco accesibles, puesto que éstos llegan a rondar los \$100-600 USD (equivalente aproximado a \$1800-10800 MXN) por gramo, lo que representa al menos 110 veces el costo de producción de un fármaco de masa molecular 10 veces más baja, costos que evitan que la producción sea escalable a grandes volúmenes (Bray, 2003; Pirtskhalava *et al.*, 2016). Esto se ha tratado de hacer reduciendo el tamaño de los PAs, o produciéndolos *de novo* (Kang *et al.*, 2014).

Para resolver estos problemas en la medida de lo posible, se ha optado por seguir otras estrategias, tales como la peptidomimética, que consiste en modificar químicamente a estos PAs de manera que mejoren características como la biodisponibilidad y la estabilidad, pero teniendo cuidado de no alterar de forma significativa ciertas características fisicoquímicas para conservar el mecanismo de acción deseado, en este caso antimicrobiano. También se denominan peptidomiméticos a los compuestos que sin derivar directamente de péptidos, presentan similitudes con cadenas laterales de aminoácidos y por tanto características fisicoquímicas, estructurales y de mecanismo de acción similares (Méndez-Samperio, 2014; Tonarelli y Simonetta, 2013).

IDENTIFICACIÓN DE CARACTERÍSTICAS DE PÉPTIDOS SIN ACTIVIDAD

Ya se mencionaron anteriormente las características que confieren la actividad antimicrobiana a un péptido de manera general. Ahora, también

es importante delimitar qué características harían que un péptido no pudiera presentar dicha actividad. Debe tomarse en cuenta que a la fecha, no existen bases de datos de péptidos que no tengan ninguna actividad antimicrobiana, lo cual puede complicar su caracterización (Porto *et al.*, 2010; Lata *et al.*, 2007; Fernandes *et al.*, 2012). Algunas aproximaciones que se han hecho para poder distinguir esto han sido utilizar péptidos obtenidos de bases constantemente curadas, como ProteinDataBank o SwissProt, o secuencias que son de tipo péptido señal o transmembranales (Thomas *et al.*, 2009; Fernandes *et al.*, 2012). Un estudio realizado apelando a esas características, encontró a nivel computacional que las características que mejor podrían definir esas diferencias son la hidrofobicidad promedio, carga neta, flexibilidad, e índices de α -helicidad y de formación de vueltas (Porto *et al.*, 2012). Otra aproximación puede ser la modificación de péptidos antimicrobianos para hacer que pierdan su actividad, ya sea por lo adición o sustitución de ciertos residuos. Esto se ha observado con los péptidos penetradores, que son péptidos que pueden internalizarse en la célula a través de la membrana plasmática. Esta característica se ha correlacionado previamente con la capacidad antimicrobiana de algunos péptidos (Rodríguez-Plaza, Morales-Nava, *et al.*, 2014; Milletti, 2012; Splith y Neundorf, 2011). Para poder hacer que un péptido penetrador pierda su actividad, se le pueden agregar residuos que contengan un carácter aniónico, y/o que lo hagan perder su estructura helicoidal al disminuir su índice de α -helicidad (Diener *et al.*, 2016; Morán-Torres, 2019).

3.2. ANTECEDENTES INMEDIATOS

ACTIVIDAD ANTIBIÓTICA EN COMPUESTOS CON OTRAS ACTIVIDADES

Además de la peptidomimética, otra aproximación que puede hacerse para buscar opciones alternativas a los PAs se da mediante la búsqueda de compuestos químicos no peptídicos (CQNPs) con actividad antimicrobiana, pero que hayan sido aprobados por la Administración de Alimentos y Medicamentos de E.U. (Food and Drug Administration, FDA) para ser utilizados en contextos no relacionados a los microbios; es decir, que sus actividades primarias (AP) reportadas sean diferentes a antimicrobianas. Esta aproximación novedosa viene del hecho de que se sabe que ciertos CQNPs que están reportados, por ejemplo, como laxantes, antihistamínicos, hormonas, antidepresivos y benzodiazepinas, pueden inducir cambios en la microbiota intestinal (Falony *et al.*, 2016). Esto impulsó el desarrollo de otro estudio, en el cual se estudió el efecto que causaban diversos CQNPs, con APs tanto antimicrobianas como no antimicrobianas, en la microbiota. En la **Figura 2** se muestra una representación esquemática del experimento principal realizado en dicho

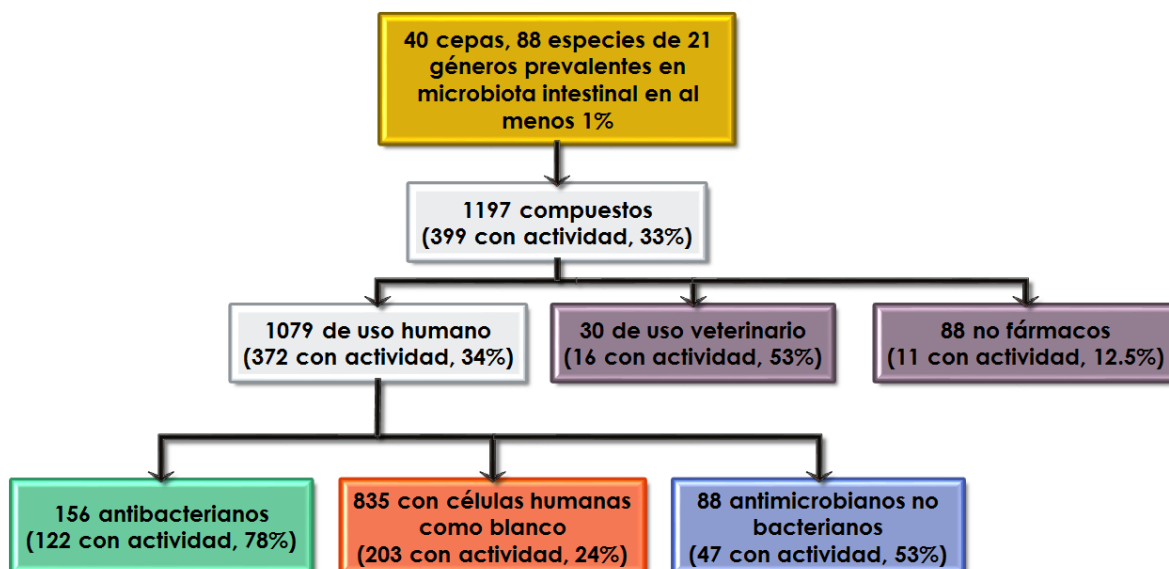


Figura 2. Representación esquemática del principal experimento realizado en un estudio donde se determinó la presencia de actividad antimicrobiana en CQNPs tanto con AP antimicrobiana como no antimicrobiana (ver texto para detalles). (Realizado a partir de Maier *et al.*, 2018).

estudio, en el cual se evaluaron 1197 CQNP de diversos tipos sobre bacterias de la microbiota intestinal. Estas bacterias correspondían a 40 cepas pertenecientes a 38 especies (estando *Escherichia coli* y *Bacteroides fragilis* representadas por 2 cepas cada una), correspondientes a 21 géneros de bacterias que se encuentran en la microbiota intestinal al menos al 1% de prevalencia, y que fueron seleccionadas además por pertenecer a individuos sanos y ser filogenéticamente diversas.

Los 1197 CQNP evaluados provienen de una base de datos llamada *Prestwick Chemical Library*, de la que se pueden separar en 3 categorías principales: Fármacos de uso veterinario (30), Compuestos no farmacológicos (metabolitos, sustancias endógenas o en fase investigativa, 88) y Fármacos destinados a humanos (1079). Estos últimos a su vez, se dividieron en 3 subcategorías: Antibacterianos (156), Antimicrobianos no bacterianos (antifúngicos, antiparasitarios y antivirales, 88) y Fármacos que tienen células humanas como blanco (CHB, donde el mecanismo de acción se da sobre proteínas, DNA o estructuras de humano, 835 compuestos).

El crecimiento de las bacterias en presencia de estos CQNP fue evaluado en medio anaerobio Gifu modificado (mGAM, el cual reproduce las condiciones nutricionales de la luz intestinal), en un periodo de 16 a 24 horas. Los resultados fueron los siguientes: en el caso de los fármacos de uso humano, se vio que 122 de los 156 antibióticos tenían actividad antimicrobiana contra al menos una de las cepas en estudio (atribuyéndose la falta de actividad de los antibióticos restantes a su incapacidad para actuar en condiciones anaerobias); 47 de los 88 antiparasitarios tenían actividad; y, notablemente, 203 de los 835 CHBs tenían actividad. Estos resultados tienen varias implicaciones: la primera, que se podría estar ante un nuevo panorama para entender el mecanismo de acción de los compuestos con actividad antimicrobiana y moduladores

de la microbiota intestinal; y la segunda, que los efectos que tienen los medicamentos en general sobre la microbiota intestinal son mayores a lo previamente proyectado, puesto que el uso sin control, aparentemente inocuo, de CHBs, podría inclusive estar causando resistencia bacteriana no sólo a los efectos antimicrobianos que éstos tienen, sino que también a los efectos que tendrían los antibióticos convencionales, por lo que sería necesario reevaluar el cómo se administran (Maier *et al.*, 2018).

NECESIDAD DE MÉTODOS SISTEMÁTICOS

Tomando en cuenta toda la información ya mencionada, se necesita utilizar métodos rápidos, sistemáticos y confiables para tratar de mitigar el problema de la resistencia microbiana, debido a que dada la amenaza que representa la aparición de superbugs (microbios resistentes a los antibióticos de uso común) para la salud mundial, y que la resistencia la pueden ocasionar tanto el mal uso de antibióticos como CHBs con actividad antibacteriana desconocida, es fundamental desarrollar estrategias que identifiquen aquellos CHBs con actividad antimicrobiana, para comenzar a planear un control también sobre el consumo de éstos. Para tratar de superar esta situación, una metodología propuesta que puede cubrir las necesidades en cuanto a rapidez, sistematicidad y confiabilidad ha sido el uso de herramientas computacionales para realizar predicciones de secuencias peptídicas con actividad antimicrobiana (PAs). Estas predicciones se pueden basar en su estructura tridimensional, o en sus propiedades fisicoquímicas. Esto se ha extendido también al estudio de los CQNPs, en donde las predicciones de CQNPs con actividades de interés se han realizado desde hace tiempo mediante algoritmos de Machine-Learning (ML) (Beltrán-Verdugo *et al.*, 2017; Fjell *et al.*, 2009; Lavecchia, 2015; Burbidge *et al.*, 2001; Napolitano *et al.*, 2013; King *et al.*, 1992).

MACHINE LEARNING

Se denomina Machine-Learning (ML) a un conjunto de metodologías por las cuales un algoritmo computacional puede, a partir de una serie de datos proporcionados, identificar patrones y características para al final mostrar el comportamiento de una serie de objetos de interés (Mullard, 2017). Esto puede ayudar a identificar las fronteras entre diferentes clases de objetos; para tal efecto, cada objeto (por ejemplo, péptidos, CQNP, etc.) debe representarse mediante un vector en el que cada elemento de éste represente algún descriptor del objeto (que pueden ser ciertas propiedades fisicoquímicas, descriptores estructurales, marcas moleculares, etc.). Existen 2 tipos principales de entrenamiento en el ML: *Aprendizaje supervisado*, en el cual al algoritmo se le proporciona una serie de datos con un comportamiento ya establecido (datos del Conjunto de Entrenamiento [TrS]), y con base en esta información, se espera que se genere un modelo que prediga el comportamiento de otra serie de datos (datos del Conjunto de Prueba [TeS]), que son del mismo tipo o similar pero en este caso se desconoce este comportamiento.

Aprendizaje no supervisado, en el cual al algoritmo sólo se le proporciona una serie de datos con un comportamiento desconocido, para que se separen los objetos en categorías de acuerdo a lo que el algoritmo pueda observar.

Adicionalmente a esta clasificación, se propone una nueva subdivisión de las metodologías de ML:

Metodologías homólogas, en las que datos de un tipo se utilizan en el TrS para predecir datos del TeS que son del mismo tipo; por ejemplo, predicción de péptidos usando péptidos.

Metodologías heterólogas, en las que datos de un tipo se utilizan en el TrS para predecir datos del TeS que son de otro tipo, o también puede ser que en un conjunto de datos se tenga un solo tipo de datos y en el otro,

varios tipos. Por ejemplo, predicción de CQNP's usando péptidos, o usando combinaciones de péptidos con CQNP's.

Un concepto adicional corresponde al *Conjunto de Descubrimiento* (DiS), que sería similar al TeS con la diferencia de que las proporciones de predicción que se obtengan no podrán ser visualizadas con porcentajes de correcto e incorrecto, puesto que de estos datos no se tiene información en lo absoluto. Normalmente estos datos se manejan una vez que ya se tiene un modelo lo suficientemente confiable previamente trabajado utilizando TrS y TeS.

Estas metodologías se han utilizado en nuestro grupo y en varios otros para predecir exitosamente varias moléculas penetradoras celulares, anticancerígenas y antimicrobianas, entre otras (Durrant y Amaro, 2015; Mullard, 2017; Veltri *et al.*, 2015; Diener *et al.*, 2016).

En este proyecto se describe por primera vez el uso de ML heterólogo para realizar reposicionamiento de CQNP's aprobados por la FDA como compuestos con actividad antimicrobiana, a partir de datos tanto de PAs como de otros CQNP's. Dentro del reposicionamiento también se puede hablar de la reclasificación de cierto tipo de sustancia para afectar un rango de blancos diferente al conocido previamente, esto es observable, por ejemplo, con antibióticos de espectro amplio o reducido.

La motivación de este trabajo se basa en una de las características del ML, es decir, se requiere conocer muchos ejemplos de moléculas con la actividad deseada para poder realizar predicciones exitosas. En las últimas dos décadas se ha acumulado información sobre PAs en la literatura que está disponible en diferentes bases de datos (Piotto *et al.*, 2012; Pirtskhalava *et al.*, 2016; Thomas *et al.*, 2009); esta información alcanza la cifra de aproximadamente 19,000 secuencias de PAs conocidas a la fecha. En comparación, el número de CQNP's conocidos con actividad antimicrobiana no es tan grande (A febrero de 2018 había 740 en la base de datos ZINC15). Con base en esta información, la propuesta de este

proyecto es entrenar modelos basados en algoritmos de ML para distinguir moléculas con propiedades fisicoquímicas típicas de PAs; estas propiedades fisicoquímicas deben ser comunes a CQNP de tal suerte que a partir de PA se puedan predecir CQNP con actividad antimicrobiana. Adicionalmente se buscaría que tengan algunas otras características típicas de los PAs, tales como baja tendencia a producir resistencia.

ANTECEDENTES EN EL GRUPO DEL USO DE ML HETERÓLOGO

Ya se ha realizado una primera aproximación a la predicción de CQNP con capacidad antimicrobiana, la cual consistió en entrenar 2 algoritmos de ML (llamados Random-Forest y Support Vector Machine) de manera que predijeran PAs y péptidos no antimicrobianos (PNAs), así como CQNP (en este caso colorantes) antimicrobianos y no antimicrobianos, a partir de una serie de datos sobre estas 4 clases de moléculas, por lo cual el aprendizaje fue de tipo supervisado. La predicción de los PNAs y de los CQNP no antimicrobianos es esencial para definir correctamente el espacio en que se encuentran los PAs y los CQNP antimicrobianos.

Para esto, se generaron 2 tipos de modelo: En el primero, el TrS consistió únicamente de datos de péptidos (3456 registros, en el que se incluían 2338 PAs y 1118 PNAs), mientras que el TeS incluía solamente datos de CQNP (119 colorantes catiónicos y anfipáticos en este caso, de los que 100 eran antimicrobianos y 19 no antimicrobianos). Esta primera aproximación dio como resultado una predicción en la que todos los CQNP se consideraron no antimicrobianos, presuntamente porque el espacio en el que residían era distinto al de los péptidos.

Esto llevó a la necesidad de entrenar un segundo modelo, en el cual se agregaron CQNP antimicrobianos y no antimicrobianos al TrS. Utilizando el algoritmo Random-Forest, se generaron 2 TrS, uno de moléculas antimicrobianas y otro de moléculas no antimicrobianas, de 150 moléculas cada uno: el 90% de cada conjunto (135 moléculas)

correspondían a una selección al azar de los péptidos correspondientes, mientras que el 10% (15 moléculas) correspondía a CQNP's al azar.

TABLA 2. Proporción de predicciones correctas obtenidas utilizando un modelo entrenado con péptidos y con CQNP's. Los porcentajes de acierto son destacables en el caso de CQNP's.

Conjuntos	Cantidad	Predicción correcta	Porcentaje
PAs	2338	2338	100%
PNAs	1118	1118	100%
Total péptidos	3456	3456	100%
CQNP's antim.	100	72	72%
CQNP's no antim.	19	17	89.47%
Total CQNP's	119	89	74.79%
Total antim.	2438	2410	98.85%
Total no antim.	1137	1135	99.82%
Total general	3575	3545	99.16%

Datos obtenidos y calculados a partir de Peláez-Coyotl, datos no publicados.

Este modelo arrojó 100% de precisión en las predicciones tanto de PAs como de PNAs, mientras que para colorantes en general, este porcentaje fue de 74.79%, tal como se observa en la **Tabla 2**. Con esto se probó que péptidos y CQNP's pueden ser evaluados con el mismo conjunto de Descriptores. Para definir si había diferencias entre la importancia de los descriptores usados por un modelo u otro, se definieron los descriptores más importantes que los algoritmos utilizan dependiendo de si se entrenan solamente con PAs y PNAs, o si también se entrenan con CQNP's. Se determinó que esa importancia sí presenta variaciones, observando que las características más importantes para distinguir solamente péptidos están relacionadas principalmente con la estructura molecular (características tipo ETA), mientras que las usadas para diferenciar conjuntos de péptidos y CQNP's tienen como características más importantes las relacionadas a electronegatividad, estado intrínseco e ionización (Peláez-Coyotl, datos no publicados).

4. JUSTIFICACIÓN

Sabiendo que existen compuestos químicos no peptídicos (CQNPs) aprobados por la FDA que no se usan como antimicrobianos, pero que podrían tener esta actividad y en consecuencia podrían afectar a la microbiota; que existen antibióticos de amplio espectro que también pudieran afectarla como efecto secundario; y que las bacterias que conforman a ésta son muy diversas y de forma práctica no es viable ensayar todos los CQNPs experimentalmente, se propone utilizar una estrategia computacional que detecte esta actividad. Sabiendo que la cantidad de datos computacionales que existen sobre CQNPs ya aprobados por la FDA no es muy abundante; y que existe una cantidad importante de datos sobre actividad antimicrobiana en péptidos que, sin embargo, no pueden ser usados clínicamente; se propone utilizar ambos tipos de sustancias para reforzar las predicciones sobre actividad antimicrobiana en la microbiota.

5. HIPÓTESIS

Moléculas con actividad antimicrobiana y de composición diferente (péptidos y CQNP) comparten un patrón de descriptores fisicoquímicos que al combinarse, pueden generar una predicción que influya en el reposicionamiento y la reclasificación de fármacos.

6. OBJETIVOS

6.1. OBJETIVO GENERAL

Identificar actividad antimicrobiana en CQNPs aprobados por la FDA, con actividades primarias (APs) diferentes ya reportadas, mediante ML heterólogo.

6.2. OBJETIVOS PARTICULARES

- 👉 Crear conjuntos de entrenamiento (TrS) y de prueba (TeS) tanto de péptidos antimicrobianos (PAs) y no antimicrobianos (PNAs) como de CQNPs antimicrobianos y con otras actividades reportadas.
- 👉 Generar y evaluar modelos de predicción de actividad de CQNPs aprobados por la FDA.
- 👉 Elegir el modelo más adecuado para luego aplicarlo a un conjunto de descubrimiento (DiS).

7. METODOLOGÍA

7.1. OBTENCIÓN DE DATOS

Péptidos

Se utilizó un conjunto de péptidos ya reportados como antimicrobianos (conjunto DSAMP), y otro de péptidos ya reportados sin actividad antimicrobiana (conjunto DSNAMP). Ambos conjuntos fueron construidos a partir de registros de 20 diferentes bases de datos (ver artículo en prensa del **Anexo 1**) y en los que ya vienen reportadas actividades diversas asociadas a cada péptido (Beltrán-Verdugo *et al.*, 2017).

CONJUNTO DE PÉPTIDOS YA REPORTADOS COMO ANTIMICROBIANOS (DSAMP): Este conjunto contenía originalmente 37251 instancias, y se le realizaron varias depuraciones, incluyendo: I) para eliminar registros repetidos, II) eliminar registros sin reporte de actividad antimicrobiana, III) conservar sólo péptidos de 10-50 aa y que sólo tuvieran los 20 aa naturales, y IV) eliminar la posibilidad de tener péptidos cíclicos; con esto se llegó a 8000 registros de péptidos con actividad. Estos registros fueron convertidos de formato FASTA (el cual muestra las secuencias ya sea de DNA o proteínas [en este caso serían proteínas], representando con letras cada nucleótido o aminoácido) a formato SMILES (que representa en una sola línea la estructura de cualquier molécula orgánica, mediante letras y símbolos que representan enlaces o ramificaciones de cierto tipo, ver ABREVIATURAS) con base en un estudio previamente reportado (Minkiewicz *et al.*, 2017).

CONJUNTO DE PÉPTIDOS YA REPORTADOS SIN ACTIVIDAD ANTIMICROBIANA (DSNAMP): El conjunto DSNAMP inicialmente constaba de 3 listas de péptidos sin reporte de actividad antimicrobiana (Fernandes *et al.*, 2012; Thomas *et al.*, 2009; Xiao *et al.*, 2013) que en total sumaban 5566 péptidos. Después de aplicar la depuración por tamaño de 10-50 aa (única que podía realizarse en este caso), se llegó a 3546 péptidos, que también fueron convertidos a formato SMILES.

Compuestos no peptídicos (CQNP)

DATOS PARA ENTRENAMIENTO Y PRUEBA (TrS, TeS): Los datos de CQNP para generar los modelos se tomaron del ya citado estudio sobre el efecto de CQNP para humanos en la microbiota intestinal (Maier *et al.*, 2018), donde 1197 fueron los compuestos evaluados (CEv).

Los nombres de estos CEv fueron buscados en la base de datos ZINC15 (<https://zinc15.docking.org/>) (Sterling y Irwin, 2015), encontrando 964 de ellos. De los 233 restantes no se encontró registro alguno, debido en su mayoría a que el nombre en ZINC15 no era exactamente el referido en el estudio. Se adaptó la búsqueda y se encontraron datos de 189 de ellos. Los 44 faltantes se buscaron de manera manual, encontrándose 34 de ellos en ZINC. Por lo que en total fueron hallados códigos para 1185 compuestos, y sólo 12 compuestos se descartaron.

Posteriormente se realizó la depuración de los datos, que consistió en eliminación de códigos ZINC repetidos, curación de datos mal anotados, y eliminación de datos de metabolitos no presentes en el estudio. También se eliminaron datos de compuestos quirales de los que no había información disponible sobre qué estereoisómero usar, o si se podían administrar como mezclas racémicas.

A partir de esto, los CEv fueron agrupados en 10 categorías distintas:

Antibacterianos con y sin actividad antimicrobiana (contra alguna de las 40 cepas usadas en el estudio), **Antiparasitarios con y sin actividad antimicrobiana, Para células humanas (CHBs) con y sin actividad antimicrobiana, No fármacos** (es decir, metabolitos o endógenos) **con y sin actividad antimicrobiana, y de uso veterinario con y sin actividad antimicrobiana.** De estas categorías, las que se usaron para construir los Conjuntos de Entrenamiento (TrS) fueron únicamente las correspondientes a fármacos que se pueden utilizar en humanos, descartando fármacos de uso veterinario y sustancias no farmacológicas; lo que llevó la cantidad de registros usados a 861,

correspondientes a 807 CQNP. En el **Apéndice 1** se puede observar una lista con todos estos compuestos. A estos registros se les extrajo el código SMILES tanto en conjunto como en las subcategorías ya mencionadas.

DATOS PARA CONJUNTO DE DESCUBRIMIENTO (DiS): Los datos de CQNP de los que se planteó obtener información novedosa se obtuvieron también a partir de la base ZINC15, en este caso utilizando los códigos ATCC (Anatomical Therapeutic Chemical Classification), que subdividen a las sustancias de acuerdo a su actividad. De aquí se obtuvieron 3727 compuestos, que se agruparon en 3 clasificaciones: Antimicrobianos (contra bacterias, hongos, protozoarios y virus) con 740 registros; Antiparasitarios eucariontes (Helminticidas, repelentes, ectoparasiticidas) con 59 registros, y compuestos No reportados (todos aquellos planteados para tener una actividad primaria [AP] diferente a la antimicrobiana), con 2928 registros. Estos datos también fueron depurados para eliminar repetidos, agrupar los compuestos en las 3 clases ya mencionadas, eliminar compuestos ya presentes en los TrS, y sólo incluir compuestos aprobados por la FDA. Con esto se llegó a 103 antimicrobianos, 8 antiparasitarios y 645 sin actividad antimicrobiana reportada, para un total de 756 compuestos.

Adicionalmente, se agregó una lista de 106 CQNP adicionales (trabajo previo en el grupo de Erika Peláez-Coyotl disponible en el siguiente link: <https://github.com/MidoriR/ColorantesPena>), donde se reportaba actividad o su ausencia para algunos compuestos); que incluye 4 compuestos aprobados por la FDA (Ticlopidina, Epinastina, Asenapina y Ácido mefenámico) para un uso distinto al antimicrobiano (por ejemplo, anti-psicóticos) que experimentalmente Peláez-Coyotl evaluó su actividad antimicrobiana; y se revisó cuáles de estos CQNP se repetían tanto con el TrS como con el DiS mediante comparación con su código ZINC, encontrándose que existía para todas las sustancias excepto 3 de ellas, que fueron descartadas; y que de los 106 compuestos, 18 ya se

encontraban en el TrS mientras que 15 se encontraban ya en el DiS, por lo que los restantes 73 (que incluyen 22 antifungales) fueron adicionados al DiS para ampliarlo a 829 compuestos. Estos compuestos se agregaron dentro de 4 clasificaciones nuevas dependiendo de lo observado (Antifungales con reporte de actividad, Antifungales sin reporte, Colorantes activos y Colorantes inactivos). También se obtuvieron sus códigos SMILES.

7.2. OBTENCIÓN DE PREDICCIONES SOBRE CQNPs

Obtención de descriptores

Los descriptores estructurales y fisicoquímicos asociados a péptidos y CQNPs se calcularon utilizando el programa PaDEL-descriptor (Yap, 2011), del cual se obtuvieron 1444 descriptores en 1 y 2 dimensiones, cuyos tipos pueden consultarse en el siguiente link como archivo de Excel:

<http://www.yapcsoft.com/dd/padeldescriptor/Descriptors.xls>

Estos descriptores se obtienen mediante el cálculo matemático o la observación experimental de una característica determinada en la representación simbólica de alguna molécula, generando un valor numérico (Yap, 2011). Algunos de estos descriptores utilizados son los siguientes: nAcid, CrippenLogP, SP-0, TopoPSA y Zagreb, entre otros.

Todos los archivos obtenidos en SMILES fueron procesados de manera que PaDEL-descriptor los reconociera, previo a la ejecución del programa.

Procesamiento de archivos con valores separados por comas (CSV)

PaDEL generó archivos con los descriptores calculados en formato CSV. Después se sustituyeron todos los valores nulos (valores de descriptores que PaDEL no pudo calcular para ciertas moléculas) por el valor 0, mediante un comando en la terminal de Ubuntu específico para ello. Después, se eliminó la primera columna (con el nombre) y se agregó al

final una columna correspondiente a la clase (es decir, si la sustancia es o no antimicrobiana). Realizado esto, los conjuntos se convirtieron a formato de relación de atributos (ARFF) para poderlos trabajar en el programa WEKA (Frank *et al.*, 2004; Mark *et al.*, 2009), en el que se pueden realizar predicciones sobre un conjunto de datos utilizando algoritmos matemáticos de aprendizaje de máquina para realizar minería y procesamiento de datos, y predicción de alguna característica basándose en otro conjunto de datos, que se utilizan para entrenar. De este procedimiento surgieron los llamados conjuntos originales.

Construcción de los conjuntos

Una vez generados los conjuntos, se procesaron y ordenaron para categorizarlos en los 3 tipos siguientes tipos de TrS:

SóloPéptidos - Péptidos antimicrobianos y no antimicrobianos (1 conjunto). Corresponden 11546 instancias.

SóloCQNPS - CQNPs antimicrobianos y no antimicrobianos, divididos en 4 conjuntos. Hay 431 instancias en 2 fracciones, y 430 en las otras 2.

Heterólogos - Péptidos antimicrobianos y no antimicrobianos; CQNPs antimicrobianos y no antimicrobianos, divididos en 4 conjuntos. Consta de 6204 instancias en 2 fracciones, y 6203 en las otras 2.

La separación en 4 conjuntos obedece a que esto permitió separar conjuntos de prueba (TeS) de los conjuntos de entrenamiento (TrS). Para realizar estos seccionamientos a los conjuntos de tipo **SóloCQNPS** y **Heterólogos**, se procesaron de las siguientes maneras:

- 👉 Eligiendo las líneas pares de los conjuntos.
- 👉 Eligiendo las líneas impares de los conjuntos.
- 👉 Eligiendo la primera mitad de los conjuntos.
- 👉 Eligiendo la segunda mitad de los conjuntos.

TABLA 3. Conjuntos de entrenamiento.

Conjunto TrS	Instancias	Descripción
SoloPeptidos	11546	8000 péptidos antimicrobianos, 3546 péptidos sin actividad antimicrobiana reportada
SoloCQNP1	431	164 CQNP1 antimicrobianos, 267 CQNP1 sin actividad antimicrobiana reportada
SoloCQNP2	430	164 CQNP2 antimicrobianos, 266 CQNP2 sin actividad antimicrobiana reportada
SoloCQNP3	430	164 CQNP3 antimicrobianos, 266 CQNP3 sin actividad antimicrobiana reportada
SoloCQNP4	431	164 CQNP4 antimicrobianos, 267 CQNP4 sin actividad antimicrobiana reportada
Heterologo1	6204	4164 compuestos antimicrobianos (4000 péptidos y 164 CQNP1), 2040 compuestos no antimicrobianos (1773 péptidos y 267 CQNP1)
Heterologo2	6203	4164 compuestos antimicrobianos (4000 péptidos y 164 CQNP2), 2039 compuestos no antimicrobianos (1773 péptidos y 266 CQNP2)
Heterologo3	6203	4164 compuestos antimicrobianos (4000 péptidos y 164 CQNP3), 2039 compuestos no antimicrobianos (1773 péptidos y 266 CQNP3)
Heterologo4	6204	4164 compuestos antimicrobianos (4000 péptidos y 164 CQNP4), 2040 compuestos no antimicrobianos (1773 péptidos y 267 CQNP4)

En las **Tablas 3** y **4** se muestran los detalles relacionados a los TrS y a los TeS, respectivamente. Una vez realizadas estas adaptaciones, los CSV se convirtieron en ARFF usando la herramienta *ARFFViewer* de WEKA.

TABLA 4. Conjuntos de prueba.

