CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN
SUPERIOR DE ENSENADA, BAJA CALIFORNIA

PROGRAMA DE POSGRADO EN CIENCIAS
EN CIENCIAS DE LA COMPUTACIÓN

Métodos para la selección de características y clasificación de péptidos antimicrobianos

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Jesús Armando Beltrán Verdugo

Ensenada, Baja California, México

2014

Tesis defendida por

Jesús Armando Beltrán Verdugo

y aprobada por el siguiente comité

Dr. Carlos Alberto Brizuela Rodríguez

Director del Comité

Dr. Israel Marck Martínez Pérez

Miembro del Comité

Dr. Hugo Homero Hidalgo Silva

Miembro del Comité

Dra. Clara Elizabeth Galindo Sánchez

Miembro del Comité

Dra. Ana Isabel Martínez García

Coordinador del Programa de Posgrado en Ciencias de la Computación

---

Dr. Jesús Favela Vara

Director de Estudios de Posgrado

Octubre, 2014

Resumen de la tesis que presenta **Jesús Armando Beltrán Verdugo** como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación.

Métodos para la selección de características y clasificación de péptidos antimicrobianos

Resumen elaborado por:

Jesús Armando Beltrán Verdugo

Los péptidos antimicrobianos (AMPs) son una alternativa potencial para combatir los patógenos resistentes a antibióticos debido a que poseen múltiples mecanismos de acción en contra de microbios tales como: bacterias, hongos y virus. Estos péptidos se encuentran en la naturaleza en casi todas las formas de vida como parte del sistema inmune. Los AMPs son una plantilla interesante para producir nuevos agentes antimicrobianos selectivos, es decir, péptidos con alta actividad antimicrobiana pero con bajos niveles de toxicidad en el organismo huésped. Las técnicas tradicionales para el diseño y optimización de péptidos pueden ser tardadas y costosas, por lo que asistirse de herramientas computacionales puede ayudar a limitar el vasto espacio de secuencias que se tienen que evaluar en el laboratorio.

Un método para la predicción de péptidos antimicrobianos (AMPs) y no antimicrobianos (no AMPs) es QSAR (*Quantitative Structure-Activity Relationship*). Este método relaciona las propiedades fisicoquímicas (descriptores moleculares) del péptido con su actividad biológica mediante un modelo matemático. Un aspecto importante para la construcción del modelo es la selección de los descriptores moleculares. Actualmente, existen miles de descriptores medibles en los péptidos, por lo que elegir los descriptores moleculares que capturen las propiedades relevantes de los AMPs se torna una tarea difícil. Las principales razones de esta dificultad son: primero, no se conoce una regla determinista que gobierne la elección de los descriptores; segundo, explorar el espacio de todos los posibles subconjuntos de descriptores no es factible, ya que el espacio de búsqueda es de tamaño 2^n (donde n es el número de descriptores).

En el presente trabajo se propone el diseño de un algoritmo para la selección de características compuesto principalmente por dos elementos: un algoritmo genético para la generación y búsqueda eficiente de los posibles subconjuntos de características y una máquina de soporte vectorial (SVM) para evaluar la calidad del subconjunto seleccionado. El algoritmo recibe como entrada un conjunto de péptidos con y sin actividad antimicrobiana, un conjunto X de características y un modelo de clasificación. La salida del algoritmo es el subconjunto de descriptores con la máxima exactitud del modelo de clasificación. Los resultados indican que con el mejor subconjunto encontrado de características se puede construir un modelo de clasificación que predice correctamente la actividad del 96 % de los péptidos de prueba. Este mismo modelo logra una exactitud de 82.3 % sobre un conjunto de casos desconocidos para el algoritmo.

Palabras Clave: **Péptidos antimicrobianos, QSAR, selección de características, clasificación de péptidos, predicción de actividad antimicrobiana, cribado virtual, SVM, algoritmo genético.**

Abstract of the thesis presented by **Jesús Armando Beltrán Verdugo** as a partial requirement to obtain the Master of Science degree in Master in Sciences in Computer Science.

Feature selection methods and classification of antimicrobial peptides

Abstract by:

Jesús Armando Beltrán Verdugo

Antimicrobial peptides (AMPs) are a promising alternative for combating pathogen that are resistant to antibiotics, because their multiple action mechanisms against microbe such as, bacteria, fungi, and virus. These peptides are in nature in almost every form of life, as a part of the defense mechanism. The AMPs are an interesting template to produce new selective antimicrobials agents, i.e., peptides with a high antimicrobial activity and a low toxicity level in the host organism. Traditional techniques for peptide design and optimization can be tedious and expensive, therefore the use of computational tools can help to reduce the sequence space that have to be evaluated in the laboratory.

QSAR (Quantitative Structure-Activity Relationship) is a method for predicting active (AMPs) and not active (non-AMPs) peptides. This method use a mathematical model to associate the peptides physicochemical properties (molecular descriptors) to their biological activity. An important aspect to build the mathematical model is the selection of molecular descriptors. Nowadays, there are thousands of proposed descriptors, therefore, to choose the ones that capture the relevant AMPs properties is a hard goal to achieve. The main reason for this are: first, it is unknown a deterministic rule that governs the descriptors selection; second, to explore the space of all possible descriptor subsets is not feasible, this is because the size of the search space is 2^n (where n is the number of descriptor).

We propose a features selection algorithm, composed by two main elements: a genetic algorithm for the generation and efficient search of the characteristics subsets, and a Support Vector Machine (SVM) to evaluate the quality of the selected subset. The algorithm receives as input a set of peptides with and without antimicrobial activity, a set X of characteristics and a classification model. The algorithm outputs the descriptor subset with the highest accuracy in the predefined classification model. The results show that the best characteristics subset achieved can develop a classification model that predicts the activity correctly over 96 % of the tested peptides. This model has an 82.3 % accuracy over a set peptides which is unknown to the algorithm.

Keywords: **Antimicrobial peptides, QSAR, feature selection, peptides clasification, prediction of antimicrobial activity, virtual screening, SVM, genetic algorithm.**

Dedicatoria

Este trabajo de tesis se lo dedico a mi esposa Linney por su apoyo, inspiración y comprensión durante mi estancia en el posgrado y en mi vida. Este trabajo también va dedicado a mis padres y hermanos por su amor, apoyo incondicional y cuyos ejemplos me han inspirado a seguir mis sueños.

Agradecimientos

Agradezco especialmente a mi director de tesis, Dr. Carlos Alberto Brizuela Rodríguez por su guía y continuas enseñanzas a lo largo de mi estancia en el posgrado. Además también quiero agradecerle por despertar en mí el interés en biocomputación y en el área de optimización. Ha sido un privilegio trabajar con alguien tan preparado, pero sobretodo, con la calidad moral del Dr. Carlos. Así mismo a los miembros del comité de tesis, Dr. Israel Marck Martínez Pérez, Dr. Hugo Homero Hidalgo Silva y Dra. Clara Elizabeth Galindo Sánchez, por su tiempo, observaciones y sugerencias a lo largo del desarrollo de este trabajo.

Agradezco a mis maestros y compañeros del posgrado de Ciencias de la Computación especialmente a los integrantes del laboratorio de biocomputación Hugo, Nelson, David, Najash, José con quienes conviví durante este par de años y de quienes recibí siempre buenas atenciones y consejos. Además también quiero agradecer a mis compañeros de la generación 2012 con quienes conviví y compartí buenas experiencias, especialmente a mis compañeros Lino, Julio, Julia, y a ti Franceli, gracias por tu apoyo incondicional .

Finalmente, a CICESE por permitirme estudiar este posgrado, y al personal administrativo por la excelente atención que siempre me brindaron y Al CONACyT por su apoyo económico para realizar mis estudios.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Lista de figuras (continuación)

Lista de figuras (continuación)

Lista de tablas

Lista de tablas (continuación)

Capítulo 1. Introducción

1.1. Motivación: resistencia a antibióticos

En los últimos 50 años, los antibióticos permitieron el tratamiento de infecciones bacterianas de manera exitosa (Hancock, 1997; Scott *et al.*, 2007). En la era antibiótica, enfermedades que en el siglo pasado eran mortales, actualmente son fácilmente curables; un ejemplo, es la enfermedad infantil de la fiebre escarlata producida por la bacteria *Streptococcus pyogenes* (Del Pozo Menéndez *et al.*, 2011), esta enfermedad era grave hace 150 años teniendo una tasa elevada de mortalidad, sin embargo en la actualidad se trata con eficacia debido a los antibióticos (Quinn, 1982). Actualmente, la efectividad de los antibióticos está alcanzando su límite y los niveles de resistencia se están incrementando a un nivel alarmante. Se entiende como resistencia a antibióticos, la resistencia que adquiere la bacteria al medicamento que en el pasado combatió eficazmente la infección causada por la misma (WHO, 2014). En general, la resistencia se desarrolla principalmente debido a los siguientes factores: las mutaciones de la bacteria a lo largo del tiempo, permitiendo mejorar su crecimiento en entornos difíciles; el uso excesivo e inapropiado de los antibióticos en el tratamiento clínico acelera la aparición de cepas resistentes a los medicamentos (WHO, 2014); asimismo, la ausencia de nuevas clases de antibióticos descubiertos en los últimos años (Leid, 2009). Un ejemplo, de bacteria resistente es *Staphylococcus aureus* resistente a meticilina (MRSA), asociado con una variedad de infecciones de moderadas a graves en los humanos y es resistente a la mayoría de los antibióticos conocidos (Dosler y Mataraci, 2013). Por todo lo anterior, surge la necesidad de desarrollar tratamientos alternos a los antibióticos tradicionales para combatir con éxito las enfermedades ocasionadas por organismos multirresistentes.

Los péptidos antimicrobianos (AMPs por sus siglas en inglés de *Antimicrobial Peptides*) son una alternativa potencial para el tratamiento de infecciones causadas por bacterias resistentes. Los AMPs se encuentran presentes en la mayoría de las formas de vida como primera línea de defensa del sistema inmune en contra de microorganismos patógenos. Los AMPs matan de manera eficiente una amplia variedad de especies (bacterias, hongos y virus) de manera directa, además son eficaces en contra de los patógenos que son resistentes a casi

todos los antibióticos convencionales (Cherkasov *et al.*, 2008). A pesar de las propiedades atractivas con las que cuentan los AMPs, estos poseen desventajas que impiden su uso como agente terapéutico: toxicidad, degradación por proteasas, amplio espectro, alto costo de producción (Fernandes *et al.*, 2012; Aoki y Ueda, 2013) son algunas de ellas. Estas desventajas, presentan oportunidades en investigación para el diseño de AMPs, teniendo como objetivo crear o identificar secuencia costo-efectivas, que tengan un alta actividad antimicrobiana sin exhibir altos niveles de toxicidad (*i.e.,* péptidos que presenten un alto índice terapéutico) (Fjell *et al.*, 2012; Jenssen *et al.*, 2006).

El proceso de diseñar y descubrir nuevos AMPs inicia con la identificación de péptidos con actividad antimicrobiana, para esto se utilizan dos técnicas: la química combinatoria, y el cribado de alto desempeño (HTS, por sus siglas en inglés de *High Throughput Screening*). Por un lado, la química combinatoria permite la síntesis rápida de un gran número de péptidos con atributos comunes (librería combinatoria de péptidos). Por otro lado, HTS se utiliza para probar miles de péptidos rápidamente de manera paralela. Los enfoques computacionales pueden resultar de gran ayuda en el proceso de descubrimiento y diseño cuando las técnicas biológicas para síntesis y prueba exhaustiva de péptidos son prohibitivamente costosas. Los enfoques computacionales facilitan la selección de péptidos al ayudar a eliminar secuencias con pobre o nula actividad en etapas tempranas del diseño (Fjell *et al.*, 2012).

Actualmente, las investigación de AMPs en cómputo van dirigidas a utilizar la gran cantidad de secuencias de AMPs e información almacenada en las bases de datos para generar conocimiento útil para el diseño de nuevos péptidos. Un objetivo deseable en el diseño de péptidos asistido por computadora (*in silico*) es desarrollar un sistema computacional capaz de evaluar automáticamente una gran cantidad de péptidos y de ese modo limitar el vasto espacio de secuencias a probar con los métodos tradicionales, esto con el fin de reducir costo y tiempo (Taboureau, 2010). Por lo anterior, un problema importante es la predicción de la actividad biológica del péptido, problema que se define como: dado un conjunto de péptidos y actividades biológicas conocidas, encontrar un modelo que asigne como salida la actividad correcta para cada péptido de entrada.

Uno de los métodos más usados para encontrar un modelo de predicción de la actividad es QSAR (*Quantitative Structure-Activity Relationship*), debido a que relaciona las propiedades fisicoquímicas cuantificables en los péptidos (descriptores moleculares) con la actividad biológica (*i.e.*, clasificar los péptidos en AMPs y no AMPs) (Fjell *et al.*, 2012; Goodarzi *et al.*, 2012). Para asociar la información del péptido con la actividad biológica, se utiliza una diversidad de modelos matemáticos. Para la detección de AMPs en la literatura se han propuesto varios modelos, tales como: redes neuronales artificiales (ANN) (Fjell *et al.*, 2009; Cherkasov *et al.*, 2008; Torrent *et al.*, 2011), máquinas de soporte vectorial (SVM) (Lata *et al.*, 2010; Torrent *et al.*, 2011; Waghu *et al.*, 2014), análisis de discriminante (DA)(Waghu *et al.*, 2014) y *random forest* (RF)(Waghu *et al.*, 2014).

En los métodos de QSAR, otro aspecto importante a considerar junto con el modelo matemático es la selección de los descriptores moleculares. Actualmente, existen miles de descriptores medibles en los péptidos (*e.g.,* el programa Dragon6 puede calcular 4885 descriptores (Helguera *et al.*, 2008)), por lo que elegir los descriptores adecuados para la identificación de AMPs se torna una tarea difícil (Goodarzi *et al.*, 2012). Una de las causas de esta dificultad, es que no se conoce una regla determinista que gobierne la elección de los descriptores. En la literatura esta selección a menudo se realiza con base a un conocimiento previo de las propiedades fisicoquímicas (cuantificadas en descriptores) que dan lugar a la actividad del péptido (Fjell *et al.*, 2012). Sin embargo, en ocasiones estas propiedades son demasiado generales y compartidas por otras moléculas (Piotto *et al.*, 2012). Por consiguiente, utilizar sólo estos descriptores no es suficiente para crear un modelo confiable capaz de predecir la actividad de nuevos péptidos.

De acuerdo con Fjell *et al.* (2012), los descriptores moleculares idealmente pueden y deben ser seleccionados de forma automática a través de un método denominado selección de características (FS) (Guyon y Elisseeff, 2003). De manera general, los métodos de FS tratan de encontrar el subconjunto de características (descriptores) que maximice algún criterio de evaluación (*e.g.,* exactitud de clasificación) dado un conjunto de características.

Con esta motivación a continuación se define una versión acotada del problema de selección de características con aplicación en la selección de descriptores moleculares para la clasificación de AMPs. Además, se plantean los objetivos de investigación de este trabajo.

1.1.1. Definición del problema

Dado un conjunto de datos $\mathcal{D}$ con un conjunto de descriptores moleculares X; y un modelo de clasificación $\mathcal{I}$. El problema consiste en encontrar $X^{'}_{opt}$ que se define a continuación:

$$X^{'}_{opt} = \underset{X' \subseteq X}{\arg\max}\, J(X', \mathcal{D}) \tag{1}$$

$$J(X', \mathcal{D}) = ACC(\mathcal{I}(\mathcal{D}')) \tag{2}$$

donde $\mathcal{D}' \subseteq \mathcal{D}$ es el conjunto de datos removiendo los valores de las variables que no estén en X'; y ACC es la exactitud del clasificador $\mathcal{I}$. Una solución es óptima si la exactitud del clasificador $ACC(\mathcal{I}(\mathcal{D}'))$ es máxima. Es importante señalar que no necesariamente $X^{'}_{opt}$ es única, esto debido a que se puede llegar a la misma exactitud utilizando diferentes conjuntos de características.

1.2. Objetivo de la investigación

1.2.1. Objetivo general

Diseñar e implementar un algoritmo de selección de propiedades fisicoquímicas dado un clasificador específico para la detección de péptidos antimicrobianos. Se espera que el algoritmo diseñado tenga un desempeño comparable con los reportados por métodos del estado del arte.

1.2.2. Objetivos específicos

- Analizar los mejores métodos de detección de péptidos antimicrobianos en cuanto a eficiencia.
- Analizar los distintos métodos para la selección de características y clasificación.
- Construir una biblioteca de casos de prueba a utilizar.
- Proponer e implementar un método de selección de características para péptidos antimicrobianos dado un clasificador específico.
- Evaluar la calidad de predicción y comparar con los métodos actuales para la detección de péptidos antimicrobianos.

1.2.3. Metodología de solución propuesta

Para abordar este problema se utilizó la metodología que se muestra en la Figura 1. Primero se realizó una revisión acerca de cómo recopilar péptidos con y sin la actividad biológica deseada con el objetivo de elaborar los casos positivos y negativos. Una vez que se obtuvieron los casos de prueba, el siguiente paso fue transformar las secuencia primaria de los péptidos a un conjunto de números (descriptores moleculares) que capturen las propiedades fisicoquímicas relevantes. Después para la selección de los descriptores moleculares relevantes se implementó un método de envoltura (*wrapper*) que se compone principalmente de un algoritmo genético y un algoritmo de aprendizaje de máquina (SVM). También, se aplicó un clasificador que relaciona las características con la actividad utilizando una SVM y el subconjunto de características resultado del algoritmo genético. Por último, se evaluó la calidad de los modelos en términos de la exactitud de predicción.

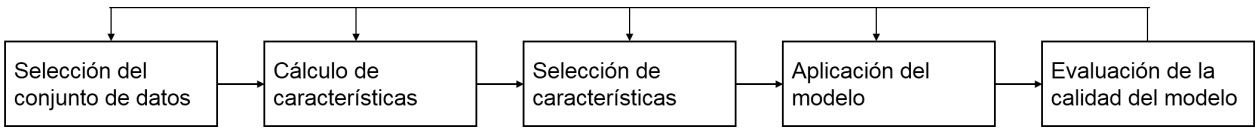


Figura 1: Metodología general propuesta.

1.3. Organización de la tesis

El presente trabajo está organizado de la siguiente manera:

En el Capítulo 2 se exponen conceptos biológicos básicos tales como péptidos, su composición, péptidos antimicrobianos y su clasificación, y las bases de datos de péptidos antimicrobianos. Por otra parte, se abordan los conceptos computacionales básicos para la comprensión del problema tratado en este trabajo. De igual modo se expone el trabajo previo relevante en la detección de AMPs.

En el Capítulo 3 se presenta el problema de selección de características (FSP), la caracterización del problema considerando los aspectos relevantes a tomar en consideración para proponer un algoritmo de selección de características. Además, se describen de forma breve algunos de los métodos para la selección de características. Finalmente, se define el problema a resolver en el presente trabajo.

En el Capítulo 4 se describen la metodología para la obtener los casos de prueba, el cálculo de características, así como las estrategias que se utilizaron para el problema de selección de características. Además se presenta el diseño y análisis del algoritmo para la selección de descriptores moleculares en AMPs.

En el Capítulo 5 se presentan los casos de prueba, los experimentos y resultados obtenidos así como una comparación con los métodos del estado del arte.

En el Capítulo 6 se exponen las conclusiones a las que se llegó, así como algunas propuestas para la continuación de este trabajo de investigación.

En el Apéndice A se muestra una clasificación de los aminoácidos según sus propiedades químicas.

En el Apéndice B se enlistan el conjunto de casos de prueba que se utilizaron.

En el Apéndice C se muestran los descriptores moleculares con los que se representó al conjunto de prueba.

En el Apéndice D se presentan los detalles de la configuración de los algoritmos implementados, así como el diseño general de los algoritmos.

En el Apéndice E se describe cómo calcular la ganancia de información y el procedimiento para el experimento de seleccionar las características utilizando el valor de ganancia de información.

Capítulo 2. Marco Teórico

2.1. Conceptos biológicos

2.1.1. Péptidos

Los péptidos son moléculas que están compuestas por cadenas cortas de aminoácidos unidos por enlaces peptídicos. La longitud de los péptidos es por lo general menor a 100 aminoácidos.

Aminoácidos

Los péptidos y proteínas están compuestos por una cadena de aminoácidos que en cada posición puede tener a uno de los 20 existentes (ver Tabla 1). La estructura general de los aminoácidos tiene una base común y un grupo R. La base común de los aminoácidos, también conocida como columna vertebral, se divide en tres partes: un grupo carboxilo (-COOH), un grupo amino (-NH_2) y un carbono-α (C^α). Por otra parte, el grupo R o cadena lateral confiere propiedades fisicoquímicas muy particulares a cada uno de los 20 aminoácidos (ver Figura 2a).

Cuando dos aminoácidos se unen para formar una cadena polipeptídica, el grupo amino de un aminoácido se une con el grupo carboxilo de otro. Al enlace resultante entre los dos aminoácidos se le conoce como enlace peptídico, mientras que a los aminoácidos presentes en la unión se les denomina residuos (ver Figura 2b).

2.1.2. Niveles estructurales de los péptidos

La estructura de los péptidos la podemos describir en diferentes niveles de organización, los cuales se clasifican de forma ascendente con respecto a la complejidad. Por tanto en cada nivel aumenta la cantidad de información que se tiene de los componentes que integran al péptido. Los niveles van desde las estructuras primarias hasta las cuaternarias.

Estructura primaria: es la estructura básica del péptido, presentando sólo información acerca de las secuencias de residuos que componen a la cadena polipeptídica (ver Figura 3a).

Tabla 1: Los 20 aminoácidos estándar y sus códigos de tres y una letra.

Aminoácido	Código de tres letras	Código de una letra
Alanina	ALA	A
Cisteína	CYS	C
Ácido aspártico	ASP	D
Ácido glutámico	GLU	E
Fenilalanina	PHE	F
Glicina	GLY	G
Histidina	HIS	H
Isoleucina	ILE	I
Lisina	LYS	K
Leucina	LEU	L
Metionina	MET	M
Asparagina	ASN	N
Prolina	PRO	P
Glutamina	GLN	Q
Arginina	ARG	R
Serina	SER	S
Treonina	THR	T
Valina	VAL	V
Triptófano	TRP	W
Tirosina	TYR	Y

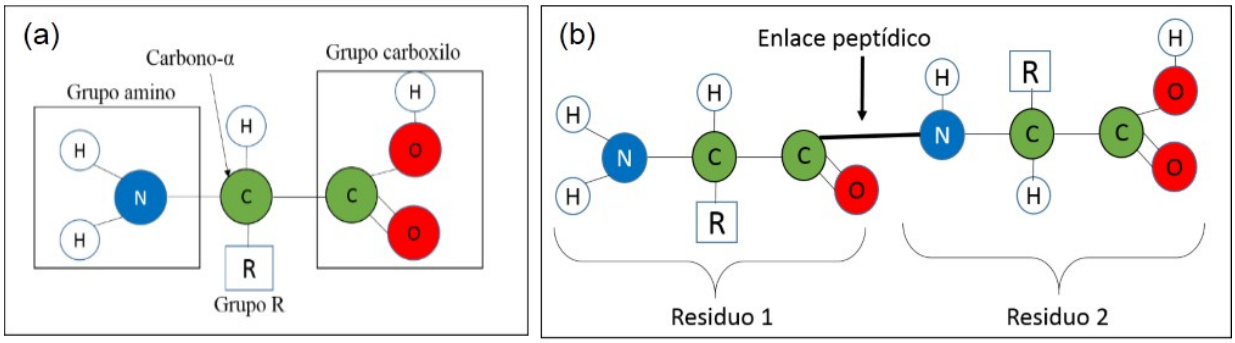


Figura 2: (a) Estructura general del aminoácido. (b) Enlace peptídico formado por las interacciones de dos aminoácidos.

Estructura secundaria: describe las regiones regulares del péptido tal como hélices α y hojas β. Las estructuras se estabilizan por enlaces de hidrógeno formados entre el átomo de oxígeno de un carboxilo y el hidrógeno del grupo amino de otro residuo de aminoácido (ver Figura 3b) (Clote y Backofen, 2000).

Estructura terciaria: este nivel estructural ofrece información acerca de la conformación nativa del péptido al interaccionar con el solvente, representando a los átomos del péptido en un espacio tridimensional (ver Figura 3d). Nos indican también cómo se agrupan espacialmente las estructuras secundarias (ver Figura 3c) (Corona de la Fuente, 2010).

Existen otros niveles estructurales más complejos que ofrecen información acerca de las interacciones de varios péptidos, pero que se escapan del alcance de este trabajo.

2.1.3. Péptidos Antimicrobianos (AMPs)

Los péptidos Antimicrobianos (AMPs por sus siglas en inglés *Antimicrobial Peptides*) son componentes esenciales en el sistema inmune para inhibir el crecimiento o establecimiento de microorganismos patógenos que invaden al organismo huésped. Los AMPs se encuentran en la mayoría de los organismos vivos como compuestos evolutivamente conservados en el sistema inmune desde hace aproximadamente 2.6 mil millones de años (Aoki y Ueda, 2013; Jenssen *et al.*, 2006).

Actualmente, los AMPs se han convertido en una alternativa potencial para el diseño de nuevos fármacos debido a que muestran una actividad microbicida hacia bacterias, hongos, parásitos y virus, además de un amplio espectro de mecanismos de ataque en contra de los patógenos (Aoki y Ueda, 2013; Jenssen *et al.*, 2006).

2.1.4. Clasificación de los AMPs

Debido a la existencia de una amplia diversidad de AMPs en cuanto a secuencias y estructuras se refiere, en la literatura se presentan diversas maneras de clasificarlos. Por ejemplo,

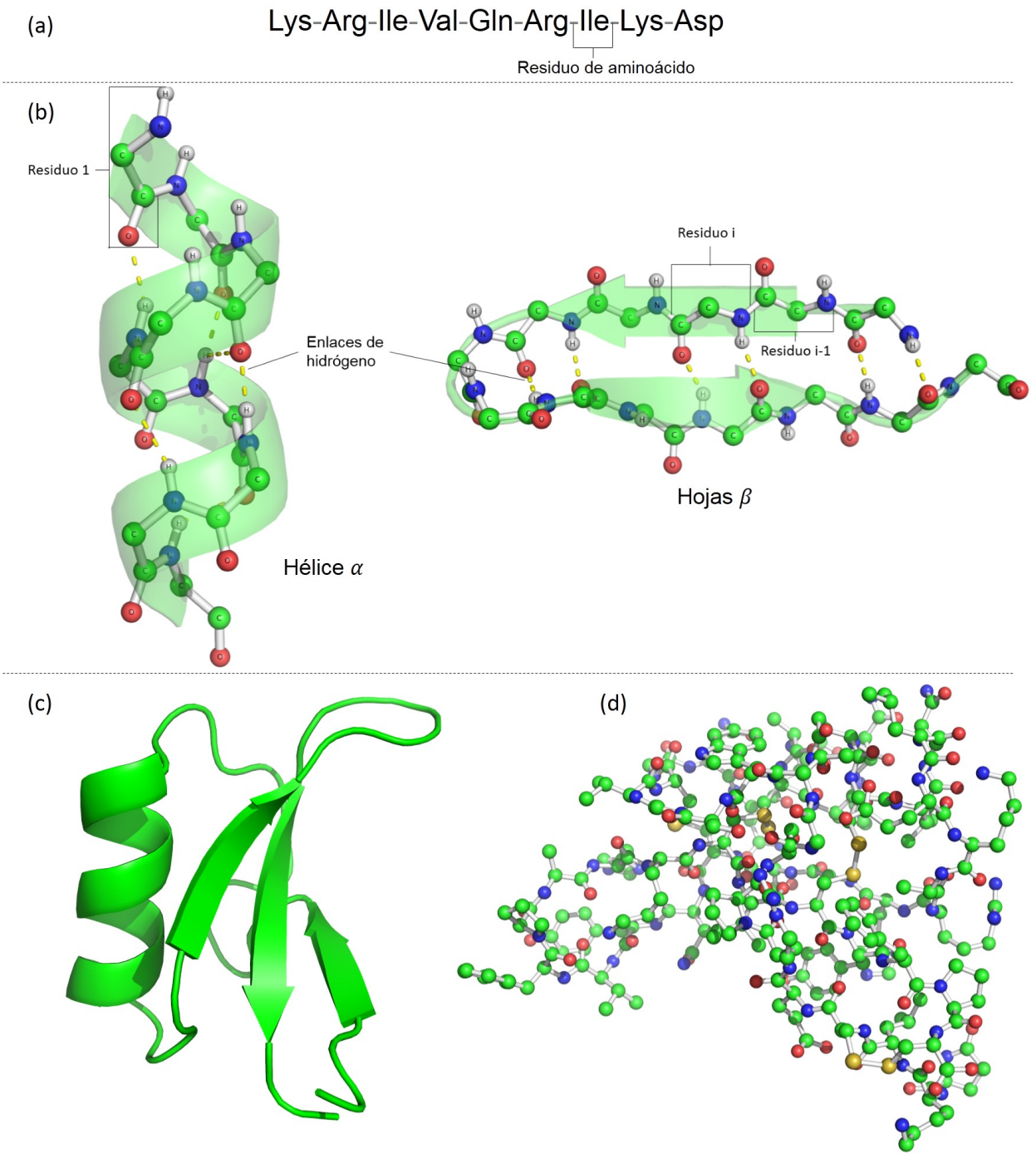


Figura 3: Niveles estructurales de los péptidos. (a) Estructura primaria usando código de tres letras por residuo. (b) Estructura secundaria: hélice α (residuos 18-25 del péptido 2K6O) y Hojas β (PDB 1LFC). (c) y (d) Estructura terciaria (PDB 1AYJ). En (c) y (d) para las estructuras de los péptido es usando un modelo de *cartoon* con ayuda del programa "PyMol".

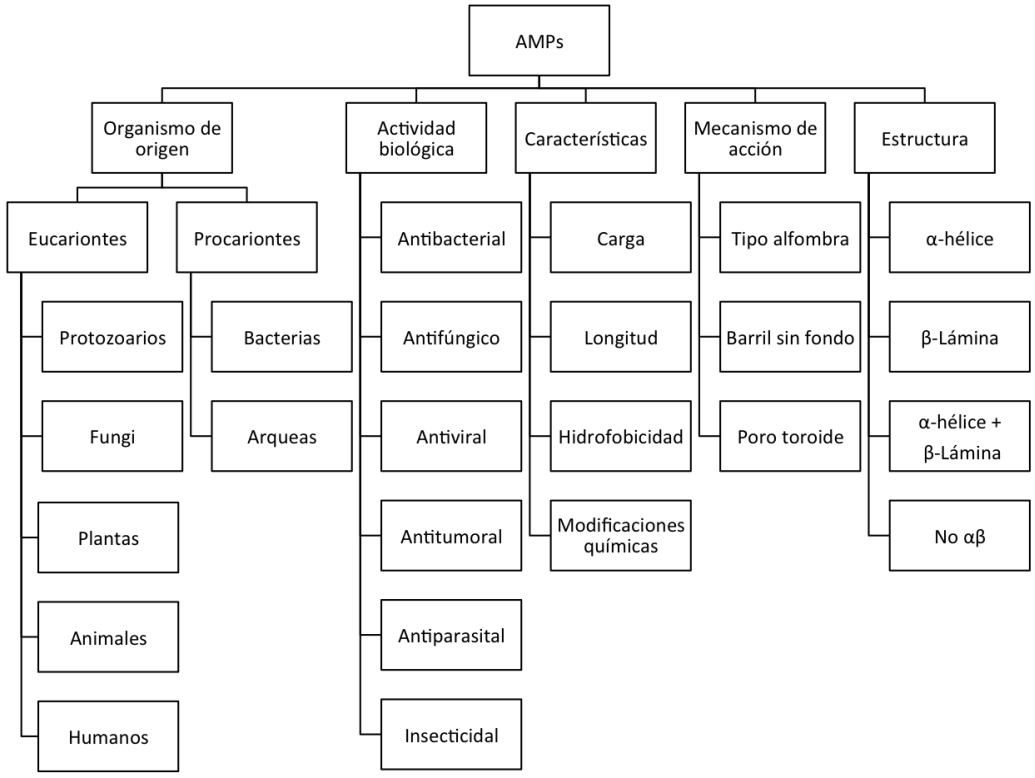


Figura 4: Clasificación de los péptidos antimicrobianos según Wang *et al.* (2010).

Wang *et al.* (2010) organiza a los AMPs según los siguientes criterios: organismo de origen, actividad biológica, propiedades fisicoquímicas, mecanismo de acción y estructura secundaria (ver Figura 4).

Organismo de origen

Adoptando la clasificación propuesta por Whittaker (1969), los AMPs se organizan de acuerdo al organismo de origen en cinco reinos de la naturaleza: 1) Procariota (bacterias y arqueas); 2) Protista (protozoarios), aquí se encuentran los organismos eucariontes unicelulares; 3) Fungi (hongos); 4) Plantae (plantas); 5) Animalia (animales). Es importante resaltar que cada reino se puede dividir en más niveles taxonómicos, tales como: *phylum*, clase, orden, familia, género y especie. Por ejemplo, el péptido *Human Lactoferricin* se clasifica de acuerdo al organismo de origen en el reino de los animales, clase de los mamíferos, orden de los primates, familia de los homínidos y especie *Homo Sapiens*.

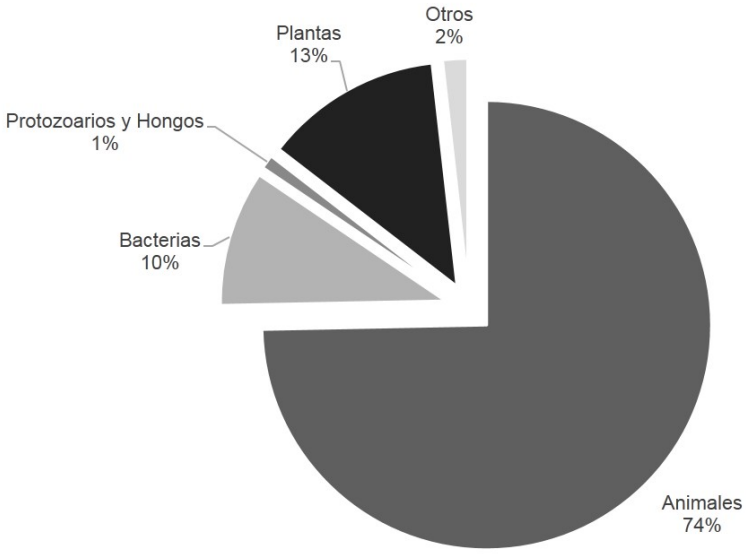


Figura 5: Organismo origen de AMPs considerando un total de 2408 péptidos de la base de datos APD (Wang *et al.*, 2009).

Muchos de los AMPs naturales han sido aislados y coleccionados en la base de datos *Antimicrobial Peptides Database* (APD, por sus siglas en inglés) (Wang *et al.*, 2009). De un total de 1920 AMPs que almacena el APD, el 74 % provienen de los animales (ver Figura 5).

Actividad biológica

La mayoría de los AMPs son eficaces en contra de una amplia gama de organismos que incluyen bacterias, hongos, virus, insectos, además de tumores y espermatozoides. También existen péptidos que tienen un reducido espectro de actividad (*e.g.,* el AMP bacteriano). Tomando en consideración la capacidad que tienen los AMP para impedir o matar a un tipo de organismo, los podemos clasificar en ocho tipos de actividad biológica: antibacteriano, antifúngico, antiviral, anticancerígeno o antitumoral, antiparasitario, insecticida, espermicida y anti VIH (Wang *et al.*, 2010).

Hasta el momento, la actividad biológica que sobresale entre los AMPs aislados es la actividad antibacteriana, seguida de la actividad antifúngica y antiviral. La mayoría de los

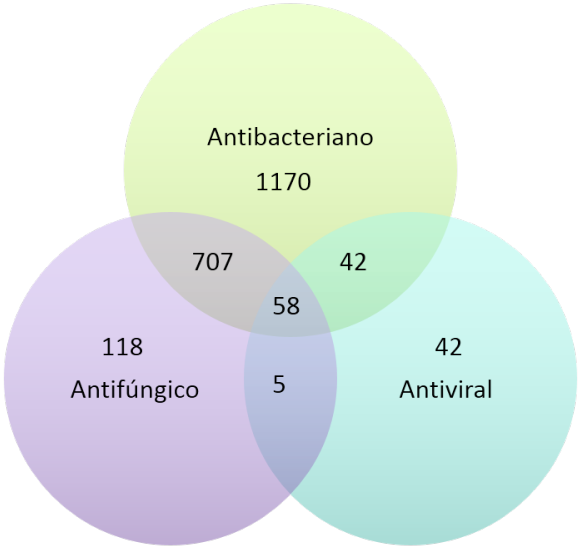


Figura 6: Actividades biológicas más abundantes en AMPs naturales en la base de datos APD (Wang *et al.*, 2009).

péptidos naturales tienen más de una actividad biológica (Figura 6). En la base de datos APD muchos péptidos comparten más de una actividad biológica. Por ejemplo, los péptidos que tienen actividad antiviral y antibacterial son 100 y de estos 58 son antifúngicos.

Características de los AMPs

Otra clasificación de los AMPs es con base en características bioquímicas o físicas, tales como: carga neta, longitud, contenido de residuos hidrófobos, entre otros.

Carga neta. Los AMP se dividen con base en la propiedad fisicoquímica de la carga neta en tres categorías: aniónicos (son aquellos AMPs con carga neta negativa); neutrales (AMPs con carga neta igual a cero); catiónicos (AMPs con carga neta positiva). Éste último es el más abundante en las principales bases de datos (CAMP, APD). Por ejemplo en APD el 88.6 % de los AMPs son catiónicos con una carga neta positiva de 4.4 en promedio (Wang *et al.*, 2010; Aoki y Ueda, 2013). Las cargas de los aminoácidos se muestran en la Sección A.3.

Longitud. Los AMPs se caracterizan por tener una longitud menor a 100 residuos, en donde la mayoría de estos péptidos tiene un tamaño de 20 a 50 residuos de longitud (Jenssen *et al.*, 2006).

Hidrofobicidad. Los AMPs se caracterizan por tener un alto porcentaje de residuos hidrófobos en sus secuencias. Por ejemplo, la mayoría de los péptidos AMP tienen de 41 % a 50 % de residuos hidrófobos, mientras que en el caso de los AMP con actividad bacteriana tiene entre 31 % y 40 % (Wang *et al.*, 2010). Los aminoácidos hidrófobos se describen en la Sección A.1.

Estructura

Con base en la posible estructura secundaria que pueden adoptar los péptidos antimicrobianos se clasifican en: familia α, familia β, familia $\alpha + \beta$, no $\alpha\beta$ (ver Figura 7).

La familia α consiste de AMPs que adoptan una estructura secundaria α-helicoidal (ver Figura 7a). Los AMPs que pertenecen a esta familia son ricos en residuos de leucina (L), glicina (G) y lisina (K) (Wang *et al.*, 2010).

Por otra parte, los AMP que pertenecen a la familia β adoptan una estructura secundaria de lámina β u hoja plegada β (ver Figura 7b) y en sus secuencias prevalecen residuos de cisteína (C), glicina (G) y arginina (R) (Wang *et al.*, 2010; Yount y Yeaman, 2004).

En la familia $\alpha + \beta$ los AMP adoptan una estructura secundaria con regiones *alpha*-helicoidales y β láminas, donde estas regiones pueden ser intercaladas o separadas, respectivamente (ver Figura 7c).

Por último, los AMPs que adoptan una estructura ni α ni β son abundantes en residuos de triptófano (W) (ver Figura 7d) (Nguyen *et al.*, 2005).

Mecanismos de acción

Los mecanismos de acción en los AMPs para actuar en contra de los microorganismos, se dividen en: péptidos que se unen a la membrana del organismo y los no orientados a la membrana (Wang *et al.*, 2010). En general, los péptidos que se unen a la membrana basan su

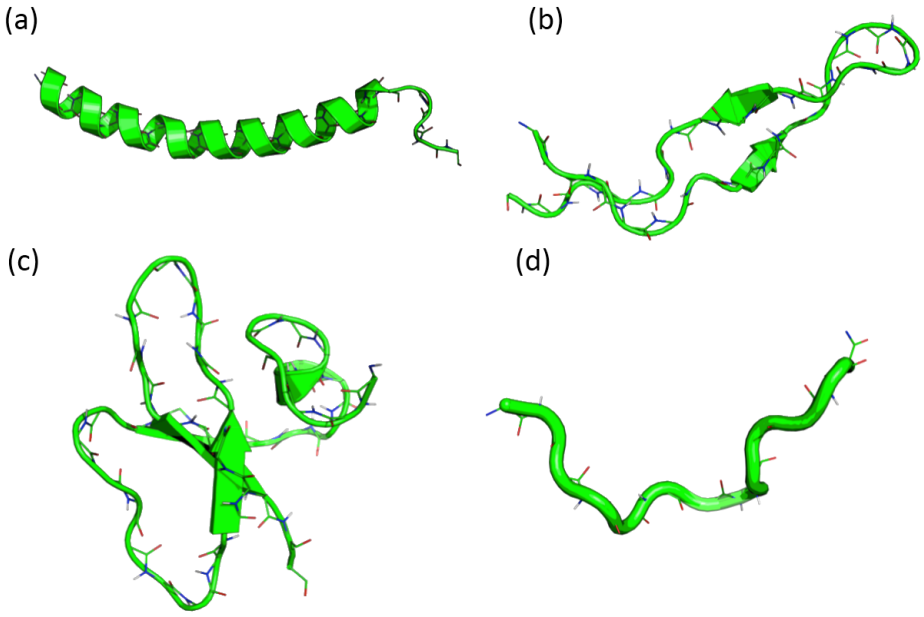


Figura 7: Ejemplos de estructuras secundarias de AMP. (a) Familia α, (*e.g.* cathelicidin LL-37; PDB 2k6O); (b) Familia β (*e.g. bovine lactoferricin* B; PDB 1LFC); (c) Familia $\alpha+\beta$ ($\beta-$*defensin2*; PDB 1FQQ); (d) no $\alpha\beta$ (*e.g. bovine indolicidin*; PDB 1G89).

mecanismo de unión en las interacciones electrostáticas; este punto de vista es soportado por la observación de muchos AMPs que conservan carga positiva y la bicapa fosfolipídica de la membrana con carga negativa, produciendo una fuerte atracción del péptido a la membrana objetivo (Figura 8).

Después de la unión péptido-membrana, se presenta una fase de conformación, donde péptidos con una estructura desordenada en el ambiente acuoso (*i.e., random coil*, ver Figura 7c) asumen una estructura anfipática α-helicoidal (Figura 7a). Por otro lado, péptidos con una estructura secundaria de β-lámina en solución acuosa (Figura 7b), mantienen la estructura al interactuar con la membrana, esto se debe a los enlaces de disulfuro (*i.e.,* enlace azufre-azufre) de la cadena principal del péptido. Con la fase de conformación inicia la introducción transversal del péptido en la bicapa lipídica mediante uno de los posibles mecanismos de acción (*i.e.,* barril sin fondo, poro toroide, modelo de alfombra). Es importante resaltar que se necesita una concentración de péptidos mínima para que se lleven a cabo los mecanismos de acción que perturbará a la célula del microorganismo objetivo (Yeaman y Yount, 2003). A continuación se describen los tres mecanismos principales de acción para la introducción

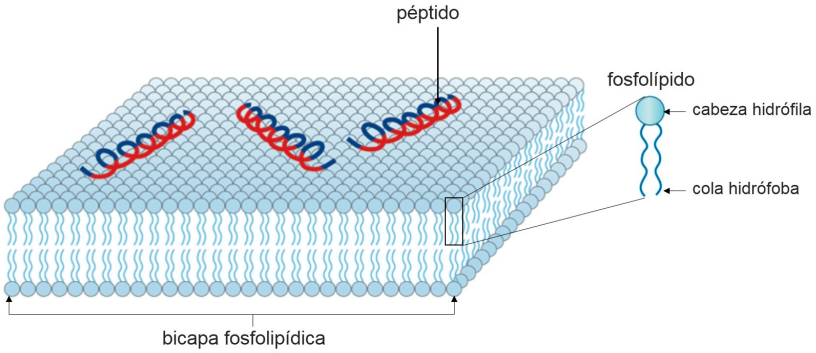


Figura 8: Interacción inicial péptido-membrana. Las cargas opuestas entre el péptido y la membrana microbiana es lo que permite la interacción. Las regiones hidrófilas del péptido se muestran en rojo y las regiones hidrófobas en azul (Brodgen, 2005).

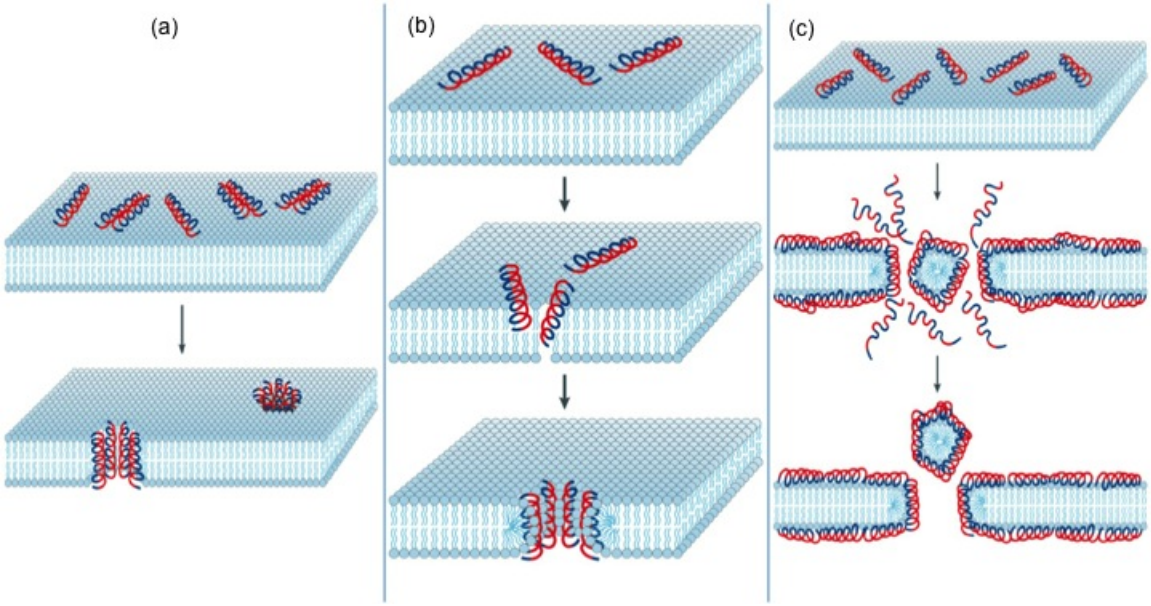


Figura 9: Mecanismos de acción para la perturbación del microorganismo objetivo. (a) Barril sin fondo. (b) Poro toroide. (c) Modelo de alfombra. (Brodgen, 2005).

transversal en la bicapa lipídica de la membrana del microorganismo.