Conjunto TeS	Instancias	Descripción
SoloPeptidos	861	328 CQNP1 antimicrobianos, 533 CQNP1 sin actividad antimicrobiana reportada
SoloCQNP1	430	164 CQNP1 antimicrobianos, 266 CQNP1 sin actividad antimicrobiana reportada. Igual al TrS de SoloCQNP2.
SoloCQNP2	431	164 CQNP2 antimicrobianos, 267 CQNP2 sin actividad antimicrobiana reportada. Igual al TrS de SoloCQNP1.
SoloCQNP3	431	164 CQNP3 antimicrobianos, 267 CQNP3 sin actividad antimicrobiana reportada. Igual al TrS de SoloCQNP4.
SoloCQNP4	430	164 CQNP4 antimicrobianos, 266 CQNP4 sin actividad antimicrobiana reportada. Igual al TrS de SoloCQNP3.
Heterologo1	430	Igual al TeS de SoloCQNP1.
Heterologo2	431	Igual al TeS de SoloCQNP2.
Heterologo3	431	Igual al TeS de SoloCQNP3.
Heterologo4	430	Igual al TeS de SoloCQNP4.

Reducción de dimensiones

Posteriormente se realizó una reducción de dimensiones de los datos mediante un análisis de componentes principales (ACP) en WEKA. Para esto fue necesario cambiar algunos valores resultantes de los descriptores

(*Infinity*) por valores numéricos (0 o 99,999,999 [99 millones]). Este proceso generó 2 nuevos conjuntos (conjuntos transformados) por cada conjunto original; a los conjuntos transformados se les aplicó el ACP, obteniendo así los conjuntos reducidos.

Generación de modelos

La selección de los mejores modelos para clasificar CQNP con actividad antibiótica contra la microbiota intestinal se realizó en 4 fases: I) Entrenamiento, II) Validación, III) Evaluación y IV) Selección. En cada fase se estimaron parámetros de rendimiento de los modelos obtenidos para con ello seleccionar al mejor modelo.

a) Fase de entrenamiento.

Los conjuntos ARFF generados anteriormente fueron procesados utilizando la herramienta Auto-WEKA, que es un algoritmo integrado a WEKA que busca la mejor combinación algoritmo-parámetros a utilizar en un determinado conjunto de datos (Thornton *et al.*, 2012; Kotthoff *et al.*, 2016). Ésta se usó para poder seleccionar el algoritmo y parametrización más adecuados para cada conjunto, generando el mejor modelo posible. Los tiempos límite utilizados para buscar los mejores modelos con Auto-WEKA fueron 10, 90, 720, 2880 y 4320 minutos.

Los conjuntos transformados se trabajaron bajo las mismas condiciones que Auto-WEKA dictaminó para su conjunto original correspondiente, mientras que los conjuntos reducidos fueron sometidos a su propio análisis por Auto-WEKA. De los 118 modelos diferentes generados, se seleccionaron aquellos cuyo porcentaje de instancias clasificadas correctamente (%CC) fuera mayor a 90%, con lo que quedaron 77 modelos. Los resultados de todas las corridas se compararon con el resultado que arrojó el algoritmo ZeroR para cada conjunto, puesto que

ZeroR elige como válida para todos los datos la clase más representada en éstos, definiendo así una clasificación al azar.

b) Fase de validación.

Una vez hecho este proceso, se utilizó WEKA para obtener la validación de los modelos asociados a los conjuntos previamente obtenidos. Para ello, a los datos de los 77 modelos elegidos anteriormente, se les realizaron 8 corridas de validación cruzada (*10-fold-cross*, 10FCV) y 4 corridas de validación por segmentación (*split*, donde un porcentaje del conjunto se usó para entrenar y otro para probar). El porcentaje utilizado para entrenar en conjuntos de **SóloPéptidos** fue de 70%, en los de **SóloCQNPS** fue de 62%, y en los **Heterólogos** fue de 67%. Para obtener un resultado representativo, las 8 10FCVs y los 4 splits generados por cada conjunto se promediaron.

c) Fase de evaluación.

Como ya se mencionó, se prepararon los Conjuntos de Entrenamiento (TrS) cada uno con su correspondiente conjunto de prueba (TeS) de manera que no se sobrelaparan sus datos (ver *Construcción de los conjuntos*). Los TeS fueron procesados para que el atributo de la clase fuera compatible con el de sus TrS y, en el caso de los TeS de conjuntos reducidos, para que la reducción de dimensiones fuera la misma. Los 77 modelos obtenidos durante el entrenamiento fueron evaluados usando su correspondiente conjunto de prueba.

Adicional a los valores obtenidos por las validaciones 10FCV y Split obtenidas anteriormente, en esta fase se guardaron los porcentajes %CC y el área bajo la curva de característica operativa de receptor (ROC), así como cuáles instancias se predecían correctamente y cuáles no para cada modelo. De esta lista se generaron barras de comportamiento, para indicar visualmente en qué zonas del conjunto hay mayor tendencia a

presentarse predicciones correctas o incorrectas, y explicar visualmente los datos numéricos de la matriz de confusión (relación de la pertenencia de los datos a su clase, comparada con la clase predicha por el algoritmo usado) arrojada para cada TeS.

d) Fase de selección.

Se esperaba que más de un modelo tuviera rendimientos adecuados en cada uno de los parámetros seleccionados. Para evaluar si estos modelos eran redundantes (es decir, que aprendieron a clasificar con fronteras muy semejantes), se evaluó si las clasificaciones de estos intersectaban, es decir, que las instancias evaluadas como correctas o incorrectas de cada uno fueran iguales en al menos 50%. Se compararon únicamente los modelos que tenían los mismos atributos.

Adicionalmente, como se anticipó que más de un modelo tendría rendimientos de clasificación muy parecidos en más de un parámetro

$$\text{CScore} = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{V_{\max_i} - V_{\text{ex}_i}}{V_{\max_i} - V_{\min_i}}}$$

Figura 3. Fórmula utilizada para calcular la distancia vectorial (CScore) para n parámetros. V_{\max_i} indica el valor máximo observado del parámetro en el i-ésimo conjunto; V_{\min_i} indica el valor mínimo, y V_{ex_i} indica el valor observado.

usado para evaluar su eficacia, se utilizó un valor único que representara a los 5 parámetros ya mencionados, se utilizó una fórmula para buscar un valor combinatorio (CScore) de manera que los parámetros fueran normalizados entre 0 y 1 para poder calcular la distancia entre los valores existentes y los valores máximos para cada parámetro (Del Río-Guerra *et al.*, 2009). Esta fórmula se muestra en la **Figura 3**.

Para usar la fórmula se utilizó el script *CombinedScore.java*, tomando en cuenta que entre menor fuera el valor arrojado para un modelo, mejor

era el desempeño de este último. Los mejores modelos se eligieron tomando en cuenta que el valor del CScore fuera menor a 0.3.

Dado que la sustitución de valores nulos por 0, y valores "Infinity" por 0 y 99 millones (ver *Construcción de los conjuntos*) no es una estrategia convencional, se procesó el conjunto de datos del mejor modelo usando procesamientos convencionales para conjuntos de datos faltantes (conservar los datos nulos tal cual; eliminar las columnas de descriptores con datos nulos, estimar los valores nulos mediante promedio o mediante análisis por vecinos más cercanos; eliminar las columnas de descriptores con datos Infinity, o estimarlos ya fuera mediante promedio o vecinos más cercanos). Todo esto se realizó en WEKA, se corrieron los conjuntos en Auto-WEKA y se obtuvieron 22 modelos a los que se les extrajeron los datos correspondientes a los 5 parámetros estadísticos elegidos, así como el valor de CScore.

Caracterización del dominio del modelo (CDM)

En este estudio se evaluó la masa molecular de péptidos y de CQNPs de los TrS, y la prevalencia de grupos funcionales en ambos tipos de datos. El estudio de los rangos de masa molecular para CQNPs se realizó obteniendo éstas como un descriptor obtenido desde PaDEL-descriptor (ver *Obtención de descriptores*). Para graficar, los datos se agruparon en bins de 300 Da.

Para la CDM por grupos funcionales, se utilizó un script de Python obtenido a partir de un estudio que halló 3080 grupos funcionales en los compuestos de la base de datos ChEMBL (Ertl, 2017). Para poder ejecutar el script, fue necesario previamente instalar el paquete "rdkitpy", el cual contiene software para trabajar datos de quimioinformática y ML. Una vez instalado, el script se ejecutó y se obtuvo para cada compuesto, cuántos y cuáles grupos funcionales se hallaron. El resultado de este conteo arrojó 274 patrones moleculares distintos, que se agruparon en 45 grupos

funcionales diferentes. Se evaluó la representatividad de cada grupo y se llegó a 35 grupos funcionales que se usarían para posteriormente obtener las gráficas correspondientes.

Importancia de atributos y visualización de los datos

Se determinó qué descriptores eran los más importantes para los conjuntos de los mejores modelos utilizando la herramienta "Select attributes" de WEKA, con el evaluador de atributos *InfoGainAttributeEval* (IGAE), que se encarga de evaluar el nivel de ganancia e importancia de información para cada atributo o descriptor al respecto de la clase evaluada en un conjunto de datos. Posteriormente se utilizó el paquete adicional *scatterPlot3D* para su visualización tridimensional. Preferentemente se escogieron los 3 descriptores más importantes de acuerdo a IGAE, pero en algunos casos dependiendo de lo observado, se llegaron a escoger atributos con importancia más baja, pero siempre de entre los 20 más importantes.

7.3. OBTENCIÓN DE PREDICCIONES SOBRE DATOS DEL DiS

Uso del mejor modelo en el conjunto de descubrimiento.

El mejor modelo seleccionado con el CScore se usó para identificar CQNPs en el conjunto de descubrimiento (DiS). Para esto fue necesario reducir las dimensiones del DiS para que fueran las mismas que las del mejor modelo (ver sección de RESULTADOS).

Se debe tomar en cuenta que hay una pérdida de información al reducir las dimensiones, aunque también que no todos los 1444 descriptores funcionan para separar los datos en todos los conjuntos, con lo que reducir una vez funciona a nivel de mantener un buen nivel de predicción, y un menor tiempo de trabajo computacional.

Se obtuvo posteriormente la matriz de confusión para el mejor modelo sobre el DiS. Una vez realizado este procedimiento, se determinaron las instancias de interés del DiS de 2 maneras:

Primero se evaluó si los compuestos del DiS sin reporte de actividad antimicrobiana podrían tenerla. Para ello se revisó cuáles datos de compuestos no antibióticos se predecían como con actividad contra la microbiota, y a qué código del Sistema de Clasificación Anatómica, Terapéutica y Química (ATCC) pertenecían sus compuestos. Como algunos datos del DiS correspondían a varios estereoisómeros del mismo compuesto, para efectos de este análisis se contó un solo isómero por sustancia, con lo que se pasó de 61 registros de FN y de 140 de FP, a 54 y 118 registros, respectivamente. Se realizó el mismo análisis observando el comportamiento del mejor modelo respecto a su TeS, para realizar comparaciones entre ambos y ver si esos tipos de compuestos eran candidatos a un análisis de reposicionamiento.

Adicionalmente se evaluaron los antibióticos predichos como con actividad contra la microbiota, puesto que tendrían que ser clasificados de amplio espectro dado que atacarían a las bacterias de la microbiota intestinal (Alcock *et al.*, 2014; Maier *et al.*, 2018). Por lo anteriormente explicado, se buscó información sobre los antibióticos del DiS en diversos estudios, para verificar si se reportaban como de espectro amplio o reducido (Sarpong y Miller, 2015; Kreitmeyr *et al.*, 2017; Di Pentima y Chan, 2010; Newman *et al.*, 2012; Newland *et al.*, 2012).

En la **Figura 4** se puede observar un diagrama de flujo que resume de manera general el procedimiento mostrado en esta sección.

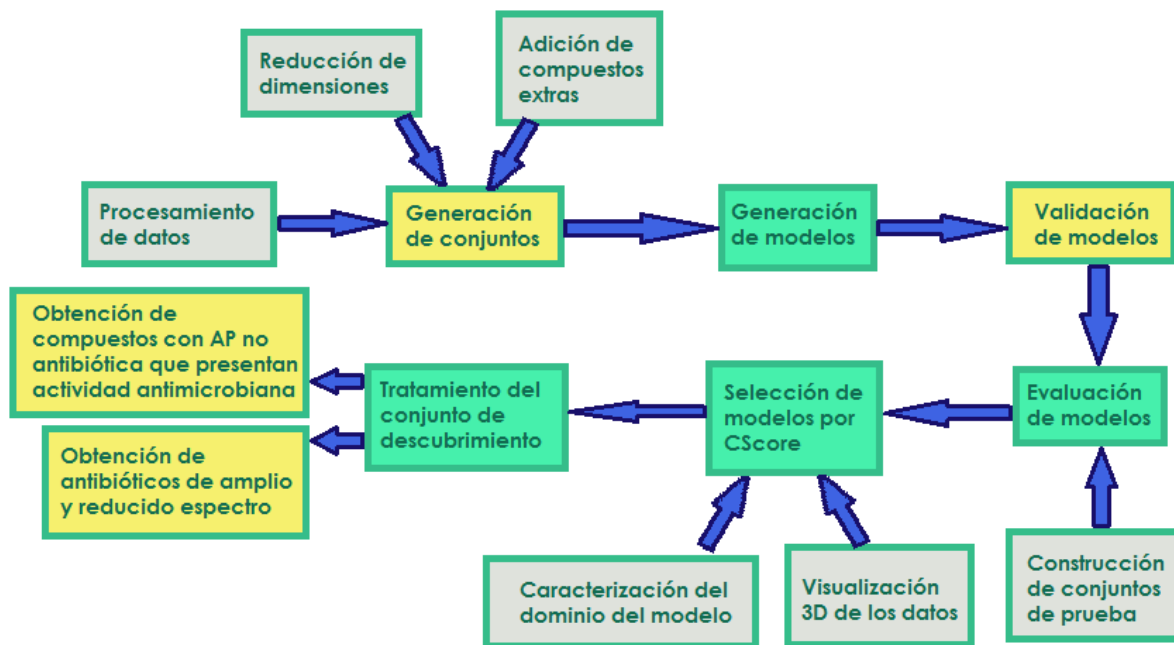


Figura 4. Diagrama de flujo que resume la sección de METODOLOGÍA. En color verde los pasos cruciales del procedimiento, en amarillo pasos más específicos, y en gris pasos no lineales necesarios para conseguir los de ambos tipos anteriores.

8. RESULTADOS

8.1. GENERACIÓN DE MODELOS

Resultados de Auto-WEKA

Después de obtenerse los conjuntos de datos para trabajar y llevar a cabo su reducción de dimensiones, se utilizó la herramienta Auto-WEKA

TABLA 5. Lista de los 135 modelos generados por Auto-WEKA antes de filtrar para eliminar modelos iguales. SP=SóloPéptidos. SNP = SóloCQNPS. Het = Heterólogo. I-0 = Conjunto reducido desde la transformación Infinity-0. I-99m = Conjunto reducido desde la transformación Infinity-99 millones. No se trabajaron conjuntos transformados (ver texto). Se observa la alta prevalencia de la combinación CRF10 en los 3 tipos de modelo.

Set	Tiempo	Combinación	Set	Tiempo	Combinación	Set	Tiempo	Combinación
SP	10	CRF10	SNP3	10	CRF10	Het2	10	CRF10
SP	90	CRF10	SNP3	90	CRF10	Het2	90	CRF10
SP	720	CRF10	SNP3	720	CRF10	Het2	720	CRF10
SP	2880	CRF10	SNP3	2880	CBagg_CLogis	Het2	2880	CJ48
SP	4320	CRF10	SNP3	4320	CBagg_CLogis	Het2	4320	CRF10
SP-I-0	10	CRF10	SNP3-I-0	10	CSMO28_CSVRBFKe	Het2-I-0	10	CRF10
SP-I-0	90	CMLPP90	SNP3-I-0	90	CSGD2,043	Het2-I-0	90	CIBk1
SP-I-0	720	CSMO15_CSVPuk	SNP3-I-0	720	CLWL_CIBk	Het2-I-0	720	CBagg_CLogis
SP-I-0	2880	CSMO15_CSVPuk	SNP3-I-0	2880	CBagg_CRF	Het2-I-0	2880	CIBk4
SP-I-0	4320	CSMO15_CSVPuk	SNP3-I-0	4320	CLWL_CIBk	Het2-I-0	4320	CBagg_CLogis
SP-I-99m	10	CRF10	SNP3-I-99m	10	CRF10	Het2-I-99m	10	CRF10
SP-I-99m	90	CMLPP90	SNP3-I-99m	90	CLWL_CIBk	Het2-I-99m	90	CBagg_CLogis
SP-I-99m	720	CMLPP90	SNP3-I-99m	720	CAAdBM1_CRF134	Het2-I-99m	720	CMLPP90
SP-I-99m	2880	CMLPP23	SNP3-I-99m	2880	CBagg_CLMT	Het2-I-99m	2880	CMLPP90
SP-I-99m	4320	CRSS_CIBk	SNP3-I-99m	4320	CLWL_CNB	Het2-I-99m	4320	CRF159
SNP1	10	CRF10	SNP4	10	CDS	Het3	10	CRF10
SNP1	90	CRF10	SNP4	90	CDS	Het3	90	CRF10
SNP1	720	CRF10	SNP4	720	CDS	Het3	720	CRF10
SNP1	2880	CRF10	SNP4	2880	CDS	Het3	2880	CRF10
SNP1	4320	CRF10	SNP4	4320	CDS	Het3	4320	CRF10
SNP1-I-0	10	CSGD0,042	SNP4-I-0	10	CSMO_CSVNoPKe	Het3-I-0	10	CRF10
SNP1-I-0	90	CAAdBM1_CSGD	SNP4-I-0	90	CIBk35	Het3-I-0	90	CIBk1E
SNP1-I-0	720	CMLPP12	SNP4-I-0	720	CSiLog0,055	Het3-I-0	720	CRF10
SNP1-I-0	2880	CSGD0,001	SNP4-I-0	2880	CKStar73	Het3-I-0	2880	CRF159
SNP1-I-0	4320	CSMO47_CSVRBFKe	SNP4-I-0	4320	CKStar65	Het3-I-0	4320	CRF159
SNP1-I-99m	10	CRF7	SNP4-I-99m	10	CRF7	Het3-I-99m	10	CRF10
SNP1-I-99m	90	CKStar34	SNP4-I-99m	90	CIBk3	Het3-I-99m	90	CRC10_CRF
SNP1-I-99m	720	CSMO67_CSVPKe	SNP4-I-99m	720	CRC11_CRT	Het3-I-99m	720	CLogis
SNP1-I-99m	2880	CRF176	SNP4-I-99m	2880	CKstar94	Het3-I-99m	2880	CRF159
SNP1-I-99m	4320	CRC15_CRT	SNP4-I-99m	4320	CRF163	Het3-I-99m	4320	CRF159
SNP2	10	CRF10	Het1	10	CRF10	Het4	10	CRF10
SNP2	90	CRF10	Het1	90	CRF10	Het4	90	CRF10
SNP2	720	CRF10	Het1	720	CRF10	Het4	720	CRF10
SNP2	2880	CRSS_CPART	Het1	2880	CRF10	Het4	2880	CRF10
SNP2	4320	CSMO_CSVNoPKe	Het1	4320	CDT	Het4	4320	CRF10
SNP2-I-0	10	CJRip	Het1-I-0	10	CRF10	Het4-I-0	10	CRF10
SNP2-I-0	90	CIBk6	Het1-I-0	90	CMLPP90	Het4-I-0	90	CMLPP90
SNP2-I-0	720	CSiLog0	Het1-I-0	720	CMLPP90	Het4-I-0	720	CMLPP90
SNP2-I-0	2880	CBagg_CIBk	Het1-I-0	2880	CSMO87_CSVPKe	Het4-I-0	2880	CRF159
SNP2-I-0	4320	CSMO15_CSVRBFKe	Het1-I-0	4320	CRF159	Het4-I-0	4320	CSMO09_CSVPuk
SNP2-I-99m	10	CJRip	Het1-I-99m	10	CRF10	Het4-I-99m	10	CRF10
SNP2-I-99m	90	CRSS_CSiLog	Het1-I-99m	90	CRF10	Het4-I-99m	90	CRF10
SNP2-I-99m	720	CLWL_CRF	Het1-I-99m	720	CMLPP90	Het4-I-99m	720	CMLPP90
SNP2-I-99m	2880	CAAdBM1_CRF12	Het1-I-99m	2880	CRF159	Het4-I-99m	2880	CRF159
SNP2-I-99m	4320	CVote	Het1-I-99m	4320	CRC20_CRF	Het4-I-99m	4320	CRF159

solamente sobre los conjuntos originales y reducidos (un total de 135 conjuntos diferentes), que a su vez generaron 135 modelos, descritos en la **Tabla 5**. Este tratamiento se hizo a 5 diferentes tiempos (10, 90, 720, 2880 y 4320 minutos) tomando en cuenta que Auto-WEKA prueba múltiples combinaciones algoritmo-parámetros, y pese a que se espera que a mayor tiempo invertido en buscar, mayor eficacia en la combinación algoritmo parámetros encontrada, podrían encontrarse casos en las que esa correlación no se cumpla. Se debe de recordar que los conjuntos reducidos se obtuvieron solamente a partir de los conjuntos transformados (que no fueron utilizados en esta parte, ver sección de METODOLOGÍA) tanto con el cambio Infinity-0 como con el cambio Infinity-99 millones, y no de los originales, por lo que aquí se hará referencia como conjuntos reducidos a aquellos que utilizan una transformación u otra. Al eliminar modelos iguales quedaron 118 modelos, conformados a su vez por 55 combinaciones algoritmo-parámetros distintas, cuya nomenclatura puede observarse en el **Apéndice 2**. Destaca observar que, por ejemplo, en algunos modelos se usaba el algoritmo RandomForest con 10 iteraciones (CRF10), mientras en otros se usaba el mismo algoritmo pero con 159 iteraciones (CRF159). Se observa que la combinación CRF10 es la que más veces aparece, estando presente en los 3 tipos de conjuntos (**SóloPéptidos**, **SóloCQNPS** y **Heterólogos**). También se observa que conjuntos de tipo **SóloPéptidos** y **Heterólogos** presentan mayor tendencia a que las combinaciones algoritmo-parámetros sean las mismas, lo que indica que las fronteras en estos espacios de datos son aprendibles de forma óptima con esas combinaciones algoritmo-parámetros.

Primera selección de modelos

Los modelos obtenidos a partir de los TrS cuyo porcentaje de instancias clasificadas correctamente (%CC) era mayor a 90% se seleccionaron y se

muestran en la **Figura 5**. Los datos numéricos asociados a esta figura se hallan en el **Apéndice 3**. Se observa que los 77 modelos seleccionados corresponden a todos los tipos de conjuntos (**SóloPéptidos**, **SóloCQNPs** y **Heterólogos**), sin haber predominio de un tipo de conjunto; pero sí se observa una tendencia en los modelos no elegidos con valores de %CC más bajos, los cuales tienden a corresponder a conjuntos de tipo **SóloCQNPs**, por lo que éstos tendrían de manera general %CC más bajo.

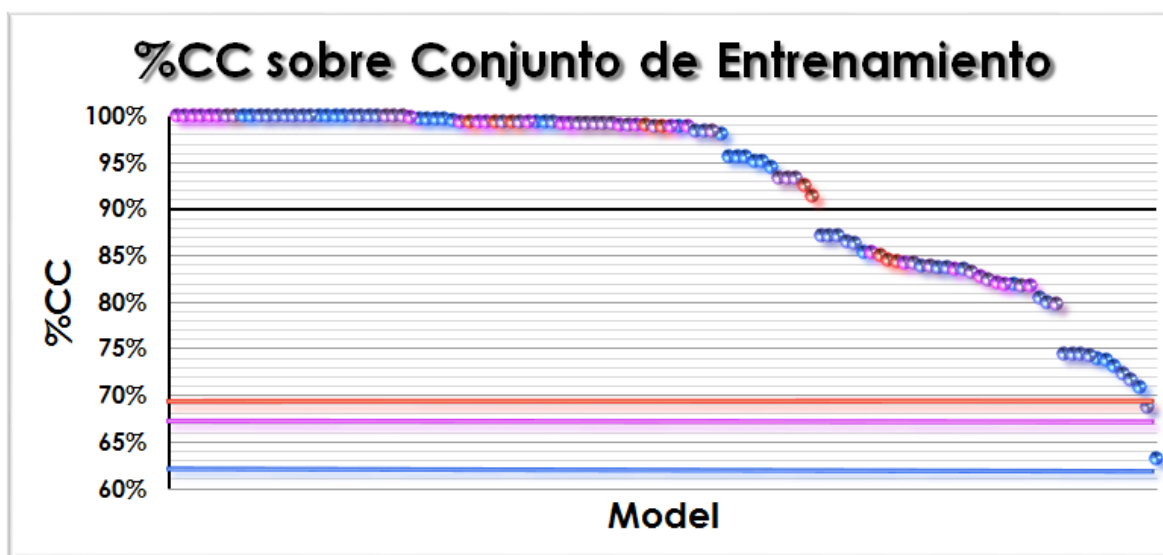


Figura 5. Evaluación de instancias clasificadas correctamente (%CC) para cada modelo en su conjunto de entrenamiento. En negro está marcada la línea (a 90%) que delimita el criterio para escoger a los primeros mejores modelos. Círculos bermellón: Modelos con **SóloPéptidos**. Círculos azules: Modelos con **SóloCQNPs**. Círculos morados: Modelos **Heterólogos**. Se observa la tendencia de los modelos **SóloCQNPs** de ser los más abundantes entre los no elegidos. Las líneas bermellón, azul y morada representan el valor obtenido con el algoritmo ZeroR para conjuntos de **SóloPéptidos**, **SóloCQNPs** y **Heterólogos** respectivamente.

En la **Figura 5** también se observa el resultado de realizar corridas de base utilizando el algoritmo ZeroR, que asigna a todas las moléculas del TrS la clase de la mayoría en el conjunto (es decir, si el 70% de los datos de un conjunto fueran antimicrobianos, ZeroR asigna a todos esa clase). En el caso de modelos con conjuntos de **SóloPéptidos**, ese valor es de 69.29%; para **SóloCQNPs** fue de 61.949%, y para **Heterólogos** fue de 67.12%. Se observa que todos los modelos tienen valores mayores a sus

correspondientes valores de ZeroR, lo que muestra que todos los modelos tienen una predicción mejor que la realizada por clase mayoritaria.

Validación de modelos

Para cada uno de los 77 modelos elegidos, se obtuvieron los resultados correspondientes a los %CC de las 8 repeticiones del 10FCV y de las 4 por split (ver METODOLOGÍA). Los resultados numéricos de estas validaciones se muestran en el **Apéndice 5**.

Se observa para los modelos con conjuntos de tipo **SóloPéptidos** y **Heterólogos**, tanto originales y transformados como reducidos, que los resultados de %CC de 10FCV no presentan diferencias notables con los de split, estando todos los resultados en un rango de 84-89%, mientras que para la mayoría de los modelos con **SóloCQNPS** sí se muestra una diferencia apreciable de rangos de porcentaje entre modelos diferentes y, en algunos modelos, entre el valor del 10FCV y el de Split; lo cual mostraría, de manera esperada, que los datos de **SóloCQNPS** sean más heterogéneos que los que incluyen péptidos, al ser estos últimos más

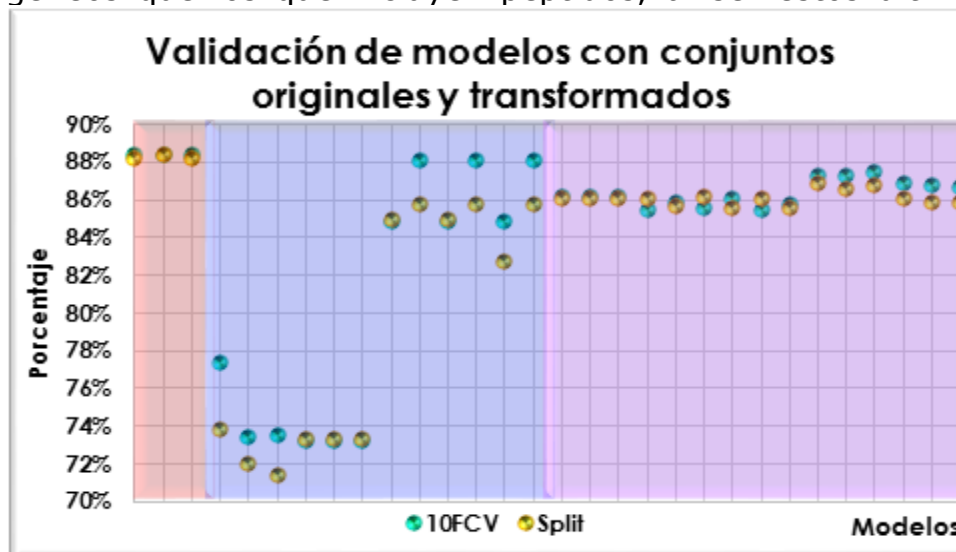


Figura 6. Comparación de valores de %CC, para validaciones por 10FCV y por Split en los modelos con conjuntos originales y transformados. El fondo indica que los conjuntos son tipo **SóloPéptidos** (rojo), **SóloCQNPs** (azul) o **Heterólogo** (morado). Conjuntos rojos y morados de manera general muestran resultados más altos.

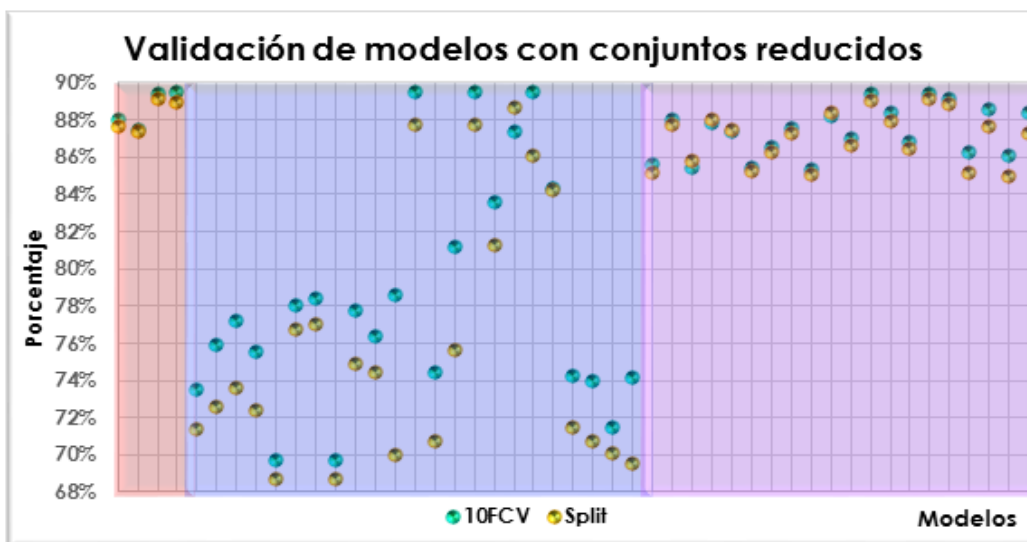


Figura 7. Comparación de valores de %CC, para validaciones por 10FCV y por Split en los modelos con conjuntos reducidos. El fondo indica que los conjuntos son tipo **SóloPéptidos** (rojo), **SóloCQNPs** (azul) o **Heterólogo** (morado). Conjuntos azules muestran resultados con mayor nivel de variación.

químicamente similares entre sí. Resalta que los conjuntos originales y transformados presentan mayor homogeneidad en valores de validación que los reducidos (Ver **Figuras 6 y 7**).

Evaluación de modelos

Se realizó la evaluación de los modelos en los conjuntos de prueba (TeS). Para representar visualmente estos resultados, se promediaron los valores de %CC observados para conjuntos de la misma clase (**SóloPéptidos**, **SóloCQNPs** y **Heterólogos**) y subclase (1-4; ver **Tabla 4**). Los resultados se muestran en la **Figura 8**, donde se ve que el utilizar los conjuntos **SóloPéptidos** para clasificar CQNPs resulta ineficaz, puesto que el valor promedio del %CC es de 39.0908%. Se obtienen mejores tendencias utilizando los conjuntos **SóloCQNPs** y **Heterólogos**, concretamente en las subclases 1 y 2 (divididos en pares y nones), en los que los valores promedio del %CC oscilan entre 71-76%. El %CC cae a valores promedio de 58-65% al usar las subdivisiones 3 y 4 (primera y

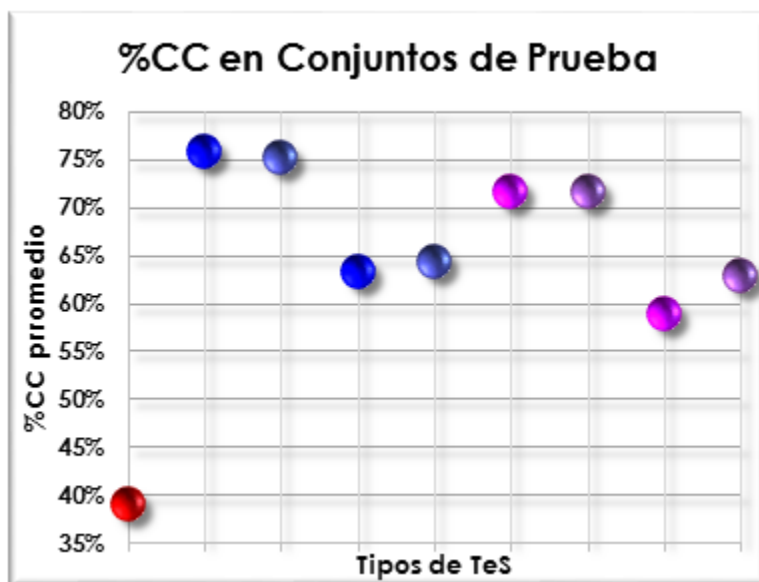


Figura 8. Promedios de todos los porcentajes de %CC de cada clase (y subclase) de modelos evaluados. 1, 2, 3 y 4 representan las subclases donde los TrS corresponden a líneas pares, impares, primera y segunda mitad del conjunto respectivamente (ver **Tabla 4**). Líneas pares e impares presentan mayor valor de %CC.

segunda mitad del conjunto), mostrando que la manera en que se subdividen los datos sí afecta el aprendizaje.

Para evaluar el comportamiento individual de cada modelo y detectar posibles tendencias de forma visual, se optó por representar estos datos mediante barras de comportamiento (ver sección de METODOLOGÍA) para el TeS de cada modelo, tal como se muestra en la **Figura 9** y en el **Apéndice 4**.

En estas barras se observa que existen tendencias diferentes para los datos. Por ejemplo, al evaluar modelos de tipo **SóloPéptidos**, la tendencia es que todos los datos sean predichos como antimicrobianos, con lo que casi todos los compuestos se clasifican como positivos, ya sea verdaderos (VP) o falsos (FP); lo que reduce todos los %CC a valores menores de 42%, por lo que estas predicciones se consideran inviables y hacen pensar en la posibilidad de que los datos de los CQNP en realidad residan en un espacio multivectorial diferente al de los péptidos.

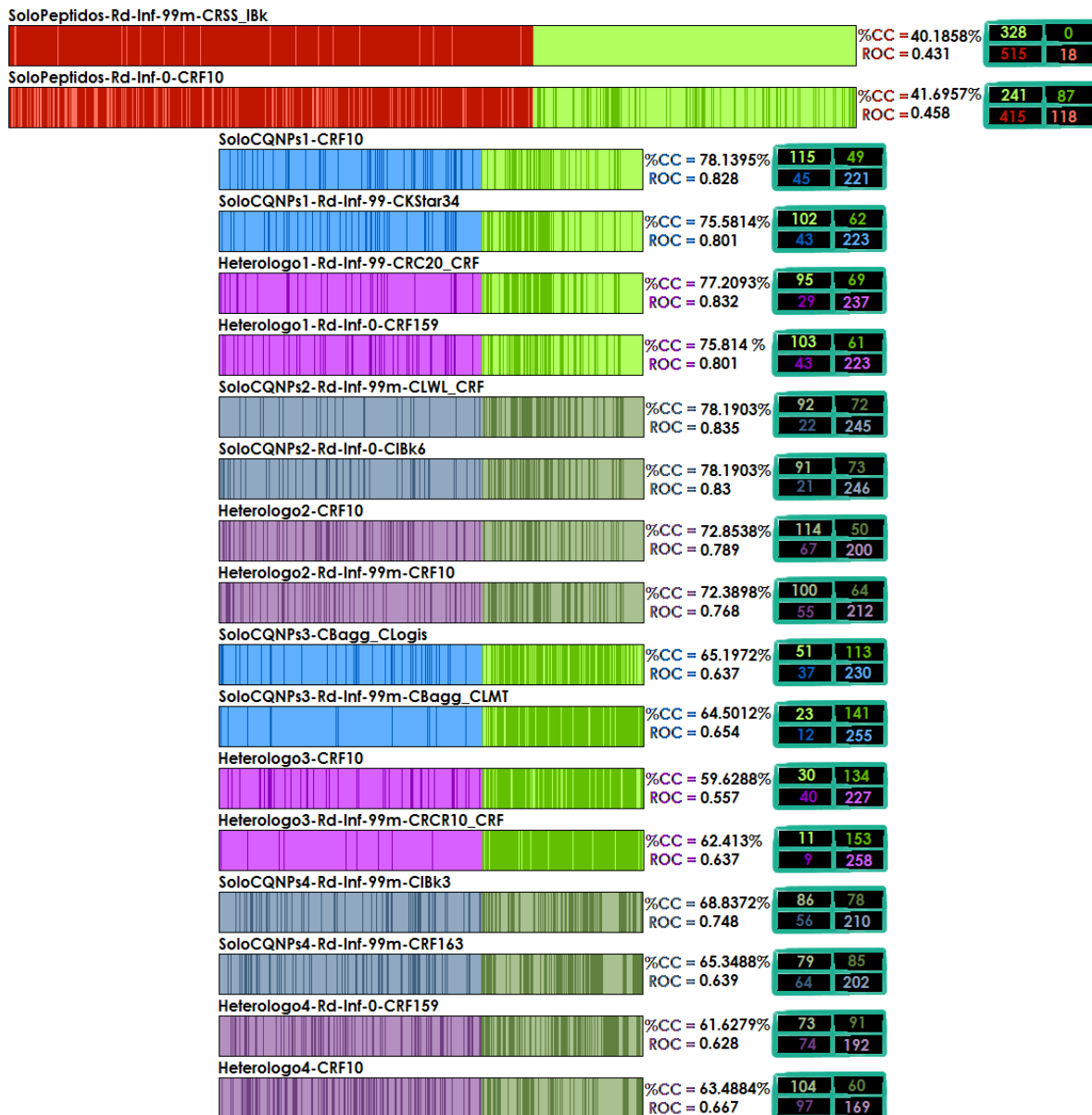


Figura 9. Barras de comportamiento de los TeS de los 2 mejores modelos por %CC de cada tipo de conjunto, ordenados por TeS iguales; mostrando también la matriz de confusión, el %CC y el valor de la curva ROC sobre TeS. Los datos color verde claro corresponden a verdaderos positivos (VP) y los oscuros a falsos negativos (FN). Los datos de otros colores claros dependiendo el tipo de conjunto, son verdaderos negativos (VN) y los oscuros falsos positivos (FP). Se observa tendencia de los subtipos 3 y 4 de **SóloCQNPs** y **Heterólogos** de presentar más clasificación como sin actividad, caso contrario a los de **SóloPéptidos**.

En todas estas predicciones la curva ROC tuvo valores menores a 0.5, por lo que todas estas clasificaciones son más pobres que la realizada

utilizando la clase mayoritaria (con el algoritmo ZeroR), con lo que se confirma la inviabilidad de estos modelos.

Al evaluar los modelos de tipo **SóloCQNPS** y **Heterólogos**, se observa que algunas de las tendencias son más favorables. Para los conjuntos de subtipo 1 y 2 (líneas pares e impares) los porcentajes %CC son mayores a 70% en la mayoría de los casos, teniendo sólo 2 conjuntos con porcentajes de 63-64% (ambos de tipo **Heterólogo**). En todos los casos la curva ROC tuvo valores mayores a 0.66 alcanzándose más de 0.8 en varios de ellos, con lo que estos modelos presentan mejor predicción que la obtenida con ZeroR. Asimismo, se observa en las matrices de confusión que la tendencia es a tener más valores de verdaderos negativos (VN) y de VP, que de sus falsos correspondientes, y en las barras se observa que los valores clasificados de manera incorrecta se dispersan de manera heterogénea a lo largo de los conjuntos, con lo que no se estaría incurriendo en un sesgo de predicción dado que las instancias del TeS están ordenadas alfabéticamente y no por clasificación de fármaco, en cuyo caso se observarían todas las instancias incorrectas en sectores específicos de las barras. Por otro lado, utilizando los conjuntos de subtipo **3** y **4** (primera y segunda mitad de los conjuntos) la tendencia muestra que los valores de %CC oscilan entre 58-64% y los valores de curva ROC suelen ser menores a 0.66, algunos de ellos incluso menores a 0.5, mostrando que estos modelos tampoco serían viables debido a lo mismo de que la tendencia es muy similar a la de ZeroR de predecir la clase mayoritaria (en este caso, ausencia de actividad). Estas diferencias entre subtipos se pueden explicar considerando un sesgo en los datos de entrenamiento (por ejemplo, no haber incluido ejemplos de los evaluados en los TeS) o bien, a que los datos presentan ejemplos de instancias que se hallan fuera de las fronteras encontradas para los modelos (datos "outliers") que se enriquecerían en los subtipos **3** y **4**.

Selección del mejor modelo

Posteriormente se buscó un criterio para identificar los mejores modelos utilizando la mayor cantidad de parámetros estadísticos evaluados que fueran de utilidad. Se utilizaron 5 parámetros que fueron los siguientes: El %CC del TeS, la curva ROC del TeS, la tasa de error estimado ajustado (EERAJ) obtenido con Auto-WEKA, y también los %CC de las validaciones 10FCV y split. En la **Figura 10** se muestran graficados estos 5 parámetros para cada modelo, los valores numéricos se hallan en el **Apéndice 5**. Como se observa, el comportamiento de los 5 parámetros es heterogéneo, no existiendo un modelo para el cual los 5 parámetros sean mejores que para todos los demás modelos; por ejemplo, los modelos con más altos valores de %CC no necesariamente son los mismos que los que tienen valores más altos en su validación por 10FCV.

Dada esta situación, se buscó una manera de condensar estos 5 parámetros en uno solo. Para esto se utilizó una fórmula para un Score

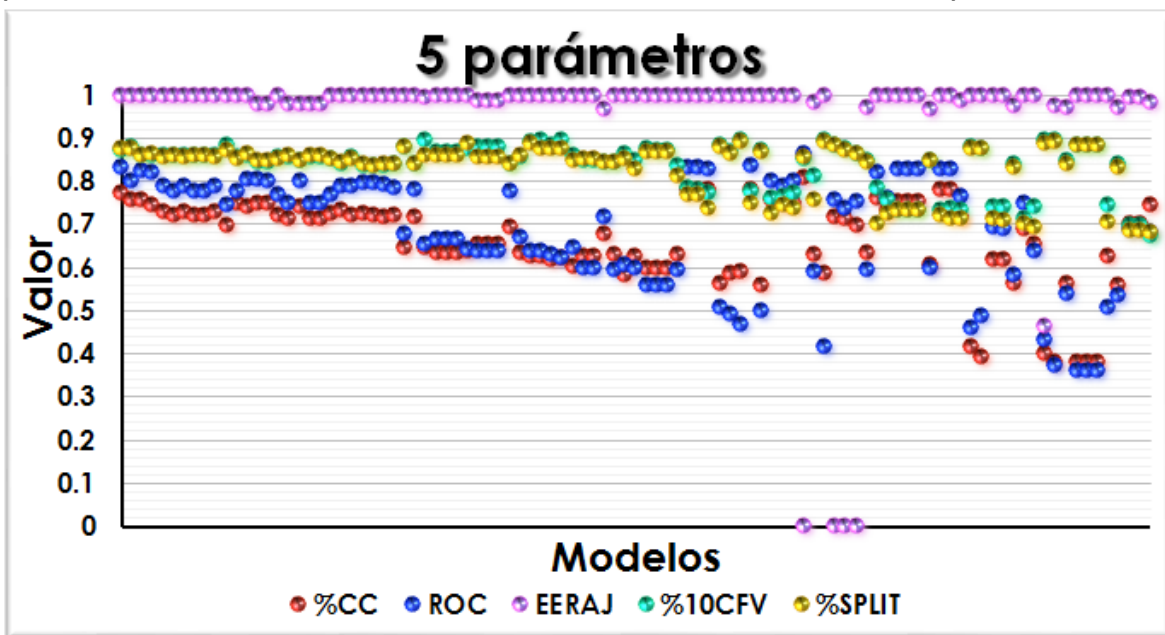


Figura 10. Valores de los 5 parámetros utilizados para encontrar el mejor modelo. EERAJ = Ratio ajustado de error estimado. Se incluyen los datos de 22 modelos obtenidos por métodos convencionales (ver texto). Los modelos de la izquierda manejan valores más similares entre sí que los de la derecha, aunque en general son heterogéneos.

combinado (CScore, obtenido mediante el script *CombinedScore.java*, ver sección de METODOLOGÍA) que requirió que los parámetros a elegir fueran normalizados a valores de 0 a 1. En la **Figura 11** se muestra un gráfico con los resultados de haber obtenido el CScore de los modelos, mientras que en el **Apéndice 6** se muestra una tabla con los datos numérico del CScore, en la que se observa que los valores obtenidos en esta evaluación oscilan entre 0.139 y 1.459. Dada la manera en que funciona el CScore, entre más bajo sea éste, mejor es el desempeño del modelo. La tendencia observada indica que los modelos de **SóloPéptidos** poseen valores de CScore más altos, los modelos de **SóloCQNPS** se dispersan entre valores de CScore intermedios y altos, y destaca que los valores más bajos (y ergo, los modelos más eficaces) corresponden a conjuntos **Heterólogos**.

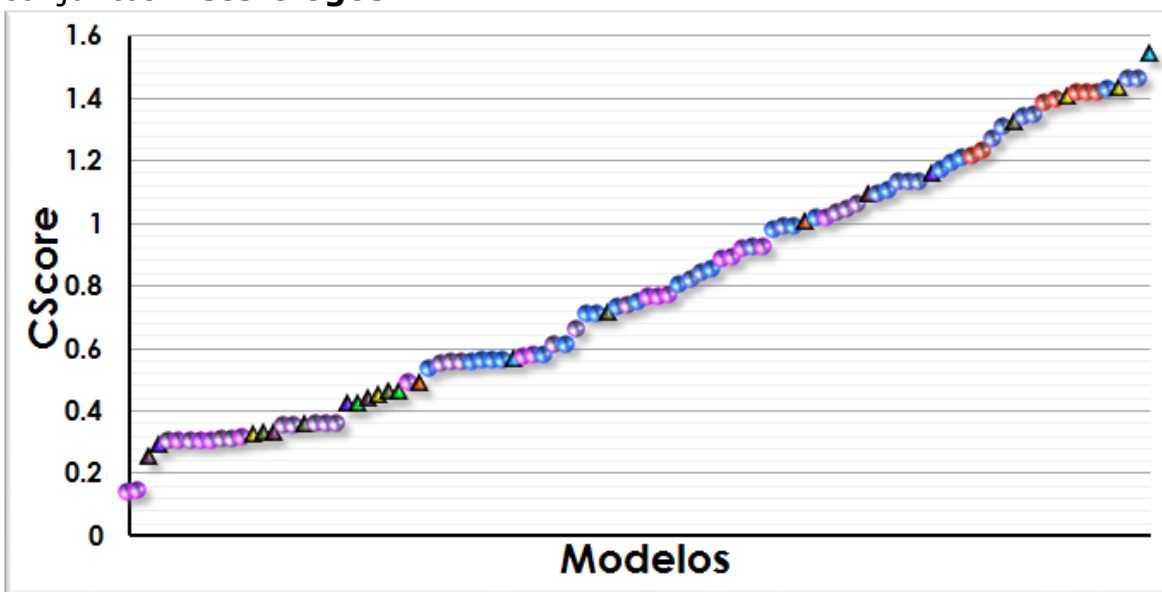


Figura 11. Valores de CScore obtenidos para todos los modelos evaluados de acuerdo con los 5 parámetros elegidos (ver texto). Los círculos rojos corresponden a modelos de **SóloPéptidos**, los azules a **SóloCQNPS** y los morados a **Heterólogos**. Los triángulos corresponden a modelos heterólogos obtenidos por métodos convencionales (ver texto). Los CScores más bajos corresponden a modelos **Heterólogos**.

En la **Tabla 6** se observan los 5 mejores modelos, destacando que todos son de tipo **Heterólogo**. El mejor modelo corresponde a la combinación

CRC20_CRF (Algoritmo RandomCommittee con 20 iteraciones, utilizando al algoritmo RandomForest) aplicada en un conjunto reducido con el cambio Infinity-99 millones.

TABLA 6. Lista de los 5 mejores modelos, junto con su valor de CScore. El mejor modelo usa el algoritmo RandomCommittee.

Modelo	Cscore
Heterologo1-Rd-Inf-99M-CRC20_CRF	0.139325155
Heterologo1-Rd-Inf-0-CRF159	0.140239399
Heterologo2-CRF10	0.299191035
Heterologo1-CRF10	0.299245774
Heterologo2-Tr-Inf-0-CRF10	0.299819626

Comparación con modelos de tipo convencional

Tomando en cuenta que la estrategia seguida para generar los conjuntos (sustituir valores nulos por 0, y sustituir valores "Infinity por 0 o 99 millones) no es convencional, se generaron 22 modelos adicionales mediante algunas técnicas convencionales de procesamiento de datos (ver sección de METODOLOGÍA) utilizando el conjunto asociado al modelo con mejor valor de CScore (**Heterólogo1**); estos 22 modelos utilizaron alguna de 11 nuevas combinaciones algoritmo-parámetros (mostradas en el **Apéndice 7**) o de las ya utilizadas anteriormente, y se obtuvieron para ellos los 5 parámetros ya mencionados (mostrados en la **Figura 11** y el **Apéndice 5**) y el CScore. En la **Figura 12** y el **Apéndice 6** se muestra que ninguno de estos modelos puede mejorar el CScore obtenido por el mejor modelo obtenido por métodos no convencionales, lo que muestra que el procedimiento originalmente utilizado es eficaz.

8.2. SIMILITUD ENTRE PÉPTIDOS Y CQNPs

Caracterización del dominio del modelo (CDM)

Observando que el mejor modelo, al ser de tipo heterólogo, utiliza datos tanto de péptidos como de no peptídicos, se realizó una primera aproximación de una búsqueda de similitudes entre ambos tipos de moléculas mediante una caracterización del dominio del modelo (CDM),

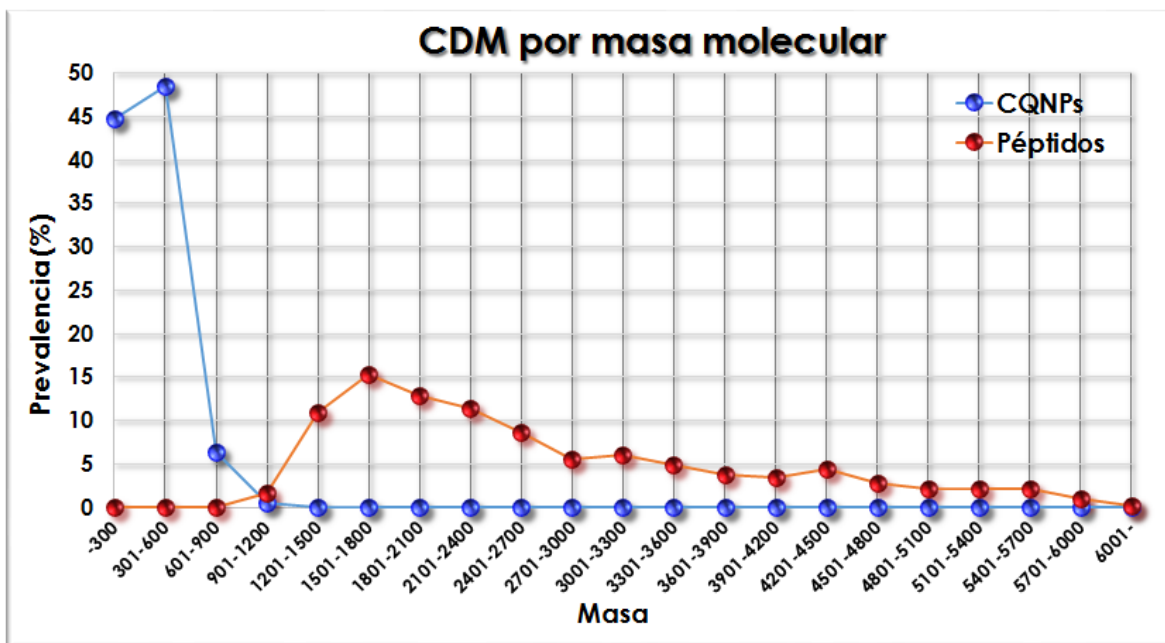


Figura 12. CDM por porcentaje de masa molecular tanto en péptidos como en CQNP. Las masas de estos últimos tienen tendencia a ser menores.

el cual se realizó con base en la observación de 2 características fácilmente observables en ambos tipos de moléculas: prevalencia de grupos funcionales, y masa molecular.

Se observa que para el caso de la CDM por masa molecular, mostrada en la **Figura 12**, prácticamente todos los CQNP están en el rango de 300 a 900 Da, con un porcentaje menor a 1% de compuestos de hasta 1200 Da; mientras que los péptidos abarcan rangos de 900 a más de 6000 Da. La CDM por grupos funcionales, mostrada en la **Figura 13**, muestra que los péptidos tienen una enorme prevalencia de grupos amida; también tienen grupos ácido, guanidinio y de los elementos oxígeno, nitrógeno y azufre, con prácticamente total ausencia de todos los demás grupos; mientras que para CQNP la prevalencia de estos 3 últimos elementos es más alta, existe mayor variedad de grupos funcionales presentes y la proporción de grupos amina y guanidinio es más baja. Por tanto, se infiere

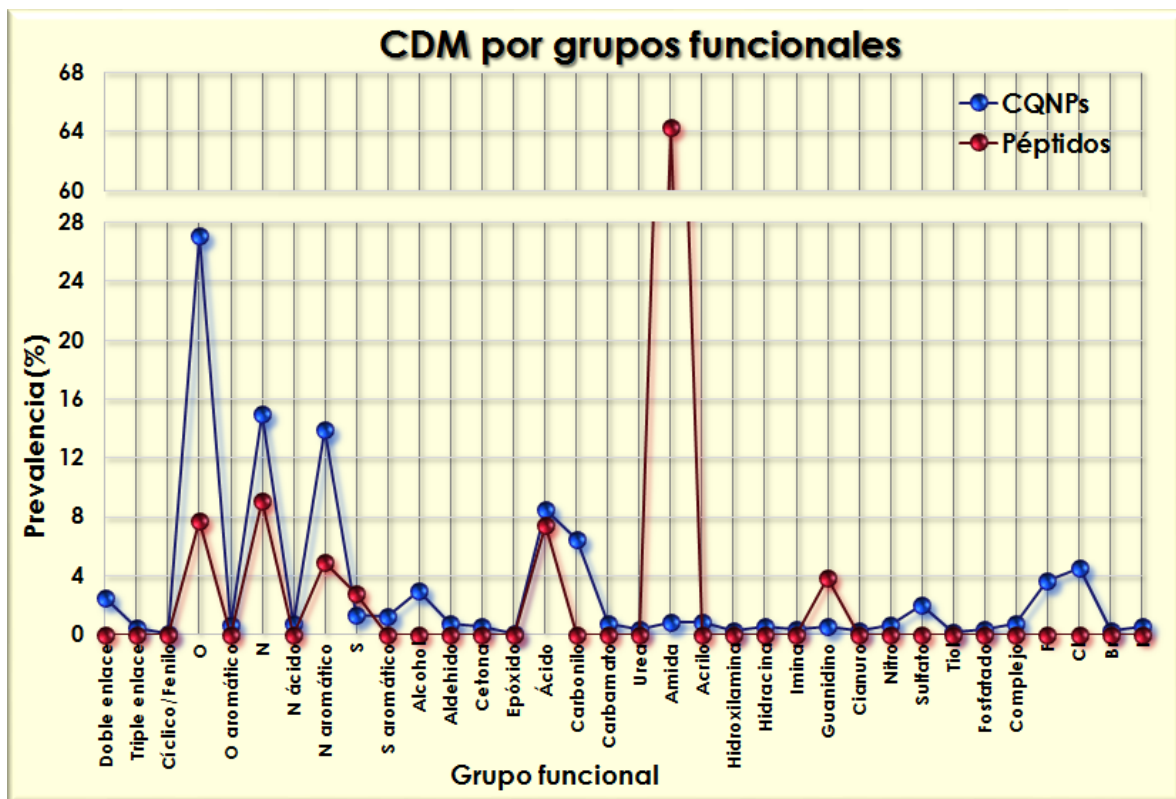


Figura 13. CDM por porcentaje de grupos funcionales en p ptidos y en CQNPs. La diversidad de grupos en estos  ltimos es mayor.

que estas caracter sticas f cilmente observables no son las que permiten la combinaci n de estos datos para su aprendizaje.

Importancia de atributos y visualizaci n de los datos

Para identificar los atributos moleculares que hac an a los datos aprendibles se realiz  la evaluaci n de los atributos por cada conjunto usando el evaluador IGAE (explicado en secci n de METODOLOG A).

La **Tabla 7** muestra los resultados de los 20 mejores atributos para el mejor modelo ya mencionado, para el mejor modelo con conjunto de **S loCQNPs**, y para el mejor modelo con **S loP ptidos**. Se observa que los conjuntos de tipo **S loP ptidos** y **Heter logos** en general comparten muchos de los mejores atributos en com n, y con un orden similar, mientras que en el caso del conjunto **S loCQNPs** s  se observa que hay diferencias en los descriptores hallados respecto a los

correspondientes a conjuntos **SóloPéptidos** y **Heterólogos**. También se observa que el descriptor MW (que en PaDEL representa a la masa molecular) no aparece entre los mejores 20 para ningún modelo, por lo que otras características diferentes serían las que definirían la aprendibilidad de estos modelos.

TABLA 7. Lista de los 20 mejores descriptores para el mejor modelo de SóloPéptidos, el mejor de SóloCQNPs y el mejor Heterólogo (mejor general).

En colores oscuros se muestran los descriptores compartidos. Abundan los descriptores tipo MIC, SP y VP.

Rank	SoloPeptidos	SoloCQNPs3	Heterologo1
1	BCUTw-1h	nHBa	BCUTw-1h
2	SP-4	SM1_Dze	SP-4
3	SP-7	SpMax7_Bhs	SP-3
4	MIC4	TopoPSA	SP-5
5	MIC5	SHBin8	SP-6
6	SP-6	ATSC0c	VP-4
7	SP-5	minHBin3	VP-5
8	AMW	SpMax6_Bhs	SP-7
9	VP-4	SM1_Dzm	VP-3
10	SP-3	GGI8	SP-2
11	VP-5	nHBin8	SP-1
12	VP-6	ETA_Eta_F	VP-6
13	VP-7	SM1_DzZ	VP-1
14	VP-3	ATS4m	SP-0
15	SP-2	hmax	VP-2
16	SP-1	SHBa	VP-0
17	MIC3	MIC5	VP-7
18	MIC2	nHBAcc2	MIC5
19	SP-0	VC-5	MIC4
20	GATS1m	maxHCsats	SPC-4

Se observa que los descriptores más abundantes son de tipo SP y VP y MIC. Los 2 primeros son de tipo conectividad por chi (χ), para lo cual es necesario calcular características como orden de enlace, o número de electrones en orbitales sigma (Pearlman y Smith, 1999). Por otra parte, los descriptores de tipo MIC se describen, de acuerdo a la información oficial proporcionada por PaDEL Descriptor, como "descriptores de índice de contenido de información modificada, con un cierto orden de enlace por simetría de vecinos, usando la masa molecular para cada átomo. Estos descriptores se encuentran más representados en modelos de tipo **SóloPéptidos** y **Heterólogos**, mientras que los descriptores para **SóloCQNPS** son más heterógeneos.

De aquí se infiere que los modelos de tipo **SóloPéptidos** y **Heterólogos** habrían aprendido de sus datos de manera similar, a diferencia de lo que habría sucedido con los modelos de tipo **SóloCQNPS**.

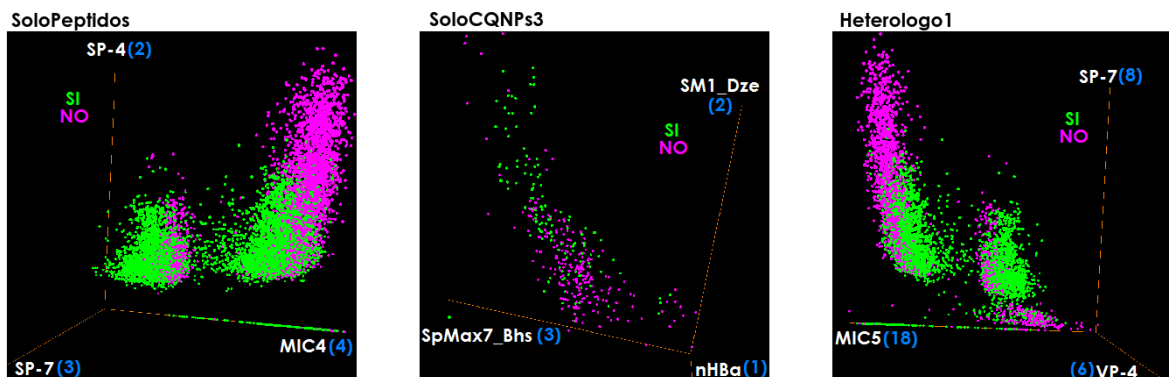


Figura 14. Visualización tridimensional del espacio en que residen los datos de los mejores modelos para péptidos, CQNPs y Heterólogos (en ese orden, izquierda a derecha). En verde los datos de compuestos con actividad antimicrobiana, y en violeta aquellos sin dicha actividad. En blanco y azul se observa el nombre de los descriptores usados para construir el espacio, así como el orden asignado por IGAE. Datos sobre péptidos y heterólogos presentan separación más definida.