Barril sin fondo (*Barrel-Stave*)

En este modelo, un conjunto de péptidos forman un anillo como de barril sin fondo alrededor de un poro acuoso. La superficie hidrófoba del péptido (color azul en los péptidos, Figura 9a) está en dirección a las regiones lipídicas de la membrana, mientras que la superficie hidrófila (región roja en los péptidos, Figura 9a) forma el revestimiento del poro (Yeaman y Yount, 2003; Zhao, 2003).

Poro toroide (*Toroid Pore*)

En este modelo se forma un poro acuoso, a diferencia del barril sin fondo el poro está compuesto por péptidos intercalados con los lípidos de la membrana. Los péptidos unidos se insertan en la membrana, donde la superficie hidrófoba de los péptidos desplaza el grupo de cabezas polares creando una brecha que induce a la deformación de la membrana curveándola (ver Figura 9b) (Yeaman y Yount, 2003).

Modelo de alfombra (*Carpet Mechanism*)

Los péptidos se unen a la superficie de la membrana celular del microorganismo objetivo, donde la membrana es cubierta por un conglomerado de péptidos como si fueran un tapiz. Después que la concentración de péptidos es alcanzada, los péptidos causan rompimiento en la membrana (Yeaman y Yount, 2003; Zhao, 2003). Este modelo sugiere que la membrana se rompe en pedazos a través de la formación de micelas (ver Figura 9c).

Las alteraciones que sufre la membrana por los mecanismo de acción (*e.g.,* adelgazamiento o la formación de poros) ocasionan la ruptura de la membrana plasmática y por tanto la pérdida de contendido celular, provocando la muerte del microorganismo (Yeaman y Yount, 2003).

2.1.5. AMP naturales y las desventajas que limitan su uso terapéutico

A pesar de las propiedades atractivas con las que cuentan los AMPs naturales, ellos poseen varias desventajas que impiden su uso como agente terapéutico. Estas desventajas se describen a continuación.

- **Toxicidad:** los AMPs pueden interactuar directamente con las células huésped y causarles lisis (*i.e.,* ruptura de la membrana celular) (Aoki y Ueda, 2013).
- **La degradación por proteasas:** la poca duración de los péptidos AMP *in vivo* probablemente es debido a la degradación que sufren por proteasas (*i.e.,* enzimas que rompen los enlaces peptídicos de las proteínas) provenientes del microorganismo huésped (Aoki y Ueda, 2013).
- **Amplio espectro:** en lo que se refiere a péptidos como antibióticos, un amplio espectro podría dañar la microbiota autóctona encargada de proveer la colonización en algunas zonas (*e.g.,* piel, tubo digestivo) para impedir que organismos patógenos se reproduzcan. Por tanto, el amplio espectro en péptidos incrementa el riesgo de enfermedades como la diarrea y otras infecciones que pueden resultar fatales (Aoki y Ueda, 2013).
- **Alto costo de producción:** los péptidos pueden costar entre 100 y 600 dólares por gramo. A consecuencia de los altos costos existen limitaciones tanto en el número de pruebas como en las variantes que se pueden realizar en los péptidos (Hancock y Sahl, 2006).

Por lo anterior, se abre la oportunidad de investigación al diseño de péptidos antimicrobianos, teniendo como objetivo crear o identificar secuencias de AMPs costo-efectivas, que tenga una alta actividad antimicrobiana sin exhibir altos niveles de toxicidad, y además cuenten con un perfil deseado de selectividad y se reduzca la proteólisis (Fjell *et al.*, 2012).

2.1.6. Bases de datos de AMPs

Las bases de datos de péptidos antimicrobianos son una herramienta útil para el registro y la administración de un gran número de secuencias de AMPs. Considerando el propósito que tienen las bases de datos, podemos clasificarlas en dos categorías: (a) **bases de datos**

Tabla 2: Catálogo de las principales bases de datos de AMPs de propósito general.

Base de datos	**Año**	**Número de AMPs**	**Tipo de AMPs**	**Sitio web**
APD	2009	2408	AMPs naturales	http://aps.unmc.edu/
CAMP	2010	5040	AMPs naturales y sintéticos	http://www.camp.bicnirrh.res.in
DAMPD	2011	1232	AMPs naturales y sintéticos	http://apps.sanbi.ac.za/dampd/
YADAMP	2012	2525	AMPs naturales y sintéticos	http://www.yadamp.unisa.it
Hemolytik	2013	3000	AMPs naturales y sintéticos	http://crdd.osdd.net/raghava/hemolytik/

Tabla 3: Catálogo de las principales bases de datos de AMPs especializadas.

Base de datos	**Año**	**Número de AMPs**	**Tipo de AMPs**	**Sitio web**
Peptaibol	2003	317	AMPs de hongos	http://peptaibol.cryst.bbk.ac.uk/
PenBase	2004	850	AMPs camarón	http://www.penbase.immunaqua.com/
Defensins	2007	363	AMPs defensinas	http://defensins.bii.a-star.edu.sg/
PhytAMP	2009	271	AMPs de plantas	http://phytamp.pfba-lab-tun.org/
Bactibase	2010	177	AMPs de bacterias (*Bacteriocin*)	http://bactibase.pfba-lab-tun.org/

de propósito general, contienen secuencias de AMPs de todo tipo (Tabla 2); (b) **bases de datos especializadas**, almacenan AMPs con propiedades comunes, tales como organismo origen, función, entre otras. (Tabla 3).

La información de los péptidos que reportan las bases de datos es muy variada, debido a que no existe una estandarización de los campos que deben contener para registrar AMPs. Sin embargo, las bases de datos comparten comúnmente los siguientes campos: nombre del péptido, secuencia primaria, actividad biológica.

Con el objetivo de mostrar los campos de mayor importancia para la presente investigación, tomamos con referencia CAMP, una de las bases de datos más importantes para la recolección de todo tipo de AMPs.

CAMP (*Collection of Anti-Microbial Peptides*)

CAMP es la base de datos de propósito general más grande con una colección de 5040 AMPs, dividendo las secuencias en experimentalmente validadas (2438 AMPs) y predichas (2438 AMPs). CAMP captura la siguiente información: nombre de la secuencia, familia del AMP, organismo origen, organismo objetivo y actividad biológica. A continuación se describen los campos de mayor importancia para la presente investigación.

Nombre del péptido: El nombre que se asigna a los AMPs es conforme a las propiedades que posee y/o organismo de donde proviene (Wang *et al.*, 2010). Por ejemplo, el AMP de nombre *Human beta defensin 2*, se le asigna este nombre porque proviene de los humanos, tiene una estructura secundaria β-lámina y el rol en el proceso biológico que desempeña es el de defender al organismo huésped (humano) cuando reconocen un componente potencialmente patógeno (ver Figura 10a).

Organismo objetivo: CAMP aparte de indicar el tipo de actividad microbicia (*i.e.,* antibacteriano, antifúngico, antiparasitario, antiviral, entre otras.) también especifica la especie a la que ataca y la concentración mínima inhibitoria (MIC) (ver Figura 10b). MIC es la menor concentración de AMPs para impedir el crecimiento de un microorganismo después de su incubación. Por lo general, MIC se mide en micromoles por mililitro ($\mu M/ml$) (Andrews, 2001).

Ontología Génica (*Gene Ontology*): Dentro de la información de los AMPs se incluyen anotaciones de Ontología Génica (GO). GO provee de un vocabulario estructurado para describir a los productos génicos (*e.g.,* péptidos) en términos del rol que desempeñan en el proceso biológico, la función molecular y su localización en la célula (ver Figura 10c) (Ashburner *et al.*, 2000).

(a)	**Title :**	Human beta defensin 2
	GenInfo Identifier :	3818537
	Source :	Homo sapiens [Human]
	Taxonomy :	Animalia, Mammals
	NCBI Taxonomy :	9606
	UniProt:	O15263
	PDB:	1E4Q, 1FD3, 1FD4, 1FQQ
	Structure Database :	CAMPST498, CAMPST20, CAMPST21, CAMPST508
	PubMed :	10603376
	Length :	39
	Activity :	Antibacterial
	Gram Nature :	Gram-ve
(b)	**Target :**	E.coli D31 (MIC = 62 microg/ml)
	Validated :	Experimentally Validated
	Pfam :	PF00711 : Defensin_beta (Beta defensin)
	InterPro :	IPR001855 : Defensin_beta-typ. IPR006080 : Defensin_beta/neutrophil.
	AMP Family :	Defensin

(c) **Gene Ontology :**

GO ID	Ontology	Definition	Evidence
GO:0005576	Cellular Component	Extracellular region	TAS
GO:0005796	Cellular Component	Golgi lumen	TAS
GO:0006935	Biological Process	Chemotaxis	TAS
GO:0042742	Biological Process	Defense response to bacterium	IEA
GO:0007186	Biological Process	G-protein coupled receptor signaling pathway	TAS
GO:0045087	Biological Process	Innate immune response	TAS

Sequence :	DPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKPP

Figura 10: Información del AMPs *Human beta defensin 2* recuperado de la base de datos CAMP. (a) Nombre del AMPs. (b) Organismo objetivo. (c) Ontología génica.

2.2. Conceptos computacionales

2.2.1. Diseño racional de AMPs

Diseñar péptidos con actividad antimicrobiana directa no es una tarea trivial, principalmente debido a la diversidad que tienen los péptidos tanto en sus secuencias como en sus estructuras. Por lo anterior, los péptidos antimicrobianos no pueden ser explicados por un simple patrón, en cambio pueden explicarse en términos de combinaciones de propiedades fisicoquímicas (*e.g.* longitud, composición de aminoácidos, anfipaticidad, entre otras).

El proceso para diseñar y descubrir nuevos AMPs inicia con la identificación de péptidos con actividad antimicrobiana, para esto se utilizan dos técnicas: la **química combinatoria**, para crear una gran cantidad de péptidos y el **cribado de alto desempeño** (HTS, por sus siglas en inglés de *High Throughput Screening*) para detectar la actividad del péptido.

Por un lado, la química combinatoria permite la síntesis rápida de un gran número de péptidos con atributos comunes que se le conoce como librería combinatoria de péptidos. Para la síntesis de los péptidos, por lo general se utiliza el método de Merrifield de fase sólida (Merrifield *et al.*, 1995).

El cribado de alto desempeño (HTS) se utiliza para probar millones de péptidos rápidamente de manera paralela y automatizada; aunque para los laboratorios pequeños podría significar probar manualmente unos miles (Wimley, 2010). Para determinar la actividad antimicrobiana del péptido, se emplea un plato con nutrientes (*i.e.,* una placa de agar) idóneo para el crecimiento de las bacterias, además se le introduce una cantidad en micromoles (μM) del péptido de interés. Si el plato se encuentra transparente, entonces significa que el péptido impidió el crecimiento de la bacteria, por lo tanto el péptido tiene actividad antimicrobiana; de otro modo, si la bacteria alcanza la fase estacionaria, el péptido es considerado con pobre o de nula actividad. En la Figura 11 se muestra un ejemplo de HTS con 96 platos, cada plato contiene una concentración de 2.2 μM de un péptido de la librería combinatoria y un caldo de nutrientes idóneo para el crecimiento de 10^3 bacterias de *E. Coli.* (Wimley, 2010).

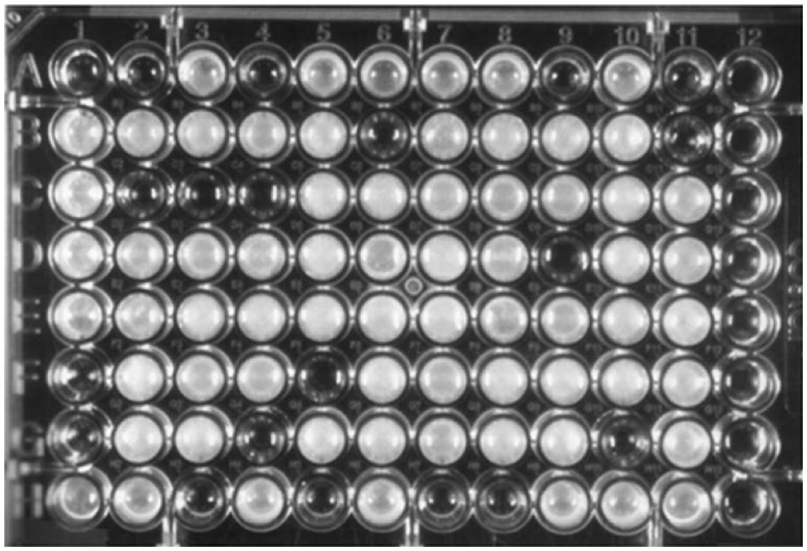


Figura 11: Cribado de alto desempeño (HTS). Cada plato contiene una concentración de 2.2 μM de un péptidos de la librería combinatoria y un caldo de nutrientes idóneo para el crecimiento de 10^3 bacterias de *E. Coli.*. Las placas opacas indican que las bacterias de *Escherichia coli* alcanzaron la fase estacionaria de crecimiento; los platos transparentes indican que el péptido inhibió el crecimiento del microbio (Wimley, 2010).

La evaluación en HTS depende de la generación de los péptidos con la técnica de química combinatoria. Por consiguiente, se debe tomar en consideración un balance entre el tamaño de la librería y el número de péptidos a evaluar. En la actualidad existen algunos retos en el uso de las técnicas que pueden impedir el balance, por ejemplo:

- El tamaño de las librerías combinatorias se incrementa muy rápido, así como su complejidad (Wimley, 2010; Fjell *et al.*, 2012). Por ejemplo, si consideramos todas las posibles combinaciones de péptidos con longitud de 6 residuos, tendríamos un total de $20^6 = 64.000.000$ secuencias (20 es número de aminoácidos que existen). Por lo tanto, crear librerías de péptidos de manera exhaustiva es prohibitivamente costoso y difícil de manejar en laboratorio cuando la longitud de los péptidos es muy grande.

- En HTS el reto es detectar sólo un pequeño subconjunto de péptidos con alta actividad en una cantidad de tiempo y esfuerzo razonable. Por otra parte, con una mayor capacidad de procesamiento e identificación, se pueden crear librerías más complejas (Wimley, 2010).

Cuando las técnicas biológicas para la síntesis y prueba exhaustiva de péptidos son prohibitivamente costosas, utilizar métodos computacionales resulta de gran ayuda. Por ejemplo, en el proceso de identificación de péptidos con actividad antimicrobiana, los métodos computacionales pueden ayudar a desechar péptidos con una actividad pobre o nula antes de evaluarse de manera experimental (Fjell *et al.*, 2012).

2.2.2. Diseño *in silico* de AMPs

Las investigaciones en el diseño *in silico* de AMPs toman una gran cantidad de secuencias e información almacenada en las bases de datos (ver Sección 2.1.6) para generar conocimiento útil para el diseño de nuevos péptidos. De acuerdo con Fjell *et al.* (2012), existen tres líneas predominantes en investigación para el diseño de AMPs: métodos basados en plantillas (*Template-based studies*), modelado biofísico, y el cribado virtual (*virtual screening*).

El **método basado en plantillas** consiste en la modificación de AMPs conocidos con la finalidad de aumentar o disminuir algunas propiedades (*e.g.,* reducir el tamaño, disminuir la toxicidad, aumentar la actividad antimicrobiana). La principal pregunta a responder en esta línea de investigación es ¿cuáles son los residuos o posiciones relevantes en los péptidos para alterar la actividad? Por lo general, en este método las secuencias son tratadas como palabras a las que se le aplican reglas gramaticales (*e.g.,* frecuencia de los aminoácidos, frecuencia de motivos) con el objetivo de identificar patrones (Fjell *et al.*, 2008; Loose *et al.*, 2006; Yount y Yeaman, 2004).

El **modelado biofísico** emplea las técnicas de dinámica molecular y perturbación de la energía libre para entender la interacción del péptido-membrana (Fjell *et al.*, 2012). La técnica de dinámica molecular se utiliza para calcular la conformación y movimientos físicos del péptido al interaccionar con la membrana del microorganismo en un determinado periodo de tiempo, describiendo las interacciones de los átomos a través de campos de fuerza intra\inter moleculares (Maccari *et al.*, 2013).

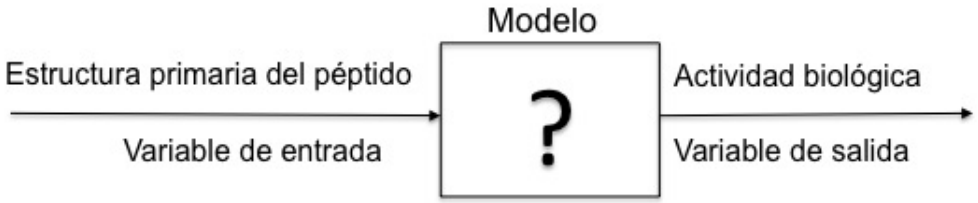


Figura 12: Problema de modelado para la predicción de la actividad biológica.

Debido al alto costo computacional de la simulación, sólo se consideran algunos átomos de la estructura del péptido, una porción de la membrana y el solvente (Fjell *et al.*, 2012).

El **cribado virtual** es una herramienta de filtrado que utiliza varias técnicas computacionales con el objetivo de reducir la librería combinatoria de péptidos, eliminando secuencias con propiedades no deseables. El cribado virtual es la línea de investigación en la que se enfoca el presente trabajo de tesis, razón por la que abordaremos más del tema en la siguiente sección.

2.2.3. Cribado Virtual (*Virtual Screening*)

El cribado virtual asiste en la examinación de grandes librerías combinatorias de péptidos con el objetivo de eliminar secuencias no deseables en etapas tempranas del diseño.

El problema más importante en el cribado virtual es la predicción de la actividad, este problema lo podemos definir como: dado un conjunto de péptidos y actividades biológicas conocidas (*e.g.,* AMP y no AMP) encontrar un modelo que asigne como salida la actividad correcta para cada péptido de entrada (ver Figura 12).

El método más usado para encontrar el modelo es conocido como **QSAR** (*Quantitative Structure-Activity Relationship*), debido a que relaciona las características estructurales químicas del péptido, descritas por los descriptores moleculares (*e.g.,* carga, hidrofobicidad) con su correspondiente actividad biológica (Goodarzi *et al.*, 2012; Fjell *et al.*, 2012). Para asociar la información del péptido (*i.e.,* descriptores moleculares) con la actividad biológica, un modelo estadístico se construye mediante algoritmos de aprendizaje de máquina. Los

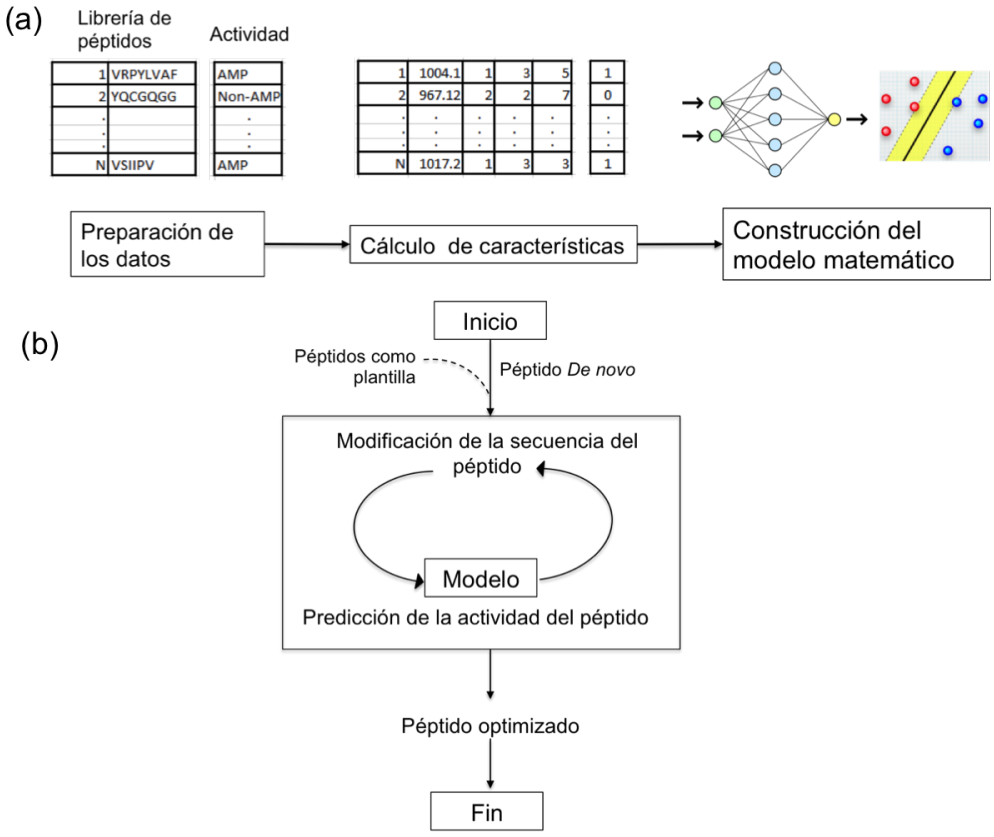


Figura 13: Diagrama general del diseño de AMP *in sillico*. (a) Construcción del modelo para la predicción de actividad antimicrobial. (b) Esquema para la generación de nuevos AMPs.

aspectos relevantes para construir el modelo estadístico son: 1) la preparación de los datos, representando a los péptidos activos (AMPs) y no activos (no AMPs); 2) la construcción del modelo estadístico que permita identificar la actividad del péptido; 3) la aplicación del modelo para el diseño de nuevos AMPs (ver Figura 13).

Preparación y representación de los péptidos

Para la construcción de un buen modelo el paso más delicado es la preparación de los datos. En esta etapa es donde se recolectan un conjunto de péptidos con una actividad biológica deseada y péptidos que carecen de la actividad (*e.g.,* AMPs y no AMPs). Al conjunto de péptidos con actividad biológica deseada, se le conoce como casos positivos y a los péptidos que carecen de la actividad reciben el nombre de casos negativos.

Para obtener los casos positivos se utilizan las bases de datos señaladas en la Sección 2.1.6, para los casos negativos se emplean las bases de datos de propósito general (Apweiler *et al.*, 2004) o la generación aleatoria de secuencias.

Después de obtener los péptidos, el siguiente paso es representarlos en términos de descriptores moleculares. **Los descriptores moleculares** son el resultado de un procedimiento lógico y matemático que transforma la información química del péptido en un número útil (Todeschini y Consonni, 2000).

Los descriptores moleculares para estudios de péptidos antimicrobianos son clasificados en dos categorías dependiendo de cómo se obtuvieron: **descriptores empíricos**, se obtienen a partir de información medida en ensayos biológicos; **descriptores calculados o basados en la estructura**, son descriptores moleculares teóricos calculados a partir de una representación molecular (Hilpert *et al.*, 2008).

En el presente trabajo nos enfocamos en los descriptores moleculares basados en la estructura. La información estructural es transformada en una representación numérica mediante un procedimiento de cómputo. Los descriptores se clasifican en diferentes niveles de dimensionalidad dependiendo de la estructura molecular que se necesite, los niveles van desde la dimensión cero hasta la cuatro (Helguera *et al.*, 2008).