Las imágenes correspondientes a la visualización tridimensional de estos 3 conjuntos pueden apreciarse en la **Figura 14**. Se puede observar que para el conjunto **SóloPéptidos**, los compuestos antimicrobianos y los no antimicrobianos están visiblemente separados, mientras en el caso del conjunto **SóloCQNPS** las tendencias de agrupación son menos claras. Esta pobre separación en los **SóloCQNPS** podría deberse a que solo se usan 3 descriptores, sin embargo dado que el resultado con Auto-WEKA mostró que **SóloCQNPS** es menos aprendible que los demás grupos, es posible afirmar que la frontera entre los antimicrobianos y los no antimicrobianos en ese tipo de conjunto no es fácil de trazar. En el caso de los **Heterólogos**, se observa una tendencia similar a la de **SóloPéptidos**, lo cual puede explicarse debido a que la mayoría de los compuestos incluidos en los conjuntos **Heterólogos** eran péptidos.

8.3. TRATAMIENTO DE DATOS NOVEDOSOS

Conjunto de Descubrimiento (DiS)

Una vez obtenido el mejor modelo, se procedió a aplicarlo sobre el Conjunto de Descubrimiento (DiS). Para esto previamente se procesaron los datos del DiS de forma idéntica a la realizada el mejor modelo (reducción de dimensiones mediante ACP de manera que el número y tipo de descriptores compuestos fuera exactamente el mismo que para el TrS reducido del modelo).

TABLA 8. Matriz de confusión asociada a los datos del DiS. Los datos en las columnas de color indican compuestos que se predice que tendrán o no actividad contra bacterias de la microbiota. Las filas de color muestran datos de si son antibióticos aprobados contra patógenos, o medicamentos diferentes a antibióticos (no antibiótico). En negro los totales.

Compuesto	Pred. Actividad	Pred. No actividad	Total AB/NAB
Antibiótico	72	61	133
No antibiótico	140	556	696
Total Pred.	212	617	829

Al ejecutar el mejor modelo sobre el DiS, se obtuvo la matriz de confusión que se muestra en la **Tabla 8**. Se puede observar que hay 829 instancias totales, de las que 212 fueron predichas como con actividad antimicrobiana, y 617 fueron predichas sin ella. También se observa que 133 de los compuestos son originalmente antibióticos, mientras que 696 no lo son (tienen otra AP). De entre ellos debe destacarse que hay 111 antimicrobianos y 645 compuestos sin ninguna actividad antimicrobiana conocida de entre los originalmente 756 buscados, así como 22 antifúngicos y 51 sin actividad antifúngica entre los 73 compuestos adicionales (ver sección de METODOLOGÍA para detalles). Se ha reportado que los antifúngicos funcionan a través de mecanismos (Calahorra *et al.*, 2018) diferentes de los de compuestos antibacterianos (por ejemplo, derivados de penicilina, sulfonamidas, etc.), por lo que se esperaba que el modelo tendiera a no predecir a estos compuestos como

antibacterianos. Por otro lado, se espera que los compuestos aprobados por la FDA tengan una actividad antimicrobiana intestinal nula o menor a la de los antifúngicos, de lo contrario sus efectos gastrointestinales secundarios serían significativos.

En este caso ya no se puede hablar de "instancias clasificadas correctamente", dado que de estos datos no se conoce información experimental sobre si afectan a la microbiota intestinal, concretamente a las 40 cepas sobre las que se realizó el estudio. Por otra parte, este modelo al identificar actividad antimicrobiana contra bacterias de la flora intestinal, puede ayudar a identificar actividades no deseadas en compuestos usados en humanos. Las distintas actividades de los compuestos aprobados por la FDA [de acuerdo al Sistema de Clasificación Anatómica, Terapéutica y Química (ATCC); ver sección de METODOLOGÍA] que fueron predichos por este modelo como antibacterianos de la flora intestinal se presentan en la **Tabla 9**.

TABLA 9. Clases ATCC más representadas de compuestos no antibióticos predichos como con actividad contra la microbiota, tanto para el TeS del mejor modelo como para el DiS. Los antineoplásicos destacan por su presencia en ambos casos.

Código ATCC para no antibióticos con actividad antimic. predicha en el TeS	Cantidad	%
Antipsicóticos	21	13.91%
Antineoplásicos varios	11	7.28%
Terapia hormonal	11	7.28%
Analgésicos y antipiréticos varios	7	4.64%
Antidepresivos	6	3.97%
Antihistamínicos sistémicos	6	3.97%
Antimetabolitos	5	3.31%

Código ATCC para no antibióticos con actividad antimic. predicha en el DiS	Cantidad	%
Antineoplásicos varios	24	20.34%
Terapia de diabetes	6	5.08%
Agentes para niveles de lípidos	5	4.24%
Agentes del sistema Renina-angiotensina	4	3.39%
Agentes para obstrucciones respiratorias	4	3.39%
Antitrombóticos	4	3.39%
Bloqueadores de canales de calcio	4	3.39%

Se observa que la clase más representada en el TeS es la de los antipsicóticos; mientras que para el DiS corresponde a la de los

antineoplásicos. El **Apéndice 8** muestra una lista de todos los CQNP del DiS que fueron predichos como con actividad antimicrobiana, así como el código ATCC de su actividad primaria.

Antibióticos de amplio o reducido espectro

Adicionalmente, tomando en cuenta que los antibióticos incluidos en el DiS son compuestos aprobados por la FDA por su eficacia contra organismos patogénicos, y que los modelos trabajados en este proyecto predicen actividad contra microorganismos no patogénicos de la microbiota intestinal, se planteó que los 59 antibióticos (de los 72 antibióticos diferentes incluidos en el DiS; ver sección de METODOLOGÍA) predichos como con actividad antimicrobiana en el DiS, tendrían que ser clasificados como de amplio espectro.

TABLA 10. Compuestos con reporte de espectro amplio (19 primeros) o reducido (3 últimos) con la predicción obtenida por el modelo. Los compuestos subrayados sin negritas tienen 2 estereoisómeros cada uno.

Compuesto	Código ZINC	Predicción	Espectro
Ceftaroline fosamil	ZINC000003989268	NO	Amplio
Fosfomicin	ZINC000001530427	NO	Amplio
Amphotericin B	ZINC000253387843	NO	Amplio
Tigecycline	ZINC000014879972	SI	Amplio
Cefibuten	ZINC000003871967	SI	Amplio
Ceftriaxone	ZINC000028467879	SI	Amplio
Ertapenem	ZINC000003918453	SI	Amplio
<u>Cefprozil</u>	ZINC000003776970	SI	Amplio
<u>Cefprozil</u>	ZINC000004474443	SI	Amplio
Azithromycin	ZINC000085537026	SI	Amplio
Telithromycin	ZINC000009574770	SI	Amplio
Amikacin	ZINC000008214483	SI	Amplio
<u>Gemifloxacin</u>	ZINC000022059926	SI	Amplio
<u>Gemifloxacin</u>	ZINC000022059930	SI	Amplio
<u>Gatifloxacin</u>	ZINC000003607120	SI	Amplio
<u>Gatifloxacin</u>	ZINC000038197764	SI	Amplio
Rifampicin	ZINC000169621223	SI	Amplio
Rifabutin	ZINC000169621215	SI	Amplio
Amoxicillin	ZINC000003830215	SI	Reducido
Phenoxymethylpenicillin	ZINC000003831282	SI	Reducido
Cephalexin	ZINC000003830500	SI	Reducido

Por lo tanto, se buscó información sobre estos compuestos en diferentes estudios (ver sección de METODOLOGÍA), encontrándose que para 19

compuestos existían reportes: 16 compuestos se reportaban como de amplio espectro y 3 como de espectro reducido, tal como muestra la **Tabla 10**. Por tanto, se sugiere que estos 3 compuestos (amoxicilina, fenoximetil-penicilina y cefalexina) sean revaluados en cuanto a su espectro de acción.

En la **Tabla 10** también se muestra que 3 compuestos con reporte de amplio espectro, no fueron predichos por el modelo de esa manera. Por lo tanto, de manera general podría decirse que de los 19 compuestos de amplio espectro encontrados, 16 fueron predichos por el modelo como tales, por lo que existiría una eficacia de 84.21% del modelo para predecir estos compuestos. La lista de compuestos puede consultarse en el **Apéndice 9**.

9. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

La identificación de compuestos antimicrobianos utilizando estrategias de Machine-Learning (ML) conlleva una gran cantidad de ventajas, como por ejemplo la mejora en el tiempo de obtención de resultados para descubrimiento de nuevos fármacos, o para encontrar efectos secundarios adversos, o actividades adicionales en compuestos ya descubiertos previamente (Durrant y Amaro, 2015). Un importante aspecto relacionado a estas técnicas consiste en buscar maneras de mejorar la fiabilidad de estas predicciones. Algunas maneras de lograr esto, incluyen el aumento en el tamaño de los conjuntos utilizados para entrenar. En este trabajo se propone que es posible utilizar datos de tipos diferentes (en este caso, péptidos y no peptídicos, los cuales suelen ser modelados de manera separada) para generar un modelo combinado (heterólogo) que aumenta la fiabilidad de las predicciones; y se muestra que efectivamente los conjuntos heterólogos producen los modelos más eficientes para clasificar compuestos antimicrobianos (ver **Figuras 10** y **11**).

Una pregunta planteada a partir de esto fue la de cuáles atributos moleculares eran relevantes para el éxito de la clasificación mejorada entre compuestos antimicrobianos y no antimicrobianos. Las **Figuras 12** y **13** muestran que los descriptores moleculares empleados en este estudio no parecen ser adecuados para separar antimicrobianos de no

$${}^m\chi_t = \sum_{i=1}^A \left[\prod_{k=1}^{m+1} (\sigma_k - h_k)^{-0.5} \right]$$

Figura 15. Ecuación por la cual se obtienen los valores de los descriptores chi tipo SP. m = orden (0 a 9), t = tipo de cálculo (de tipo camino en este caso), A = átomos pesados (no hidrógenos) en la molécula, k = átomos en fragmentos moleculares medidos, σ_k = electrones en orbital sigma, y h_k = átomos de hidrógeno unidos.

antimicrobianos entre péptidos o CQNP. La **Tabla 7** muestra que las mejores clasificaciones se logran con subconjuntos de descriptores, que pueden ser agrupados como SP, VP, MIC, y BCUTw-1h. Puede observarse que los descriptores de tipo SP y VP utilizan conectividad por chi (χ), que es un concepto que incluye características más específicas que la masa o que los grupos funcionales. En la **Figura 15** se observa una de las ecuaciones que llevan a la obtención de los valores numéricos de estos descriptores, mostrando que se utilizan características como orden de enlace, o número de electrones en orbitales sigma (Pearlman y Smith, 1999). Por otra parte, los descriptores de tipo MIC se describen, de acuerdo a la información oficial proporcionada por PaDEL Descriptor, como "descriptores de índice de contenido de información modificada, con un cierto orden de enlace por simetría de vecinos, usando la masa molecular para cada átomo", lo cual da a entender que la masa molecular, a pesar de no ser usada directamente, influye en el resultado de estos cálculos. Finalmente, el descriptor BCUTw-1h revisa diversas subestructuras de una sola molécula, calculando el valor más alto medible para una matriz de adyacencia medida subconjuntos consistentes en los números atómicos de ciertos átomos y a los enlaces que tienen con otros átomos adyacentes (U.S. Environmental Protection Agency, 2008). De esta información se obtiene que las características que darían similitud a péptidos y CQNP son propiedades derivadas de la masa y los vecinos moleculares. Los resultados muestran que la combinación de estos descriptores es adecuada para la clasificación, por lo que se espera que otros descriptores que combinen a la masa y vecinos podrían llegar a mejorar las eficiencias de clasificación reportadas en este trabajo. Estos descriptores adicionales ya se sabe que existen en paquetes adicionales diferentes a PaDEL, tales como ChemDes (Dong *et al.*, 2015), ChemoPy (Cao *et al.*, 2013), PyBel, BlueDesc o RDKit, entre otros. Al menos hay ya

reporte del posible uso de 3679 descriptores (Dong *et al.*, 2015), lo que podría ser una opción a probar a futuro.

Se pueden comparar el valor de curva ROC del mejor modelo en este estudio, con uno de los mejores modelos de péptidos (Beltrán-Verdugo *et al.*, 2018): El mejor modelo en este estudio tuvo un valor de curva ROC de 0.832; el cual es comparable con el 0.85 en la curva ROC del modelo de péptidos. Esto indica que la aprendibilidad de un modelo heterólogo es tan buena como la de los mejores modelos para péptidos.

La comparación con el modelo previamente obtenido en el laboratorio es un tanto más complicada de realizar, puesto que los procedimientos realizados no fueron exactamente los mismos. Sin embargo, con base en el resultado de predicción de CQNP, se puede observar que el mejor modelo obtenido en aquél trabajo, con el algoritmo RandomForest, fue de 74.79% (ver sección de ANTECEDENTES INMEDIATOS), mientras que el mejor modelo usado en este trabajo (con el algoritmo RandomCommittee) presenta un porcentaje de 77.21% de acierto en ese rubro. Es de destacar que el mejor modelo que utiliza el algoritmo RandomForest (y segundo mejor modelo en general en este trabajo) presenta 75.81% de acierto.

Otro aspecto importante de este trabajo son los descriptores obtenidos para clasificar mejor los compuestos antimicrobianos, tanto péptidos como CQNP, que afectan a la microbiota intestinal. Aunque el objetivo de este trabajo no era identificar descriptores comunes para péptidos como CQNP (dado que éstos son previamente calculados mediante ciertos programas, ver sección de METODOLOGÍA), se buscó cuáles descriptores eran relevantes para poder diferenciar entre compuestos antimicrobianos y no antimicrobianos. Los resultados indican que la solución a este problema requiere la transformación y posterior reducción de 86 descriptores moleculares calculados, lo que sugiere que otros

descriptores, probablemente asociados a éstos 86, pueden mejorar el desempeño del mejor modelo de este trabajo.

Para efectos de mejorar este desempeño, es importante hacer notar que los péptidos utilizados en el estudio no se usaron para evaluar las 40 cepas sobre las que se trabajó en el estudio sobre actividad de CQNP en microbiota (Maier et al., 2018), aunque se sabe que algunos de ellos tienen efecto sobre al menos un microorganismo perteneciente a esas cepas (*E. coli*). Por otro lado, los CQNP trabajados tenían actividad antimicrobiana contra al menos una de estas 40 cepas. Dado esto, una alternativa para mejorar el rendimiento de los mejores modelos sería incluir antibióticos dirigidos a los microorganismos más comunes de la microbiota; aunque eso requeriría datos experimentales con los que actualmente no se cuenta.

El *reposicionamiento de fármacos* consiste en que un compuesto ya aprobado por alguna organización de salud (como la FDA) para su uso en el tratamiento de alguna enfermedad o padecimiento, sea propuesto para ser utilizado en el tratamiento de alguna otra enfermedad no relacionada a aquella para la que ya se aprobó (Novac, 2013).

De acuerdo a lo mostrado en la **Tabla 9**, las predicciones sugieren que existen antimicrobianos contra la flora intestinal principalmente entre los compuestos antipsicóticos y antineoplásicos. De los primeros existen reportes que hablan sobre su efecto en bacterias de la microbiota intestinal (Flowers et al., 2017; Bahr et al., 2015), lo cual correlaciona con las observaciones realizadas en el estudio experimental del que deriva este trabajo, en donde se reporta que la clase ATCC de los antipsicóticos está sobrerrepresentada, y que casi todas las subclases de antipsicóticos mostraban alguna actividad contra la microbiota (Maier et al., 2018). En el DiS se reporta que son 3 los antipsicóticos que pudieran tener esta actividad (tioridazina, cariprazina y brexpiprazol) y de los 3 existe reporte

de que afectan a la microbiota intestinal (Dinan y Cryan, 2018; Qureshi *et al.*, 2018).

Con respecto a los compuestos anticáncer, se ha reportado que en general estos compuestos pueden ser antimicrobianos (Imamura *et al.*, 1997; Mir *et al.*, 2003), aunque a la fecha no existen reportes sobre si los antineoplásicos del DiS afectan a la microbiota intestinal.

Estos resultados sugieren que es necesaria una vigilancia en estos compuestos por su posible capacidad de alterar la flora intestinal. Como ya se ha mencionado, los tratamientos con antibióticos duran un promedio de 5 días, mientras que los tratamientos con antipsicóticos o antidiabéticos suelen ser de por vida, o al menos durante más de un año (Ho *et al.*, 2011; Montagnani y Gonnelli, 2013). Es posible que las dosis utilizadas de manera farmacológica para combatir estos males, sean escasas para un combate de microbios que no se esté visualizando (lo cual sería equivalente a consumir dosis incompletas de antibiótico). Esto podría inducir a la generación de resistencia a antibióticos como efecto secundario.

A la fecha, no se han reportado estudios de ML que ayuden a identificar antibióticos de amplio espectro. En este trabajo, la definición de antibióticos de amplio espectro incluye a aquellos que atacan tanto a bacterias patogénicas como a aquellas que no lo son. Por lo tanto, los resultados de los antibióticos que se predice que tendrían actividad antimicrobiana en el DiS serían antibióticos que atacarían a la microbiota intestinal y serían clasificados como de amplio espectro. De acuerdo a los resultados obtenidos, poco más de la mitad de los compuestos antibióticos evaluados podrían representar antibióticos de amplio espectro, incluyendo a 2 compuestos (amoxicilina y cefalexina) de los que el reporte es que son de espectro reducido (Sarpong y Miller, 2015; Kreitmeyr *et al.*, 2017) pero ya existe información de que afectan a la microbiota intestinal (NIH, 2018). Por otro lado, también existen

antibióticos no predichos por el modelo pero que tienen reporte de actividad. Un ejemplo es la Ceftarolina fosamil, compuesto recientemente aprobado por la FDA para tratar infecciones cutáneas y neumónicas; predicho como sin actividad contra la microbiota y que tiene reporte de escasa actividad contra la microbiota (File Jr. et al., 2012).

Para poder destacar la significancia de los resultados de los antibióticos de amplio espectro, fue necesario observar el comportamiento de los compuestos adicionales (antifúngicos). Siendo que todos estos modelos fueron entrenados para revisar actividad sobre bacteria, se esperaba que estos compuestos no tuvieran actividad antibacteriana dado que las características de los hongos son diferentes a las de las bacterias (expectativa-CA). Se comparó esta hipótesis con aquella de que los antibióticos aprobados por la FDA no deberían tener actividad contra la microbiota, pues de lo contrario habría efectos secundarios de malestares gastrointestinales en los pacientes (expectativa-FDA). Para abordar la importancia de estos hallazgos, ambas hipótesis fueron comparadas. Si los antibióticos aprobados por la FDA tienen menos probabilidad de actuar contra la microbiota que los compuestos adicionales, entonces la expectativa-FDA es menor a la expectativa-CA. Y de hecho, se observó que los antibióticos aprobados por la FDA tienen aproximadamente el doble de probabilidad de no actuar contra la microbiota, que los antifúngicos. Esto puede deberse a la alteración de las interacciones entre bacterias y hongos intestinales, las cuales aún no han sido bien esclarecidas.

Por lo tanto, los resultados indican que incluso cuando los antibióticos aprobados por la FDA son más seguros (no actúan contra la microbiota) que el grupo de control conformado por los compuestos adicionales, se identificó que algunos de estos compuestos necesitan ser reevaluados como promotores potenciales de resistencia microbiana por ser de amplio espectro.

10. CONCLUSIONES

- 👉 Los modelos con las mejores clasificaciones para compuestos antimicrobianos de la flora intestinal son de tipo **Heterólogo**.
- 👉 Se identificaron 118 compuestos con posible actividad antimicrobiana, que están aprobados por la FDA para otras actividades; de un total de 829 compuestos.
- 👉 Se identificaron 56 antibióticos que podrían ser de amplio espectro; 3 de ellos se reportaban como de espectro reducido y deberían reevaluarse; de un total de 829 compuestos.

11. PERSPECTIVAS

- 👉 Comprobar algunas de las predicciones realizadas mediante curvas de crecimiento de bacterias de la microbiota en presencia de algunos de estos compuestos.
- 👉 Estudiar el espectro de los 3 antibióticos mencionados para su posible reevaluación.

12. REFERENCIAS BIBLIOGRÁFICAS

- Agarwal, S., Yewale, V.N. y Dharmapalan, D.** (2015) Antibiotics use and misuse in children: A knowledge, attitude and practice survey of parents in India. *J. Clin. Diagnostic Res.*, **9**, SC21–SC24.
- Ageitos, J.M., Sánchez-Pérez, A., Calo-Mata, P. y Villa, T.G.** (2017) Antimicrobial peptides (AMPs): Ancient compounds that represent novel weapons in the fight against bacteria. *Biochem. Pharmacol.*, **133**, 117–138. Available at: <http://dx.doi.org/10.1016/j.bcp.2016.09.018>.
- Alcock, J., Maley, C.C. y Aktipis, C.A.** (2014) Is eating behavior manipulated by the gastrointestinal microbiota? Evolutionary pressures and potential mechanisms. *BioEssays*, **36**, 940–949.
- Bahr, S.M., Tyler, B.C., Wooldridge, N., et al.** (2015) Use of the second-generation antipsychotic, risperidone, and secondary weight gain are associated with an altered gut microbiota in children. *Transl. Psychiatry*, **5**.
- Beisswenger, C. y Bals, R.** (2005) Functions of Antimicrobial Peptides in Host Defense and Immunity. *Curr. Protein Pept. Sci.*, **6**, 255–264. Available at: <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2037&volume=6&issue=3&spage=255>.
- Beltrán-Verdugo, J.A., Aguilera-Mendoza, L. y Brizuela-Rodríguez, C.A.** (2017) Feature weighting for antimicrobial peptides classification: a multi-objective evolutionary approach. *Bioinforma. Biomed.*, 276–283.
- Bray, B.L.** (2003) Large-scale manufacture of peptide therapeutics by chemical synthesis. *Nat. Rev. Drug Discov.*, **2**, 587–593.
- Buckwold, F.J. y Ronald, A.R.** (1979) Antimicrobial misuse-effects and suggestions for control. *J. Antimicrob. Chemother.*, **5**, 129–136.
- Burbidge, R., Trotter, M., Buxton, B. y Holden, S.** (2001) Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.*, **26**, 5–14.
- Calahorra, M., Sánchez-Sánchez, N.S. y Peña-Díaz, A.** (2018) Influence of phenothiazines, phenazines and phenoxazine on cation transport in *Candida albicans*. *J. Appl. Microbiol.*, **125**, 1728–1738.
- Calva, J.** (1996) Antibiotic use in a periurban community in Mexico: A household and drugstore survey. *Soc. Sci. Med.*, **42**, 1121–1128. Available at: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=e-med6&AN=23055360>.

- Chavolla-Canal, A.J., González-Mercado, M.G. y Ruiz-Larios, Ó.A.** (2016) Prevalencia de bacterias aisladas con resistencia antibiótica extendida en los cultivos de orina durante 8 años en un hospital de segundo nivel en México. *Rev. Mex. Urol.*, **76**, 213–217. Available at: <http://dx.doi.org/10.1016/j.uromx.2016.04.003>.
- Col, N.F. y O'Connor, R.W.** (1987) Estimating worldwide current antibiotic usage: report of Task Force 1. *Rev. Infect. Dis.*, **9**, S232–S243. Available at: <papers2://publication/uuid/661208D3-0639-41BD-BA45-B36D5D7239FE>.
- Diener, C., Garza-Ramos Martínez, G., Moreno-Blas, D., Castillo-González, D.A., Corzo, G., Castro-Obregón, S. y Río-Guerra, G. Del** (2016) Effective Design of Multifunctional Peptides by Combining Compatible Functions. *PLoS Comput. Biol.*, **12**, 1–19.
- Dinan, T.G. y Cryan, J.F.** (2018) Schizophrenia and the Microbiome: Time to Focus on the Impact of Antipsychotic Treatment on the Gut Microbiota. *World J. Biol. Psychiatry*, **19**, 568–570.
- Durrant, J.D. y Amaro, R.E.** (2015) Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.*, **85**, 14–21. Available at: <http://doi.wiley.com/10.1111/cbdd.12423>.
- Ertl, P.** (2017) An algorithm to identify functional groups in organic molecules. *J. Cheminform.*, **9**, 1–7.
- Falony, G., Joossens, M., Vieira-Silva, S., et al.** (2016) Population-level analysis of gut microbiome variation. *Science (80-.)*, **352**, 560–564.
- Fernandes, F.C., Rigden, D.J. y Franco, O.L.** (2012) Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. *Biopolymers*, **98**, 280–287.
- Fjell, C.D., Jenssen, H., Hilpert, K., Cheung, W.A., Panté, N., Hancock, R.E.W. y Cherkasov, A.** (2009) Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.*, **52**, 2006–2015.
- Flowers, S.A., Evans, S.J., Ward, K.M., McInnis, M.G. y Ellingrod, V.L.** (2017) Interaction Between Atypical Antipsychotics and the Gut Microbiome in a Bipolar Disease Cohort. *Pharmacotherapy*, **37**, 261–267.
- Frank, E., Hall, M., Trigg, L., Holmes, G. y Witten, I.H.** (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.
- Ho, B.-C., Andreasen, N.C., Ziebell, S., Pierson, R. y Magnotta, V.** (2011) Long-term Antipsychotic Treatment and Brain Volumes. *Arch. Gen. Psychiatry*, **68**, 128.
- Imamura, N., Nishijima, M., Takadera, T., Adachi, K., Sakai, M. y Sano, H.**

- (1997) New Anticancer Antibiotics Pelagiomycins, Produced by a New Marine Bacterium *Pelagibacter variabilis*. *J. Antibiot. (Tokyo)*, **50**, 8–12.
- Kang, S.-J., Park, S.-J., Mishig-Ochir, T. y Lee, B.-J.** (2014) Antimicrobial peptides: therapeutic potentials. *Expert Rev. Anti. Infect. Ther.*, **12**, 1477–1486. Available at:
<http://www.tandfonline.com/doi/full/10.1586/14787210.2014.976613>.
- Kardas, P., Devine, S., Golembesky, A. y Roberts, C.** (2005) A systematic review and meta-analysis of misuse of antibiotic therapies in the community. *Int. J. Antimicrob. Agents*, **26**, 106–113.
- King, R.D., Muggleton, S., Lewis, R.A. y Sternberg, M.J.E.** (1992) Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 11322–6. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/1454814><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC50542>.
- Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F. y Leyton-Brown, K.** (2016) Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.*, **17**, 1–5.
- Kreitmeyr, K., Both, U. Von, Pecar, A., Borde, J.P., Mikolajczyk, R. y Huebner, J.** (2017) Pediatric antibiotic stewardship: successful interventions to reduce broad-spectrum antibiotic use on general pediatric wards. *Infection*, **45**, 493–504.
- Lata, S., Sharma, B.K. y Raghava, G.P.S.** (2007) Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, **8**, 1–10.
- Lavecchia, A.** (2015) Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today*, **20**, 318–331.
- Li, Y.** (2014) China’s misuse of antibiotics should be curbed. *BMJ*, **348**, 1–2. Available at: <http://dx.doi.org/doi:10.1136/bmj.g1083>.
- Mahlapuu, M., Håkansson, J., Ringstad, L. y Björn, C.** (2016) Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. *Front. Cell. Infect. Microbiol.*, **6**, 1–12. Available at:
<http://journal.frontiersin.org/article/10.3389/fcimb.2016.00194/full>.
- Maier, L., Pruteanu, M., Kuhn, M., et al.** (2018) Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, **555**, 623–628. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/29555994>.
- Mark, H., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P. y Witten, I.H.** (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explor.*, **11**, 10–18.

- Méndez-Samperio, P.** (2014) Peptidomimetics as a new generation of antimicrobial agents: Current progress. *Infect. Drug Resist.*, **7**, 229–237.
- Milletti, F.** (2012) Cell-penetrating peptides: Classes, origin, and current landscape. *Drug Discov. Today*, **17**, 850–860.
- Minkiewicz, P., Iwaniak, A. y Darewicz, M.** (2017) Annotation of peptide structures using SMILES and other chemical codes-practical solutions. *Molecules*, **22**, 1–17.
- Mir, M.A., Majee, S., Das, S. y Dasgupta, D.** (2003) Association of chromatin with anticancer antibiotics, mithramycin and chromomycin A3. *Bioorganic Med. Chem.*, **11**, 2791–2801.
- Mojsoska, B. y Jenssen, H.** (2015) Peptides and peptidomimetics for antimicrobial drug design. *Pharmaceuticals*, **8**, 366–415.
- Montagnani, A. y Gonnelli, S.** (2013) Antidiabetic therapy effects on bone metabolism and fracture risk. *Diabetes, Obes. Metab.*, **15**, 784–791. Available at: <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L369444069%5Cnhttp://dx.doi.org/10.1111/dom.12077%5Cnhttp://rug.on.worldcat.org/atoztitles/link/?sid=EMBASE&issn=14628902&id=doi:10.1111%2Fdom.12077&atitle=Antidiabetic+therapy+effec>.
- Morán-Torres, R.U.** (2019) *DISEÑO Y EVALUACIÓN DE PÉPTIDOS PENETRADORES CELULARES SELECTIVOS*. Tesis de maestría, UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO.
- Mullard, A.** (2017) The Drug-Maker’s Guide To the Galaxy: How Machine Learning and Big Data Are Helping Chemists Search the Vast Chemical Universe for Better Medicin. *Nature*, **549**, 445–447. Available at: http://www.nature.com/news/the-drug-maker-s-guide-to-the-galaxy-1.22683?WT.mc_id=FBK_NatureNews&sf117719919=1.
- Nagarajan, K., Marimuthu, S.K., Palanisamy, S. y Subbiah, L.** (2018) Peptide Therapeutics Versus Superbugs: Highlight on Current Research and Advancements. *Int. J. Pept. Res. Ther.*, **24**, 19–33. Available at: <http://dx.doi.org/10.1007/s10989-017-9650-0>.
- Napolitano, F., Zhao, Y., Moreira, V.M., Tagliaferri, R., Kere, J., D’Amato, M. y Greco, D.** (2013) Drug Repositioning: A Machine-Learning Approach through Data Integration. *J. Cheminform.*, **5**, 1–9.
- Newland, J.G., Stach, L.M., Lurgio, S.A. De, et al.** (2012) Impact of a prospective-audit-with-feedback antimicrobial stewardship program at a children’s hospital. *J. Pediatric Infect. Dis. Soc.*, **1**, 179–186.
- Newman, R.E., Hedican, E.B., Herigon, J.C., Williams, D.D., Williams, A.R. y**

- Newland, J.G.** (2012) Impact of a Guideline on Management of Children Hospitalized With Community-Acquired Pneumonia. *Pediatrics*, **129**, e597–e604. Available at: <http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.2011-1533>.
- Novac, N.** (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci.*, **34**, 267–272.
- Pearlman, R.S. y Smith, K.M.** (1999) Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.*, **39**, 28–35.
- Pentima, M.C. Di y Chan, S.** (2010) Impact of antimicrobial stewardship program on vancomycin use in a pediatric teaching hospital. *Pediatr. Infect. Dis. J.*, **29**, 707–711.
- Piotto, S.P., Sessa, L., Concilio, S. y Iannelli, P.** (2012) YADAMP: Yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents*, **39**, 346–351.
- Pirtskhalava, M., Gabrielian, A., Cruz, P., et al.** (2016) DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.*, **44**, D1104–D1112.
- Polanco-González, C.** (2009) *Determinación de secuencias peptídicas antibacterianas*. Tesis de doctorado, Universidad Nacional Autónoma de México.
- Porto, W.F., Fernandes, F.C. y Franco, O.L.** (2010) An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, **6268 LNBI**, 59–62.
- Porto, W.F., Pires, Á.S. y Franco, O.L.** (2012) CS-AMPPred: An Updated SVM Model for Antimicrobial Activity Prediction in Cysteine-Stabilized Peptides. *PLoS One*, **7**, e51444.
- Qureshi, N.A., Al-Dossari, D.S., Salem, S.O., Alharbi, F.K., Alkhamees, O.A. y Alsanad, S.M.** (2018) Antipsychotic Medications and Weight Gain : Etiologies , Predictors and Adverse Clinical Consequences. *Int. Neuropsychiatr. Dis. J.*, **11**, 1–19.
- Río-Guerra, G. Del, Koschützki, D. y Coello, G.** (2009) How to identify essential genes from molecular networks? *BMC Syst. Biol.*, **3**, 102.
- Rodríguez-Plaza, J.G., Morales-Nava, R., Diener, C., et al.** (2014) Cell penetrating peptides and cationic antibacterial peptides: Two sides of the same coin. *J. Biol. Chem.*, **289**, 14448–14457.
- Rodríguez-Plaza, J.G., Rivas-Santiago, B., Hernández-Pando, R. y Río-Guerra, G. Del** (2014) Prospective Tuberculosis Treatment: Peptides, Immunity and

- Autophagy. *J. Mol. Genet. Med.*, **08**, 1000128. Available at:
<http://www.omicsonline.com/open-access/prospective-tuberculosis-treatment-peptides-immunity-and-autophagy-1747-0862.1000128.php?aid=31157>.
- Rossolini, G.M.** (2016) Multidrug-Resistant and Extremely Drug-Resistant Bacteria: Are We Facing the End of the Antibiotic Era? *J. Siena Acad. Sci.*, **7**. Available at:
<http://pagepressjournals.org/index.php/jsas/article/view/6409>.
- Rousounides, A., Papaevangelou, V., Hadjipanayis, A., Panagakou, S., Theodoridou, M., Syrogiannopoulos, G. y Hadjichristodoulou, C.** (2011) Descriptive study on parents' knowledge, attitudes and practices on antibiotic use and misuse in children with upper respiratory tract infections in Cyprus. *Int. J. Environ. Res. Public Health*, **8**, 3246–3262.
- Sarpong, E.M. y Miller, G.E.** (2015) Narrow- and broad-spectrum antibiotic use among U.S. children. *Health Serv. Res.*, **50**, 830–846.
- Splith, K. y Neundorf, I.** (2011) Antimicrobial peptides with cell-penetrating peptide properties and vice versa. *Eur. Biophys. J.*, **40**, 387–397.
- Sterling, T. y Irwin, J.J.** (2015) ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, **55**, 2324–2337.
- Tanwar, J., Das, S., Fatima, Z. y Hameed, S.** (2014) Multidrug resistance: An emerging crisis. *Interdiscip. Perspect. Infect. Dis.*, **2014**, e541340. Available at:
<http://www.hindawi.com/journals/ipid/2014/541340/abs/>,
<http://www.hindawi.com/journals/ipid/2014/541340/abs/%5Cnhttp://downloads.hindawi.com/journals/ipid/2014/541340.pdf%5Cnhttp://www.hindawi.com/journals/ipid/2014/541340/%5Cnhttp://www.ncbi.nlm.nih.gov/>.
- Thomas, S., Karnik, S., Barai, R.S., Jayaraman, V.K. y Idicula-Thomas, S.** (2009) CAMP: A useful resource for research on antimicrobial peptides. *Nucleic Acids Res.*, **38**.
- Thornton, C., Hutter, F., Hoos, H.H. y Leyton-Brown, K.** (2012) Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. *KDD*, 1–9. Available at: <http://arxiv.org/abs/1208.3719>.
- Tonarelli, G. y Simonetta, A.** (2013) Péptidos antimicrobianos de organismos procariotas y eucariotas como agentes terapéuticos y conservantes de alimentos. *Fabrib*, **17**, 137–177. Available at:
<http://bibliotecavirtual.unl.edu.ar/ojs/index.php/FABICIB/article/download/4316/6549%5Cnhttp://bibliotecavirtual.unl.edu.ar/publicaciones/index.php/FABICIB/article/view/4316>.
- U.S. Environmental Protection Agency** (2008) Description of the Molecular

Descriptors Appearing in the Toxicity Estimation Software Tool. , 6–8.

- Uhlig, T., Kyprianou, T., Martinelli, F.G., Oppici, C.A., Heiligers, D., Hills, D., Calvo, X.R. y Verhaert, P.** (2014) The emergence of peptides in the pharmaceutical business: From exploration to exploitation. *EuPA Open Proteomics*, **4**, 58–69. Available at: <http://dx.doi.org/10.1016/j.euprot.2014.05.003>.
- Veltri, D., Kamath, U. y Shehu, A.** (2015) Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **14**, 1–1. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7172462>.
- Wang, G.** (2015) Improved Methods for Classification, Prediction and Design of Antimicrobial Peptides. *Comput. Pept.*, 1–333.
- Xiao, X., Wang, P., Lin, W.Z., Jia, J.H. y Chou, K.C.** (2013) IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **436**, 168–177.
- Yap, C.W.** (2011) PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.
- Yeung, A.T.Y., Gellatly, S.L. y Hancock, R.E.W.** (2011) Multifunctional cationic host defence peptides and their clinical applications. *Cell. Mol. Life Sci.*, **68**, 2161–2176.
- <https://github.com/MidoriR/ColorantesPena>, consultada el 10 de octubre de 2018 a las 17:00.

13. APÉNDICES

Apéndice 1. Lista de compuestos utilizados en el estudio del cual se obtuvo el conjunto de entrenamiento (TrS). En verde compuestos antibacterianos, en azul otros antiparasitarios, y en rojo compuestos para células humanas. En tonos claros compuestos sin actividad contra alguna de las 40 cepas estudiadas, y en color oscuro aquellos con actividad. En **negrita** compuestos con 2 estereoisómeros, y en **mayúscula cursiva** aquellos con 4, lo que totaliza 861 compuestos.

Chlorphenesin	Sulfameter	Anthralin	Clebopride
Nystatin	Sulfamethazine	Antipyrine	Clemastine
Rimantadine	Sulfamethizole	Aripiprazole	Clobetasol
Albendazole	Sulfamethoxazole	Ascorbic acid	Clocortolone
Bephenium	Sulfamethoxyipyridazine	Asenapine	Clodronate
Diethylcarbamazine	Sulfanilamide	Astenile	Clofibrate
Flubendazol	Sulfaphenazole	Atorvastatin	Clonidine
Levamisole	Sulfapyridine	Azacyclonol	Clonixin
Mebendazole	Sulfathiazole	Balsalazide	Clargiline
Oxantel	Sulfisoxazole	Beclomethasone	Clozapine
Pyrantel	Viomycin	Bemegride	Colchicin
Ciclopirox	Sulfabenzamide	Benazepril	Cortisol
Fluconazole	Baclofen	Benfotiamine	Cortisone
Griseofulvin	Doxepin	Benoxinate	Cromolyn
Liranaftate	Dropropizine	Benperidol	Cyanocobalamin
Naftifine	Homatropine	Benzamil	Cyclizine
Terconazole	Isoproterenol	Benzocaine	Cyproterone
Tiabendazole	Meloprolol	Benzonate	Cytarabine
Voriconazole	Octopamine	Benzthiazide	Danazol
Etanidazole	Propranolol	Benztropine	Debrisoquin
Meglumine	Syneprhine	Bethistine	Decamethonium
Proguanil	Trihexyphenidyl	Betamethasone	Deferoxamine
Abacavir	Velnacrine	Betazole	Deflazacort
Acyclovir	3-alpha-Hydroxy-5-beta-androstan-17-one	Bezafibrate	Desipramine
Didanosine	5-fluorouracil	Biotin	Desloratadine
Emtricitabine	Acarbose	Bisacodyl	Dexamethasone
Famciclovir	Aceclofenac	Bretylum	Dexfenfluramine
Ganciclovir	Acefylline	Brinzolamide	Dextromethorphan
Lamivudine	Acemetacin	Bromhexine	Diazoxide
Moroxydine	Acepromazine	Bromopride	Dibenzepine
Penciclovir	Acetazolamide	Bucladesine	Dibucaine
Podophyllotoxin	Acetohexamide	Buflomedil	Dichlorphenamide
Stavudine	Acetylsalicylic acid	Bumetanide	Diclofenac
Valacyclovir	Acipimox	Buspiron	Diflorasone
Zalcitabine	Acitretin	Busulfan	Diflunisal
Kanamycin A	Adamantamine	Butacaine	Digitoxigenin
2-Aminobenzenesulfonamide	Adiphenine	Butalbital	Dihydroergotamine
4-aminosalicylic acid	Adrenosterone	Butylscopolamine	Dilazep
Ampicillin	Alcuronium	Caffeine	Diltiazem
Dapsone	Alendronate	Calcipotriene	Dimethadione
D-cycloserine	Alfaxalone	Candesartan	Dinoprost
Ethambutol	Allopurinol	Canrenoic acid	Diphepanil
Ethionamide	Alprostadi	Canrenone	Diphenhydramine
Isoniazid	Altretamine	Captopril	Diphenidol
Methenamine	Alverine	Carbachol	Diphenylpyraline
Neomycin	Ambroxol	Carbamazepine	Dipyridamole
Paromomycin	Amcinonide	Carbetapentane	Dipyron
Phthalylsulfathiazole	Amidopyrine	Carbimazole	Dizocilpine
Piromidic acid	Amifostine	Celecoxib	Docetaxel
Prothionamide	Aminocaproic acid	Chenodiol	Dolasetron
Pyrazinamide	Aminohippuric acid	Chlorambucil	Domperidone
Streptomycin	Amiodarone	Chlormadinone	Dopamine
Succinylsulfathiazole	Amityriptiline	Chloropyramine	Dorzolamide
Sulfacetamide	Amrinone	Chlorothiazide	Dosulepin
Sulfadiazine	Anastrozole	Chlorpropamide	Doxofylline
Sulfadimethoxine	Androsterone	Chlorzoxazone	Droperidol
Sulfaguandine	Anetholtrithion	Clostazol	Dyclonine
Sulfamerazine	Antazoline	Cimetidine	Dydrogesterone

Edrophonium	Hymecromone	Meticrane	Pentolinium	Selegiline
Emedastine	Hyoscyamine	Metoclopramide	Pentoxifylline	Sildenafil
Enalapril	Ibandronate	Metrizamide	Pentylentetrazol	Spaglumic
Epiandrosterone	Ibudilast	Metypalone	Pergolide	Spironolactone
Epitiofanol	Ibuprofen	Mevastatin	Perindopril	Sulfasalazine
Equilin	Imatinib	Mianserine	Phenacetin	Suxibuzone
Escitalopram	Imipramine	Miglitol	Phenazopyridine	Tacrine
Eserine	Indomethacin	Milrinone	Phenelzine	Telenzepine
Eseroline	Iobenguane	Minaprine	Phenformin	Tenoxicam
Estradiol-17 beta	Iopamidol	Minoxidil	Phentermine	Tetracaine
Estriol	Iproniazide	Mirtazapine	Phentolamine	Theobromine
Estrone	Irinotecan	Mizolastine	Phenylbutazone	Theophylline
Ethamivan	Irsogladine	Moclobemide	Picotamide	Thiamine
Ethamsylate	Isocarboxazid	Molsidomine	Pilocarpine	Thiopropazine
Ethinylestradiol	Isometheptene	Monobenzone	Pinacidil	Tiaprude
Ethisterone	Isopropamide	Moxisylyte	Piracetam	Tibolone
Ethynodiol	Isosorbide mononitrate	Moxonidine	Pirenzepine	Ticlopidine
Etidronic acid	Isotretinoin	Nabumetone	Piretanide	Timolol
Etifenin	Isoxicam	N-Acetyl-L-leucine	Piribedil	Tizanidine
Etofilline	Itopride	Nalbuphine	Piroxicam	Tolazamide
Etoricoxib	Ketanserin	Nalmefene	Pralidoxime	Tolazoline
Exemestane	Ketotifen	Naloxone	Pramipexole	Tolbutamide
Felbinac	Lamotrigine	Naltrexone	Pramoxine	Tolmetin
Fenofibrate	Lanatoside C	Nandrolone	Pravastatin	Tomoxetine
Fenspiride	Letrozole	Naphazoline	Prazosin	Topiramate
Finasteride	Levalbuterol	Naproxen	Prednicarbate	Topotecan
Fipexide	Levobunolol	Neostigmine	Prednisolone	Torsemide
Flavoxate	Levocabastine	Niacin	Prednisone	Tranexamic acid
Fludrocortisone	Levodopa	Nialamide	Primidone	Tranilast
Flumetasone	Levonordefrin	Nicergoline	Probenecid	Tranlycypromine
Flunisolide	Lidocaine	Nicotinamide	Probucof	Trapidil
Fluocinolone	Liothyronine	Nifedipine	Procainamide	Triamcinolone
Fluocinonide	Lisinopril	Nifenazone	Procaine	Triamterene
Flurandrenolide	L-Methyldopa	Niflumic acid	Procarbazine	Triflusal
Flutamide	Losartan	Nizatidine	Progesterone	Trimetazidine
Fluticasone	Lovastatin	Nocodazole	Propantheline	Trimethadione
Fluvoxamine	Maprotiline	Nomegestrol	Proparacaine	Trimethobenzamide
Fomepizole	Meclofenoxate	Norethindrone	Propofol	Trioxsalen
Formestane	Medrysone	Norethynodrel	Propoxycaine	Tripelennamine
Fosfosal	Mefenamic acid	Norgestrel	Propylthiouracil	Tripolidine
Furosemide	Mefexamide	Nortriptyline hydrochloride	Proscillaridin A	Tropisetron
Gabapentin	Megestrol	Olanzapine	Pyridostigmine	Tyloxapof
Galanthamine	Melatonin	Olmesartan	Pyridoxine	Urapidil
Gallamine	Meloxicam	Olopatadine	Pyrilamine	Urosiol
Gemfibrozil	Mephentermine	Opipramol	Pyrithyldione	Valproic acid
Gestrinone	Meprylcaine	Oxandrolone	Quinapril	Vatalanib
Glibenclamide	Mesalamine	Oxcarbazepine	Ramipril	Vincamine
Glimepiride	Mesna	Oxolamine	Ranitidine	Vinpocetine
Glipizide	Mestranol	Oxymetazoline	Remoxipride	Vorinostat
Guaiacol	Metaraminol	Ozagrel	Repaglinide	Xylometazoline
Guanabenz	Metformin	Pancuronium	Retinoic acid	Yohimbine
Guanethidine	Methapyrilene	Panthenol	Rimexolone	Zaleplon
Guanfacine	Methazolamide	Papaverine	Risperidone	Zomepirac
Halcinonide	Methimazole	Pargyline	Ritodrine	Zonisamide
Homochlorcyclizine	Methylatropine	Paroxetine	Rivastigmine	AVERMECTIN B1A
Hydralazine	Methyldopate	Pempidine	Rofecoxib	Ivermectin
Hydrochlorothiazide	Methylergometrine	Penbutolol	Ropinirole	Miconazole
Hydroflumethiazide	Methylprednisolone	Penicillamine	Roxatidine	Seritaconazole

Apéndice 1 (Cont).

Sulconazole	Cefoperazone	Sparfloxacin	Darifenacin	Moricizine
Tioconazole	Ceforanide	Spectinomycin	Daunorubicin	Nefazodone
Secnidazole	Cefotaxime	Sulbactam	Demecarium	Nicorandil
Disulfiram	Cefotetan	Tazobactam	Deptropine	Nilutamide
Fenbendazole	Cefotiam	Tetracycline	Desmethylocyclobenzaprine	Nimesulide
Hycanthone	Cefoxitin	Thiamphenicol	Diacerein	Norgestimate
Niclosamide	Ceftazidime	Tinidazole	Dicumarol	Oxaprozin
Niridazole	Cefuroxime	Tobramycin	Dicyclomine	Oxethazaine
Pyrvinium	Cephalothin	Trimethoprim	Dienestrol	Paclitaxel
Butenafine	Chloramphenicol	Troleandomycin	Digoxin	Pemirolast
Clotrimazole	Chloroxine	Vancomycin hydrochloride	Dimethisoquin	Perphenazine
Flucytosine	Chlortetracycline	Benzethonium	Doxorubicin	Phenindione
Haloprogin	Cinoxacin	Chlorhexidine	Entacapone	Pimethixene
Ketoconazole	Ciprofloxacin	Dequalinium	Epirizole	Pimozide
Oxiconazole	Clarithromycin	Hexachlorophene	Erlotinib	Pizotifen
Tolnaftate	Clavulanate	Thimerosal	Estradiol Valerate	Pranlukast
Amodiaquin	Clindamycin	Thonzonium	Estopipate	Pridinol
Artemisinin	Clofazimine	Triclosan	Ethacrynic acid	Prochlorperazine
Atovaquone	Cloxacillin	Atenolol	Ethaverine	Promazine
Clioquinol	Demeclocycline	Diethylstilbestrol	Etofenamate	Protriptyline
Diloxanide	Dicloxacillin	Lansoprazole	Etomidate	Quetiapine
Nitrofuril	Dirithromycin	Omeprazole	Etoposide	Raloxifene
Pentamidine	Doxycycline	8-Azaguanine	Etreinate	Reserpine
Pyrimethamine	Enoxacin	Acamprosate	Famotidine	Riluzole
Efavirenz	Fleroxacin	Acetaminophen	Fenbufen	Sertindole
Idoxuridine	Flucloxacillin	Acetylcysteine	Fentiazac	Sertraline
Imiquimod	Furazolidone	Alclometasone	Floxuridine	Simvastatin
Ribavirin	Fusidic acid	Alfacalcidol	Fludarabine	Sipiperone
Saquinavir	Imipenem	Alfadolone	Flufenamic acid	Stanozolol
Trifluridine	Lincomycin	Ambrisentan	Flunarizine	Streptozotocin
Vidarabine	Linezolid	Amiloride	Fluorometholone	Suloctidil
Zidovudine	Loracarbef	Amoxapine	Fluphenazine	Sumatriptan
JOSAMYCIN	Lymecycline	Aniracetam	Fluspirilen	Tamoxifen
RIFABUTIN	Mafenide	Apomorphine	Folic acid	Telmisartan
RIFAMPICIN	Meclocycline	Aprepitant	Gefitinib	Temozolomide
RIFAPENTINE	Meropenem	Astemizole	Gemcitabine	Testosterone
RIFAXIMIN	Methacycline	Azacytidine-5	Gliqidone	Thiethylperazine
TICARCILLIN	Metronidazole	Azathioprine	Haloperidol	Thioguanosine
Erythromycin	Minocycline	Benzbromarone	Idebenone	Thyroxine
Gentamicin	Moxifloxacin	Benzylamine	Isosorbide dinitrate	Tiratricol
Nadifloxacin	Nafcillin	Bosentan	Lacidipine	Tolfenamic acid
Oflloxacin	Nalidixic acid	Bromperidol	Leflunomide	Toremifene
Ornidazole	Nifuroxazide	Bufexamac	Lidoflazine	Trazodone
Phenethicillin	Nitrofurantoin	Butamben	Loperamide	Trifluoperazine
Spiramycin	Norfloxacin	Carbenoxolone	Loratadine	Triflupromazine
Talampicillin	Oxacillin	Chlorotrianisene	Loxapine	Vanoxerine
7-aminocephalosporanic acid	Oxolinic acid	Chlorpromazine	Lynestrenol	Vardenafil
Azlocillin	Oxytetracycline	Chlorprothixene	Mebhydrolin	Vecuronium
Aztreonam	Pefloxacin	Cinnarizine	Meclofenamic acid	Vinburnine
Benzathine	Pipemidic	Cladribine	Mercaptopurine	Zafirlukast
Benzylpenicillin	Piperacillin	Clemizole	Metergoline	Ziprasidone
Cefaclor	Pivampicillin	Cloflilium	Methantheline	Zotepine
Cefadroxil	Pivmecillinam	Clomiphene	Methiazole	Zuclopenthixol
Cefazolin	Ribostamycin	Clomipramine	Methotrexate	
Cefdinir	Roxithromycin	Cyclobenzaprine	Methotrimeprazine	
Cefepime	Rufloxacin	Cyproheptadine	Mifepristone	
Cefixime	Sarafloxacin	Dacarbazine	Mometasone	
Cefmetazole	Sisomicin	Dantrolene	Montelukast	

Apéndice 1 (Cont).

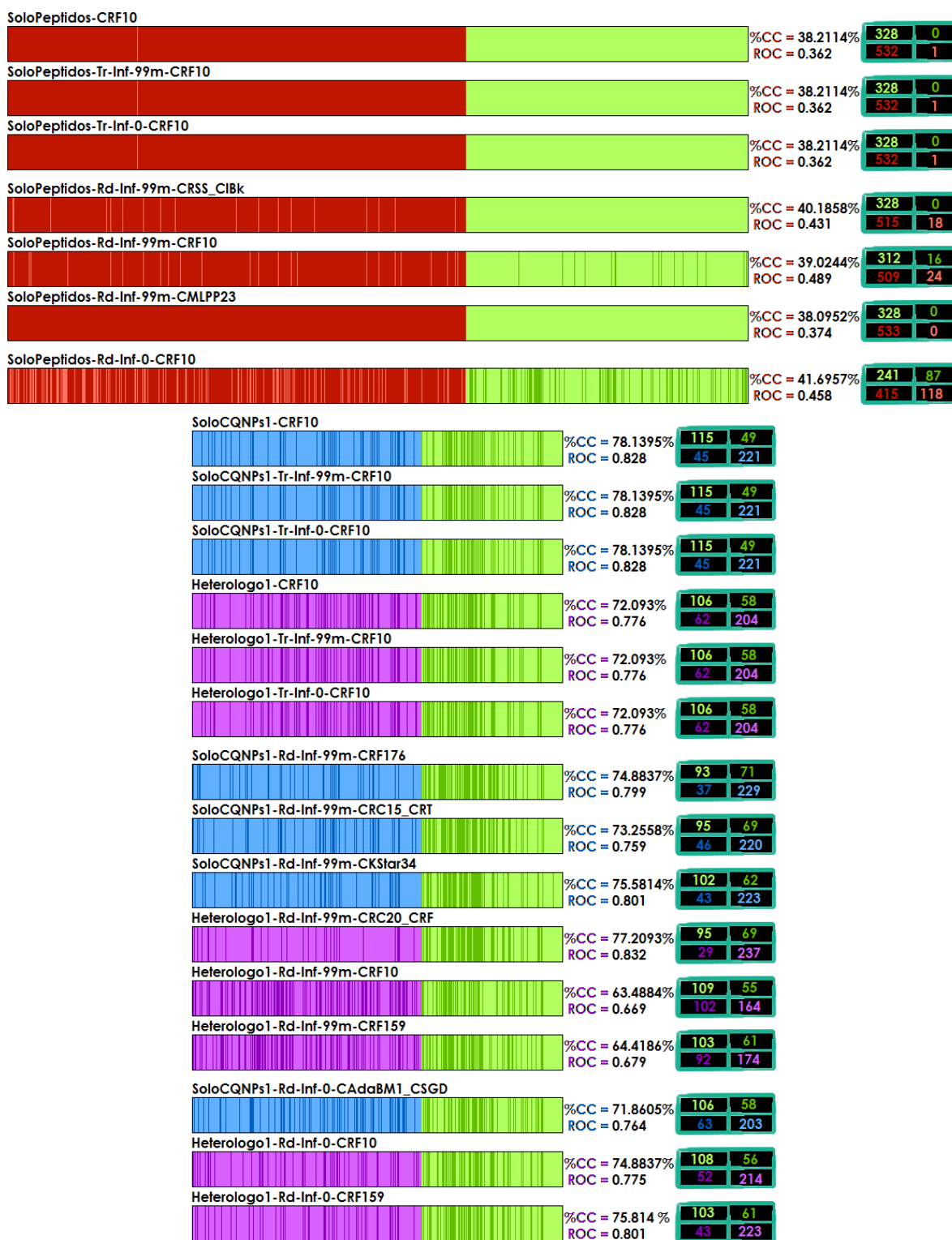
Apéndice 2. Lista de abreviaturas utilizadas para denotar las 55 diferentes combinaciones algoritmo-parámetros de los modelos trabajados.





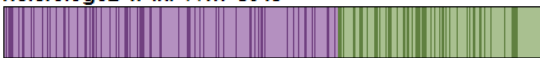
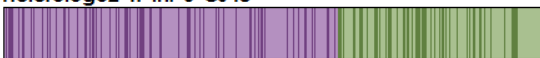

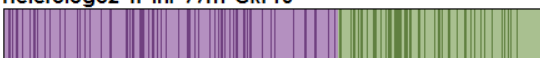
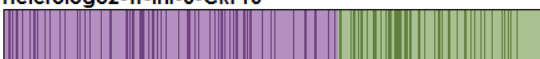
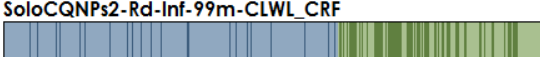
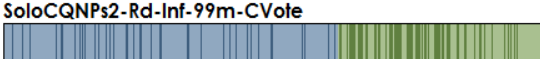

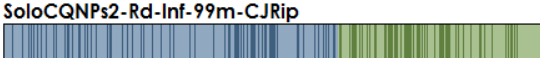



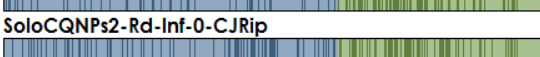
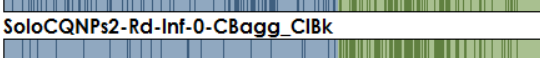

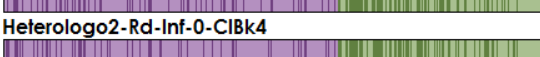
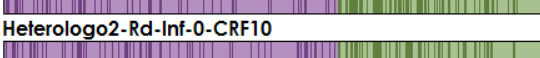
Combinación	Significado
CAdaBM1_CRF12	Clasificador AdBoostM1, que utiliza al clasificador Random Forest con 12 iteraciones.
CAdaBM1_CRF134	Clasificador AdBoostM1, que utiliza al clasificador Random Forest con 134 iteraciones.
CAdaBM1_CSGD	Clasificador AdBoostM1, que utiliza al clasificador SGD.
CBagg_CIBk	Clasificador Bagging, que utiliza al clasificador IBk.
CBagg_CLMT	Clasificador Bagging, que utiliza al clasificador LMT.
CBagg_CLogis	Clasificador Bagging, que utiliza al clasificador Logistic.
CBagg_CRF	Clasificador Bagging, que utiliza al clasificador Random Forest.
CDS	Clasificador Decision Stump.
CDT	Clasificador Decision Table.
CIBk1	Clasificador Ibk usando 1 vecino.
CIBk1E	Clasificador Ibk usando 1 vecino y validado con el error cuadrado de la media en vez del error absoluto de la media.
CIBk3	Clasificador Ibk usando 3 vecinos.
CIBk4	Clasificador Ibk usando 4 vecinos.
CIBk6	Clasificador Ibk usando 6 vecinos.
CIBk35	Clasificador Ibk usando 35 vecinos.
CJ48	Clasificador J48.
CJRip	Clasificador JRip.
CKStar34	Clasificador Kstar usando una mezcla global de valor 34.
CKStar65	Clasificador Kstar usando una mezcla global de valor 65.
CKStar73	Clasificador Kstar usando una mezcla global de valor 73.
CKstar94	Clasificador Kstar usando una mezcla global de valor 94.
CLogis	Clasificador Logistic.
CLWL_CIBk	Clasificador LWL, que utiliza al clasificador IBk.
CLWL_CNB	Clasificador LWL, que utiliza al clasificador Naive Bayes.
CLWL_CRF	Clasificador LWL, que utiliza al clasificador Random Forest.
CMLPP12	Clasificador MultiLayer Perceptron, usando un ratio de aprendizaje de actualización de pesos de 0.12...
CMLPP23	Clasificador MultiLayer Perceptron, usando un ratio de aprendizaje de actualización de pesos de 0.23...
CMLPP90	Clasificador MultiLayer Perceptron, usando un ratio de aprendizaje de actualización de pesos de 0.90...
CRC10_CRF	Clasificador Random Committee con 10 iteraciones, que utiliza al clasificador Random Forest.
CRC11_CRT	Clasificador Random Committee con 11 iteraciones, que utiliza al clasificador Random Tree.
CRC15_CRT	Clasificador Random Committee con 15 iteraciones, que utiliza al clasificador Random Tree.
CRC20_CRF	Clasificador Random Committee con 20 iteraciones, que utiliza al clasificador Random Forest.
CRF7	Clasificador Random Forest usando 7 iteraciones.
CRF10	Clasificador Random Forest usando 10 iteraciones.
CRF159	Clasificador Random Forest usando 159 iteraciones.
CRF163	Clasificador Random Forest usando 163 iteraciones.
CRF176	Clasificador Random Forest usando 176 iteraciones.
CRSS_CIBk	Clasificador Random Subspace, que utiliza al clasificador IBk.
CRSS_CPART	Clasificador Random Subspace, que utiliza al clasificador PART.
CRSS_CSiLog	Clasificador Random Subspace, que utiliza al clasificador Simple Logistic.
CSGD0,001	Clasificador SGD, usando un ratio de aprendizaje de 0.001...
CSGD0,042	Clasificador SGD, usando un ratio de aprendizaje de 0.042...
CSGD2,043	Clasificador SGD, usando un ratio de aprendizaje de 2.043...
CSiLog0	Clasificador Simple Logistic, sin utilizar poda por peso.
CSiLog0,055	Clasificador Simple Logistic, usando un valor beta de poda por peso de 0.055...
CSMO_CSVNoPKe	Clasificador Supporting Machine Operator, usando el núcleo Support Vector Normalized PolyKernel.
CSMO09_CSVPuk	Clasificador Supporting Machine Operator, con una complejidad de 1.09... y usando el núcleo Support Vector Puk.
CSMO15_CSVPuk	Clasificador Supporting Machine Operator, con una complejidad de 1.15... y usando el núcleo Support Vector Puk.
CSMO15_CSVRBFKe	Clasificador Supporting Machine Operator, con una complejidad de 1.15... y usando el núcleo Support Vector RBF-Kernel.
CSMO28_CSVRBFKe	Clasificador Supporting Machine Operator, con una complejidad de 1.28... y usando el núcleo Support Vector RBF-Kernel.
CSMO47_CSVRBFKe	Clasificador Supporting Machine Operator, con una complejidad de 1.47... y usando el núcleo Support Vector RBF-Kernel.
CSMO67_CSVPKe	Clasificador Supporting Machine Operator, con una complejidad de 0.67... y usando el núcleo Support Vector PolyKernel.
CSMO87_CSVPKe	Clasificador Supporting Machine Operator, con una complejidad de 0.87... y usando el núcleo Support Vector PolyKernel.
CVote	Clasificador Vote.
CZeroR	Clasificador ZeroR.