- **Dimensión cero (0D):** este tipo de descriptores contienen información derivada de la frecuencia de los residuos en el péptido. Algunos ejemplos de los descriptores 0D son el número de aminoácidos hidrófobos, longitud del péptido, carga neta, peso molecular, número de átomos.
- **Dimensión uno (1D):** contienen información acerca de fragmentos del péptido, sin embargo, son independientes de información de la estructura de la molécula (*i.e.,* sólo se utiliza la estructura primaria del péptido). Ejemplos de los descriptores 1D son la distancia entre dos residuos de triptófano (Trp) y el momento hidrofóbico (Hilpert *et al.*, 2008).

- **Dimensión dos (2D):** se les conoce como grafos invariantes o descriptores topológicos y contienen información derivada de un grafo molecular. En el grafo molecular sólo se representa la estructura atómica del péptido pero es independiente de la conformación que adopta éste. Un ejemplo de descriptor 2D es el índice de Wiener que mide las distancias que existe entre todos los átomos del péptido (Helguera *et al.*, 2008).

- **Dimensión tres (3D):** para el cálculo de estos descriptores se necesita la estructura tridimensional del péptido. Algunos ejemplos de descriptores 3D son el volumen y área de superficie.

Con las coordenadas geométricas de los átomos del péptido (estructura terciaria) podemos calcular desde los descriptores 0D hasta los 3D, sin embargo, la mayoría de los péptidos que se encuentran en las bases de datos, tienen una representación solo de estructura primaria. Predecir la estructura terciaria a partir de la secuencia del péptido (*Protein Sequence-Structure Alignment*) es un problema NP-difícil (Lathrop *et al.*, 1998).

Actualmente existen softwares comerciales y gratuitos que ofrecen el cálculo de miles de descriptores sobre moléculas naturales y muchos de ellos son personalizables de acuerdo con el tipo de molécula (e.g. AMPs) (Fjell *et al.*, 2012). En la Tabla 4 se muestra la lista de softwares para el cálculo de descriptores que se han utilizado para AMPs, cada software se encuentra clasificado por tipo de licencia en comercial y libre.

Construcción del modelo

Una vez que los descriptores moleculares se calculan en los péptidos, el siguiente paso es usar estas medidas para predecir otra propiedad de interés (e.g., actividad antimicrobiana, toxicidad) de manera no trivial.

Los modelos para predecir AMPs se organizan de acuerdo con el tipo de variable de salida en dos categorías: **modelos de regresión**, sirven para predecir la actividad del péptido, utilizando la actividad como una variable continua (*e.g.,* predecir la mínima concentración inhibitoria (MIC)); **clasificación**, sirve para predecir la actividad de un péptido como activa o inactiva (Hilpert *et al.*, 2008; Duda *et al.*, 2000), es decir, la variable de salida es binaria.

Tabla 4: Lista de softwares para el cálculo de descriptores moleculares.

Nombre del paquete	Número de descriptores	Tipo de licencia	sitio web
Dragon 6	4885 descriptores (0D,1D,2D,3D)	Comercial	http://www.talete.mi.it/
ADMEWORKS ModelBuilder	400 descriptores (2D y 3D)	Comercial	http://www.fqs.pl/
MOE	Miles de descriptores (1D,2D,3D)	Comercial	http://www.chemcomp.com/MOE-Cheminformatics_and_QSAR.htm
PEDES	32 descriptores (0D y 1D)	Libre	
PaDEL-Descriptor	1875 descriptores (1D,2D y 3D)	Libre	http://padel.nus.edu.sg/

Para propósito del presente trabajo sólo nos ocuparemos de los modelos de clasificación. En los modelos de clasificación, los enfoques de aprendizaje de máquina más utilizados son: las redes neuronales artificiales (ANN por sus siglas en inglés de *Artificial Neural Network*) y las máquinas de soporte vectorial (SVM por sus siglas en inglés *Support Vector Machine*) debido al poder predictivo que tienen.

Red neuronal artificial (ANN)

ANN es un modelo matemático basado en algunas propiedades biológicas de las redes neuronales. La red consiste en tres capas: un conjunto de nodos de entrada conectados en forma de red con los nodos de la capa oculta. Cada nodo de la capa oculta, toma los valores de los nodos de entrada y los transforma en una suma (Hilpert *et al.*, 2008; Duda *et al.*, 2000). El nodo de salida toma la suma de cada nodo de la capa oculta y lo transforma en un valor de salida entre 0 y 1 (ver Figura 14).

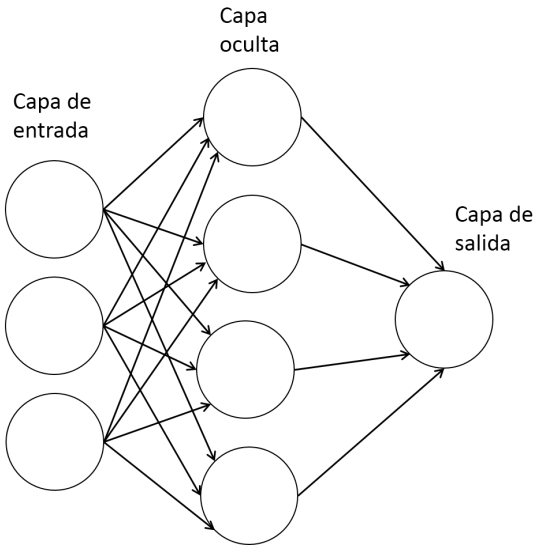


Figura 14: Estructura de una red neuronal artificial.

Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (SVM) son un sistema de aprendizaje que usualmente se utiliza para clasificar elementos de dos clases. Las SVM aprenden de un conjunto de datos de entrenamiento para hacer predicciones de nuevos elementos.

Para lograr la clasificación de los elementos en sus respectivas clases, las SVM tienen que encontrar el hiperplano óptimo que separe a las dos clases (ver Figura 15). Se dice que un hiperplano es óptimo si la distancia entre los elementos más cercanos al hiperplano de ambas clases es maximal. A los patrones más cercanos se les conoce como vectores de soporte.

Para el entrenamiento de la SVM se supone que cada dato es representado por un vector, denotado por x_i que tiene un conjunto de n características como codificación del elemento i, y una etiqueta y_i que indica la clase a la que pertenece. Por ejemplo, la Figura 15 muestra la representación del péptido i en un vector $x_i = (8.08, 0.17)$ con las características de carga $(Z(pH7))$ e hidrofobicidad (H_k), además el péptido i pertenece a la clase de los AMP representado por $y_i = 1$.

Para predecir nuevos elementos las SVM utilizan el hiperplano como una regla simple de clasificación: todos los elementos que estén arriba del hiperplano son clasificados como

miembros de una clase, caso contrario son clasificados como miembro de la otra clase. Por ejemplo, en la Figura 15 los elementos que están arriba del hiperplano son etiquetados como 1 (*i.e.,* elementos de la clase AMP) y los elementos que están abajo son etiquetados como -1 (*i.e.,* elementos de la clase NoAMP).

Formulación matemática de las SVM

Dado un conjunto de datos de entrenamiento etiquetados $\mathcal{D} = \cup_{i=1}^{p}\{(\mathbf{x_i}, y_i)|\mathbf{x_i} \in \Re^d, y_i = \pm 1\}$ deseamos encontrar el hiperplano que maximice la distancia entre los hiperplanos que pasan por los vectores de soporte. La ecuación del hiperplano es

$$w^t x + b = 0 \tag{3}$$

donde w es un vector de pesos que indica la orientación del hiperplano, x es un punto localizado en el hiperplano y b (*bias)* es la distancia que existe entre el origen y el hiperplano. Los hiperplanos que pasan por los vectores de soporte son llamados hiperplanos canónicos, los cuales son $w^t x + b = -1$ y $w^t x + b = +1$. La distancia que existe entre los dos hiperplanos es igual a $2/ \parallel w \parallel$, como la mitad de la distancia entre dos hiperplanos canónicos es el margen tenemos que $\gamma = 1/ \parallel w \parallel$. Por lo tanto maximizar el margen es equivalente a minimizar:

$$\frac{1}{2} \parallel w \parallel \tag{4}$$

sujeto a la restricción:

$$y_i(w^t \mathbf{x_i} + b) \geq 1 \; i = 1, \ldots, p \tag{5}$$

La restricción exige que todos los objetos sean clasificados correctamente. En nuestro caso:

$$w^t \mathbf{x_i} + b \leq -1 \text{ si } y_i = -1$$

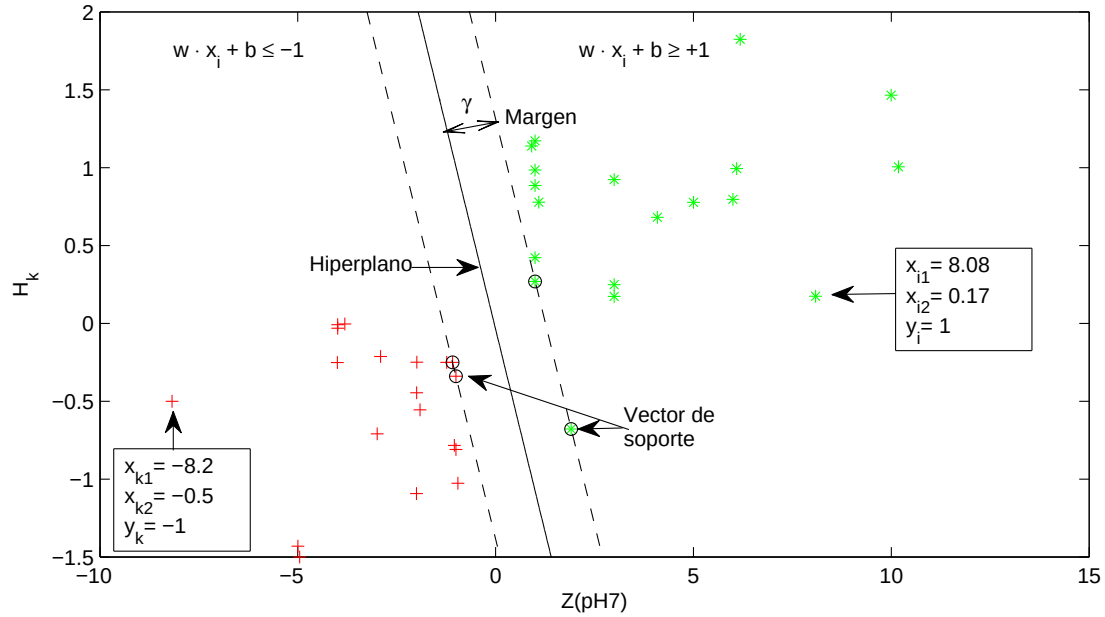


Figura 15: SVM consiste en encontrar el hiperplano óptimo, es decir el hiperplano con la distancia máxima entre los patrones más cercanos (vectores de soporte).

Tabla 5: Matriz de confusión, contiene información acerca de la predicción del clasificador y el valor observado en los datos.

		Predicho	
		Negativo	Positivo
Actual	Casos negativos	TN	FP
	Casos positivos	FN	TP

$$w^t \mathbf{x_i} + b \geq +1 \text{ si } y_i = +1.$$

Medidas de calidad para evaluar los métodos de aprendizaje de máquina

Para evaluar la calidad de los modelos se han propuesto varias medidas, la mayoría usan en esencia la comparación entre la predicción del clasificador y el valor observado en los datos. Cuando el modelo acierta en la etiqueta de un elemento que pertenece a los casos positivos se le conoce como verdadero positivo (TP), sin embargo cuando no lo reconoce tiene un falso negativo (FN). De otra manera, cuando el clasificador se equivoca en la predicción de un elemento que pertenece a los casos negativos se le conoce como falso positivo (FP) y cuando no se equivoca se tiene un verdadero negativo (TN) (ver Tabla 5).

Exactitud (*Accuracy*)

La exactitud es una medida que nos dice el número de predicciones que son correctas sin importar la clase a la que pertenecen los elementos (Lata *et al.*, 2007). La exactitud está definida por la siguiente ecuación:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} 100\,\% \tag{6}$$

Los valores de ACC van desde 0 a 100 porciento, valores cercanos al 100 % indican un mejor desempeño del clasificador.

Coeficiente de Correlación de Matthews (MCC)

El coeficiente de correlación de Matthews (MCC) es usado para evaluar el desempeño de un clasificador binario. MCC toma en cuenta los valores de TP, FP, FN, TN y a diferencia del ACC es considerado como una medida balanceada que puede ser usada aún cuando las clases tengan tamaños muy diferentes. El MCC es definido como:

$$MCC = \frac{(TP \text{ x } TN) - (FN \text{ x } FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{7}$$

Los valores de MCC van de -1 a 1, valores más cercano al 1 indican mejor desempeño del clasificador.

Sensibilidad y Especificidad

La **sensibilidad** indica la fracción de los casos positivos que son predichos correctamente por el clasificador (Lata *et al.*, 2007). La sensibilidad está dada por la siguiente ecuación:

$$Sens = \frac{TP}{TP + FN} \tag{8}$$

La **especificidad** denota la fracción de los casos negativos que son predichos correcta-

mente (Lata *et al.*, 2010). La especificidad está dada por la siguiente ecuación:

$$Espec = \frac{TN}{TN + FP} \tag{9}$$

Los valores de la sensibilidad y especificidad van desde 0 a 1. Valores más cercanos al 1 indican mejor desempeño del clasificador con respecto a los casos positivos o negativos, respectivamente.

Trabajo previo en aprendizaje de máquina para la clasificación de AMP

Con el objetivo de acelerar el proceso racional en el descubrimiento de nuevos AMPs se han propuesto varios algoritmos de aprendizaje de máquina. Los métodos propuestos emplean los AMPs como casos positivos sin distinguir el tipo de actividad antimicrobiana (*i.e.,* antibacterial, antiviral, antifúngico). Por otro lado, los casos negativos están conformados por péptidos que no exhiben la actividad antimicrobiana, sin embargo, para este tipo de péptido no existe una base de datos.

Fjell *et al.* (2009) y Cherkasov *et al.* (2008) crearon un sistema de votación basado en 30 redes neuronales artificiales (ANNs) para la identificación de AMPs que combaten las superbacterias resistentes a múltiples antibióticos (*e.g., Staphylococcus aureus* resistente a la meticilina). Cada ANN da como salida un 1 si el péptido tiene una alta actividad antimicrobiana y un 0 en otro caso. Para entrenar la red se utilizó un conjunto de 1433 péptidos de longitud 9 generados aleatoriamente, eligiendo al 5 % de los mejores péptidos con respecto a la concentración inhibitoria (IC50) como casos positivos y el otro 95 % como casos negativos. Para la prueba se utilizaron 99577 péptidos logrando el algoritmo un desempeño de 94 % en precisión y un MCC de 0.88 (ver Tabla 6).

Torrent *et al.* (2011) presentan ANN y SVM como dos enfoques de aprendizaje de máquina basados en ocho características fisicoquímicas medibles en AMPs para la identificación de péptidos activos. Los datos para realizar las pruebas y entrenamiento se extrajeron de las

bases de datos CAMP y Uniprot para obtener los casos positivos y negativos, respectivamente. En el caso de la ANN se utilizaron 1074 péptidos para el entrenamiento, 537 péptidos para la validación y prueba. Por otra parte, para la SVM se utilizaron 1611 péptidos para entrenamiento y 537 péptidos para prueba. La exactitud que tuvieron ambos enfoques fue de 89.4 % y 75 %, respectivamente.

ANFIS (Fernandes *et al.*, 2012) es un sistema que combina lógica difusa y ANN para la identificación de AMPs de longitud variable. Para la selección de las características fisicoquímicas utilizan una capa difusa que selecciona el par de características con mejor desempeño de acuerdo con la heurística de Jang (1996). Las características sirven como entrada para la ANN, el desempeño que tiene es de 96.7 % de precisión y un MCC de 0.94 (ver Tabla 6).

Lata *et al.* (2010) desarrollaron un método basado en SVM y en la composición de aminoácidos que presentan los péptidos con el objetivo de identificar la actividad antimicrobiana. Para los casos positivos, ellos seleccionaron aleatoriamente 999 AMPs de la base de datos APD, para los casos negativos extrajeron 999 proteínas no secretoras de SwissProt. El desempeño del algoritmo fue de 91.64 % de precisión y un MCC de 0.843 (ver Tabla 6).

Por último, Thomas *et al.* (2010) y Waghu *et al.* (2014) utilizaron un conjunto de AMPs experimentalmente validados para desarrollar una herramienta de predicción de actividad antimicrobiana basada en los métodos de aprendizaje de máquina tales como: *Random Forest* (RF), Análisis de Discriminante (DA), ANN y SVM. Los modelos de predicción tienen una precisión de 93.2 % (RF), 87.5 % (DA), 86.3 % (ANN) y 91.5 % (SVM).

Si bien todos estos trabajos obtienen resultados interesantes ninguno de ellos aborda el problema central que da origen a esta tesis, que es la selección del subconjunto de características que se necesitan para lograr una clasificación óptima.

Tabla 6: Métodos de aprendizaje de máquina para la predicción de AMPs

Método	Bases de datos		Número de descriptores	Desempeño			Referencia
	Conjunto de datos positivos	Conjunto de datos negativos		Entrenamiento	Validación	Prueba	
ANN	Generación aleatoria	Generación aleatoria	44		MCC=0.88		(Fjell *et al.*, 2009; Cherkasov *et al.*, 2008)
ANN	CAMP	Uniprot	8	MCC=0.79	MCC=0.797	MCC=0.74	(Torrent *et al.*, 2011)
SVM	CAMP	Uniprot	8		ACC=75 %		(Torrent *et al.*, 2011)
ANFIS	APD2	PDB	8		MCC=0.94		(Fernandes *et al.*, 2012)
SVM	APD	SwissProt				MCC=0.84	(Lata *et al.*, 2010)
DA	CAMP	Uniprot	64	MCC=0.75	ACC=87.5	MCC=0.74	(Thomas *et al.*, 2010)
RF	CAMP	Uniprot	64	MCC=0.82	ACC=92.5	MCC=0.84	(Waghu *et al.*, 2014)
ANN	CAMP	Uniprot	64	MCC=0.72	ACC=86.3	MCC=0.72	(Waghu *et al.*, 2014)
SVM	CAMP	Uniprot	64	MCC=0.91	ACC=91.5	MCC=0.83	(Waghu *et al.*, 2014)

Capítulo 3. Definición del problema

3.1. Introducción

Por lo general, los métodos para la predicción de AMPs usan un conjunto de péptidos con actividad conocida para generar reglas que se puedan aplicar a péptidos con actividad desconocida. Para representar cada péptido se utilizan descriptores moleculares debido a que cuantifican las propiedades fisicoquímicas de la molécula. Actualmente, el número de descriptores medibles en los péptidos se encuentra en el orden de los miles (*e.g.,* Dragon6 puede calcular 4885 descriptores), por lo que elegir los descriptores adecuados para la identificación de AMPs, se torna en una tarea difícil (Goodarzi *et al.*, 2012).

No se conoce una regla determinista que gobierne la elección de los descriptores (Yasri y Hartsough, 2001). En muchos de los modelos QSAR, la selección de los descriptores se realiza de manera empírica, es decir, se seleccionan en base a un conocimiento previo relacionado con el impacto que tiene el descriptor en la actividad del péptido (Hellberg *et al.*, 1987; Jenssen *et al.*, 2007; Wang *et al.*, 2011a). Sin embargo, en ocasiones estas características pueden ser demasiado generales y no compartidas por todos los AMPs (Fjell *et al.*, 2012). Por ejemplo, las propiedades fisicoquímicas implicadas en las funciones básicas de los AMP (*e.g.,* carga y anfipaticidad) son demasiado vagas y desafortunadamente compartidas por otro grupo de polipéptidos, tales como las histonas (Piotto *et al.*, 2012). Por lo tanto, utilizar estas características no es suficiente para crear un modelo confiable para predecir la actividad de nuevos péptidos.

Otra manera de seleccionar los descriptores es de forma automática mediante los métodos de selección de características (Fjell *et al.*, 2012). Cabe recalcar que elegir los descriptores adecuados es una de las tareas más importantes en el desempeño del modelo de clasificación.

En las siguientes secciones se presenta el problema de selección de características (FSP por las siglas en inglés de *Feature Selection Problem*), en donde se organiza de la siguiente forma: primero se describe la notación a utilizar en la definición del problema, después se presenta

el FSP de manera formal. Enseguida se definen aspectos relevantes a tomar en consideración para proponer un algoritmo al FSP. Por último, se presenta el FSP en AMPs, el cual es el objetivo de este trabajo de tesis.

3.2. Problema de selección de características (FSP)

El problema de selección de características en términos generales se define como, dado un conjunto de características candidatas, seleccionar un subconjunto con respecto a alguno de los siguientes enfoques (Molina *et al.*, 2002):

1. El subconjunto de características que maximice el criterio de evaluación.
2. El subconjunto de características de menor tamaño que satisfaga la restricción o el criterio de evaluación.
3. El subconjunto con el mejor compromiso entre el tamaño y el valor del criterio de evaluación.

Encontrar el subconjunto óptimo con respecto al criterio de evaluación entre el conjunto de características disponibles es un problema NP-difícil (Amaldi y Kann, 1998). Por consiguiente, realizar una evaluación exhaustiva de todos los posibles subconjuntos no es factible, incluso para tan solo una centena de descriptores.

A continuación se describe formalmente el problema y los elementos importantes en el proceso de selección de características.

3.2.1. Definición matemática de FSP

Primero se presenta la notación a utilizar en la definición del FSP.

Conjunto de características

Sea X el conjunto de características con cardinalidad $\mid X \mid = n$, es decir

$$X = \{X_1, X_2, \ldots, X_n\} ,$$

donde X_i es la i-ésima característica.

Instancia

Sea $\mathbf{x_r}$ la variable que representa una instancia de X, entonces

$$\mathbf{x_r} = (x_{r1}, x_{r2}, x_{r3}, \ldots, x_{rn}) \,,$$

donde $\mathbf{x_r}$ es un vector n-dimensional, tal que x_{rj} denota el valor de la característica X_j.

Espacio de búsqueda

Sea $\mathcal{H}$ el espacio de todos los subconjuntos que se pueden formar a partir de X, entonces

$$\mathcal{H} = \mathcal{P}(X) - \emptyset \,,$$

donde $\mathcal{P}(X)$ es el conjunto potencia de X de aquí que el tamaño de $\mathcal{H}$ es de $2^n - 1$.

Conjunto de datos

Sea $\mathcal{D}$ el conjunto de datos de tamaño $|\mathcal{D}| = p$, entonces

$$\mathcal{D} = \bigcup_{i=1}^{p} \{(\mathbf{x_i}, y_i) | \mathbf{x_i} \in \mathbb{R}^n,\ y_i \in \{0, 1\}\} \,,$$

donde $\mathbf{x_i}$ es la i-ésima ocurrencia del conjunto de características X y es un vector en un espacio n-dimensional; y_i es la etiqueta o la clase a la que pertenece $\mathbf{x_i}$. $\mathcal{D}$ contiene tanto casos positivos (*i.e.,* $y_i = 1$) como casos negativos (*i.e.,* $y_i = 0$).

Problema de selección de características (FPS)

El FSP tiene como entrada un conjunto de datos $\mathcal{D}$, que es el conjunto de datos con X características $|X| = n$; y J el criterio de evaluación. El FSP consiste en encontrar X'_{opt}, que se define a continuación.

$$X^{'}_{opt} = \arg\max_{X' \subseteq X} J(X', \mathcal{D}) \,, \tag{10}$$

donde $J(X', \mathcal{D})$ es la función que evalúa al subconjunto X' usando el conjunto de datos $\mathcal{D}$; $X^{'}_{opt}$ es igual a encontrar el subconjunto X' para el cual J alcanza su máximo valor.