Apéndice 3. Datos numéricos asociados a la **Figura 5**, con el mismo código de color. Los datos en negritas corresponden a modelos que cumplen el criterio de %CC mayor a 90%.

Combinación	Instancias	Edición	%CC	Combinación	Instancias	Edición	%CC
CRF159	Heter1	Rd-Inf-0	100.0000%	CRF10	Heter1	Rd-Inf-99m	98.8717%
CRF159	Heter1	Rd-Inf-99m	100.0000%	CLWL_CNB	SCQNP3	Rd-Inf-99m	98.8372%
CRF159	Heter3	Rd-Inf-0	100.0000%	CRF10	Heter3	Rd-Inf-99m	98.8070%
CIBk1E	Heter3	Rd-Inf-0	100.0000%	CJRip	SCQNP2	Rd-Inf-0	98.3721%
CRF159	Heter3	Rd-Inf-99m	100.0000%	CJRip	SCQNP2	Rd-Inf-99m	98.3721%
CRC10_CRF	Heter3	Rd-Inf-99m	100.0000%	CRF10	Heter4	Rd-Inf-0	98.3398%
CRF159	Heter4	Rd-Inf-0	100.0000%	CBagg_CLMT	SCQNP3	Rd-Inf-99m	98.1395%
CRF159	Heter4	Rd-Inf-99m	100.0000%	CBagg_CLogis	SCQNP3	Original	95.5814%
CKStar34	SCQNP1	Rd-Inf-99m	100.0000%	CBagg_CLogis	SCQNP3	Tr-Inf-0	95.5814%
CRF176	SCQNP1	Rd-Inf-99m	100.0000%	CBagg_CLogis	SCQNP3	Tr-Inf-99m	95.5814%
CRF10	SCQNP2	Original	100.0000%	CAAdBM1_CSGD	SCQNP1	Rd-Inf-0	95.1276%
CRF10	SCQNP2	Tr-Inf-0	100.0000%	CRC15_CRT	SCQNP1	Rd-Inf-99m	95.1276%
CRF10	SCQNP2	Tr-Inf-99m	100.0000%	CBagg_CRF	SCQNP3	Rd-Inf-0	94.4186%
CIBk6	SCQNP2	Rd-Inf-0	100.0000%	CJ48	Heter2	Original	93.2774%
CLWL_CRF	SCQNP2	Rd-Inf-99m	100.0000%	CJ48	Heter2	Tr-Inf-0	93.2774%
CAAdBM1_CRF12	SCQNP2	Rd-Inf-99m	100.0000%	CJ48	Heter2	Tr-Inf-99m	93.2774%
CVote	SCQNP2	Rd-Inf-99m	100.0000%	CRSS_CIBk	SP	Rd-Inf-99m	92.5082%
CLWL_CIBk	SCQNP3	Rd-Inf-0	100.0000%	CMLPP23	SP	Rd-Inf-99m	91.3130%
CLWL_CIBk	SCQNP3	Rd-Inf-99m	100.0000%	CRSS_CPART	SCQNP2	Original	87.2093%
CRF10	SCQNP3	Rd-Inf-99m	100.0000%	CRSS_CPART	SCQNP2	Tr-Inf-0	87.2093%
CAAdBM1_CRF134	SCQNP3	Rd-Inf-99m	100.0000%	CRSS_CPART	SCQNP2	Tr-Inf-99m	87.2093%
CKStar73	SCQNP4	Rd-Inf-0	100.0000%	CRF7	SCQNP4	Rd-Inf-99m	86.5429%
CKStar65	SCQNP4	Rd-Inf-0	100.0000%	CRC11_CRT	SCQNP4	Rd-Inf-99m	86.3109%
CIBk3	SCQNP4	Rd-Inf-99m	100.0000%	CRF7	SCQNP1	Rd-Inf-99m	85.3828%
CRF163	SCQNP4	Rd-Inf-99m	100.0000%	CDT	Heter1	Tr-Inf-99m	85.2837%
CIBk1	Heter2	Rd-Inf-0	99.9678%	CSMO15_CSVpuk	SP	Rd-Inf-0	84.9732%
CIBk4	Heter2	Rd-Inf-0	99.9678%	CMLPP90	SP	Rd-Inf-0	84.5314%
CRF159	Heter2	Rd-Inf-99m	99.9678%	CMLPP90	SP	Rd-Inf-99m	84.3149%
CRC20_CRF	Heter1	Rd-Inf-99m	99.8549%	CDT	Heter1	Tr-Inf-0	84.2521%
CRF10	SCQNP3	Original	99.7674%	CMLPP90	Heter2	Rd-Inf-99m	84.1690%
CRF10	SCQNP3	Tr-Inf-0	99.7674%	CRSS_CstLog	SCQNP2	Rd-Inf-99m	83.9535%
CRF10	SCQNP3	Tr-Inf-99m	99.7674%	CMLPP90	Heter4	Rd-Inf-99m	83.8491%
CSMO28_CSVRBFKe	SCQNP3	Rd-Inf-0	99.7674%	CSMO_CSVNoPKe	SCQNP2	Tr-Inf-0	83.7209%
CRF10	SCQNP1	Original	99.5360%	CSMO_CSVNoPKe	SCQNP2	Tr-Inf-99m	83.7209%
CRF10	Heter1	Tr-Inf-99m	99.4036%	CDT	Heter1	Original	83.5912%
CRF10	SP	Tr-Inf-99m	99.3937%	CSMO_CSVNoPKe	SCQNP2	Original	83.4884%
CRF10	Heter1	Tr-Inf-0	99.3875%	CSMO09_CSVpuk	Heter4	Rd-Inf-0	83.1883%
CRF10	Heter3	Tr-Inf-0	99.3874%	CMLPP90	Heter1	Rd-Inf-0	82.7369%
CRF10	SP	Original	99.3851%	CBagg_CLogis	Heter2	Rd-Inf-99m	82.3473%
CRF10	Heter4	Tr-Inf-0	99.3553%	CSMO87_CSVPKe	Heter1	Rd-Inf-0	82.0761%
CRF10	SP	Tr-Inf-0	99.3504%	CMLPP90	Heter1	Rd-Inf-99m	81.9310%
CRF10	Heter2	Original	99.3068%	CSGD2,043	SCQNP3	Rd-Inf-0	81.8605%
CRF10	Heter3	Tr-Inf-99m	99.3068%	CBagg_CLogis	Heter2	Rd-Inf-0	81.7830%
CRF10	SCQNP1	Tr-Inf-0	99.3039%	CLogis	Heter3	Rd-Inf-99m	81.7830%
CRF10	SCQNP1	Tr-Inf-99m	99.3039%	CSMO15_CSVRBFKe	SCQNP2	Rd-Inf-0	80.4651%
CBagg_CIBk	SCQNP2	Rd-Inf-0	99.3023%	CIBk35	SCQNP4	Rd-Inf-0	80.0464%
CRF10	Heter3	Original	99.2262%	CMLPP90	Heter4	Rd-Inf-0	79.7711%
CRF10	Heter1	Original	99.1941%	CDS	SCQNP4	Original	74.4780%
CRF10	Heter2	Tr-Inf-0	99.1939%	CDS	SCQNP4	Tr-Inf-0	74.4780%
CRF10	Heter4	Tr-Inf-99m	99.1779%	CDS	SCQNP4	Tr-Inf-99m	74.4780%
CRF10	Heter2	Tr-Inf-99m	99.1778%	CSMO_CSVNoPKe	SCQNP4	Rd-Inf-0	74.2459%
CRF10	Heter4	Original	99.1296%	CSMO47_CSVRBFKe	SCQNP1	Rd-Inf-0	74.0139%
CRF10	Heter2	Rd-Inf-99m	99.1295%	CSGD0,001	SCQNP1	Rd-Inf-0	73.7819%
CRF10	Heter1	Rd-Inf-0	99.1135%	CSGD0,042	SCQNP1	Rd-Inf-0	73.0858%
CRF10	Heter2	Rd-Inf-0	99.0972%	CSiLog0,055	SCQNP4	Rd-Inf-0	72.3898%
CRF10	Heter3	Rd-Inf-0	99.0327%	CSiLog0	SCQNP2	Rd-Inf-0	71.6279%
CRF10	SP	Rd-Inf-0	98.9780%	CSMO67_CSVPKe	SCQNP1	Rd-Inf-99m	70.7657%
CRF10	Heter4	Rd-Inf-99m	98.9201%	CMLPP12	SCQNP1	Rd-Inf-0	68.6775%
CRF10	SP	Rd-Inf-99m	98.9174%	CKStar94	SCQNP4	Rd-Inf-99m	63.1090%

Apéndice 4. Barras de comportamiento asociadas a los TeS de todos los modelos evaluados. Las acotaciones son las mismas que en la **Figura 9.**

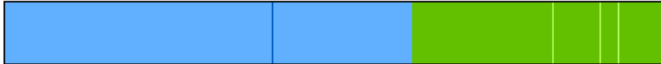








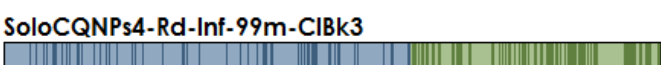
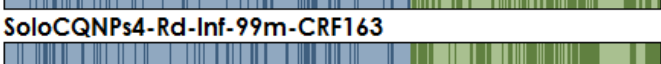
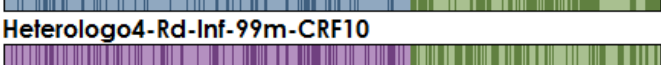
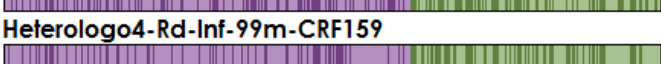
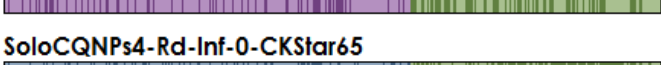
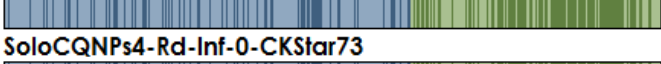
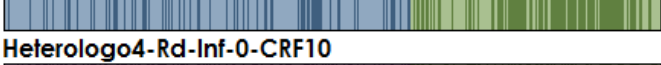
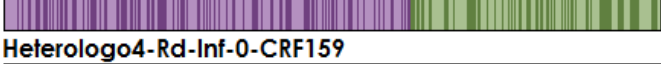


SoloCQNPs2-CRF10		%CC = 75.174% ROC = 0.826	109 52	55 215
SoloCQNPs2-Tr-Inf-99m-CRF10		%CC = 75.174% ROC = 0.826	109 52	55 215
SoloCQNPs2-Tr-Inf-0-CRF10		%CC = 75.406% ROC = 0.827	109 51	55 216
Heterologo2-CJ48		%CC = 71.2297% ROC = 0.75	109 69	55 198
Heterologo2-Tr-Inf-99m-CJ48		%CC = 71.2297% ROC = 0.75	109 69	55 198
Heterologo2-Tr-Inf-0-CJ48		%CC = 71.2297% ROC = 0.75	109 69	55 198
Heterologo2-CRF10		%CC = 72.8538% ROC = 0.789	114 67	50 200
Heterologo2-Tr-Inf-99m-CRF10		%CC = 72.8538% ROC = 0.789	114 67	50 200
Heterologo2-Tr-Inf-0-CRF10		%CC = 72.8538% ROC = 0.789	114 67	50 200
SoloCQNPs2-Rd-Inf-99m-CLWL_CRF		%CC = 78.1903% ROC = 0.835	92 22	72 245
SoloCQNPs2-Rd-Inf-99m-CVote		%CC = 76.1021% ROC = 0.818	99 38	65 229
SoloCQNPs2-Rd-Inf-99m-CAdaBM1_CRF12		%CC = 74.71% ROC = 0.791	109 54	55 213
SoloCQNPs2-Rd-Inf-99m-CJRip		%CC = 70.3016% ROC = 0.691	105 69	59 198
Heterologo2-Rd-Inf-99m-CRF10		%CC = 72.3898% ROC = 0.768	100 55	64 212
Heterologo2-Rd-Inf-99m-CRF159		%CC = 71.6937% ROC = 0.757	89 47	75 220
SoloCQNPs2-Rd-Inf-0-CIBk6		%CC = 78.1903% ROC = 0.83	91 21	73 246
SoloCQNPs2-Rd-Inf-0-CJRip		%CC = 70.3016% ROC = 0.691	105 69	59 198
SoloCQNPs2-Rd-Inf-0-CBagg_CIBk		%CC = 78.1903% ROC = 0.83	91 21	73 246
Heterologo2-Rd-Inf-0-CIBk1		%CC = 69.8376% ROC = 0.752	95 61	69 206
Heterologo2-Rd-Inf-0-CIBk4		%CC = 71.4617% ROC = 0.736	95 54	69 213
Heterologo2-Rd-Inf-0-CRF10		%CC = 72.1578% ROC = 0.768	109 65	55 202

Apéndice 4 (Cont).

SoloCQNPs3-CBagg_CLogis		%CC = 65.1972% ROC = 0.637	51 37	113 230
SoloCQNPs3-Tr-Inf-99m-CBagg_CLogis		%CC = 65.1972% ROC = 0.637	51 37	113 230
SoloCQNPs3-Tr-Inf-0-CBagg_CLogis		%CC = 65.1972% ROC = 0.637	51 37	113 230
SoloCQNPs3-CRF10		%CC = 62.413% ROC = 0.598	28 26	136 241
SoloCQNPs3-Tr-Inf-99m-CRF10		%CC = 62.413% ROC = 0.598	28 26	136 241
SoloCQNPs3-Tr-Inf-0-CRF10		%CC = 62.413% ROC = 0.598	28 26	136 241
Heterologo3-CRF10		%CC = 59.6288% ROC = 0.557	30 40	134 227
Heterologo3-Tr-Inf-99m-CRF10		%CC = 59.6288% ROC = 0.557	30 40	134 227
Heterologo3-Tr-Inf-0-CRF10		%CC = 59.6288% ROC = 0.557	30 40	134 227
SoloCQNPs3-Rd-Inf-99m-CRF10		%CC = 62.877% ROC = 0.593	28 24	136 243
SoloCQNPs3-Rd-Inf-99m-CAdaBM1_CRF134		%CC = 63.8051% ROC = 0.641	19 11	145 256
SoloCQNPs3-Rd-Inf-99m-CBagg_CLMT		%CC = 64.5012% ROC = 0.654	23 12	141 255
SoloCQNPs3-Rd-Inf-99m-CLWL_CIBk		%CC = 62.181% ROC = 0.62	16 15	148 252
SoloCQNPs3-Rd-Inf-99m-CLWL_CNB		%CC = 63.109% ROC = 0.593	24 19	140 248
Heterologo3-Rd-Inf-99m-CRC10_CRF		%CC = 62.413% ROC = 0.637	11 9	153 258
Heterologo3-Rd-Inf-99m-CRF10		%CC = 58.7007% ROC = 0.493	37 51	127 216
Heterologo3-Rd-Inf-99m-CRF159		%CC = 58.4687% ROC = 0.414	20 35	144 232

Apéndice 4 (Cont).

SoloCQNP3-Rd-Inf-0-CSMO28_CSVRBFKe		%CC = 62.413% ROC = 0.507	3 1	161 266
SoloCQNP3-Rd-Inf-0-CBagg_CRF		%CC = 63.109% ROC = 0.591	8 3	156 264
SoloCQNP3-Rd-Inf-0-CLWL_CIBk		%CC = 62.645% ROC = 0.636	16 13	148 254
Heterologo3-Rd-Inf-0-CIBk1E		%CC = 56.3805% ROC = 0.508	39 63	125 204
Heterologo3-Rd-Inf-0-CRF10		%CC = 55.6845% ROC = 0.499	25 52	139 215
Heterologo3-Rd-Inf-0-CRF159		%CC = 58.9327% ROC = 0.466	16 29	148 238
Heterologo4-CRF10		%CC = 63.4884% ROC = 0.667	104 97	60 169
Heterologo4-Tr-Inf-99m-CRF10		%CC = 63.4884% ROC = 0.667	104 97	60 169
Heterologo4-Tr-Inf-0-CRF10		%CC = 63.4884% ROC = 0.667	104 97	60 169
SoloCQNP4-Rd-Inf-99m-CIBk3		%CC = 68.8372% ROC = 0.748	86 56	78 210
SoloCQNP4-Rd-Inf-99m-CRF163		%CC = 65.3488% ROC = 0.639	79 64	85 202
Heterologo4-Rd-Inf-99m-CRF10		%CC = 58.3721% ROC = 0.608	93 108	71 158
Heterologo4-Rd-Inf-99m-CRF159		%CC = 61.6279% ROC = 0.628	73 74	91 192
SoloCQNP4-Rd-Inf-0-CKStar65		%CC = 61.6279% ROC = 0.688	68 69	96 197
SoloCQNP4-Rd-Inf-0-CKStar73		%CC = 61.6279% ROC = 0.693	68 69	96 197
Heterologo4-Rd-Inf-0-CRF10		%CC = 58.3721% ROC = 0.608	93 108	71 158
Heterologo4-Rd-Inf-0-CRF159		%CC = 61.6279% ROC = 0.628	73 74	91 192

Apéndice 4 (Cont).

Apéndice 5. Valores de los 5 parámetros estadísticos obtenidos correspondientes a las **Figuras 6, 7 y 10**. Se incluyen los datos de los modelos convencionales.

Modelo	%CC	ROC	EERAJ	%10CFV	%SPLIT
Heterologo1-Rd-Inf-99M-CRC20_CRF	0.772093	0.832	0.99955	0.873267	0.874817
Heterologo1-Rd-Inf-0-CRF159	0.75814	0.801	1	0.879735	0.877015
HeterologoConvencional-Nulos-ImputadosConPromedio-CAdaBM1_CJ48	0.755245	0.822	1	0.859684875	0.86150475
HeterologoConvencional-Nulos-Borrados--Infinity-ImputadosConPromedio-CRF159	0.74359	0.82	1	0.860752875	0.8651685
Heterologo2-CRF10	0.728538	0.789	0.99785	0.858919	0.856619
Heterologo1-CRF10	0.72093	0.776	0.9975	0.861501	0.861016
Heterologo2-Tr-Inf-0-CRF10	0.728538	0.789	0.99785	0.860088	0.855154
Heterologo1-Tr-Inf-99M-CRF10	0.72093	0.776	0.9975	0.86153	0.860283
Heterologo1-Tr-Inf-0-CRF10	0.72093	0.776	0.9975	0.861098	0.860283
Heterologo2-Tr-Inf-99M-CRF10	0.728538	0.789	0.99785	0.857912	0.855887
Heterologo4-Rd-Inf-99M-CRF159	0.695349	0.743	1	0.88188	0.872863
Heterologo1-Rd-Inf-0-CRF10	0.748837	0.775	0.99825	0.856242	0.851857
HeterologoConvencional-Nulos-Infinity-Borrados-CRF159	0.741259	0.806	1	0.86073275	0.86480225
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRF92	0.748252	0.805	0.9782	0.843422625	0.849536
HeterologoConvencional-Nulos-ImputadosConPromedio-CRF92	0.750583	0.8	0.97905	0.844954125	0.8491695
Heterologo2-Rd-Inf-0-CRF10	0.721578	0.768	0.99785	0.854446	0.851979
Heterologo2-Tr-Inf-0-CJ48	0.712297	0.75	0.9792	0.855367	0.861383
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRSS_CREPT	0.741259	0.799	0.97925	0.846586375	0.8491695
Heterologo2-CJ48	0.712297	0.75	0.9792	0.854849	0.86065
Heterologo2-Tr-Inf-99M-CJ48	0.712297	0.75	0.9792	0.854849	0.86065
Heterologo2-Rd-Inf-99M-CRF10	0.723898	0.768	0.9982	0.853055	0.850146
HeterologoConvencional-Nulos-Borrados--Infinity-ImputadosConPromedio-CRF10	0.731935	0.788	0.99715	0.83782025	0.84208625
HeterologoConvencional-Nulos-Borrados-CAdaBM1_CPART	0.722611	0.787	1	0.853840875	0.85210075
HeterologoConvencional-Nulos-ImputadosConPromedio-CRF10	0.724942	0.796	0.99715	0.834918625	0.8415975
HeterologoConvencional-Nulos-Infinity-Borrados-CRF10	0.722611	0.798	0.997	0.834596125	0.83573525
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRF10	0.717949	0.793	0.99705	0.837256125	0.8415975
HeterologoConvencional-Nulos-Borrados-CRF10	0.72028	0.786	0.99685	0.838969	0.8407425
Heterologo1-Rd-Inf-99M-CRF159	0.644186	0.679	1	0.878264	0.880313
HeterologoConvencional-Nulos-PorDefault-CRF10	0.715618	0.779	0.997	0.8387675	0.83915475
SoloCQNP3-Rd-Inf-99M-CBagg_CLMT	0.645012	0.654	0.99423	0.895058	0.86043
Heterologo4-CRF10	0.634884	0.667	0.9973	0.868271	0.860772
Heterologo4-Tr-Inf-0-CRF10	0.634884	0.667	0.9973	0.86813	0.858329
Heterologo4-Tr-Inf-99M-CRF10	0.634884	0.667	0.9973	0.86678	0.858452
SoloCQNP3-Rd-Inf-99M-CAdaBM1_CRF134	0.638051	0.641	1	0.873837	0.886503
SoloCQNP3-CBagg_CLogis	0.651972	0.637	0.986296	0.880523	0.857362
SoloCQNP3-Tr-Inf-99M-CBagg_CLogis	0.651972	0.637	0.986296	0.880523	0.857362
SoloCQNP3-Tr-Inf-0-CBagg_CLogis	0.651972	0.637	0.986296	0.880523	0.857362
HeterologoConvencional-Nulos-Borrados--Infinity-ImputadosConVecinosCercanos-CRF10	0.692308	0.778	0.9971	0.838445125	0.83939925
Heterologo1-Rd-Inf-99M-CRF10	0.634884	0.669	0.99775	0.853844	0.85798
Heterologo3-Rd-Inf-99M-CRC10_CRF	0.62413	0.637	1	0.891464	0.888007
SoloCQNP3-Rd-Inf-0-CLWL_CIBk	0.62645	0.636	1	0.894767	0.877301
Heterologo4-Rd-Inf-0-CRF159	0.616279	0.628	1	0.885195	0.876404
SoloCQNP3-Rd-Inf-99M-CLWL_CIBk	0.62181	0.62	1	0.895142	0.877301
Heterologo4-Rd-Inf-99M-CRF10	0.602326	0.645	0.9974	0.860211	0.84917
SoloCQNP3-CRF10	0.62413	0.598	0.999279	0.848256	0.849694
SoloCQNP3-Tr-Inf-0-CRF10	0.62413	0.598	0.999279	0.848256	0.849694
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CLMT	0.675991	0.718	0.96685	0.845236125	0.84208625
SoloCQNP3-Rd-Inf-99M-CLWL_CNB	0.63109	0.593	0.996394	0.843314	0.842025
Heterologo4-Rd-Inf-0-CRF10	0.583721	0.608	0.9974	0.86277	0.851002
SoloCQNP3-Tr-Inf-99M-CRF10	0.62413	0.598	0.999279	0.848256	0.827055

Modelo	%CC	ROC	EERAJ	%10CFV	%SPLIT
Heterologo3-Tr-Inf-99M-CRF10	0.596288	0.557	0.9976	0.874254	0.867489
Heterologo3-CRF10	0.596288	0.557	0.9976	0.872663	0.868466
Heterologo3-Tr-Inf-0-CRF10	0.596288	0.557	0.9976	0.873207	0.865657
SoloCQNP3-Rd-Inf-99M-CRF10	0.62877	0.593	1	0.835465	0.812884
SoloCQNP2-Rd-Inf-0-CBagg_CIBk	0.781903	0.83	0.997836	0.784012	0.769939
SoloCQNP2-Rd-Inf-0-CIBk6	0.781903	0.83	0.997836	0.780523	0.766871
SoloCQNP1-CRF10	0.781395	0.828	0.997841	0.773202	0.737805
Heterologo3-Rd-Inf-0-CIBk1E	0.563805	0.508	1	0.883323	0.878725
Heterologo3-Rd-Inf-99M-CRF10	0.587007	0.493	0.9976	0.867806	0.86407
Heterologo3-Rd-Inf-0-CRF159	0.589327	0.466	1	0.894225	0.890104
SoloCQNP2-Rd-Inf-99M-CLWL_CRF	0.781903	0.835	1	0.777035	0.748466
Heterologo3-Rd-Inf-0-CRF10	0.556845	0.499	0.99795	0.870184	0.866023
SoloCQNP1-Rd-Inf-99M-CKStar34	0.755814	0.799	1	0.759281	0.725612
SoloCQNP2-Rd-Inf-99M-CAdaBM1_CRF12	0.7471	0.79	1	0.763663	0.743865
SoloCQNP1-Rd-Inf-99M-CRF176	0.748837	0.799	1	0.771462	0.736281
HeterologoConvencional-Nulos-PorDefault-CRF239	0.808858	0.863	0	0.85075775	0.8564975
SoloCQNP3-Rd-Inf-0-CBagg_CRF	0.63109	0.591	0.982689	0.811337	0.756135
Heterologo3-Rd-Inf-99M-CRF159	0.584687	0.414	1	0.893882	0.891426
Heterologo2-Rd-Inf-99M-CRF159	0.716937	0.757	0	0.88165	0.883366
Heterologo2-Rd-Inf-0-CIBk4	0.714617	0.736	0	0.875746	0.872496
Heterologo2-Rd-Inf-0-CIBk1	0.698376	0.752	0	0.865267	0.862482
HeterologoConvencional-Nulos-ImputadosConPromedio-CBagg44_Clogis	0.634033	0.594	0.9707	0.847352125	0.8442845
SoloCQNP2-Rd-Inf-99M-CVote	0.761021	0.818	1	0.785465	0.699387
SoloCQNP1-Rd-Inf-99M-CRC15_CRT	0.732558	0.759	0.99928	0.75493	0.724086
SoloCQNP2-Tr-Inf-0-CRF10	0.75406	0.827	1	0.732558	0.733129
SoloCQNP2-CRF10	0.75174	0.826	1	0.732558	0.733129
SoloCQNP2-Tr-Inf-99M-CRF10	0.75174	0.826	1	0.732558	0.733129
HeterologoConvencional-Nulos-Borrados--Infinity-ImputadosConPromedio-CBagg44_Clogis	0.606061	0.597	0.96625	0.849629125	0.84709325
SoloCQNP1-Tr-Inf-0-CRF10	0.781395	0.828	0.997841	0.734339	0.719512
SoloCQNP1-Tr-Inf-99M-CRF10	0.781395	0.828	0.997841	0.735209	0.713415
SoloCQNP1-Rd-Inf-0-CAdaBM1_CSGD	0.718605	0.764	0.984888	0.734629	0.713415
SoloPeptidos-Rd-Inf-0-CRF10	0.416957	0.458	0.997797	0.879807	0.876371
SoloPeptidos-Rd-Inf-99M-CRF10	0.390244	0.489	0.997556	0.874578	0.873701
SoloCQNP4-Rd-Inf-0-CKStar73	0.616279	0.693	1	0.741879	0.714939
SoloCQNP4-Rd-Inf-0-CKStar65	0.616279	0.688	1	0.739269	0.707317
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CLogis0,000000100	0.564103	0.582	0.97445	0.838767625	0.829873
SoloCQNP4-Rd-Inf-99M-CIBk3	0.688372	0.748	1	0.714327	0.70097
SoloCQNP4-Rd-Inf-99M-CRF163	0.653488	0.639	1	0.741299	0.695122
SoloPeptidos-Rd-Inf-99M-CRSS_CIBk	0.401858	0.431	0.462758	0.895083	0.889073
SoloPeptidos-Rd-Inf-99M-CMLPP23	0.380952	0.374	0.973057	0.893805	0.891021
HeterologoConvencional-Nulos-Infinity-Borrados-CBagg81_Clogis	0.564103	0.538	0.96995	0.8466065	0.8406205
SoloPeptidos-Tr-Inf-99M-CRF10	0.382114	0.362	0.997958	0.884029	0.882073
SoloPeptidos-CRF10	0.382114	0.362	0.997958	0.883553	0.881856
SoloPeptidos-Tr-Inf-0-CRF10	0.382114	0.361	0.997958	0.883347	0.883805
SoloCQNP3-Rd-Inf-0-CSMO28_CSVBFKe	0.62413	0.507	0.999279	0.744186	0.707056
HeterologoConvencional-Nulos-Infinity-Borrados-CLogis0,000000656	0.559441	0.536	0.9708	0.840843125	0.83280425
SoloCQNP2-Rd-Inf-99M-CJRip	0.703016	0.691	0.994951	0.696802	0.687117
SoloCQNP2-Rd-Inf-0-CJRip	0.703016	0.691	0.994951	0.696802	0.687117
HeterologoConvencional-Nulos-Borrados--Infinity-ImputadosConVecinosCercanos-CSMO30_CSVBFKe	0.74359	0.68	0.98095	0.67437125	0.68136325

Apéndice 5 (Cont).

Apéndice 6. Valores de CScore obtenidos para elegir los mejores modelos. EERAJ = Ratio ajustado de error estimado. Los colores correlacionan con los datos de la **Figura 16** y del apéndice anterior.