Se puede pensar que seleccionar todas las características del conjunto X da como resultado el máximo valor de la función J, sin embargo, en la práctica se ha demostrado que esto no es siempre el caso. Principalmente porque dentro del conjunto X pueden existir características irrelevantes que agreguen ruido a la información útil. Por lo general, esta situación ocurre cuando el tamaño del conjunto X es muy grande y el número de instancias de X es muy pequeño, manifestándose en el denominado fenómeno "del pico"(*peaking phenomenon*), donde el empleo de un número grande de características produce una peor exactitud en el desempeño del clasificador que cuando se usa un número pequeño de características (Sima y Dougherty, 2008). El fenómeno de pico ha sido demostrado para la clasificación discreta por Hughes (1968).

3.2.2. Caracterización del FSP

El FSP puede ser tratado como un problema de búsqueda en el espacio de las posibles soluciones $\mathcal{H}$ (Molina *et al.*, 2002). Para la caracterización del FSP como un problema de búsqueda es necesario tomar en consideración lo siguiente: ¿Cómo buscar en el espacio de los posibles subconjuntos de características?, ¿cómo evaluar la calidad de los posibles subconjuntos de características?

Por lo anterior, es necesario definir una estrategia de búsqueda y una medida de evaluación para la caracterización del problema. A continuación se describen a detalle estos dos aspectos.

Estrategia de búsqueda

Un algoritmo de búsqueda es responsable de dirigir el proceso de selección de características usando una estrategia específica. En general, las estrategias de búsqueda sólo visitan una parte del espacio $\mathcal{H}$, debido que para un conjunto de datos con n características el espacio de búsqueda es de 2^n. Esto implica que cuando n crece considerablemente se convierte prohibitivamente costoso explorar exhaustivamente el espacio de soluciones. De acuerdo con Molina *et al.* (2002) existen tres tipos de estrategias de búsquedas: exponencial, secuencial y estocástica. Sólo la primer estrategia de búsqueda es exacta y el resto son heurísticas, estas se describen en forma sucinta a continuación.

Búsqueda exponencial

La búsqueda exponencial o completa es una búsqueda óptima debido a que garantiza encontrar la mejor solución (Dash y Liu, 1997). La búsqueda exhaustiva (*i.e.,* recorre todas las posibles soluciones del espacio $\mathcal{H}$) es un tipo de búsqueda exponencial. Además, existen otras técnicas tales como ramificación y poda (*Branch and bound*) que permiten reducir el espacio del búsqueda, sin comprometer la posibilidad de encontrar el óptimo (Liu y Yu, 2005). El costo computacional de la búsqueda exponencial es de $T(n) = O(2^n)$.

Búsqueda secuencial

La idea general de la búsqueda secuencial es seleccionar una característica para agregarla o eliminarla del subconjunto de características. La búsqueda secuencial es más eficiente con respecto a la exponencial, sin embargo, no garantiza el resultado óptimo (*i.e.,* es una heurística). Las técnicas principales en la búsqueda secuencial son: selección hacia delante (SFS), selección hacia atrás (SBS), y selección bidireccional (Molina *et al.*, 2002; Liu y Yu, 2005). En cada iteración, la técnica SFS agrega a la solución la característica que aumenta el criterio de evaluación. La técnica de SBS elimina en cada iteración la característica que hace más pequeño el criterio de evaluación en cada iteración. En general el costo computacional que tiene la estrategia de búsqueda secuencial es de $T(n) = O(n^2)$ (Liu y Yu, 2005).

Búsqueda estocástica

La búsqueda estocástica es una heurística que, a diferencia de las búsquedas secuencial y exponencial, utiliza la aleatoriedad para evitar quedarse atrapado en mínimos locales. Este tipo de estrategias puede dar en muchas ocasiones el subconjunto óptimo en un tiempo computacional razonable. Ejemplos de búsqueda estocástica son el recocido simulado (*simulated annealing*) y el algoritmo genético (Molina *et al.*, 2002; Dash y Liu, 1997).

Criterio de evaluación

Un criterio de evaluación es una medida que determina la calidad de los subconjuntos que se producen por la estrategia de búsqueda, esta medida se define a continuación:

Sea $J(X')$ la variable que representa el criterio de evaluación para el subconjunto X', entonces

$$J : X' \subseteq X \to \mathbb{R}\,, \tag{11}$$

donde $\mathbb{R}$ es el conjunto de los reales, valores grandes de J indican que el subconjunto X' tiene mucha relevancia, caso contrario indican poca relevancia.

Existen muchos enfoques para evaluar la calidad de un subconjunto de características, la mayoría coinciden en medir la capacidad de las características para separar las clases. El criterio de evaluación J puede ser categorizado basándose en la dependencia que tiene con el algoritmo de aprendizaje.

Los **métodos de filtrado** son independientes del algoritmo de aprendizaje de máquina y reducen el conjunto de características basado en criterios de evaluación tales como: distancia entre clases, ganancia de información y dependencia entre las características (Dash y Liu, 1997).

Los **métodos de envoltura** (*wrapper*) utilizan un algoritmo de aprendizaje de máquina (*e.g.,* clasificador) para evaluar la calidad de los subconjuntos (Kohavi y John, 1997). Como las características son seleccionadas por el clasificador que después será usado para predecir nuevos elementos, el nivel de precisión es más alto que el de los métodos de filtrado. Sin embargo, el tiempo computacional requerido es muy costoso comparado con los métodos de filtrado. Algunos de los criterios de evaluación para los métodos de envoltura son la probabilidad de error del clasificador y las medidas de calidad descritas en la Sección 2.2.3.

Para propósito del presente trabajo sólo utilizaremos los métodos de envoltura y sus criterios de evaluación.

3.3. Problema de selección de características en AMPs

Nuestro problema se enfoca en encontrar un subconjunto de descriptores moleculares útiles para la construcción de un buen predictor de AMPs y péptidos antibacterianos. Para esto es necesario seleccionar el conjunto de péptidos representativo de las diferentes clases (*i.e.,* AMP, noAMP, antibacteriano y no antibacteriano), para después representar cada péptido en términos de descriptores moleculares.

El algoritmo de selección de características recibirá como entrada los péptidos representados en descriptores moleculares y dará como salida el subconjunto de descriptores óptimo y la exactitud del clasificador. Después, con el mejor subconjunto se creará un modelo QSAR con el que se examinará un conjunto de péptidos con actividad desconocida para determinar cuáles son antimicrobianos.

A continuación se define formalmente el problema, así como el criterio de evaluación para medir la calidad de los subconjuntos.

3.3.1. Definición formal del problema

Dado un conjunto de datos $\mathcal{D}$ con un conjunto de características X; y un modelo de clasificación $\mathcal{I}$. El problema consiste en encontrar $X^{'}_{opt}$ que se define a continuación.

$$X^{'}_{opt} = \underset{X' \subseteq X}{\arg\,\max} J(X', \mathcal{D}) \tag{12}$$

$$J(X', \mathcal{D}) = ACC(\mathcal{I}(\mathcal{D}'))\,, \tag{13}$$

donde $\mathcal{D}' \subseteq \mathcal{D}$ es el conjunto de datos removiendo los valores de las variables que no estén en X'; y ACC es la exactitud del clasificador $\mathcal{I}$. Una solución es óptima si la exactitud del clasificador $ACC(\mathcal{I}(\mathcal{D}'))$ es máxima. Es importante señalar que no necesariamente $X^{'}_{opt}$ es única, esto debido a que se puede llegar a la misma exactitud utilizando diferentes conjuntos de características.

Capítulo 4. Materiales y Métodos

Un diagrama esquemático de la metodología general que se utilizó en esta tesis se muestra en la Figura 16. Iniciamos con la recopilación y preparación de los datos en donde se seleccionaron péptidos con y sin la actividad biológica deseada. Enseguida a cada péptido recolectado se le calcularon sus descriptores moleculares (*e.g.,* hidrofobicidad, peso molecular, carga), lo que involucra transformar la secuencia primaria del péptido en un conjunto de números que capturen las propiedades fisicoquímicas relevantes. Después, se aplicaron un algoritmo genético y un algoritmo de aprendizaje máquina para seleccionar las características relevantes para la identificación de péptidos con una actividad biológica deseada. También, se aplicó un clasificador que relaciona las características con la actividad utilizando una SVM y el subconjunto de características resultado del algoritmo genético. Por último, se evaluó la calidad del modelo en términos de la exactitud de predicción.

En este capítulo se presenta el encadenamiento de procesos propuestos para el diseño del modelo para la identificación de péptidos con una actividad deseada, para cada actividad se describen los métodos y materiales utilizados (Figura 16).

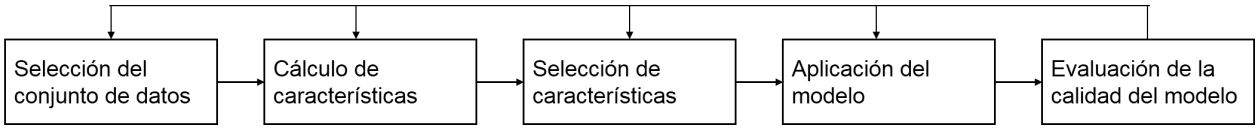
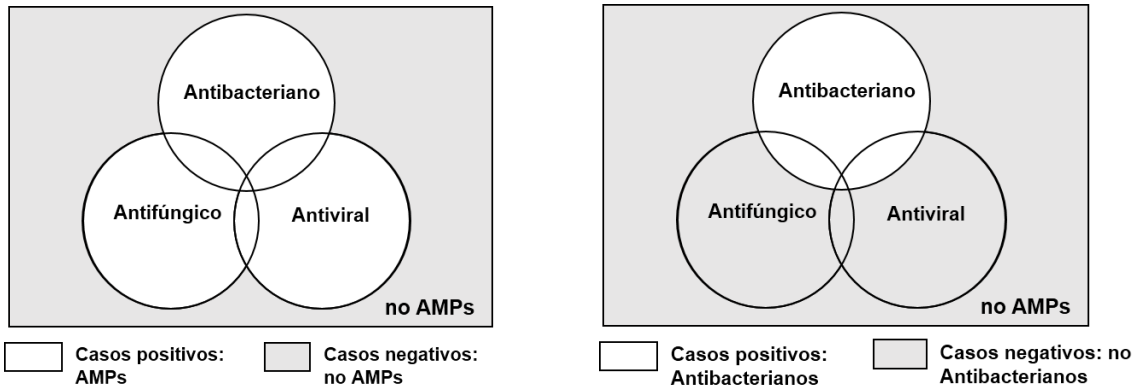


Figura 16: Metodología general propuesta.



(a) Conjunto de datos: modelo AMPs

(b) Conjunto de datos: modelo Antibac.

Figura 17: Conjunto de péptidos con y sin la actividad biológica deseada.

4.1. Selección del conjunto de datos

El objetivo principal del presente trabajo es crear dos modelos: el primero de nombre modelo AMP para la identificación de péptidos con actividad antimicrobiana; el segundo modelo denominado Antibac, para identificar AMPs con actividad específica en contra de una clase particular de microbios, las bacterias. Para la construcción de los modelos es necesario la recopilación de secuencias de péptidos con y sin la actividad deseada. En la Figura 17 se muestra un diagrama de Venn para representar el conjunto de péptidos utilizados como casos positivos y negativos dado un modelo. Por ejemplo, para el modelo Antibac, se consideran como casos positivos el conjunto de péptidos con actividad antibacteriana y como casos negativo los péptidos con actividad antifúngica, antiviral o sin activdad antimicrobiana (ver Figura 17b).

A continuación, se presenta la metodología para obtener los conjuntos de datos para los modelos AMP y Antibac, respectivamente.

4.1.1. Conjunto de datos para el modelo AMP

En esta sección se describe la obtención y preparación de los datos para la predicción de la actividad antimicrobiana en los péptidos. La metodología se obtuvo a partir de la revisión de la literatura de los principales métodos para extraer péptidos con y sin actividad antimicrobiana (Wang *et al.*, 2011b; Lata *et al.*, 2007; Joseph *et al.*, 2012).

Casos de prueba positivos: péptidos antimicrobianos

Para crear los casos positivos de péptidos con actividad antimicrobiana se utilizó la base de datos *Collection of AntiMicrobial Peptides* (CAMP) (Waghu *et al.*, 2014), seleccionando sólo las secuencias con anotación experimentalmente validada. Después de obtener las secuencias, se eliminaron aquellas que contienen aminoácidos no estándares, tales como: B, J, O, U, X y Z. Por último, con el objetivo de tener un conjunto de prueba no redundante se eliminan las secuencias de péptidos que tienen una identidad del 50 % o más, utilizando el programa BlastClust (Dondoshansky y Wolf, 2002). Al final el conjunto de péptidos con actividad antimicrobiana está formado por 1702 secuencias (ver Anexo B, tablas 28 y 29). En la Figura 18 presentamos un diagrama esquemático de la metodología para obtener los péptidos antimicrobianos.

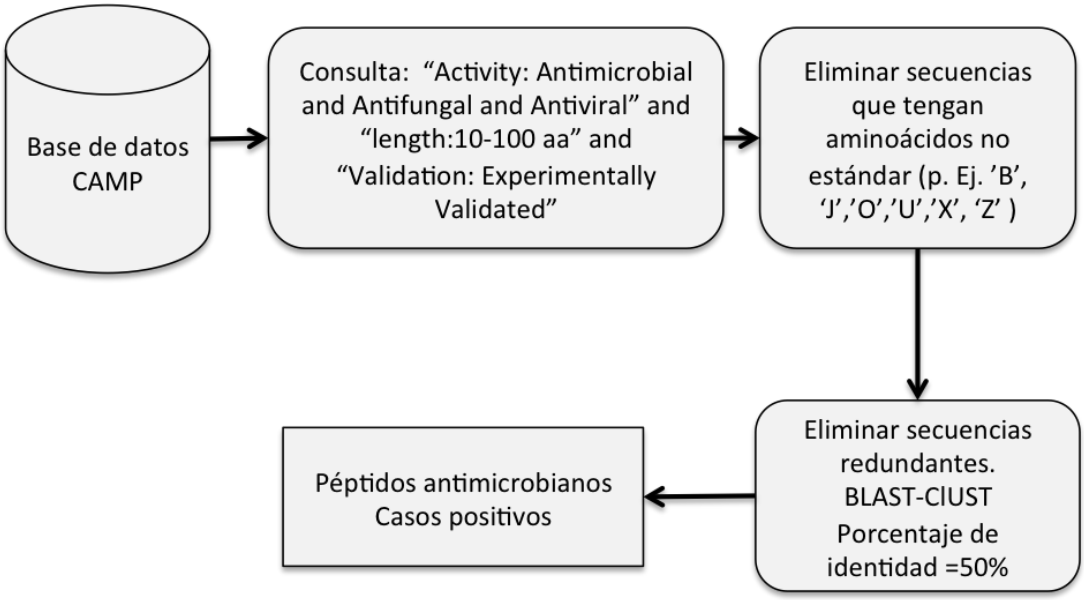


Figura 18: Metodología para la obtención de los casos positivos (AMPs).

Casos de prueba negativos: péptidos no antimicrobianos

Al no existir un base de datos que contenga únicamente péptidos con actividad no antimicrobiana, fue necesario recurrir a bases de datos de propósito general tal como Uniprot (almacena proteínas y péptidos de todo tipo) (Apweiler *et al.*, 2004). Los pasos para construir el conjunto de péptidos sin actividad antimicrobiana fueron los siguientes:

1. Solicitar a la base de datos macromoléculas del tipo proteínas sin ADN, ARN y no mezclas (*i.e.,* híbridos de ADN y ARN). Además, las secuencias no deben contener la anotación de actividad antimicrobiana y tener una longitud de 10 a 100 residuos.

2. Eliminar las proteínas de membrana (*i.e.,* proteínas que interaccionen con membranas biológicas) y proteínas extracelulares. La razón para esto es que las proteínas de membrana tienen propiedades similares a los AMPs y los péptidos antimicrobianos por lo general son secretados por las células. Para eliminar este tipo de proteínas usamos el programa *Phobius web server* (Käll *et al.*, 2007).

3. Eliminar secuencias que tengan aminoácidos no estándar.

4. Crear un conglomerado de péptidos con BlasClust (Dondoshansky y Wolf, 2002) utilizando un 50 % de identidad, con el objetivo de eliminar secuencias redundantes para obtener finalmente el conjunto de péptidos que servirán como casos de prueba negativos.

El conjunto resultante de péptidos sin actividad antimicrobiana al aplicar la metodología es de 1884 secuencias. En el Apéndice B tablas 30 y 31 se muestran los identificadores de los péptidos que forman parte de los casos negativos.

4.1.2. Conjunto de datos para el modelo Antibac

En esta sección se describe la obtención y preparación de los datos para la predicción de la actividad antibacteriana en los péptidos. El conjunto de casos positivos está compuesto por péptidos con la actividad antibacteriana y el conjunto de casos negativos se compone de los péptidos con actividad antifúngica, antiviral o sin actividad antimicrobiana (ver subsección 4.1.1).

Casos de prueba positivos: péptidos antibacterianos

Para crear los casos positivos de péptidos con actividad antibacteriana se utilizó la base de datos *Collection of AntiMicrobial Peptides* (CAMP) (Waghu *et al.*, 2014), seleccionando sólo los péptidos con longitud de 10 a 100 aminoácidos. Después se descartaron las secuencias con aminoácidos no estándares. Por último, con el objetivo de tener un conjunto de prueba no redundante se creó un conglomerado de péptidos con BlastClust (Emmanouilidis *et al.*, 2000). El conjunto resultante de péptidos con actividad antibacteriana es de 2214 secuencias.

Casos de prueba negativos: péptidos no antibacterianos

Para crear el conjunto de casos negativos se utilizaron las secuencias obtenidas de la subsección 4.1.1 y la base de datos CAMP (Waghu *et al.*, 2014). A CAMP se le solicitaron secuencias de longitud de 10 a 100 aminoácidos que tuvieran la actividad antifúngica o antiviral, y sin la anotación de actividad antimicrobiana. Después se descartaron las secuencias con aminoácidos no estándares. Por último al tener pocas secuencias de péptidos con actividad antifúngica y antiviral (*i.e.,* 323 secuencias), no se realizó el conglomerado.

El conjunto resultante de péptidos sin actividad antibacteriana es de 2207 secuencias (323 con actividad antifúngica y antiviral y 1884 péptidos no antimicrobianos).

4.2. Cálculo de características: Descriptores moleculares

Con el objetivo de encontrar similaridades entre los AMPs, se representa cada péptido en términos de descriptores moleculares. Los descriptores moleculares permiten transformar la información química estructural del péptido en un vector numérico mediante un proceso de cómputo. Para realizar el cálculo primero se necesita representar a los péptidos en una estructura molecular adecuada dependiendo del nivel de la dimensionalidad de los descriptores que se deseen obtener (ver Sección 2.2.3). En el presente trabajo utilizamos un grafo molecular para representar la información estructural de los péptidos debido a que podemos derivar una gran cantidad de descriptores de manera sencilla a partir de éste (*i.e.,* se pueden calcular descriptores desde 0D hasta 2D).

En esta sección se describe la metodología para calcular los descriptores moleculares a partir de la secuencia primaria de los péptidos. Primero se transforma la secuencia del péptido en un grafo molecular; segundo, se almacena el grafo molecular en un archivo Mol; por último, se transforma el grafo en un vector de características. A continuación se describe a detalle el procedimiento para el cálculo de características.

4.2.1. Grafo topológico molecular

Un grafo G es un par ordenado de los conjuntos disjuntos (V, E) tal que $E \subseteq V^2$ y $V \neq \emptyset$. V es el conjunto vértices y E es el conjunto de aristas. Una arista $(i, j) \in E$ representa la unión del vértice i con el vértice j (Bollobas, 2004).

Cuando una molécula como el caso de un péptido es representada en forma de grafo recibe el nombre de **grafo molecular**, donde los átomos son los vértices y los enlaces son las aristas. Para simplificar la representación del péptido en el grafo se eliminan los átomos de hidrógeno (ver Figura 19).

La idea de representar los péptidos por medio de grafos moleculares es tener acceso a la información estructural independiente de la conformación del péptido, por ejemplo; el tipo de enlace entre dos átomos, las distancias que existen entre todos los átomos del péptido, entre otros. Para acceder eficientemente a la información del péptido, los grafos moleculares son representados por varias matrices topológicas tales como: matriz de adyacencia, matriz de distancia y matriz de conexión. Por otra parte, con el objetivo de compartir y almacenar las matrices topológicas de las estructuras de los péptidos se utilizan los archivos Mol. A continuación se describen las matrices y el formato estándar para su almacenamiento.

Matriz de adyacencia

La matriz de adyacencia es una matriz cuadrada que contiene información acerca de los átomos que se encuentran contiguos. Se supone que los vértices son numerados de manera arbitraria como $1, 2, \ldots, |V|$. La matriz de adyacencia A de un grafo G es una matriz cuadrada y simétrica de tamaño $|A| = |V|$ x $|V|$, tal que para cada elemento $a_{i,j}$ toma uno de los siguientes valores:

$$a_{i,j} = \begin{cases} 1, & \text{si } (i, j) \in E, \\ 0, & \text{en otro caso } v. \end{cases} \tag{14}$$

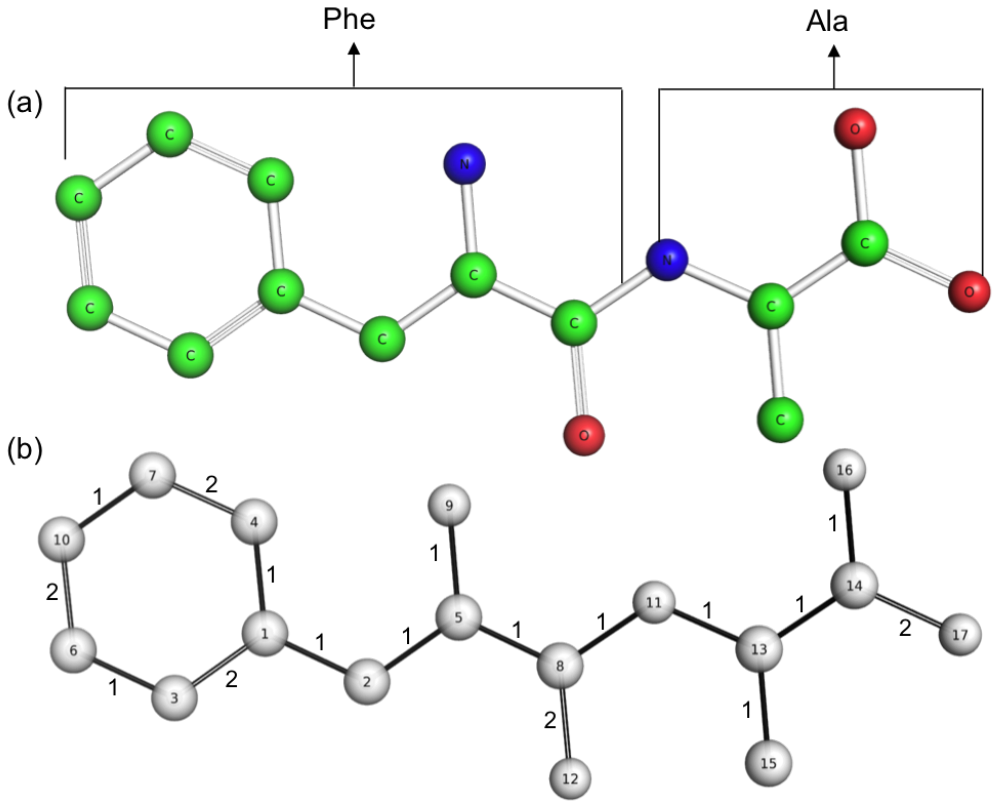


Figura 19: (a) Estructura 2D del péptido Phe-Ala; (b) Representación del péptido en grafo molecular con identificador del átomos y tipo de enlace entre los átomos.