Modelo	Cscore	Modelo	Cscore
Heterologo1-Rd-Inf-99M-CRC20_CRF	0.139325155	Heterologo3-Tr-Inf-99M-CRF10	0.763618875
Heterologo1-Rd-Inf-0-CRF159	0.140239399	Heterologo3-CRF10	0.764048461
HeterologoConvencional-Nulos-ImputadosConPromedio-CAdaBM1_CJ48	0.249711376	Heterologo3-Tr-Inf-0-CRF10	0.765790467
HeterologoConvencional-Nulos-Borrados-Infinity-ImputadosConPromedio-CRF159	0.292864208	SoloCQNP3-Rd-Inf-99M-CRF10	0.803019789
Heterologo2-CRF10	0.299191035	SoloCQNP2-Rd-Inf-0-CBagg_CIBk	0.817086346
Heterologo1-CRF10	0.299245774	SoloCQNP2-Rd-Inf-0-CIBk6	0.84008085
Heterologo2-Tr-Inf-0-CRF10	0.299819626	SoloCQNP1-CRF10	0.851109418
Heterologo1-Tr-Inf-99M-CRF10	0.300964013	Heterologo3-Rd-Inf-0-CIBk1E	0.882736841
Heterologo1-Tr-Inf-0-CRF10	0.302195804	Heterologo3-Rd-Inf-99M-CRF10	0.890952726
Heterologo2-Tr-Inf-99M-CRF10	0.304350358	Heterologo3-Rd-Inf-0-CRF159	0.914757994
Heterologo4-Rd-Inf-99M-CRF159	0.309156758	SoloCQNP2-Rd-Inf-99M-CLWL_CRF	0.918807758
Heterologo1-Rd-Inf-0-CRF10	0.314328442	Heterologo3-Rd-Inf-0-CRF10	0.921331117
HeterologoConvencional-Nulos-Infinity-Borrados-CRF159	0.322249634	SoloCQNP1-Rd-Inf-99M-CKStar34	0.977967284
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRF92	0.3268929	SoloCQNP2-Rd-Inf-99M-CAdaBM1_CRF12	0.988744993
HeterologoConvencional-Nulos-ImputadosConPromedio-CRF92	0.326968865	SoloCQNP1-Rd-Inf-99M-CRF176	0.988964352
Heterologo2-Rd-Inf-0-CRF10	0.349283913	HeterologoConvencional-Nulos-ParDefault-CRF239	1.002547431
Heterologo2-Tr-Inf-0-CJ48	0.352940035	SoloCQNP3-Rd-Inf-0-CBagg_CRF	1.011826133
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRSS_CREPT	0.354325123	Heterologo3-Rd-Inf-99M-CRF159	1.015308218
Heterologo2-CJ48	0.355933941	Heterologo2-Rd-Inf-99M-CRF159	1.029328079
Heterologo2-Tr-Inf-99M-CJ48	0.355933941	Heterologo2-Rd-Inf-0-CIBk4	1.043998547
Heterologo2-Rd-Inf-99M-CRF10	0.356036937	Heterologo2-Rd-Inf-0-CIBk1	1.056795997
HeterologoConvencional-Nulos-Borrados-Infinity-ImputadosConPromedio-CRF10	0.422658329	HeterologoConvencional-Nulos-ImputadosConPromedio-CBagg44_CLogis	1.089480126
HeterologoConvencional-Nulos-Borrados-CAdaBM1_CPART	0.424288598	SoloCQNP2-Rd-Inf-99M-CVote	1.092370956
HeterologoConvencional-Nulos-ImputadosConPromedio-CRF10	0.436857782	SoloCQNP1-Rd-Inf-99M-CRC15_CRT	1.100658876
HeterologoConvencional-Nulos-Infinity-Borrados-CRF10	0.452142368	SoloCQNP2-Tr-Inf-0-CRF10	1.130202133
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CRF10	0.459371201	SoloCQNP2-CRF10	1.130605868
HeterologoConvencional-Nulos-Borrados-CRF10	0.461413946	SoloCQNP2-Tr-Inf-99M-CRF10	1.130605868
Heterologo1-Rd-Inf-99M-CRF159	0.48630407	HeterologoConvencional-Nulos-Borrados-Infinity-ImputadosConPromedio-CBagg44_CLogis	1.156368258
HeterologoConvencional-Nulos-ParDefault-CRF10	0.489583125	SoloCQNP1-Tr-Inf-0-CRF10	1.168569077
SoloCQNP3-Rd-Inf-99M-CBagg_CLMT	0.534255528	SoloCQNP1-Tr-Inf-99M-CRF10	1.187253819
Heterologo4-CRF10	0.548584722	SoloCQNP1-Rd-Inf-0-CAdaBM1_CSGD	1.208991232
Heterologo4-Tr-Inf-0-CRF10	0.552149813	SoloPeplidos-Rd-Inf-0-CRF10	1.213455315
Heterologo4-Tr-Inf-99M-CRF10	0.553692165	SoloPeplidos-Rd-Inf-99M-CRF10	1.22690989
SoloCQNP3-Rd-Inf-99M-CAdaBM1_CRF134	0.555294694	SoloCQNP4-Rd-Inf-0-CKStar73	1.266365543
SoloCQNP3-CBagg_CLogis	0.559395124	SoloCQNP4-Rd-Inf-0-CKStar65	1.302484989
SoloCQNP3-Tr-Inf-99M-CBagg_CLogis	0.559395124	HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CLogis0.000000100	1.323952656
SoloCQNP3-Tr-Inf-0-CBagg_CLogis	0.559395124	SoloCQNP4-Rd-Inf-99M-CIBk3	1.337230082
HeterologoConvencional-Nulos-Borrados-Infinity-ImputadosConVecinosCercanos-CRF10	0.565612332	SoloCQNP4-Rd-Inf-99M-CRF163	1.341033792
Heterologo1-Rd-Inf-99M-CRF10	0.572064138	SoloPeplidos-Rd-Inf-99M-CRSS_CIBk	1.383346972
Heterologo3-Rd-Inf-99M-CRC10_CRF	0.574417135	SoloPeplidos-Rd-Inf-99M-CMLPP23	1.395233016
SoloCQNP3-Rd-Inf-0-CLWL_CIBk	0.575639771	HeterologoConvencional-Nulos-Infinity-Borrados-CBagg81_CLogis	1.405443627
Heterologo4-Rd-Inf-0-CRF159	0.607675235	SoloPeplidos-Tr-Inf-99M-CRF10	1.412528388
SoloCQNP3-Rd-Inf-99M-CLWL_CIBk	0.608233099	SoloPeplidos-CRF10	1.412660438
Heterologo4-Rd-Inf-99M-CRF10	0.659598559	SoloPeplidos-Tr-Inf-0-CRF10	1.413910978
SoloCQNP3-CRF10	0.708832432	SoloCQNP3-Rd-Inf-0-CSMO28_CSVRBFKe	1.423829116
SoloCQNP3-Tr-Inf-0-CRF10	0.708832432	HeterologoConvencional-Nulos-Infinity-Borrados-CLogis0.000000656	1.429429085
HeterologoConvencional-Nulos-ImputadosConVecinosCercanos-CLMT	0.710074836	SoloCQNP2-Rd-Inf-99M-CJ Rip	1.459804387
SoloCQNP3-Rd-Inf-99M-CLWL_CNB	0.727255224	SoloCQNP2-Rd-Inf-0-CJ Rip	1.459804387
Heterologo4-Rd-Inf-0-CRF10	0.734475227	HeterologoConvencional-Nulos-Borrados-Infinity-ImputadosConVecinosCercanos-CSMO30_CSVRBFKe	1.543382527
SoloCQNP3-Tr-Inf-99M-CRF10	0.748323942		

Apéndice 7. Lista de abreviaturas utilizadas para denotar las 11 diferentes combinaciones algoritmo-parámetros nuevas de los modelos convencionales.

Combinación	Significado
CAdaBM1_CJ48	Clasificador AdBoostM1, que utiliza al clasificador J48.
CAdaBM1_CPART	Clasificador AdBoostM1, que utiliza al clasificador PART.
CBagg44_CLogis	Clasificador Bagging usando el 44% del TrS, que utiliza al clasificador Logistic.
CBagg81_CLogis	Clasificador Bagging usando el 81% del TrS, que utiliza al clasificador Logistic.
CLMT	Clasificador LMT.
CLogis0.000000100	Clasificador Logistic con un valor de arista de 1×10^{-7} .
CLogis0.000000656	Clasificador Logistic con un valor de arista de 6.56×10^{-7} .
CRF239	Clasificador Random Forest usando 239 iteraciones.
CRF92	Clasificador Random Forest usando 92 iteraciones.
CRSS_CREPT	Clasificador Random Subspace, que utiliza al clasificador REPTree.
CSMO30_CSVRBFKe	Clasificador Supporting Machine Operator, con una complejidad de 1.30... y usando el núcleo Support Vector RBF-Kernel.

Apéndice 8. Lista de todos los CQNP del Discovery Set (DiS) que fueron predichos por el mejor modelo, como portadores de actividad antimicrobiana. Al lado de cada compuesto está el código ATCC correspondiente a su actividad primaria.

CQNP	ATCC	Axitinib	L01X (Antineoplásicos varios)
Alosetron	A03A (Agentes para fallas de funcionalidad gastrointestinal)	Bosutinib	L01X (Antineoplásicos varios)
Nabilone	A04A (Antieméticos-náusea)	Cabozantinib	L01X (Antineoplásicos varios)
Ondansetron	A04A (Antieméticos-náusea)	Carfilzomib	L01X (Antineoplásicos varios)
Tetrahydrocannabinol	A04A (Antieméticos-náusea)	Crizotinib	L01X (Antineoplásicos varios)
Alvimopan	A06A (Antiestreñimiento)	Dabrafenib	L01X (Antineoplásicos varios)
Lactulose	A06A (Antiestreñimiento)	Erimodegib	L01X (Antineoplásicos varios)
Orlistat	A08 (Terapia de obesidad)	Everolimus	L01X (Antineoplásicos varios)
Alogliptin	A10 (Terapia de diabetes)	Ibrutinib	L01X (Antineoplásicos varios)
Canagliflozin	A10 (Terapia de diabetes)	Lenvatinib	L01X (Antineoplásicos varios)
Dapagliflozin	A10 (Terapia de diabetes)	Leptomycin B	L01X (Antineoplásicos varios)
Empagliflozin	A10 (Terapia de diabetes)	Lysodren	L01X (Antineoplásicos varios)
Linagliptin	A10 (Terapia de diabetes)	Nintedanib	L01X (Antineoplásicos varios)
Sitagliptin	A10 (Terapia de diabetes)	Olaparib	L01X (Antineoplásicos varios)
a-Tocopherol	A11 (Vitaminas)	Palbociclib	L01X (Antineoplásicos varios)
Riboflavin	A11 (Vitaminas)	Ponatinib	L01X (Antineoplásicos varios)
Nitisinone	A16A (Otros agentes alimenticios y metabólicos)	Radicalol	L01X (Antineoplásicos varios)
Tetrahydrobiopterin	A16A (Otros agentes alimenticios y metabólicos)	Regorafenib	L01X (Antineoplásicos varios)
Clopidogrel	B01A (Antitrombóticos)	Romidepsin	L01X (Antineoplásicos varios)
Prasugrel	B01A (Antitrombóticos)	Sorafenib	L01X (Antineoplásicos varios)
Vorapaxar	B01A (Antitrombóticos)	Sunitinib	L01X (Antineoplásicos varios)
Warfarin	B01A (Antitrombóticos)	Trametinib	L01X (Antineoplásicos varios)
Adenosine	C01E (Otros agentes de terapia cardiaca)	Vandetanib	L01X (Antineoplásicos varios)
Regadenoson	C01E (Otros agentes de terapia cardiaca)	Enzalutamide	L02 (Terapia hormonal)
Allitridin	C02 (Antihipertensivos)	Fulvestrant	L02 (Terapia hormonal)
Doxazosin	C02 (Antihipertensivos)	Plerixafor	L03 (Inmunoestimulantes)
Riociguat	C02 (Antihipertensivos)	Bafilomycin	L04 (Inmunosupresores)
Clevidipine	C08 (Bloqueadores de canales de calcio)	Rapamycin	L04 (Inmunosupresores)
Nicardipine	C08 (Bloqueadores de canales de calcio)	Flurbiprofen	M01 (Antiinflamatorios y antirreumáticos)
Nimodipine	C08 (Bloqueadores de canales de calcio)	Sulindac	M01 (Antiinflamatorios y antirreumáticos)
Nisoldipine	C08 (Bloqueadores de canales de calcio)	Utric acid	M01 (Antiinflamatorios y antirreumáticos)
Aliskiren	C09 (Agentes del sistema Renina-Angiotensina)	Atracurium	M03 (Relajantes musculares)
Azilsartan medoxomil	C09 (Agentes del sistema Renina-Angiotensina)	Estazolam	M05B (Agentes para problemas de hueso)
Eprosartan	C09 (Agentes del sistema Renina-Angiotensina)	Quazeparn	M05B (Agentes para problemas de hueso)
Valsartan	C09 (Agentes del sistema Renina-Angiotensina)	Secobarbital	M05B (Agentes para problemas de hueso)
Ezetimibe	C10 (Agentes para niveles de lípidos)	Rasagiline	N04B (Antiparkinsonianos dopaminérgicos)
Lomitapide	C10 (Agentes para niveles de lípidos)	Brexiprazole	N05A (Antipsicóticos)
Pitavastatin	C10 (Agentes para niveles de lípidos)	Cariprazine	N05A (Antipsicóticos)
Rosuvastatin	C10 (Agentes para niveles de lípidos)	Thioridazine	N05A (Antipsicóticos)
Zaragozic acid A	C10 (Agentes para niveles de lípidos)	Alprazolam	N05B (Ansiolíticos)
Pimecrolimus	D11A (Tratamientos dermatológicos varios)	Chlordiazepoxide	N05B (Ansiolíticos)
Tacrolimus	D11A (Tratamientos dermatológicos varios)	Hydroxyzine	N05B (Ansiolíticos)
Bromocriptine	G02C (Ginecológicos varios)	Fluoxetine	N06A (Antidepresivos)
Cabergoline	G02C (Ginecológicos varios)	Dimethyl fumarate	N07X (Agentes varios para sistema nervioso)
Flibanserin	G02C (Ginecológicos varios)	Tetrabenazine	N07X (Agentes varios para sistema nervioso)
Fesoterodine	G04BD (Terapia de frecuencia y retención urinaria)	Indacaterol	R03 (Agentes para obstrucciones respiratorias)
Oxybutynin	G04BD (Terapia de frecuencia y retención urinaria)	Nedocromil	R03 (Agentes para obstrucciones respiratorias)
Dutasteride	G04C (Terapia de hiperplasia prostática benigna)	Olodaterol	R03 (Agentes para obstrucciones respiratorias)
Sildenafil	G04C (Terapia de hiperplasia prostática benigna)	Roflumilast	R03 (Agentes para obstrucciones respiratorias)
Doxercalciferol	H05 (Terapia paratiroidea)	Fexofenadine	R06A (Antihistamínicos sistémicos)
Berberine	J02 (Antimicrobianos antifúngicos sistémicos en general)	Meclizine	R06A (Antihistamínicos sistémicos)
Sinefungin	J02 (Antimicrobianos antifúngicos sistémicos en general)	Ivacaftor	R07A (Agentes respiratorios varios)
Carmustine	L01A (Agentes Alquilantes)	Tafuprost	S01E (Antiglaucoma)
2-amino-mercaptopurine	L01B (Antimetabolitos)	Travoprost	S01E (Antiglaucoma)
Capecitabine	L01B (Antimetabolitos)	Glutathione	V03AB (Anfidotos)
Decitabine	L01B (Antimetabolitos)	Folinic acid	V03AF (Agentes protectores contra la quimioterapia)
Cabazitaxel	L01C (Alcaloides de plantas)	2-Di-I-ASP	V08A (Agentes de contraste)
Teniposide	L01C (Alcaloides de plantas)	Nile Blue	V08A (Agentes de contraste)
Afatinib	L01X (Antineoplásicos varios)	Octadecyl Rhodamine B-ate	V08A (Agentes de contraste)
		Ioflupane	V09AB (Agentes radioactivos para diagnóstico)

Apéndice 9. Lista de todos los antibióticos que podrían clasificarse como de amplio espectro. Se muestra un solo registro por compuesto, los compuestos amarillos tienen 2 registros (correspondientes a 2 estereoisómeros), y el anaranjado tiene 4 registros.

Rifaximin	<u>Itraconazole</u>	<u>Quinacrine</u>
<u>Econazole</u>	Posaconazole	<u>Chloroquine</u>
Bifonazole	Rifampicin	<u>Hydroxychloroquine</u>
Efinaconazole	Rifabutin	<u>Mefloquine</u>
Inosine	Rifapentine	Artemether
<u>Butoconazole</u>	Bedaquiline	Besifloxacin
Tigecycline	Indinavir	<u>Praziquantel</u>
Amoxicillin	Nelfinavir	Dichlorophen
Phenoxymethylpenicillin	Atazanavir	Lindane
Ceftibuten	Tipranavir	Biopermethrin
Ceftriaxone	Baraclude	Permethrin
Ertapenem	Telbivudine	Cispermethrin
Doripenem	Zanamivir	Transpermethrin
Cephalexin	Oseltamivir	Antimycin A
<u>Cefprozil</u>	Raltegravir	Monensin
Azithromycin	Elvitegravir	Nonylacridine orange
Telithromycin	Dolutegravir	Azure B
Amikacin	Daclatasvir	Azure C
<u>Gemifloxacin</u>	Sofosbuvir	Rhodamine 6G
<u>Gatifloxacin</u>	Ixabepilone	

14. ANEXOS

Anexo 1 (Página siguiente): Artículo publicado en la revista *Molecules*.

Article

Heterologous Machine Learning for the Identification of Antimicrobial Activity in Human-Targeted Drugs

Rodrigo A. Nava Lara ¹, Longendri Aguilera-Mendoza ², Carlos A. Brizuela ², Antonio Peña ³ and Gabriel Del Rio ^{1,*}

¹ Department of biochemistry and structural biology, Instituto de Fisiología Celular, UNAM, Mexico City 04510, Mexico; rnava@email.ifc.unam.mx

² Computer Science Department, CICESE Research Center, Ensenada, Baja California 22860, Mexico; longendri@gmail.com (L.A.-M.); cbrizuel@cicese.mx (C.A.B.)

³ Department of genetics, Instituto de Fisiología Celular, UNAM, Mexico City 04510, Mexico; apd@ifc.unam.mx

* Correspondence: gdelrio@ifc.unam.mx; Tel.: +52-55-5622-5663

Academic Editor: Julio Caballero

Received: 1 February 2019; Accepted: 14 March 2019; Published: 31 March 2019



Abstract: The emergence of microbes resistant to common antibiotics represent a current treat to human health. It has been recently recognized that non-antibiotic labeled drugs may promote antibiotic-resistance mechanisms in the human microbiome by presenting a secondary antibiotic activity; hence, the development of computer-assisted procedures to identify antibiotic activity in human-targeted compounds may assist in preventing the emergence of resistant microbes. In this regard, it is worth noting that while most antibiotics used to treat human infectious diseases are non-peptidic compounds, most known antimicrobials nowadays are peptides, therefore all computer-based models aimed to predict antimicrobials either use small datasets of non-peptidic compounds rendering predictions with poor reliability or they predict antimicrobial peptides that are not currently used in humans. Here we report a machine-learning-based approach trained to identify gut antimicrobial compounds; a unique aspect of our model is the use of heterologous training sets, in which peptide and non-peptide antimicrobial compounds were used to increase the size of the training data set. Our results show that combining peptide and non-peptide antimicrobial compounds rendered the best classification of gut antimicrobial compounds. Furthermore, this classification model was tested on the latest human-approved drugs expecting to identify antibiotics with broad-spectrum activity and our results show that the model rendered predictions consistent with current knowledge about broad-spectrum antibiotics. Therefore, heterologous machine learning rendered an efficient computational approach to classify antimicrobial compounds.

Keywords: machine-learning; antimicrobial peptide; non-peptidic antimicrobial compound; antimicrobial activity

1. Introduction

Drug-resistant microbes are one of the most important challenges for modern medicine [1] considering the increased rate in morbidity and mortality associated with antibiotic-resistant pathogens [2]. It is now commonly accepted that misuse of antibiotics is a major factor that promotes microbial resistance to these agents [3]; such is the case of broad-spectrum antibiotics that tend to promote resistance and are now prescribed in very restricted situations [4]. Furthermore, it has been noted that many non-antibiotic human-targeted drugs alter the gut microbiome in patients taking such drugs [5,6]. This alteration has been shown to be the consequence of a non-reported collateral antimicrobial activity, suggesting that microbe resistance to an antibiotic may emerge as

a consequence of using those human-targeted drugs [7]. Furthermore, some antibiotics may have not been tested against the gut microbiome and may as well promote the emergence of resistant microbes. Since the experimental validation of antimicrobial activity for the gut microbiome requires tests on hundreds/thousands of cultivable and non-cultivable microorganisms and the number of new human-targeted drugs may include dozens of compounds, it is relevant to develop efficient computational strategies for the identification of secondary antimicrobial activity of human-targeted drugs. In the present work we present a computational strategy aimed to improve the identification of compounds with antimicrobial activity using machine-learning-based approaches.

Previous computational approaches to identify antibiotics using Quantitative Structure-Activity Relationships (QSAR) [8,9] and machine-learning-based [10,11] procedures have been reported. In these computational approaches, non-peptidic chemical compounds (from now on referred to as NPCC) are represented by chemical descriptors (e.g., *LogP*, molecular weight, polarizability) and each compound is labeled as antibiotic or non-antibiotic; then a clustering algorithm separates antibiotics from non-antibiotics. An important limitation of these previous studies is that the number of chemical compounds used to train the models is limited (less than one thousand NPCC have been described with antimicrobial activity) and the reliability of these models requires further improvement. Alternatively, antimicrobial peptides now accumulate in more than 10,000 in different databases [12–14], and several computational models have been reported to effectively classify antimicrobial from non-antimicrobial peptides [15–17]. Although peptides represent an important new focus to develop pharmaceuticals, most human-targeted drugs are NPCC; therefore computational models to identify antimicrobial activity in these compounds should focus on NPCC. The need to use common molecular descriptors between polypeptides and NPCC has been previously noted for protein-ligand recognition and protein folding, as a fundamental aspect to deal with induced-fit or conformer selection mechanisms for molecular recognition [18]; the aim of this work though, is not to find common descriptors to peptides and NPCC since there are already packages that solve this problem (see below). Here we propose that combining peptides and NPCC increases the training set size and this should improve the reliability of the computational models. The present work tests this proposal and validates the idea that heterologous (NPCC and peptides) training sets render the best classifying models. We then show how this improved model may assist in the identification of broad-spectrum antibiotics on FDA-approved NPCC.

2. Results

2.1. Training and Testing Gut Antimicrobial Classifiers

Building data sets to combine peptides and NPCC required the use of molecular descriptors common to both types of compounds; in our case, we used 1444 descriptors calculated by PadelDescriptor (see Methods). Then, to identify the best machine-learning model to classify gut antimicrobials, three groups of training sets were used (see Table 1). The first group included only peptides (TrOnlyPeptides), the second group comprises 4 sets and included only NPCC (TrNPCC1-4) and the third group combined these two previous sets (TrHeterologous1-4) resulting in a total of 9 training sets (see Table 1); this rendered a total of 45 training sets. These 45 sets were further processed to substitute any null or "Infinity" values using three different approaches, and a reduction of dimensions was performed via principal-component analysis (PCA, see Methods). This procedure rendered a total of 50 Training Sets; all these sets are included in Supplemental Tables S1(A–E)–S9(A–E).

Nine testing sets were built using the NPCC recently reported by Maier et al. [7] with and without gut antimicrobial activity (see Table 2). The same processing of these testing sets was performed as in the case of the training sets (see above), rendering again a total of 50 data sets (see Supplemental Tables S10(A–E)–S18(A–E)). Please note that in both training and testing sets all peptides included were tested against only one gut microbe assayed against the NPCC used in these sets and that although there are

many more peptides than NPCC in our training and testing sets, this imbalance is not relevant to find the border between antimicrobials and non-antimicrobials compounds.

Table 1. Training data sets.

Training Set	Entries	Description
TrOnlyPeptides	11,546	8000 antimicrobial peptides, 3546 peptides with no known antimicrobial activity
TrNPCC1	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity
TrNPCC2	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity
TrNPCC3	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity
TrNPCC4	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity
TrHeterologous1	6204	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2040 no antimicrobial compounds (1773 peptides and 267 non-peptidic compounds)
TrHeterologous2	6203	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2039 no antimicrobial compounds (1773 peptides and 266 non-peptidic compounds)
TrHeterologous3	6203	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2039 no antimicrobial compounds (1773 peptides and 266 non-peptidic compounds)
TrHeterologous4	6204	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2040 no antimicrobial compounds (1773 peptides and 267 non-peptidic compounds)

The original NPCC from Maier et al. [7], here referred to as OnlyNonPeptides, was used to build TrNPCC1 by taking only the odd listed compounds, TrNPCC2 by taking even listed compounds, TrNPCC3 and TrNPCC4, included the first and second half of the data set respectively. The OnlyPeptides data set was divided to generate TrHeterologous1, TrHeterologous2, TrHeterologous3 and TrHeterologous4 by taking the odds listed peptides, even listed peptides, first and second half, respectively. Then, these TrHeterologous1-4 data sets with peptides were combined with the TrNPCC1-4 to complete these sets.

Table 2. Testing data sets.

Testing Set	Entries	Description
TeOnlyPeptides	861	328 antimicrobial and 533 non-antimicrobial non-peptides
TeNPCC1	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity. Same as TrNPCC2.
TeNPCC2	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity. Same as TrNPCC1.
TeNPCC3	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity. Same as TrNPCC4.
TeNPCC4	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity. Same as TrNPCC3.
TeHeterologous1	430	Same as TeNPCC1.
TeHeterologous2	431	Same as TeNPCC2.
TeHeterologous3	431	Same as TeNPCC3.
TeHeterologous4	430	Same as TeNPCC4.

The original NPCC from Maier et al. [7], here referred to as OnlyNonPeptides, was used to build all Testing Sets. TeOnlyPeptides was built taking all the 861 listed compounds. TeNPCC1 and TeHeterologous1 were built by taking only the even listed compounds. TeNPCC2 and TeHeterologous2 included only the odd listed compounds. TeNPCC3 and TeHeterologous3 included the second half of OnlyNonPeptides, TeNPCC4 and TeHeterologous4 included the first half of the data set. Testing sets were built so they were the complement of the compounds listed for their Training sets, so, for example, if a training set was built using the even listed compounds (e.g., TrNPCC1), its Testing set would be built with the odd listed compounds (e.g., TeNPCC1). Heterologous Testing Sets were the same as OnlyNonPeptides Testing sets, due to the fact that the interest compounds are of non-peptidic nature.

Five different statistical parameters (adjusted estimated error rate on the training set (AEER); correctly classified instances in the training set after splitting 33% for testing (%Split); 10-fold cross-validation (%10FCV); correctly classified instances on the testing set (%CC); area under the receiver operator characteristic curve on the testing set (AUROC)) that evaluated the performance on either the training or testing sets (see Methods) were used to identify the best classifier.

As shown in Figure 1, the best models included heterologous compounds (peptides and NPCC): circles in Figure 1 represent heterologous training sets and accumulate on the upper part of Figure 1, that is, those models with highest statistical parameters evaluating the model performance (the actual data in this figure for these models are included in Supplemental Table S19). Treating the training set rendering the best model with the K-nearest neighbor or mean-imputation approaches did not improve the performance of the best model (see Supplemental Tables S6G, S6I, S6J and its corresponding test set in Table S15G; supplemental Tables S6K, S6L and their corresponding test sets in Supplemental Tables S15I and S15J).

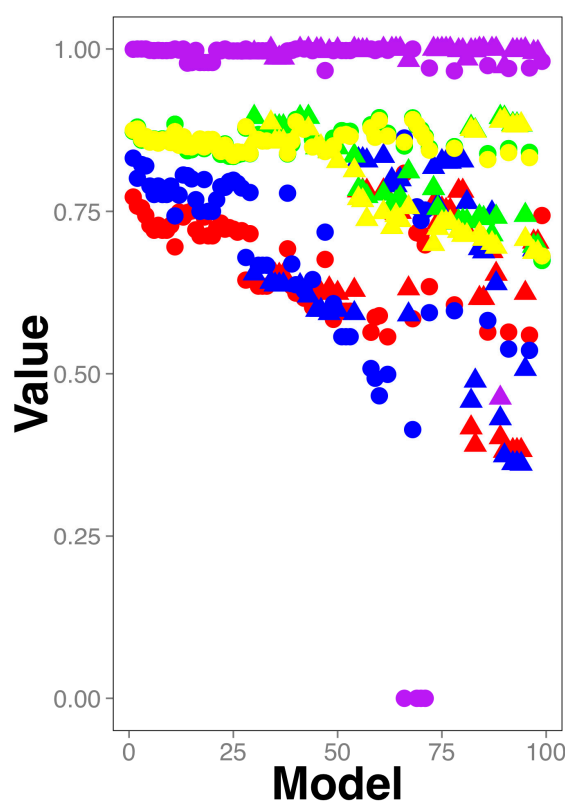


Figure 1. Classifiers performance. Five statistical parameters (yellow circle: Correctly classified instances in the test set after splitting 33% for testing (%Split); green circle: 10-fold cross-validation (%10FCV); red circle: Correctly classified instances on the training set; blue circle: AUROC on the testing set; purple circle: AUROC on the training set) were used to evaluate the performance of 100 models. The models using heterologous data are represented by circles, triangles are used otherwise. The actual data of this plot can be found in Supplementary Table S19.

Yet, none of these models surpassed the others in all 5 parameters. To aid in the visualization of this aspect of our results, Figure 1 displays the values in descending order from left to right; therefore, the models on the left side of the plot have better scores than those on the right. For instance, models using heterologous (represented by circles) testing sets (the red and blue circles, corresponding with the parameters correctly classified instances and AUROC, respectively) have better performance than those models using heterologous training sets on the left side of Figure 1, have better performance than those models using heterologous testing sets on the right side of the plot, yet those on the right side including either heterologous or non-heterologous training sets (green and yellow circles or triangles) have better scores than those models using heterologous or non-heterologous training sets on the left side of the plot. Please note that the statistical parameter adjusted estimated error rate is the value that AutoWeka optimizes, hence for all the reported models is close to 1.0 and consequently does not contribute to differentiate the performance of the models. This statistical parameter is shown in Figure 1 to note

side of the plot, yet, those on the right side including either heterologous or non-heterologous training sets (green and yellow circles or triangles) have better scores than those models using heterologous or non-heterologous training sets on the left side of the plot. The models in the middle of the plot have on the other hand, intermediate performances. Please note that the statistical parameter adjusted estimated error rate is the value that AutoWeka optimizes, hence for all the reported models is close to 1.0 and consequently does not contribute to differentiate the performance of the models. This statistical parameter is shown in Figure 1 to note that all models have similar error rates, yet different statistical parameters, hence, the best model obtained from AutoWeka cannot be selected simply by considering the error rate value reported.

Thus, to aid in the identification of the best models, we used a previous score developed by our group that takes into account multiple statistical parameters, the Combined Score or simply *CScore* [19]:

Molecules 2019, 24, x 6 of 14

$$CScore = \frac{1}{5} \sum_{n=1}^5 \sqrt{\frac{MaxS_n - S_{i,n}}{MaxS_n - MinS_n}} \quad (1)$$

Thus, to aid in the identification of the best models, we used a previous score developed by our group that takes into account multiple statistical parameters, the Combined Score or simply *CScore*, where $MaxS_n$ and $MinS_n$ represent the maximum and minimum scores for a given statistical parameter n over all models; $S_{i,n}$ is the score observed for a given statistical parameter n and model i ; n represents the index of the statistical parameter to evaluate (in our case were 5 parameters: AEER, %Split, %10FCV, %CC and AUROC). Thus, formula 1 calculates *CScore* for each model.

CScore averages the difference of each statistical parameter to its best value (e.g. true-positive rate best value is 1, so the difference between the observed true positive rate and 1 is included in the *CScore*), therefore the lower the *CScore* value the better the classifying model. Figure 2 (and Supplementary Table S20) shows that the five best models are those using heterologous training sets (the ones below the 0.3 line in Figure 2). Furthermore, we noticed that the top 5 best models (the ones below the 0.3 line in Figure 2). Furthermore, we noticed that the top 5 best models more than 70% of their classifications hence, these were mainly redundant (see Figure 3).

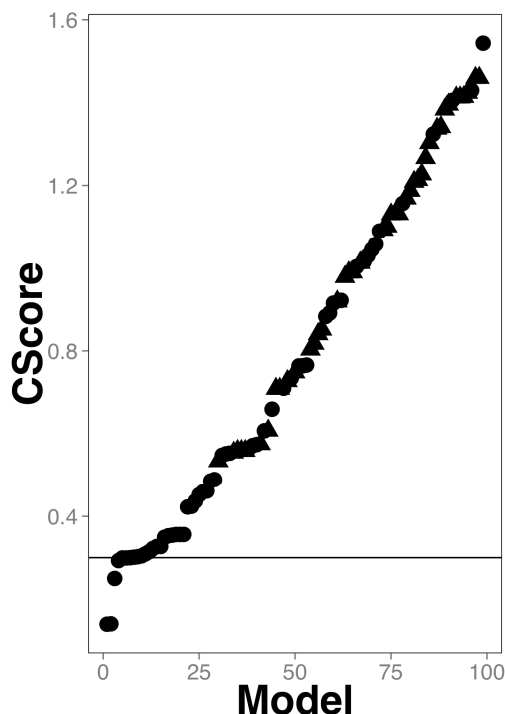


Figure 2. Classifiers combined scores. A circle represents each model; the best model has the lowest *CScore*. The line represents the *CScore* = 0.3, that separates the top 5 models from the rest. The models using heterologous data are represented as circles; triangles are used otherwise. The actual data of this plot can be found in Supplementary Table S20.

Figure 2. Classifiers combined scores. A circle represents each model; the best model has the lowest *C*Score. The line represents the *C*Score = 0.3, that separates the top 5 models from the rest. The models using heterologous data are represented as circles; triangles are used otherwise. The actual data of this plot can be found in Supplementary Table S20.

6 of 13

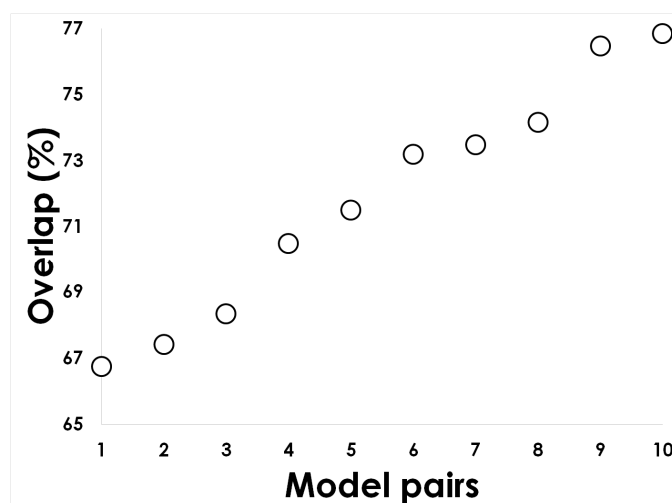


Figure 3. Classifiers overlap. The predictions of antimicrobial compounds of the top 5 models were compared to quantify their overlap. The image shows the 10 pairs of models generated from these 5 top models. The comparison was performed on the discovery set (see Methods) because not every model had the same testing set. Therefore, we selected the best model based on the lowest *C*Score; such model was built using the RandomCommittee algorithm (see Supplemental Table S21 for the algorithm parameters) on the TrHeterologous1 set (see Table 1) that included 86 molecular PCA-reduced descriptors and achieved the following performance: AEEK: 0.99955; %Split: 87.4; %10FCV: 87.3; %CC: 77.2; AUROC: 0.83 (model named TrHeterologous1-Reduced-With-99M-GRC20-CRF in Supplemental Table S21; the corresponding data set for this model is reported in Supplemental Table S6E).

2.2. Identifying Broad-Spectrum Antibiotics among FDA-Approved Compounds
We used the best model to predict NPCC with expected gut antimicrobial activity among FDA-approved drugs. The motivation to perform this prediction is not for testing purposes, as in the case of the training and testing sets used before. Hence, the set of compounds used in this prediction stage is referred to as the discovery set, because we aimed to discover potential compounds with gut antimicrobial activity. We used 756 FDA-approved compounds included in the ZINC database (see Methods) that were not part of the training or testing sets; these compounds included 111 antimicrobials and 645 compounds without any known antimicrobial activity; we also added 73 NPCC that included 22 antifungal compounds and 51 without any reported antifungal activity (see Supplementary Table S22). We have previously reported that these 22 antifungals work through a mechanism (alter calcium intake [20]) different from antibacterial compounds (e.g., penicillin derivatives, sulphonamides, etc), thus we expected our model to predict few of these compounds as antibacterials. FDA-approved compounds on the other hand are expected not to have, or to have minor, gut antimicrobial activity otherwise their secondary gastrointestinal effects

2.2. Identifying Broad-Spectrum Antibiotics among FDA-Approved Compounds

We used the best model to predict NPCC with expected gut antimicrobial activity among FDA-approved drugs. The motivation to perform this prediction is not for testing purposes, as in the case of the training and testing sets used before. Hence, the set of compounds used in this prediction stage is referred to as the discovery set, because we aimed to discover potential compounds with gut antimicrobial activity. We used 756 FDA-approved compounds included in the ZINC database (see Methods) that were not part of the training or testing sets; these compounds included 111 antimicrobials and 645 compounds without any known antimicrobial activity; we also added 73 NPCC that included 22 antifungal compounds and 51 without any reported antifungal activity (see Supplementary Table S22). We have previously reported that these 22 antifungals work through a mechanism (alter calcium intake [20]) different from antibacterial compounds (e.g., penicillin derivatives, sulphonamides, etc), thus we expected our model to predict few of these compounds as antibacterials. FDA-approved compounds on the other hand are expected not to have, or to have minor, gut antimicrobial activity otherwise their secondary gastrointestinal effects would be significant. We would expect that FDA-approved drugs would be less likely predicted to act against non-athogenic gut microbes than antifungals. To evaluate the reliability of our predictions using the discovery set, we considered that antibiotic compounds against the non-pathogenic gut flora among the FDA-approved drugs should be considered broad-spectrum antibiotics; please note that our classifier was not trained to predict this class of antibiotics, yet the combination of the predictions of our classifier on the FDA-approved drugs would render this information. The definition of broad-spectrum antibiotics is somehow arbitrary, for instance, it is considered that antibiotics that act on G(+) and G(−) are broad-spectrum antibiotics for some authors, while those acting against pathogenic and non-pathogenic microorganisms are classified as broad-spectrum antibiotics by others [21,22]. The list of broad-spectrum antibiotics was obtained from five recent works (see Methods), including 19

broad-spectrum and 3 narrow-spectrum antibiotics (see Supplementary Table S22). We were able to identify 72 true positives (FDA-approved antibiotics against pathogenic microbes predicted to act against non-pathogenic gut microbes) in the discovery set that we predicted should be considered as broad-spectrum antibiotics (see Table 3).

Table 3. Confusion matrix for the discovery set.

	Predicted Gut Antimicrobial	Predicted No Antimicrobial
Pathogenic antimicrobial	72	61
No antimicrobial	140	556

The actual data for this table can be found in Supplementary Table S22.

From these 72 antimicrobials, only 16 had been annotated as broad-spectrum antibiotics and 3 as narrow-spectrum antibiotics (see Supplemental Table S22). Hence, we propose that these 3 annotated narrow-spectrum antibiotics should be considered more likely as broad-spectrum antibiotics (see Table 4).

Table 4. True pathogenic antimicrobials predicted by the best classifier on the discovery set.

Compound Name	Annotation
Amoxicillin	Narrow spectrum
Phenoxymethylpenicillin	Narrow spectrum
Cephalexin	Narrow spectrum

On the other hand, among the 61 false negatives, 3 compounds were annotated as broad-spectrum antibiotics (see Supplemental Table S22). This annotation is consistent with our predictions, since these antibiotics directed towards pathogenic microorganisms are unlikely to affect the non-pathogenic gut microbes. Furthermore, 17 out of the 22 antifungal compounds were predicted as antimicrobials.

Thus, in total we were able to correctly identify 16 out of the 19 known broad-spectrum antibiotics and we suggest that 3 of the annotated narrow-spectrum antibiotics should be re-evaluated; hence, the reliability to identify broad-spectrum antibiotics was 84.2%. Furthermore, our results suggest that 56 (61 true negatives less 5 antifungals) (50.4%) out of 111 antibiotics approved by the FDA included in our discovery set are unlikely to affect gut microbes. In comparison, 5 (22.7%) out of 22 antifungals were predicted not to act against the gut microbes (see Supplemental Table S21). Thus, it is twice as much less likely that FDA-approved antibiotics would be toxic against gut microbes than antifungals.

3. Discussion

The identification of antimicrobial compounds assisted by machine-learning techniques has multiple advantages, such as reduction of the invested time to develop novel pharmaceuticals or to flag molecules that could have secondary antimicrobial activity [17]. An important aspect of these techniques is how to improve the reliability of these predictions. One way to achieve this is to increase the number of examples in the training and testing sets. In this work we propose that it is possible to use chemical compounds of different nature (peptides and NPCC) that are commonly modeled separately as antimicrobials to improve the reliability of the predictions. Here we show that indeed, the training sets that rendered the best classifiers of antimicrobial compounds were heterologous, those including NPCC and peptides (see Figures 1 and 2). We can compare our best classifier with previous works in terms of the learnability of our classes, that is, how well gut antimicrobial compounds are differentiated from non-antimicrobial gut compounds. In that sense, the numeric performance achieved by the best classifier on the testing set (AUC = 0.83) is comparable with the performance achieved with one of the best antimicrobial peptide classifiers (AUC = 0.85) recently reported [23], indicating that the learnability of heterologous training sets is as good as those of only peptides.

Another important aspect of our work is the molecular descriptors obtained to best classify gut antimicrobial compounds that included both peptides and NPCC. Although our goal was not to identify common descriptors for NPCC and peptides (these are already calculated by available packages, see Methods), we did look for those descriptors that are relevant to learn the difference between antimicrobials from non-antimicrobials. Our results indicate that the solution to this problem requires the transformation of 86 computed molecular descriptors, suggesting that other molecular descriptors, most likely associated to these 86 descriptors, may improve the current best-model performance.

In terms of improving the performance reported in this work, it is worth mentioning that we used peptides that were not tested by Maier et al. [7] yet, these peptides had reported antibiotic activity against at least one microorganism (*Escherichia coli*) found in the gut and tested by Maier and collaborators. On the other hand, the NPCC included in our work had antibiotic activity against at least one of the 40 gut microorganisms tested by Maier and collaborators. Hence, one alternative approach to improve the performance of classifiers aimed at identifying gut microorganisms would be to include antibiotics that target more common gut microorganisms; that would require further experimental data that is not currently available at present.

To the best of our knowledge, no previous machine-learning efforts to assist in the identification of broad-spectrum antibiotics have been reported; here the definition of broad-spectrum antibiotics was restricted to those acting against both pathogenic and non-pathogenic microorganisms. Hence, using a classifier trained to identify gut non-pathogenic antimicrobial compounds to predict this activity in FDA-approved antibiotics targeted against pathogenic microorganisms represents a way to identify broad-spectrum antibiotics. Our results suggest that half of the FDA-approved antibiotics are likely to have antimicrobial activity against the gut microorganisms indicating that these require further testing or investigation. For instance, two annotated narrow-spectrum antibiotics, amoxicillin and cephalexin, that were predicted to alter gut microbes are known to affect the gastrointestinal flora [24]. On the other hand, the broad-spectrum antibiotic ceftaroline fosamil recently approved by the FDA to treat bacterial pneumonia and skin infections, which was not predicted to affect the gut flora, was reported to have minor gastrointestinal effects during clinical trials [25].

How significant is our finding that almost half of the FDA-approved antibiotics are predicted to have a broad-spectrum activity? To address this question, we included in the discovery set a group of antifungal compounds. All microorganisms used to train our models were bacteria, hence we expected that these antifungals that act through a mechanism different from those reported for bacteria would be unlikely predicted to act against bacteria; let's refer to this negative prediction as *expectation-antifungal*. On the other hand, most FDA-approved antibiotics should unlikely present antibiotic activity against gut microbes, otherwise these would frequently have secondary gastrointestinal effects on patients; let's refer to this negative prediction as *expectation-FDA*. Then, to address the significance of our findings about broad-spectrum antibiotics requires evaluating *expectation-antifungal* and *expectation-FDA*; if FDA-approved drugs are less likely to act on gut microbes than antifungals then $expectation-FDA < expectation-antifungal$. Indeed, we observed that FDA-approved antibiotics are twice as much less likely to act against gut microbes than antifungals. Thus, our results indicate that even when FDA-approved antibiotics are safer (do not act against non-pathogenic resident gut bacteria) than our control group (antifungals), we identified some of these compounds that need to be re-assessed as potential promoters of resistance among microbes for their potential broad-spectrum activity.

In summary, we report a computational approach to use heterologous antimicrobial compounds (peptides and non-peptides) to improve the discriminatory power of machine-learning approaches. We show that training a classifier to identify antibiotics against the gut flora using heterologous training sets correctly anticipate adverse gastrointestinal reactions in patients receiving these antibiotics.

4. Materials and Methods

4.1. Materials

Peptides included in the training sets were obtained from the non-redundant data set of 20 public databases (see Table 5). Testing sets were derived from the work reported by Maier and collaborators (see Supplemental Tables S10–S18). Finally, a discovery set containing 750 FDA-approved drugs for treating human infectious diseases and 76 antifungal drugs was built from the ZINC database [26]. Molecular descriptors were computed with PadelDescriptor [27]. For every training and test set, we performed five different approaches to process the molecular descriptors for each peptide and/or NPCC. These included: no processing; eliminate every null value; substitute every “Infinity” value for 0 or 99,999,999; reduction of the dimensionality applying a principal component analysis implemented in WEKA package (see below). Since the substitution of Infinity values for 0 or 99,999,999 is not a conventional strategy, we performed an imputation of the Infinity and null values using the K nearest neighbor or mean imputation approaches, but only on the best model data set for comparison. That is, from the 9 training sets we generated a total of 45 training sets following the different approaches described before; the same applies to the 9 testing sets. For the discovery set only the transformation applied to the best classifier was performed.

Table 5. Antimicrobial peptide databases used in the present study.

Database	Focused on	Reference
BACTIBASE	Bacteriocins	[28]
Bagel	Bacteriocins	[29]
CAMP	General and Patented AMPs	[14]
DADP	Anuran AMPs	[30]
DAMPD	General AMPs *	[31]
DBAASP	General AMPs	[13]
Defensins	Defensins	[32]
HIPdb	Anti-HIV peptides	[33]
LAMP	General and Patented AMPs	[34]
MilkAMP	AMPs of dairy origin	[35]
PhytAMP	Plant AMPs	[36]
PenBase	Penaeidin AMPs	[37]
Peptaibol	Peptaibols	[38]
RAPD	Recombinant AMPs	[39]
AMPer	Eukaryotic AMPs	[40]
UniprotKb	General AMPs	[41]
YADAMP	General AMPs	[42]
AMSDb	Eukaryotic AMPs	[43]
APD	General AMPs	[44]
AVPdb	Antiviral peptides	[45]

* AMPs stands for Antimicrobial Peptides.

4.2. Method

To identify the best model to classify gut antimicrobial compounds, we followed a systematic method previously reported by our group [46]. Briefly, given the training sets, 52 different machine-learning algorithms implemented in WEKA [47] and their parameters were systematically

analyzed to identify the algorithm, parameters and molecular descriptors that renders the lowest possible error in classification; this systematic analysis was performed by the Bayesian optimization algorithm implemented in AutoWEKA [48]. We ran AutoWEKA against any training set for 10, 90, 720, 2880 and 4320 minutes to identify when the optimization has reached a plateau in the classification error. Afterwards, a 10-fold cross validation and 67% split tests were performed in WEKA. Finally, these classifiers were evaluated against their corresponding testing sets. Two statistical parameters were chosen to evaluate the performance of the classifiers during the testing, including: Area under the ROC curve and correctly classified instances on the testing set. Therefore, a total of 5 statistical parameters were used to define the best classifiers, three for the training phase (adjusted estimated error rate on the training set; correctly classified instances in the training set after splitting 33% for testing; 10-fold cross-validation) and two for the testing phase (AUROC and correctly classified instances).

To identify the intersection set between the top 5 classifiers, we compared the predictions of these classifiers rendering 10 possible pairs of predictions on the discovery set; we used this set because not every classifier had the same testing set. The best model was identified using a combined score (see formula 1): the model with the lowest combined score was chosen. The model then was used to predict gut antimicrobial compounds in the discovery set using WEKA command line (see Supplemental File S1). To annotate as broad-spectrum or narrow-spectrum antibiotics, we used five different previous works that classified antibiotic action [22,49–52].

Supplementary Materials: The following are available online at <http://bis.ifc.unam.mx:8080/ironbios/heteroml/>, File S1: Script to execute the best model to predict antimicrobials on FDA-approved drugs, Table S1A–E: Training sets in ARF format for TrOnlyPeptides, Table S2A–E: Training sets in ARFF format for TrNPCC1, Table S3A–E: Training sets in ARFF format for TrNPCC2, Table S4A–E: Training sets in ARFF format for TrNPCC3, Table S5A–E: Training sets in ARFF format for TrNPCC4, Table S6A–L: Training sets in ARFF format for TrHeterologous1, Table S7A–E: Training sets in ARFF format for TrHeterologous2, Table S8A–E: Training sets in ARFF format for TrHeterologous3, Table S9A–E: Training sets in ARFF format for TrHeterologous4, Table S10A–E: Testing sets in ARF format for TeOnlyPeptides, Table S11A–E: Testing sets in ARF format for TeNPCC1, Table S12A–E: Testing sets in ARF format for TeNPCC2, Table S13A–E: Testing sets in ARF format for TeNPCC3, Table S14A–E: Testing sets in ARF format for TeNPCC4, Table S15A–J: Testing sets in ARF format for TeHeterologous1, Table S16A–E: Testing sets in ARF format for TeHeterologous2, Table S17A–E: Testing sets in ARF format for TeHeterologous3, Table S18A–E: Testing sets in ARF format for TeHeterologous4, Table S19: Parameter values for all models tested, Table S20: CScore values for all model tested, Table S21: Best models algorithms and corresponding parameters and Table S22: Discovery set.

Author Contributions: Conceptualization, G.D.R.; methodology, G.D.R., C.A.B. and R.A.N.L.; software, G.D.R. and R.A.N.L.; validation, R.A.N.L. and G.D.R.; formal analysis, R.A.N.L.; investigation, G.D.R. and R.A.N.L.; resources, G.D.R., C.A.B. and A.P.; data curation, R.A.N.L. and L.A.-M.; writing—original draft preparation, G.D.R.; writing—review and editing, C.A.B., A.P., L.A.-M. and R.A.N.L.; visualization, G.D.R. and R.A.N.L.; supervision, G.D.R. and C.A.B.; project administration, G.D.R.; funding acquisition, G.D.R.

Funding: This research was funded by the following agencies: CONACyT (gran numbers FOIN-219 and CB-252316), Programa de Apoyo a Proyectos de Investigación y Tecnológicos of the UNAM (grant numbers IG100416 and IN208014) and the Instituto de fisiología celular at UNAM. RN is enrolled as a graduate student of the Programa de maestría en ciencias bioquímicas at UNAM and received support from CONACyT with the fellowship number 631360.

Acknowledgments: To María Teresa Lara Ortiz for her technical assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Akova, M. Epidemiology of antimicrobial resistance in bloodstream infections. *Virulence* **2016**, *7*, 252–266. [CrossRef]
2. Aslam, B.; Wang, W.; Arshad, M.I.; Khurshid, M.; Muzammil, S.; Nisar, M.A.; Alvi, R.F.; Aslam, M.A.; Qamar, M.U.; Salamat, M.K.F.; et al. Antibiotic resistance: A rundown of a global crisis. *Infect. Drug Resist.* **2018**, *11*, 1645–1658. [CrossRef]

3. Roger, P.-M.; Montera, E.; Lesselingue, D.; Troadec, N.; Charlot, P.; Simand, A.; Rancezot, A.; Pantaloni, O.; Guichard, T.; Dautezac, V.; et al. Risk factors for unnecessary antibiotic therapy: A major role for clinical management. *Clin Infect Dis.* **2018**. [CrossRef] [PubMed]
4. Jackson, M.A.; Goodrich, J.K.; Maxan, M.-E.; Freedberg, D.E.; Abrams, J.A.; Poole, A.C.; Sutter, J.L.; Welter, D.; Ley, R.E.; Bell, J.T.; et al. Proton pump inhibitors alter the composition of the gut microbiota. *Gut* **2016**, *65*, 749–756. [CrossRef] [PubMed]
5. Rogers, M.A.M.; Aronoff, D.M. The influence of non-steroidal anti-inflammatory drugs on the gut microbiome. *Clin. Microbiol. Infect.* **2016**, *22*, 178.e1–178.e9. [CrossRef]
6. Flowers, S.A.; Evans, S.J.; Ward, K.M.; McInnis, M.G.; Ellingrod, V.L. Interaction between Atypical Antipsychotics and the Gut Microbiome in a Bipolar Disease Cohort. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* **2017**, *37*, 261–267. [CrossRef] [PubMed]
7. Maier, L.; Pruteanu, M.; Kuhn, M.; Zeller, G.; Telzerow, A.; Anderson, E.E.; Brochado, A.R.; Fernandez, K.C.; Dose, H.; Mori, H.; et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **2018**, *555*, 623–628. [CrossRef] [PubMed]
8. González-Díaz, H.; Prado-Prado, F.J.; Santana, L.; Uriarte, E. Unify QSAR approach to antimicrobials. Part 1: Predicting antifungal activity against different species. *Bioorg. Med. Chem.* **2006**, *14*, 5973–5980. [CrossRef]
9. Rath, E.C.; Gill, H.; Bai, Y. Identification of potential antimicrobials against *Salmonella typhimurium* and *Listeria monocytogenes* using Quantitative Structure-Activity Relation modeling. *PLoS ONE* **2017**, *12*, e0189580. [CrossRef]
10. Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F.J.; Salabert-Salvador, M.T.; Díaz-Villanueva, W.; Castro-Bleda, M.J.; Villanueva-Pareja, A. Artificial Neural Networks and Linear Discriminant Analysis: A Valuable Combination in the Selection of New Antibacterial Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031–1041. [CrossRef]
11. Nguyen, M.; Long, S.W.; McDermott, P.F.; Olsen, R.J.; Olson, R.; Stevens, R.L.; Tyson, G.H.; Zhao, S.; Davis, J.J. Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* **2018**. Available online: <http://www.ncbi.nlm.nih.gov/pubmed/30333126> (accessed on 19 January 2019). [CrossRef] [PubMed]
12. Wang, Z.; Wang, G. APD: The Antimicrobial Peptide Database. *Nucleic Acids Res.* **2004**, *32*, D590–D592. [CrossRef] [PubMed]
13. Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H.L.; Squires, R.B.; Hurt, D.E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B.; et al. DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **2016**, *44*, D1104–D1112. [CrossRef] [PubMed]
14. Wagh, F.H.; Barai, R.S.; Gurung, P.; Idicula-Thomas, S. CAMP R3: A database on sequences, structures and signatures of antimicrobial peptides: Table 1. *Nucleic Acids Res.* **2016**, *44*, D1094–D1097. [CrossRef] [PubMed]
15. Del Rio, G.; Castro-Obregon, S.; Rao, R.; Ellerby, H.M.; Bredesen, D.E. APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. *FEBS Lett.* **2001**, *494*, 213–219. [CrossRef]
16. Toropova, M.A.; Veselinović, A.M.; Veselinović, J.B.; Stojanović, D.B.; Toropov, A.A. QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. *Comput. Biol. Chem.* **2015**, *59*, 126–130. [CrossRef]
17. Durrant, J.D.; Amaro, R.E. Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.* **2015**, *85*, 14–21. [CrossRef]
18. Battisti, A.; Zamuner, S.; Sarti, E.; Laio, A. Toward a unified scoring function for native state discrimination and drug-binding pocket recognition. *Phys. Chem. Chem. Phys.* **2018**, *20*, 17148–17155. [CrossRef]
19. Del Rio, G.; Koschützki, D.; Coello, G. How to identify essential genes from molecular networks? *BMC Syst. Biol.* **2009**, *3*, 102. [CrossRef] [PubMed]
20. Calahorra, M.; Sánchez, N.S.; Peña, A. Influence of phenothiazines, phenazines and phenoxazine on cation transport in *Candida albicans*. *J. Appl. Microbiol.* **2018**, *125*, 1728–1738. [CrossRef] [PubMed]
21. Acar, J. Broad- and narrow-spectrum antibiotics: An unhelpful categorization. *Clin. Microbiol. Infect.* **1997**, *3*, 395–396. [CrossRef]
22. Sarpong, E.M.; Miller, G.E. Narrow- and Broad-Spectrum Antibiotic Use among U.S. Children. *Health Serv. Res.* **2015**, *50*, 830–846. [CrossRef] [PubMed]

23. Beltran, J.A.; Aguilera-Mendoza, L.; Brizuela, C.A. Optimal selection of molecular descriptors for antimicrobial peptides classification: An evolutionary feature weighting approach. *BMC Genomics* **2018**, *19*, 672. [CrossRef]
24. NIH DailyMed. 26/November 2018. Available online: <https://dailymed.nlm.nih.gov/dailymed/index.cfm> (accessed on 19 January 2019).
25. File, T.M.; Wilcox, M.H.; Stein, G.E. Summary of Ceftaroline Fosamil Clinical Trial Studies and Clinical Safety. *Clin. Infect. Dis.* **2012**, *55*, S173–S180. [CrossRef] [PubMed]
26. Sterling, T.; Irwin, J.J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef]
27. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
28. Hammami, R.; Zouhir, A.; Le Lay, C.; Ben Hamida, J.; Fliss, I. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiol.* **2010**, *10*, 22. [CrossRef]
29. De Jong, A.; van Heel, A.J.; Kok, J.; Kuipers, O.P. BAGEL2: Mining for bacteriocins in genomic data. *Nucleic Acids Res.* **2010**, *38*, W647–W651. [CrossRef]
30. Novković, M.; Simunić, J.; Bojović, V.; Tossi, A.; Juretić, D. DADP: The database of anuran defense peptides. *Bioinformatics* **2012**, *28*, 1406–1407. [CrossRef] [PubMed]
31. Seshadri Sundararajan, V.; Gabere, M.N.; Pretorius, A.; Adam, S.; Christoffels, A.; Lehvälaiho, M.; Archer, J.A.C.; Bajic, V.B. DAMPD: A manually curated antimicrobial peptide database. *Nucleic Acids Res.* **2012**, *40*, D1108–D1112. [CrossRef]
32. Seebah, S.; Suresh, A.; Zhuo, S.; Choong, Y.H.; Chua, H.; Chuon, D.; Beuerman, R.; Verma, C. Defensins knowledgebase: A manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.* **2007**, *35*, D265–D268. [CrossRef]
33. Qureshi, A.; Thakur, N.; Kumar, M. HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides. *PLoS ONE* **2013**, *8*, e54908. [CrossRef] [PubMed]
34. Zhao, X.; Wu, H.; Lu, H.; Li, G.; Huang, Q. LAMP: A Database Linking Antimicrobial Peptides. *PLoS ONE* **2013**, *8*, e66557. [CrossRef] [PubMed]
35. Théolier, J.; Fliss, I.; Jean, J.; Hammami, R. MilkAMP: A comprehensive database of antimicrobial peptides of dairy origin. *Dairy Sci. Technol.* **2014**, *94*, 181–193. [CrossRef]
36. Hammami, R.; Ben Hamida, J.; Vergoten, G.; Fliss, I. PhytAMP: A database dedicated to antimicrobial plant peptides. *Nucleic Acids Res.* **2009**, *37*, D963–D968. [CrossRef] [PubMed]
37. Gueguen, Y.; Garnier, J.; Robert, L.; Lefranc, M.; Mougnot, I.; Lorgèril, J.; Janech, M.; Gross, P.S.; Warr, G.W.; Cuthbertson, B.; et al. PenBase, the shrimp antimicrobial peptide penaeidin database: Sequence-based classification and recommended nomenclature. *Dev. Comp. Immunol.* **2006**, *30*, 283–288. [CrossRef]
38. Whitmore, L.; Wallace, B.A. The Peptaibol Database: A database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.* **2004**, *32*, D593–D594. [CrossRef]
39. Li, Y.; Chen, Z. RAPD: A database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol. Lett.* **2008**, *289*, 126–129. [CrossRef]
40. Fjell, C.D.; Hancock, R.E.W.; Cherkasov, A. AMPer: A database and an automated discovery tool for antimicrobial peptides. *Bioinformatics* **2007**, *23*, 1148–1155. Available online: <http://www.ncbi.nlm.nih.gov/pubmed/17341497> (accessed on 23 January 2019). [CrossRef]
41. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.
42. Piotto, S.P.; Sessa, L.; Concilio, S.; Iannelli, P. YADAMP: Yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents* **2012**, *39*, 346–351. [CrossRef] [PubMed]
43. Tossi, A.; Sandri, L. Molecular diversity in gene-encoded, cationic antimicrobial polypeptides. *Curr. Pharm. Des.* **2002**, *8*, 743–761. [CrossRef] [PubMed]
44. Wang, G.; Li, X.; Wang, Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087–D1093. [CrossRef]
45. Qureshi, A.; Thakur, N.; Tandon, H.; Kumar, M. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.* **2014**, *42*, D1147–D1153. [CrossRef] [PubMed]

46. Corral-Corral, R.; Beltrán, J.; Brizuela, C.; Del Rio, G. Systematic Identification of Machine-Learning Models Aimed to Classify Critical Residues for Protein Function from Protein Structure. *Molecules* **2017**, *22*, 1673. [[CrossRef](#)]
47. Witten, I.H.; Ian, H.; Frank, E.; Hall, M.A.; Mark, A. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2011; 629p.
48. Kotthoff, L.; Thornton, C.; Hoos, H.H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
49. Newland, J.G.; Stach, L.M.; De Lurgio, S.A.; Hedican, E.; Yu, D.; Prasad, P.A.; Jackson, M.A.; Myers, A.L.; Zaoutis, T.E. Impact of a Prospective-Audit-With-Feedback Antimicrobial Stewardship Program at a Children’s Hospital. *J. Pediatric Infect. Dis. Soc.* **2012**, *1*, 179–186. [[CrossRef](#)]
50. Newman, R.E.; Hedican, E.B.; Herigon, J.C.; Williams, D.D.; Williams, A.R.; Jason, G. Newland. Impact of a Guideline on Management of Children Hospitalized With Community-Acquired Pneumonia. *Pediatrics* **2012**, *129*, e597–e604. [[CrossRef](#)]
51. Di Pentima, M.C.; Chan, S. Impact of Antimicrobial Stewardship Program on Vancomycin Use in a Pediatric Teaching Hospital. *Pediatr. Infect. Dis. J.* **2010**, *29*, 707–711. [[CrossRef](#)]
52. Kreitmeyr, K.; von Both, U.; Pecar, A.; Borde, J.P.; Mikolajczyk, R.; Huebner, J. Pediatric antibiotic stewardship: Successful interventions to reduce broad-spectrum antibiotic use on general pediatric wards. *Infection* **2017**, *45*, 493–504. [[CrossRef](#)] [[PubMed](#)]

Sample Availability: All data used in this study are available as supplemental data.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).