En la Tabla 7 se muestra un ejemplo de la matriz de adyacencia para el péptido Phe-Ala (ver Figura 19).

Matriz de distancia

La matriz de distancia D contiene información acerca de la longitud del camino más corto entre un par de vértices en el grafo G. La matriz de distancia D de un grafo G es una matriz cuadrada y simétrica de tamaño $|D| = |V|$ x $|V|$, tal que para cada elemento d_{ij} de la matriz puede tomar uno de los siguientes valores:

$$d_{i,j} = \begin{cases} d(v_i, v_j), & \text{si } i \neq j, \\ 0, & \text{si } i = j\ . \end{cases} \tag{15}$$

Tabla 7: Matriz de adyacencia para el grafo de la Figura 19.

a_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3			0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4				0	0	0	1	0	0	0	0	0	0	0	0	0	0
5					0	0	0	1	1	0	0	0	0	0	0	0	0
6						0	0	0	0	1	0	0	0	0	0	0	0
7							0	0	0	1	0	0	0	0	0	0	0
8								0	0	0	1	1	0	0	0	0	0
9									0	0	0	0	0	0	0	0	0
10										0	0	0	0	0	0	0	0
11											0	0	1	0	0	0	0
12												0	0	0	0	0	0
13													0	1	1	0	0
14														0	0	1	1
15															0	0	0
16																0	0
17																	0

donde $d_{i,j}$ es el número de aristas en el camino más corto en el grafo G entre el vértice i y el vértice j. La matriz de distancia para el grafo de la Figura 19 se muestra en la Tabla 8.

Matriz de conexión

La matriz de conexión C sirve para capturar el tipo de enlace que existe entre un par de átomos. La matriz C es simétrica de tamaño $|V|$ x $|V|$, tal que cada elemento $c_{i,j}$ en la matriz toma uno de los siguientes valores:

$$c_{i,j} = \begin{cases} 2, & \text{si } (i,j) \in E \text{ y el enlace es doble} \\ 1, & \text{si } (i,j) \in E \text{ y el enlace es simple} \\ 0, & \text{si } (i,j) \notin E \ . \end{cases} \tag{16}$$

En la Tabla 9 se muestra un ejemplo de la matriz de conexión para el péptido Phe-Ala (ver Figura 19).

Tabla 8: Matriz de distancia para el grafo de la Figura 19.

d_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	1	1	1	2	2	2	3	3	3	4	4	5	6	6	7	8
2		0	2	2	1	3	3	2	2	4	3	3	4	5	5	6	6
3			0	2	3	1	3	4	4	2	5	5	6	7	7	8	8
4				0	3	3	1	4	4	2	5	5	6	7	7	8	8
5					0	4	4	1	1	5	2	2	3	4	4	5	5
6						0	2	5	5	1	6	6	7	8	8	9	9
7							0	5	5	1	6	6	7	8	8	9	9
8								0	2	6	1	1	2	3	3	4	4
9									0	6	3	3	4	5	5	6	6
10										0	7	7	8	9	9	10	10
11											0	2	1	2	2	3	3
12												0	3	4	4	5	5
13													0	1	1	2	2
14														0	2	1	1
15															0	3	3
16																0	2
17																	0

Tabla 9: Matriz de conexión para el grafo de la Figura 19.

c_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3			0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4				0	0	0	2	0	0	0	0	0	0	0	0	0	0
5					0	0	0	1	1	0	0	0	0	0	0	0	0
6						0	0	0	0	2	0	0	0	0	0	0	0
7							0	0	0	1	0	0	0	0	0	0	0
8								0	0	0	1	2	0	0	0	0	0
9									0	0	0	0	0	0	0	0	0
10										0	0	0	0	0	0	0	0
11											0	0	1	0	0	0	0
12												0	0	0	0	0	0
13													0	1	1	0	0
14														0	0	1	2
15															0	0	0
16																0	0
17																	0

Archivo Mol

Con el objetivo de compartir y almacenar las matrices topológicas de las estructuras de los péptidos, utilizamos los archivos Mol. Mol es un formato para los archivos de texto que fue desarrollado por MDL *Information System* con el objetivo de estandarizar la información molecular (MDL, 2005). A continuación se describen los campos del archivo Mol de mayor importancia para la presente investigación.

Un archivo Mol está compuesto por un encabezado y una tabla de conexiones. El **encabezado** sirve para identificar la molécula, contiene información tal como: nombre de la molécula, fecha, comentarios. La **tabla de conexión** contiene información que describe la relación estructural y propiedades de una colección de átomos. La tabla de conexión se divide en dos secciones: en la primera sección se declara la lista de átomos y coordenadas 2D, estas coordenadas se calculan a partir de las distancias relativas entre los átomos de la matriz de distancia; en la segunda sección se declara la lista y el tipo de enlace entre los átomos, en esta parte se combinan la matriz de adyacencia y conexión en un solo bloque.

En la Figura 20 se muestra un ejemplo del formato que siguen los registros de los grafos moleculares en un archivo Mol, el ejemplo corresponde al péptido Phe-Ala (Figura 19).

4.2.2. Cálculo de descriptores en péptidos

En esta sección se describe cómo transformar un grafo molecular a un vector numérico de características. Para realizar la transformación es necesario un proceso de cómputo que recibe como entrada un grafo y da como salida el vector de características. En el presente trabajo utilizamos los programas JPeDes (*Java Peptide Descriptors*) y PaDel-Descriptor, para el cálculo de los descriptores moleculares.

Se calcularon un conjunto de 770 descriptores para cada péptido del conjunto de datos descrito en la Sección 4.1 usando PaDel-Descriptor (Yap, 2011). Los descriptores calculados son del tipo 1D y 2D. Por otra parte, para el cálculo de descriptores que dependen de la composición de los aminoácidos (descriptores OD) y de los que dependen de la secuencia (descriptores 1D) se utilizó el programa JPeDes. JPeDes es un software que se basa en el

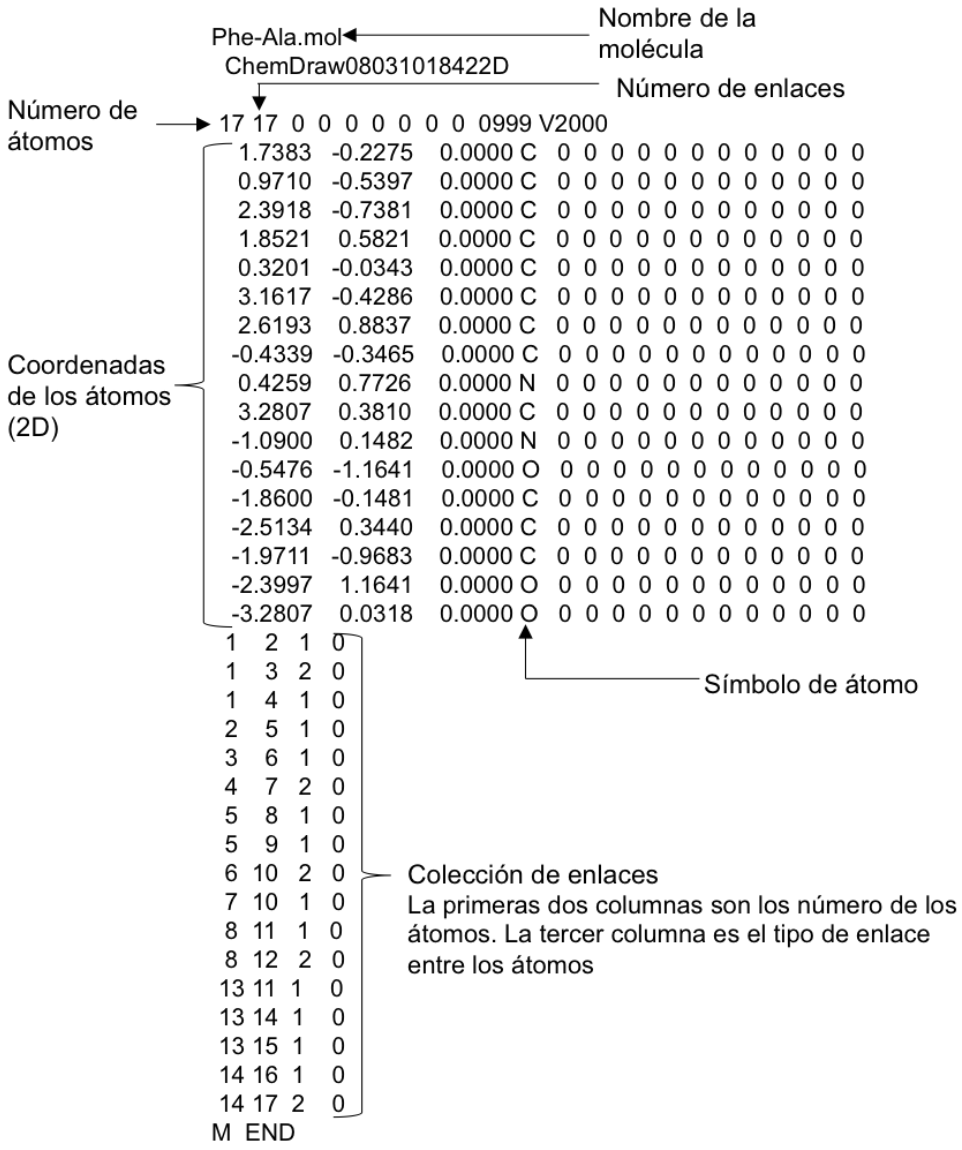


Figura 20: Formato MOL para el registro de una estructura molecular 2D. El ejemplo corresponde al péptido Phe-Ala de la Figura 19.

Descriptores moleculares

	ID	ALogP	ALogp2	AMR	apol	nAtom	nHeavy	nH	nC	nN	nO
Péptido 1 →	"CAMPSQ1022"	-11.043	121.96	414.92	267.2	256	119	137	78	19	22
	"CAMPSQ1006"	-38.291	1466.2	1192	765.4	727	340	387	220	59	58
	"CAMPSQ1052"	-38.395	1474.2	1164.6	715.3	661	336	325	207	61	62
	"CAMPSQ1077"	-26.485	701.44	517.85	341	325	163	162	98	28	37
Péptido *i* →	"CAMPSQ1121"	-7.7316	59.778	365.17	226.7	211	100	111	71	15	14
	"CAMPSQ1081"	-28.41	807.15	801.15	535.1	514	242	272	153	44	45
	"CAMPSQ1135"	-28.543	814.7	754.91	478.4	446	227	219	142	41	43
	"CAMPSQ114"	-15.895	252.66	617.25	391	372	167	205	116	31	20
	"CAMPSQ1165"	-65.102	4238.2	1725.3	1110	1043	514	529	334	85	95
	"CAMPSQ1170"	-71.054	5048.7	2133.3	1371	1310	603	707	395	105	100
	"CAMPSQ1166"	-34.038	1158.6	847.92	552.7	519	255	264	166	44	45
	"CAMPSQ1175"	-18.31	335.26	509.13	330.5	313	149	164	98	26	25

Figura 21: Ejemplo de péptidos representados como descriptores moleculares.

programa PeDes (Japelj, 2005) para el cálculo de 28 descriptores moleculares. La lista de los descriptores utilizados en esta investigación se muestran en el Apéndice C.

La salida de ambos programas consiste en un archivo CSV (*comma-separate values*) que contiene una tabla donde cada renglón es un péptido y cada columna es un descriptor molecular. En la Figura 21 se muestra un extracto del archivo, el cual muestra registros de péptidos representados como descriptores moleculares.

4.3. Selección de características

Para resolver el problema de selección de características en la clasificación de péptidos antimicrobianos en esta sección se describe el diseño de un algoritmo genético utilizando el método de envoltura (*wrapper).* El método de envoltura está compuesto principalmente por dos elementos: una **estrategia de búsqueda** para la generación de los posibles subconjuntos de características, en este caso se propone como estrategia un algoritmo genético; un **algoritmo de inducción** para evaluar la calidad del subconjunto seleccionado, donde la estrategia de búsqueda utiliza como caja negra el algoritmo de inducción (ver Figura 22).

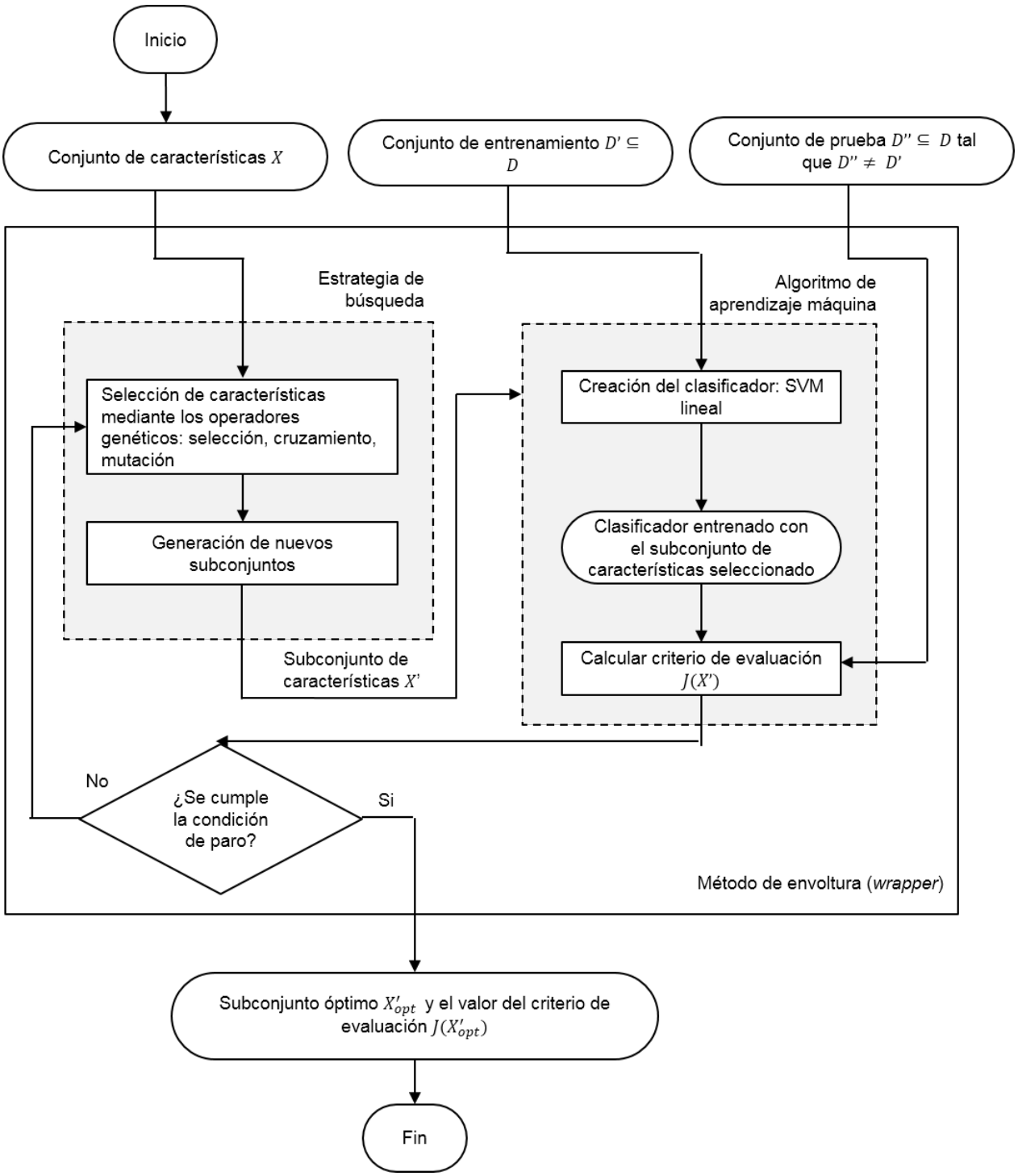


Figura 22: Diagrama general para el método de envoltura. El algoritmo de aprendizaje máquina es usado como caja negra por la estrategia de búsqueda.

4.3.1. Algoritmo de inducción

En aprendizaje de máquina un algoritmo de inducción es típicamente presentado con un conjunto de casos de entrenamiento $\mathcal{D}$, donde cada caso describe un vector $\mathbf{x} \in \Re^n$ de valores para las características X y una etiqueta de la clase $y \in Y$ (Kohavi y John, 1997). La tarea del algoritmo de inducción es producir un clasificador $I : X \to Y$ que sea útil para etiquetar correctamente casos desconocidos.

En el presente trabajo para construir el clasificador binario utilizamos una máquina de soporte vectorial lineal (SVM) y un conjunto de datos de entrenamiento, los cuales fueron descritos en la Sección 4.1.

4.3.2. Estrategia de búsqueda

Sea $X = \{X_1, X_2, \cdots, X_n\}$ un conjunto de características medibles en los péptidos, sea $\mathcal{H} = \mathcal{P}(X) - \emptyset$ el espacio de búsqueda, donde $\mathcal{P}(X)$: conjunto potencia de X; nos interesa determinar cuál es el subconjunto $X' \in \mathcal{H}$ que maximiza el criterio de evaluación J. Encontrar el subconjunto óptimo X' de característica entre un total de $2^n - 1$ posibles soluciones es un problema NP-difícil (Amaldi y Kann, 1998). Por lo anterior, es necesario una estrategia de búsqueda que explore eficientemente el espacio $\mathcal{H}$. En la literatura se han propuesto varias estrategias tales como: ramificación y poda (*Branch and bound*) (Liu y Yu, 2005), selección hacia delante (SFS) (Molina *et al.*, 2002), algoritmos genéticos (GAs)(Huang *et al.*, 2007; Pavan *et al.*, 2006), entre otros.

En el presente trabajo se propone como estrategia de búsqueda un algoritmo genético por las siguientes razones:

- GA es una de las técnicas más populares usadas para la selección de características en modelos QSAR (Goodarzi *et al.*, 2012; Pavan *et al.*, 2006, 2005). Además, de acuerdo con Kudo y Sklansky (2000), los GA son apropiados para problemas de selección de características de gran escala (*i.e.,* problemas con más de 50 características) debido a

que tiene altas posibilidades de encontrar la mejor solución comparado con otros algoritmos de selección.

- GA es una heurística que, a diferencia de los métodos secuenciales, es capaz de escapar de óptimos locales.

- GA es capaz de devolver una solución válida (*i.e.,* un subconjunto de características) en cada iteración del algoritmo (Huang *et al.*, 2007).

Algoritmo Genético

Se propone un algoritmo genético, donde cada individuo en la población representa un subconjunto de características. El objetivo es encontrar al subconjunto que satisfaga la siguiente expresión:

$$G_{opt} = \arg\max_{G \in \mathcal{G}} Fitness(G) \,, \tag{17}$$

donde G es la representación de un subconjunto de características en el espacio donde se llevará la búsqueda evolutiva (espacio del genotipo). En el Algoritmo 1, podemos observar los pasos llevados a cabo para obtener la solución óptima G_{opt}.

En las siguientes subsecciones se describen los principales pasos del algoritmo genético para la selección de características.

Algoritmo 1 Algoritmo genético para la selección de características.

Entrada: datos de entrenamiento $\mathcal{D}$ con características X, $|X| = n$,
J medida de evaluación a maximizar,
n_g número máximo de generaciones,
n_{gwi} número de generaciones sin mejora,
n_i número de individuos en la población I
n_P número de padres,
p_c probabilidad de cruzamiento,
p_m probabilidad de mutación

Salida: subconjunto de características X' y el valor del criterio de evaluación $J(X')$

1: Generar una población I inicial aleatoria de tamaño n_i
2: Calcular la aptitud para cada individuo
3: **repetir**
4: Seleccionar a los padres P de la población I
5: Aplicar operador de cruzamiento a P con una probabilidad p_c para generar los hijos O
6: Aplicar operador de mutación a O con una probabilidad p_m
7: Calcular la aptitud para cada individuo en O
8: Seleccionar a los sobrevivientes de $I + O$ para la siguiente generación
9: **hasta que** el número de generaciones sea igual n_g o el número de generaciones sin mejora sea igual a n_{gwi}

Representación de un subconjunto de características

Dado un conjunto de características $X = \{X_1, ..., X_n\}$, un individuo es un subconjunto $X_G \subseteq X$ representado por el vector G, entonces,

$$G = (g_1, g_2, g_3, ..., g_m) \text{ de donde } X_G = \{X_{g_1}, X_{g_2}, ..., X_{g_m}\},$$

tal que,

$$m \leq n,$$
$$g_i \neq g_j,\ i \neq j\ \forall i \in \{1, 2, \ldots, m\}$$
$$g_i = k, \text{ para } 1 \leq k \leq n, \text{ si } X_k \text{ es parte de la solución}$$
$$g_1 < g_2 < \ldots < g_m\ .$$

Esta representación permite que cada característica esté como un entero. Por ejemplo: si tenemos el conjunto de características X conformado por $X = \{MW, Nres, H_k, IP, Z(7)\}$ y

nuestra posible solución factible es el subconjunto X_G que está compuesto por $X_G = \{MW, IP, Z(7)\}$ la representación que toma en el algoritmo genético es como la que aparece en la Figura 23.

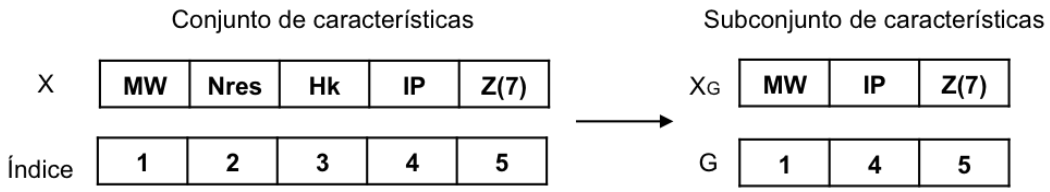


Figura 23: Representación de una solución factible en el algoritmo genético para la selección de características.

Función objetivo

La función objetivo está definida por:

$$Fitness(G) = J(X_G, D') \,,$$

donde X_G corresponde al subconjunto de características codificadas en el genotipo G, y $D' \subseteq D$ es el conjunto de entrenamiento removiendo las variables que no estén en X_G, es decir, $\mathcal{D}' = \bigcup_{i=1}^{p}\{(\mathbf{x_i}, y_i)|\mathbf{x_i} \in R^{|X_G|},\ y_i \in \{0,1\}\}$. $\mathbf{x_i} =< x_{i1}, ..., x_{i|X_G|} >$ es un vector de números reales que toma el subconjunto de características $X_G = \{X_{g1}, ..., X_{gm}\}$ tal que, $X_{g1} = x_{i1}, ..., X_{gm} = x_{i|X_G|}$. Un ejemplo del conjunto de entrenamiento se muestra en la Tabla 10.

Para definir la función de evaluación J es necesario introducir primero algunas definiciones básicas. Los conjuntos de prueba están formados por un grupo de casos positivos y un grupo de casos negativos. Cuando el predictor acierta en la etiqueta de un elemento que pertenece a los casos positivos se le conoce como verdadero positivo (TP), sin embargo cuando no lo reconoce se tiene un falso negativo (FN). De otra manera, cuando el predictor se equivoca en la clasificación de un elemento que pertenece a los casos negativos se le conoce como falso

positivo (FP) y cuando no se equivoca se tiene un verdadero negativo (TN). A partir de las comparaciones entre el valor esperado y el arrojado por el predictor se definen las siguientes medidas de calidad:

$ACC(\mathcal{I}(\mathcal{D}')) = \frac{TP+FN}{TP+FN+TN+FP}100$, es la exactitud del clasificador $\mathcal{I}$,

$MCC(\mathcal{I}(\mathcal{D}')) = \frac{(TP \text{ x } TN)-(FN \text{ x } FP)}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$, coeficiente de correlación de Matthews del clasificador $\mathcal{I}$ con los datos de entrenamiento $\mathcal{D}'$,

$\mathcal{I}$, es una máquina de soporte vectorial (SVM) lineal.

Con base en las medidas de calidad definidas previamente se propone la siguiente función de evaluación:

$$J(X_G, \mathcal{D}') = ACC(\mathcal{I}(\mathcal{D}')) + MCC(\mathcal{I}(\mathcal{D}')) + 1 - \frac{\mid X_G \mid}{\mid X \mid} \tag{18}$$

El intervalo de valores que puede tomar la función J es [0,102), donde la exactitud (ACC) en J tiene mayor importancia y los otros dos términos en la función sirven como criterios de desempate entre los individuos que tienen la misma exactitud. El término MCC da un mayor peso a individuos que tengan la especificidad y sensibilidad similares. Por otra parte, el término $1 - (|X_G|/|X|)$ sirve para dar un mayor peso a individuos que tengan un menor número de características.

Selección de padres

La selección de padres es un proceso estocástico para elegir a los individuos que posteriormente podrán cruzarse. En la selección de padres se utilizó la estrategia de torneo binario que consiste en seleccionar al azar $n_i/2$ parejas de individuos, donde n_i es el tamaño de la población. Para cada pareja se selecciona al individuo que tenga la mejor aptitud (ver Algoritmo 2).

Tabla 10: Ejemplo para el conjunto de datos de entrenamiento

Conjunto de entrenamiento $\mathcal{D}$

X

He	Hk	Z(pH5)	Z(pH7)	Z(pH9)	IP	Clase
-0.31	-1.15	-4.95	-5.9	-6.19	3.67	0
-0.18	-0.44	6.79	4.18	3.63	10.43	1
-0.21	-0.67	9.2	5.19	2.81	9.89	1
-0.01	0.41	0.21	-0.26	-5.11	5.97	0
-0.14	-0.25	0.21	-0.25	-5.11	5.97	0

Conjunto de entrenamiento $\mathcal{D}'$

$G =< 1, 2, 6 >$ $X_G = \{He, Hk, IP\}$

He	Hk	IP	Clase
-0.31	-1.15	3.67	0
-0.18	-0.44	10.43	1
-0.21	-0.67	9.89	1
-0.01	0.41	5.97	0
-0.14	-0.25	5.97	0

Algoritmo 2 Torneo binario sin remplazo para la selección de padres

Entrada: arreglo de individuos I de tamaño $n_i < i_1, i_2, ..., i_n >$
μ número de padres a seleccionar
Salida: P arreglo de padres de tamaño μ

```
1: P[1,...,μ] nuevo arreglo
2: padre_actual = 1;
3: mientras padre_actual ≤ μ hacer
4:    Generar un número k en el intervalo [1, n_i]
5:    Generar un número l en el intervalo [1, n_i]
6:    si I[k].aptitud ≥ I[l].aptitud entonces
7:       P[padre_actual] = I[k]
8:    si no
9:       P[padre_actual] = I[l]
10:   fin si
11:   padre_actual++
12: fin mientras
```

El Algoritmo 2 de selección de padres tiene un tiempo de ejecución en el peor de los casos de $T(n_i) = O(n_i)$. Lo anterior se debe a que el proceso de seleccionar un padre tiene un tiempo de ejecución de $O(1)$, dado que se necesita elegir μ padres, donde $\mu \leq n_i$, entonces el tiempo de ejecución para seleccionar μ padres es en el peor de los casos de $O(n_i)$.

Cruzamiento

En este paso, se decidió utilizar el operador de cruzamiento SSOCF (*Subset size-Oriented Common Feature*) (Emmanouilidis *et al.*, 2000) debido a que nos permite mantener bloques informativos comunes en los padres, es decir, los padres p_i y p_j heredan a los hijos las características que ambos tienen en común. Por otra parte, las características no compartidas son seleccionadas para heredarse a los hijos con una probabilidad $Prob(h_{p_i}) = (n_{p_i} - n_c)/n$, donde n_{p_i} es el número de características del padre p_i, n_c son las características comunes y n es el número total de características. Un ejemplo de este cruzamiento se muestra en las figuras 24 y 25, el pseudocódigo se describe en el Algoritmo 3.

El tiempo de ejecución para el algoritmo de cruzamiento CFC (ver Algoritmo 4) se describe a continuación. Primero se supone que el tamaño de un padre $tamaño(p_i) = O(n)$ donde n es el número de características disponibles. Para seleccionar los elementos comunes entre el padre p_1 y padre p_2, el tiempo de ejecución en el peor de los casos (*i.e.,* cuando $p_1 = p_2$) es de $O(n)$ (ver Algoritmo 4, pasos 6-20). Por otra parte, el tiempo de ejecución para copiar los elementos que no tienen en común los padres p_1 y p_2 a los hijos h_1 y h_2, respectivamente, es de $O(n)$ (ver pasos 27-40). Por último en los pasos 41-42 del algoritmo se realiza un ordenamiento ascendente en función de las características de los hijos h_1 y h_2 con un tiempo de ejecución de $O(n \log n)$. Por lo anterior el tiempo de ejecución del algoritmo *CommunFeatureCrossover* es de $O(n \log n)$.

Dado que el Algoritmo SSOC (ver Algoritmo 3) ejecuta $\mu/2$ veces el Algoritmo 4, donde $\mu \leq n_i$; entonces el tiempo de ejecución en el peor de los casos para SSOC es de $T(n_i) = O(n_i n \log n)$.

Mutación

La mutación es un operador de explotación, es decir, permite realizar pequeñas variaciones en los cromosomas de los hijos con una probabilidad p_m con el objetivo de encontrar mejores soluciones. Si el individuo es seleccionado para la mutación entonces elegimos k números para agregar o eliminar en su cromosoma, esto dependiendo si los números están presentes o

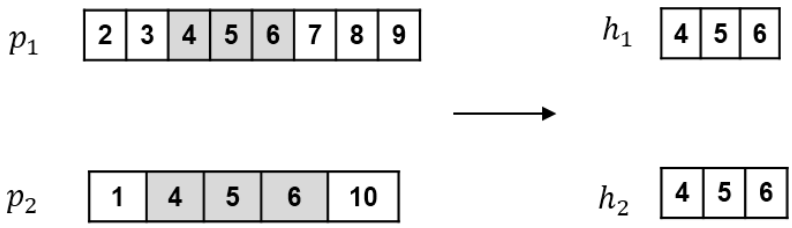


Figura 24: Algoritmo CFC. Pasos del 6 al 20: los padres heredan a los hijos las características que ambos tiene en común.

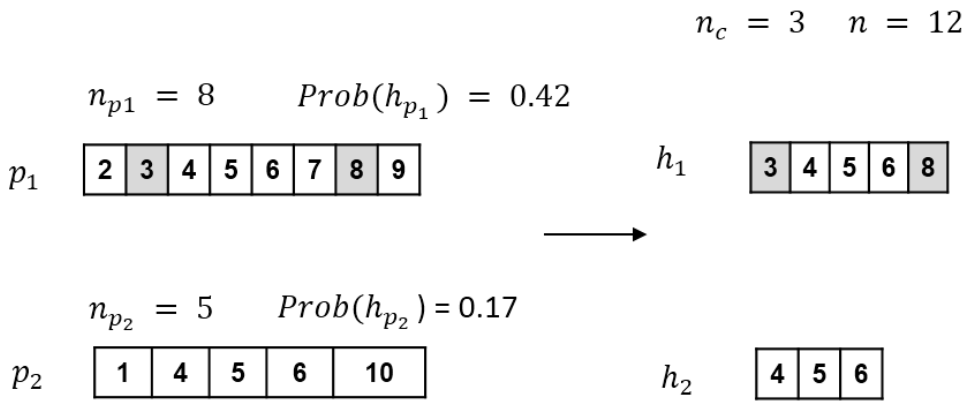


Figura 25: Algoritmo CFC. Pasos del 21 al 44: los padres heredan a los hijos las características que ambos no tiene en común con una probabilidad $Prob(h_{p_i})$. En este ejemplo, el h_2 no hereda más característica debido a que la probabilidad del h_{p_2} es muy pequeña.

ausentes en el cromosoma. Con el objetivo de que el cromosoma sufra pequeñas variaciones, k toma el tamaño desde 1 hasta el 10 % de las n características totales (ver Figura 26).

Dado que el tamaño de un hijo h_i es a lo más $O(n)$, donde n es el número de características, entonces el tiempo de ejecución en el peor de los casos para el Algoritmo 6 (INDEL) es de $O(n)$. Esto es debido a que el tiempo ejecución en el peor de los casos para eliminar un elemento del hijo h_i es de $O(n)$. Por otra parte, el Algoritmo 5 (k-INDELs) ejecuta el Algoritmo 6 k veces con una probabilidad p_m, entonces el tiempo de ejecución para realizar el proceso de mutación a un hijo es de $O(p_m kn)$. Por último, tomando en cuenta que el proceso de mutación se realiza para λ hijos entonces el tiempo de ejecución en el peor de los casos para el algoritmo de mutación es de $O(\lambda p_m kn)$.

Algoritmo 3 SSOCF para el cruzamiento de los padres.

Entrada: P arreglo de $\mu < p_1, p_2, ..., p_\mu >$
n número total características
p_c probabilidad de cruzamiento
Salida: O arreglo de hijos de tamaño λ, donde $\lambda = \mu$

```
1: O[1,...,λ] es un nuevo arreglo
2: para i ← 2 hasta P.length hacer
3:    Generar un número r de una distribución uniforme en el intervalo [0,1)
4:    si r < p_c entonces
5:       communFeatureCrossover(P[i − 1], P[i], O, n)
6:    si no
7:       O[i − 1] = P[i − 1]
8:       O[i] = P[i]
9:    fin si
10: fin para
```

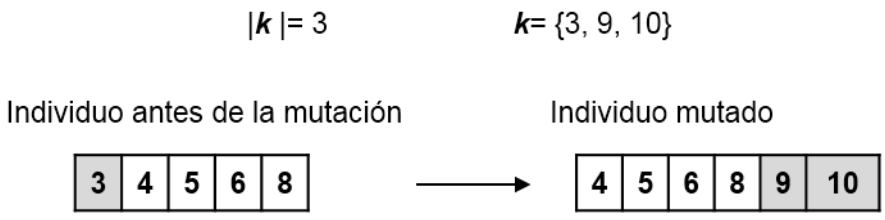


Figura 26: Operador de mutación k-INDELs.

Algoritmo 4 CommunFeatureCrossover (CFC)

Entrada: padres p_{i-1}, p_i seleccionados para cruzamiento
arreglo O de hijos de tamaño λ
n número total características
Salida: o_{i-1}, o_i hijos i, $i-1$

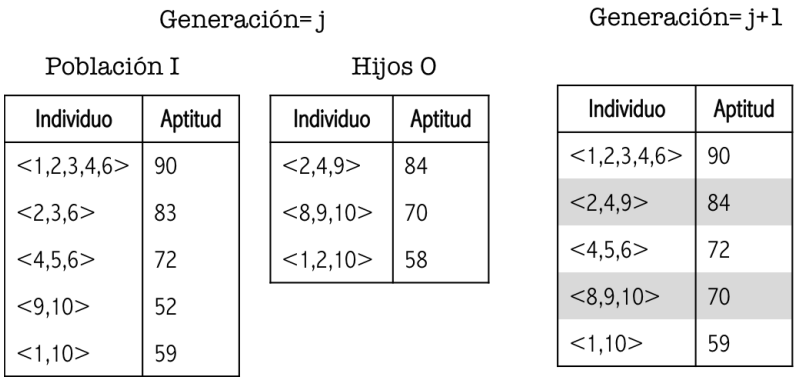
1: $p_1 = \text{mín}(p_{(i-1)}.n_p, p_i.n_p)$
2: $p_2 = \text{máx}(p_{(i-1)}.n_p, p_i.n_p)$
3: $j = 1$
4: $auxaleloj = 1$
5: // Agregar a h_1 y h_2 los elementos comunes de p_1 y p_2
6: **para** $aleloi \leftarrow 1$ hasta $p1.n_p$ **hacer**
7: **para** $aleloj \leftarrow auxaleloj$ hasta $p2.n_p$ **hacer**
8: **si** p_1.genotype[$aleloi$] =p_2.genotype[$aleloj$] **entonces**
9: h_1.genotype[i] =p_1.genotype[$aleloi$]
10: h_2.genotype[i] =p_2.genotype[$aleloj$]
11: $auxaleloj = aleloj + 1$
12: $aleloj = p2.n_p$
13: $j++$
14: **si no**
15: **si** p_2.genotype[$aleloj$] > p_1.genotype[$aleloi$] **entonces**
16: $aleloj = p2.n_p$
17: **fin si**
18: **fin si**
19: **fin para**
20: **fin para**
21: $c = j$
22: $n_c = j - 1$ // número de características comunes
23: $prob1 = (p_1.n_p - n_c)/n$
24: $prob2 = (p_2.np - n_c)/n$
25: p_1.genotype = eliminarAlelosComunes(p_1.genotype, $h1$.genotype)
26: p_2.genotype = eliminarAlelosComunes(p_2.genotype, $h2$.genotype)
27: **para** $aleloi \leftarrow 1$ hasta $p1.n_p$ **hacer**
28: Generar un número r de una distribución uniforme en el intervalo [0,1)
29: **si** $r < prob1$ **entonces**
30: h_1.genotype[j] =p_1.genotype[$aleloi$]
31: $j++$
32: **fin si**
33: **fin para**
34: **para** $aleloj \leftarrow 1$ hasta $p2.n_p$ **hacer**
35: Generar un número r de una distribución uniforme en el intervalo [0,1)
36: **si** $r < prob2$ **entonces**
37: h_2.genotype[c] =p_2.genotype[$aleloj$]
38: $c++$
39: **fin si**
40: **fin para**
41: Ordenar(h_1)
42: Ordenar (h_2)
43: O[$i-1$] = h_1
44: O[i] = h_2

Algoritmo 5 K-INDELs para la mutación de los hijos

Entrada: hijo h_i seleccionado para mutación
n número total características
p_m probabilidad de mutación
Salida: hijo h_i mutado
1: Generar un número entero k en el intervalo $(0, 0.10n]$
2: **para** $i \leftarrow o$ hasta k **hacer**
3: INDEL(h_i, p_m, n)
4: **fin para**

Algoritmo 6 INDEL para la mutación de un alelo en el cromosoma del hijo

Entrada: hijo h_i seleccionado para mutación
n número total características
p_m probabilidad de mutación
Salida: hijo h_i mutado
1: Generar un número entero r de una distribución uniforme en el intervalo $[0, 1)$
2: **si** $r < p_m$ **entonces**
3: Generar un número entero j en un intervalo $(0, n]$
4: **si** h_i.genotype contiene el número j **entonces**
5: Eliminar el número j de h_i.genotype
6: **si no**
7: Agregar el número j a h_i.genotype
8: **fin si**
9: **fin si**



Individuo	Aptitud
<1,2,3,4,6>	90
<2,3,6>	83
<4,5,6>	72
<9,10>	52
<1,10>	59

Individuo	Aptitud
<2,4,9>	84
<8,9,10>	70
<1,2,10>	58

Individuo	Aptitud
<1,2,3,4,6>	90
<2,4,9>	84
<4,5,6>	72
<8,9,10>	70
<1,10>	59

Figura 27: selección de los sobrevivientes.

Selección de los sobrevivientes

Para la selección de los n_i individuos que sobrevivirán en la siguiente generación de la población I ($|I| = n_i$) y los hijos O ($|O| = \lambda$), se utilizó el método de reemplazar al peor. Este método consiste en ordenar de manera descendente a los individuos I y O de acuerdo a su aptitud, cada individuo o_i de O recorre la población I en orden descendente, si existe un individuo j que tenga una aptitud menor a la de o_i, entonces o_i reemplaza a j en la población I (Ver Figura 27).

El tiempo de ejecución en el peor de los casos para seleccionar a los sobrevivientes es de $O(n_i)$.

Análisis del algoritmo genético

En la Tabla 11 se muestra el tiempo de ejecución para cada procedimiento que forma parte del algoritmo genético, donde el peor tiempo de ejecución es de $O(n_i n \log n)$. Si se toma en cuenta que este procedimiento se ejecuta un número de generaciones n_g, entonces el tiempo de ejecución del algoritmo genético en el peor de los casos es $O(n_g n_i n \log n)$.

Tabla 11: Tiempo de ejecución en el peor de los casos para cada uno de los procedimientos que conforman el algoritmo genético para la selección de características.

Algoritmo	Tiempo de ejecución
Inicializar a la población	$O(n_i n \log n)$
Seleccionar a los padres	$O(n_i)$
Cruzamiento	$O(n_i n \log n)$
Mutación	$O(p_m k n)$
Seleccionar los sobrevivientes	$O(n_i)$

Capítulo 5. Pruebas y resultados

En este capítulo se presenta las pruebas realizadas a los algoritmos propuestos en el Capítulo 4, Sección 4.1, en donde se describen las configuraciones de los algoritmos, el hardware y software de implementación. Además, se muestran los resultados y análisis a las soluciones del algoritmo propuesto, incluyendo la comparación de nuestra mejor solución con los resultados obtenidos por métodos del estado del arte descritos en la Sección 2.2.3.

5.1. Conjunto de prueba y validación

Para realizar los experimentos se utilizaron los conjuntos de datos AMP y Antibac, recuperados mediante la metodología descrita en la Sección 4.1. Cada conjunto de datos está compuesto por 3000 péptidos (1500 AMPs y 1500 no AMPs; 1500 antibacterianos y 1500 no antibacterianos) de los cuales, se seleccionaron de manera aleatoria el 90 % para entrenamiento y 10 % para pruebas. Por otro lado, para evaluar la calidad de los resultados se utilizaron dos conjuntos de validación: el primero, para la validación del modelo AMP, este conjunto está compuesto por 202 AMPs y 384 no AMPs; el segundo conjunto es para la validación del modelo Antibac, compuesto por 714 péptidos antibacterianos y 707 péptidos sin actividad antibacteriana conocida. En el Apéndice B, se muestran los identificadores que forman parte del conjunto de entrenamiento, prueba y validación.

Los descriptores moleculares se calcularon para todos los péptidos bajo estudio (*i.e.*, 3000 del conjunto AMP y 3000 del conjunto Antibac). Para calcular los descriptores constitucionales y dependientes de la secuencia (*i.e.*, 0D y 1D) se utilizó el programa JPeDes, como resultado se obtuvieron los conjuntos de datos AMP y Antibac representados por 28 características (los descriptores se muestran en el Apéndice C, Tabla 33), por convención llamaremos a los datos representados de esta forma AMP_A y Antibac_A. Por otra parte, para calcular los descriptores moleculares de dimensión 0D, 1D y 2D se utilizó el programa PaDel-Descriptor (Yap, 2011). En esta parte, es importante mencionar que existen características similares para ambas clases (*i.e.*, actividad biológica deseada y no deseada), por lo tanto se eliminaron las características con ganancia de información igual a cero. Como resul-

Tabla 12: Conjuntos de prueba, entrenamiento y validación para el Algoritmo 1.

Conjunto de datos	**Número de características**	**Entrenamiento**	**Prueba**	**Validación**
AMP_A	28	2700	300	586
AMP_B	253			
AMP_A+B	278			
Antibac_A	28			1421
Antibac_B	315			
Antibac_A+B	337			

tado de este procedimiento obtuvimos los conjuntos de datos AMP y Antibac representados por 253 y 315 características, respectivamente (los índices de los descriptores se muestran en el Apéndice C, Tabla 32). De aquí en adelante llamaremos a estos conjuntos AMP_B y Antibac_B.

Además, se combinaron las características disponibles en los programas JPeDes y PaDel-Descriptor (Yap, 2011) para representar el conjunto AMP y Antibac en características, excluyendo las características con valor de ganancia de información igual a cero con respecto a la actividad biológica. Como resultado se obtuvieron los conjuntos de datos AMP y Antibac representados por 278 y 337 características, por convención llamaremos a estos conjuntos AMP_A+B y Antibac_A+B, respectivamente (los índices de los descriptores se muestran en el Apéndice C, Tabla 34).

Es importante señalar en este punto que los conjuntos AMP_A, AMP_B y AMP_A+B son los mismos, lo único que cambia son los descriptores moleculares con los que se representan. De igual modo ocurre con Antibac_A, Antibac_B y Antibac_A+B.

Un resumen del conjunto de datos que se utilizarán en los siguientes experimentos se muestra en la Tabla 12.

5.2. Configuración de los algoritmos

En esta sección se describen los parámetros de configuración para los algoritmos presentados en el Capítulo 4, Sección 4.3.

Tabla 13: Configuración del algoritmo genético para el problema de selección de características.

Parámetros	**Valor**
Inicialización de la población	Aleatoria
Selección de padres	Torneo binario
Cruzamiento	SSOCF
Mutación	k-INDELs
Selección de los sobrevivientes	Reemplazar al peor
Generaciones sin mejora	10 % del número de generaciones

Los experimentos se realizaron bajo el sistema operativo Windows 7, versión Home Premium 64-bits, en una computadora con procesador Intel (TM) Core (R) i7 de 3.6 GHz de velocidad y memoria RAM de 8 GB.

5.2.1. Algoritmo genético para la selección de características (GAFS)

El algoritmo genético para el problema de selección de características GAFS (ver Algoritmo 1) se codificó e implementó en Java 1.7.0 usando NetBeans 7.3.1. Las especificaciones del algoritmo se explican a continuación: la inicialización de la población es de manera aleatoria; la selección de padres se realizó mediante torneo binario para elegir a los mejores individuos; para el operador de cruzamiento se utilizó el algoritmo SSOCF (*subset size-oriented common feature*) con una probabilidad p_c; la mutación se realizó mediante el algoritmo k-INDELs con una probabilidad p_m; por último la estrategia para la selección de los sobrevivientes se realizó mediante el reemplazo del peor individuo de la población (ver Tabla 13).

Con el objetivo de encontrar la mejor configuración para los parámetros del algoritmo genético se aplicó un proceso de prueba y error. Este proceso consiste en ejecutar un determinado número de veces el algoritmo genético utilizando diferentes configuraciones, posteriormente, se selecciona la configuración con el mejor resultado en promedio para un conjunto de datos específico. En el presente trabajo, se ejecutó 4 veces el algoritmo genético dada una configuración, los parámetros que se variaron por ejecución fueron: número de generaciones n_g, número de individuos n_i, número de padres μ, probabilidad de cruzamiento p_c, probabilidad de mutación p_m. La lista de las ejecuciones del algoritmo genético con cada configuración se muestra en la Sección D.1 del Apéndice D.

En la Tabla 14 se muestra la configuración que obtuvo los mejores resultados para cada conjunto de datos, esta configuración se seleccionó para realizar los experimentos que siguen.

Tabla 14: Parámetros de configuración para el algoritmo genético.

	Conjunto de datos		
Parámetros	AMP_A Antibac_A	AMP_B	AMP_A+B Antibac_B Antibac_A+B
Número máximo de generaciones n_g	500	600	550
Número de generaciones sin mejora n_{gwi}	50	60	55
Número de individuos en la población n_i	100	200	300
Número de padres μ	100	200	300
Número de hijos λ	100	200	300
Probabilidad de cruzamiento p_c	0.8	0.8	0.8
Probabilidad de mutación p_m	0.3	0.8	0.5

5.2.2. Máquina de soporte vectorial

La máquina de soporte vectorial se implementó utilizando las librerías para Java de Weka 3.6.10 (Hall *et al.*, 2009) y LIBSVM 3.18 (Chang y Lin, 2011). En la Tabla 15 se muestran los parámetros de configuración utilizados para los programas Weka y LIBSVM.

5.3. Ganancia de información

La ganancia de información (GI) es un criterio estadístico que mide qué tan bien una característica separa las clases dado un conjunto de datos. En este caso la GI se utilizó para jerarquizar las características y seleccionar las k mejores de acuerdo con un umbral θ para

Tabla 15: Parámetros de configuración para la máquina de soporte vectorial.

Librería	**Parámetro**	**Valor**
LIBSVM 3.18	Tipo de SVM:	C_SVM
	Tipo de Kernel:	Lineal
	Normalizar:	si
	Shrinking:	no
	Debug:	no
	Costo	10
Weka 3.6.10	Entrenamiento:	90 %
	Prueba:	10 %

construir el clasificador.

Los objetivos de este experimento son:

- Evaluar el poder predictivo que tienen las k mejores características para clasificar los péptidos con actividad biológica deseada, por ejemplo: AMP y no AMPs, antibacterianos y no antibacterianos.
- Comparar el mejor resultado obtenido con ganancia de información tras variar el umbral θ de ganancia de información con el mejor el mejor resultado obtenido con el algoritmo genético.

El procedimiento de este experimento se muestra a detalle en el Apéndice E.

5.4. Resultados

A continuación se muestran los resultados obtenidos de los experimentos realizados.

5.4.1. Ganancia de información

Para el experimento de ganancia de información (GI) (ver subsección 5.3) los resultados son los siguientes.

En las figuras 28, 29, 30, 31, 32 y 33 se muestra el comportamiento de la exactitud y números de características en función del umbral de GI para los conjuntos AMP y Antibac representados en características. Cada gráfica tiene en el eje principal Y (margen izquierdo) la exactitud del clasificador y en el eje secundario Y (margen derecho) el número de características. Los resultados muestran que conforme aumenta el umbral θ de ganancia de información, el número de descriptores moleculares seleccionados disminuye, obteniendo sólo aquellos descriptores que tienen un mayor poder predictivo de manera individual con respecto a la clase (*i.e.,* AMP y no AMP o antibacteriano y no antibacteriano). En lo que se refiere a la exactitud del clasificador este es variable con respecto a θ, por lo tanto elegir los descriptores con mayor ganancia de información no necesariamente asegura el mejor resultado en la predicción de las clases.

Tabla 16: Resultado de las mejores soluciones obtenidas utilizando ganancia de información para el conjunto de datos AMP.

Conjunto de datos	Umbral θ	Número de características	Exactitud (ACC)
AMP_A	0.002	25	90.33
AMP_B	0.022	92	90.33
AMP_A+B	0.019	122	91

Tabla 17: Resultado de las mejores soluciones obtenidas utilizando ganancia de información para el conjunto de datos Antibac.

Conjunto de datos	Umbral θ	Número de características	Exactitud (ACC)
Antibac_A	0.02	19	83.67
Antibac_B	0.02	144	90
Antibac_A+B	0.027	141	89.67

Los mejores resultados en exactitud que se obtuvieron son: 91 % para la predicción de AMPs utilizando el conjunto de datos AMP_A+B (ver Tabla 16); 90 % de exactitud para la predicción de péptidos antibacterianos utilizando el conjunto Antibac_B (ver Tabla 17).

5.4.2. Algoritmo genético para la selección de características (GAFS)

Para los conjuntos de datos AMP y Antibac representados en distintos grupos de características (*i.e.,* AMP_A, AMP_B, AMP_A+B y Antibac_A, Antibac_B, Antibac_A+B) se ejecutó el algoritmo de selección de características (GAFS) 30 veces. En las tablas 18 y 19 se muestran la calidad promedio de las mejores soluciones (subconjuntos) encontradas por GAFS para cada conjunto de prueba. Los criterios que se tomaron en cuenta para la mejor solución encontrada por ejecución son: número de características y aptitud. En general, las soluciones de GAFS son muy diversas con respecto al número de características seleccionadas, sin embargo, las soluciones presentan una aptitud similar.

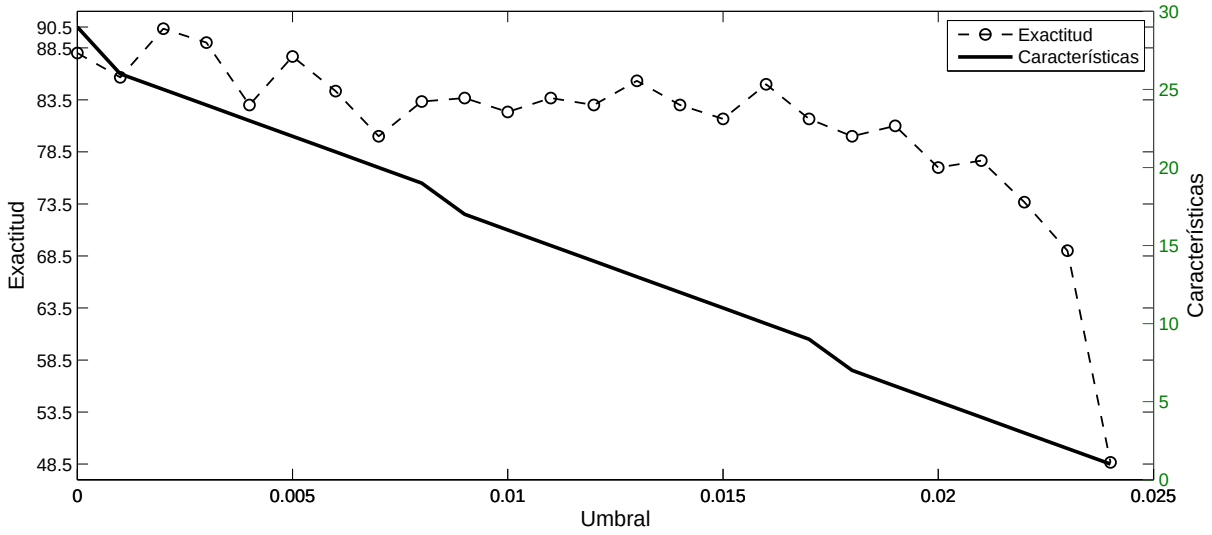


Figura 28: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos AMP_A.

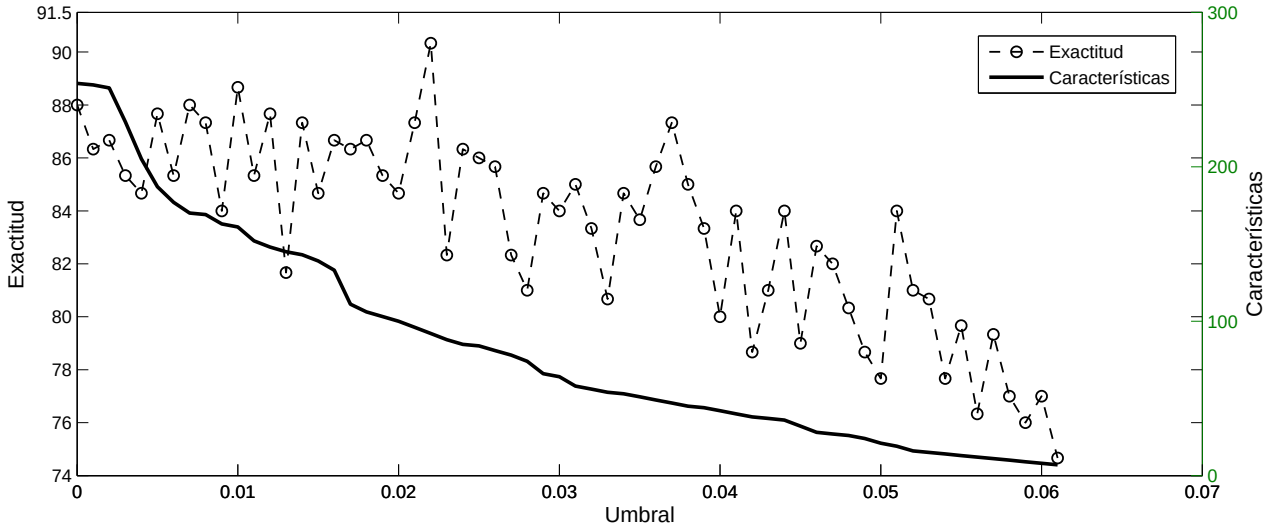


Figura 29: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos AMP_B.

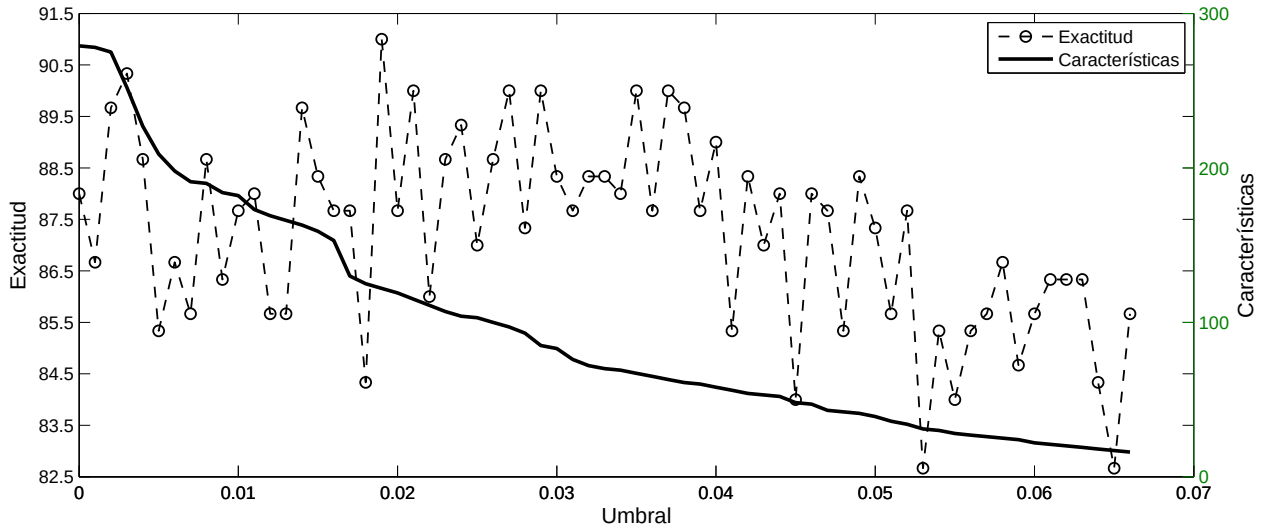


Figura 30: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos AMP_A+B.

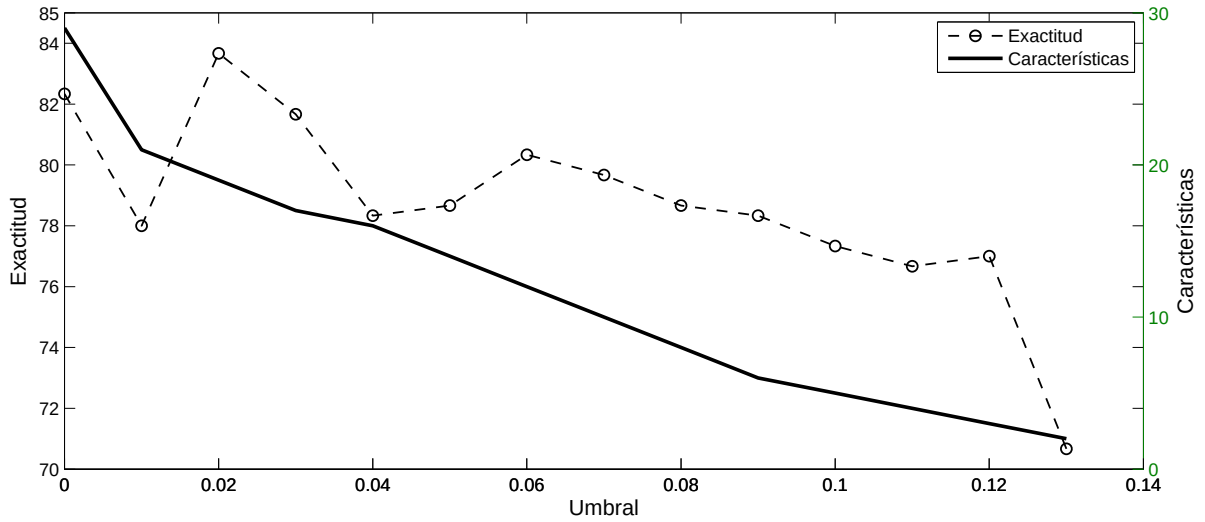


Figura 31: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos Antibac_A.

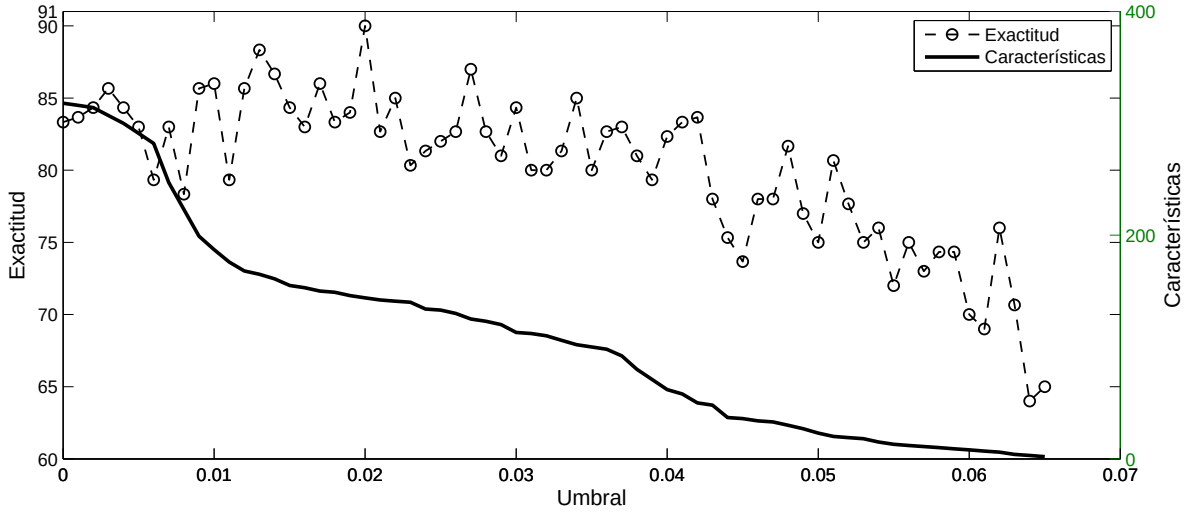


Figura 32: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos Antibac_B.

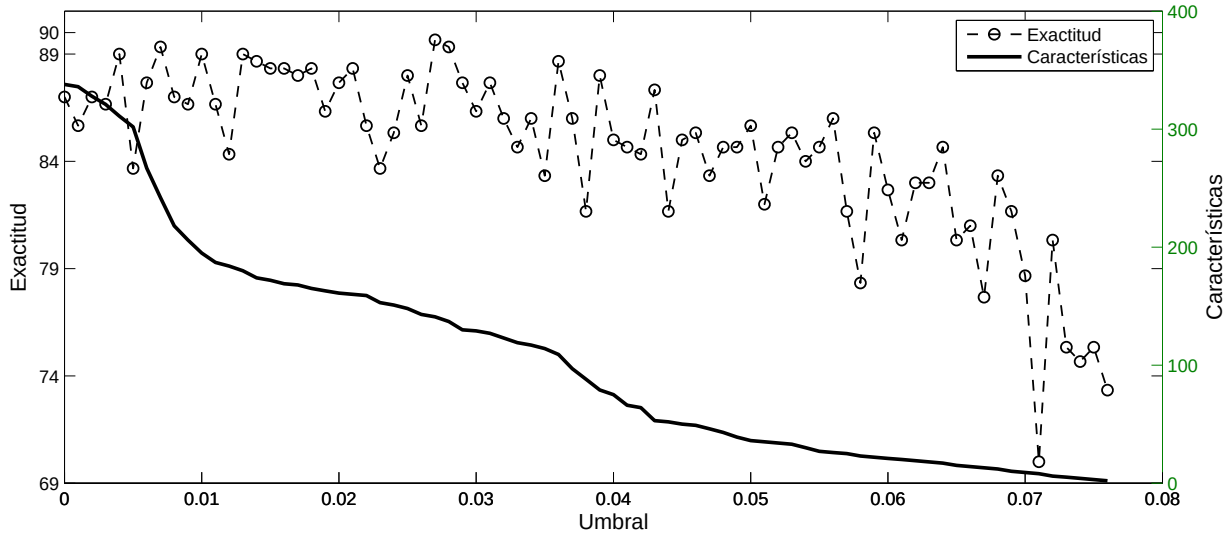


Figura 33: Exactitud y número de características en función del umbral de ganancia de información para el conjunto de datos Antibac_A+B.

Por otra parte, para evaluar si GAFS cumple con el objetivo de disminuir las características y mejorar la aptitud, se compararon los resultados antes y después de seleccionar las características (ver figuras 34 y 35).

Los resultados para el conjunto AMP muestran que el algoritmo disminuye un 40 % el número de características con respecto al total y logra un aumento de 6 % de aptitud con respecto a la aptitud obtenida al utilizar todas las características. El algoritmo GAFS para el conjunto Antibac disminuye en un 50 % el número de características y en lo que respecta a la aptitud se logra un aumento de al menos un 7 % con respecto a la aptitud obtenida al utilizar todas las características. Por lo tanto, GAFS obtiene una mejor aptitud comparado con la aptitud lograda usando todas las características.

Además se realizó una comparación entre las soluciones promedio de GAFS y las mejores soluciones de ganancia de información. Las soluciones de GAFS superan a las de ganancia de información para los conjuntos de datos AMP y Antibac (ver figuras 34 y 35). Por lo tanto, es mejor seleccionar un subconjunto de descriptores que juntos tengan un buen desempeño predictivo, aunque por separado estos descriptores no sean útiles con respecto al criterio de ganancia de información.

La mejor aptitud encontrada para el algoritmo GAFS para el conjunto AMP, fue usando la representación de características AMP_A+B. Se obtuvo un 97.59 de aptitud y se seleccionaron 94 características (ver Tabla 20). El comportamiento general de GAFS para encontrar el mejor individuo para el conjunto AMP se muestra en las figuras 36, 37 y 38.

Para el conjunto de Antibac, la representación de características Antibac_A+B obtuvo la mejor aptitud en GAFS con 94.51 y 128 características seleccionadas (ver Tabla 21). El comportamiento general de GAFS para encontrar el mejor individuo para el conjunto Antibac se muestra en las figuras 39, 40 y 41.

Tabla 18: Calidad promedio de las mejores soluciones en términos de la función de aptitud del algoritmo GAFS para el conjunto de datos AMP.

Conjunto de datos	**Número total de características**	**Nro. promedio de características seleccionadas**	**Desv. std. promedio de características seleccionadas**	**ACC promedio (Aptitud)**	**Desv. std. aptitud promedio**
AMP_A	28	11.1	5.67	91.4 (92.5)	0.6
AMP_B	253	117.65	43.41	92.05 (93.16)	0.59
AMP_A+B	278	125.1	55.37	94.47 (95.62)	0.62

Tabla 19: Calidad promedio de las mejores soluciones en términos de la función de aptitud del algoritmo GAFS para el conjunto de datos Antibac.

Conjunto de datos	**Número total de características**	**Número promedio de características seleccionadas**	**Desv. std. promedio de características seleccionadas**	**ACC promedio (Aptitud)**	**Desv. std. aptitud promedio**
Antibac_A	28	10.67	6.39	89.2 (90.50)	0.91
Antibac_B	315	157.35	61.73	90.5 (91.63)	0.52
Antibac_A+B	337	164.21	75.59	93 (94.05)	0.29

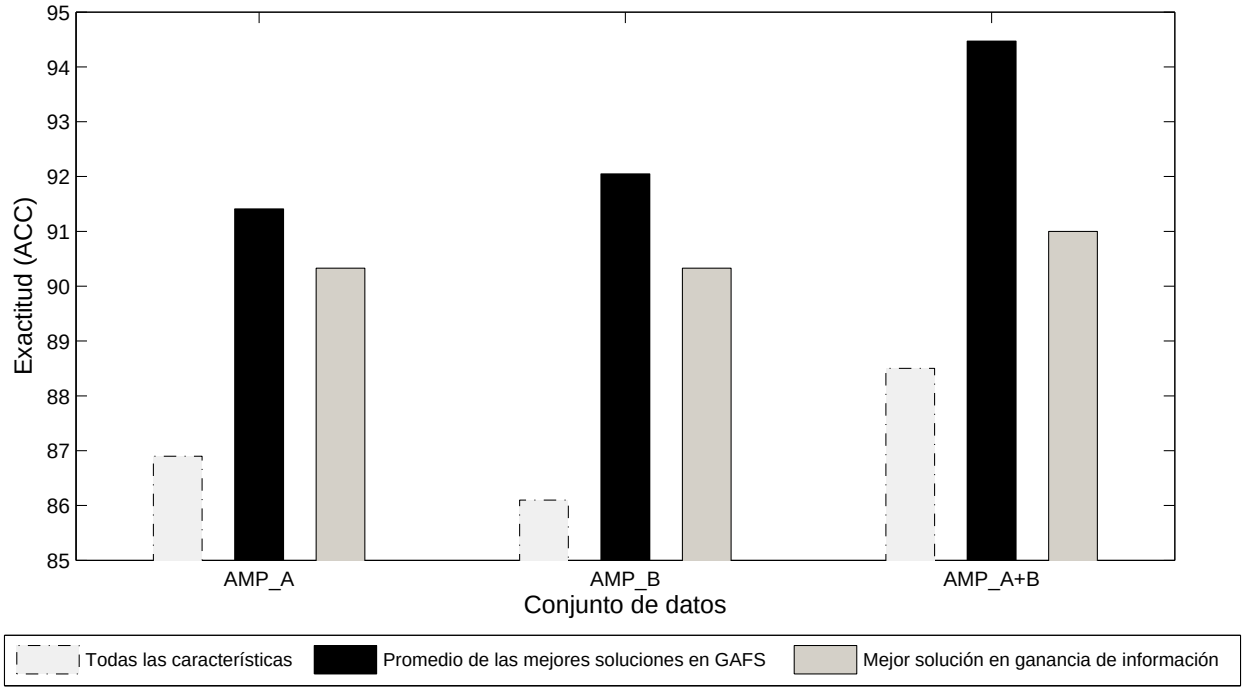


Figura 34: Comparación entre los conjuntos de datos antes y después de aplicar el algoritmo de selección de características GAFS para el conjunto de datos AMP.

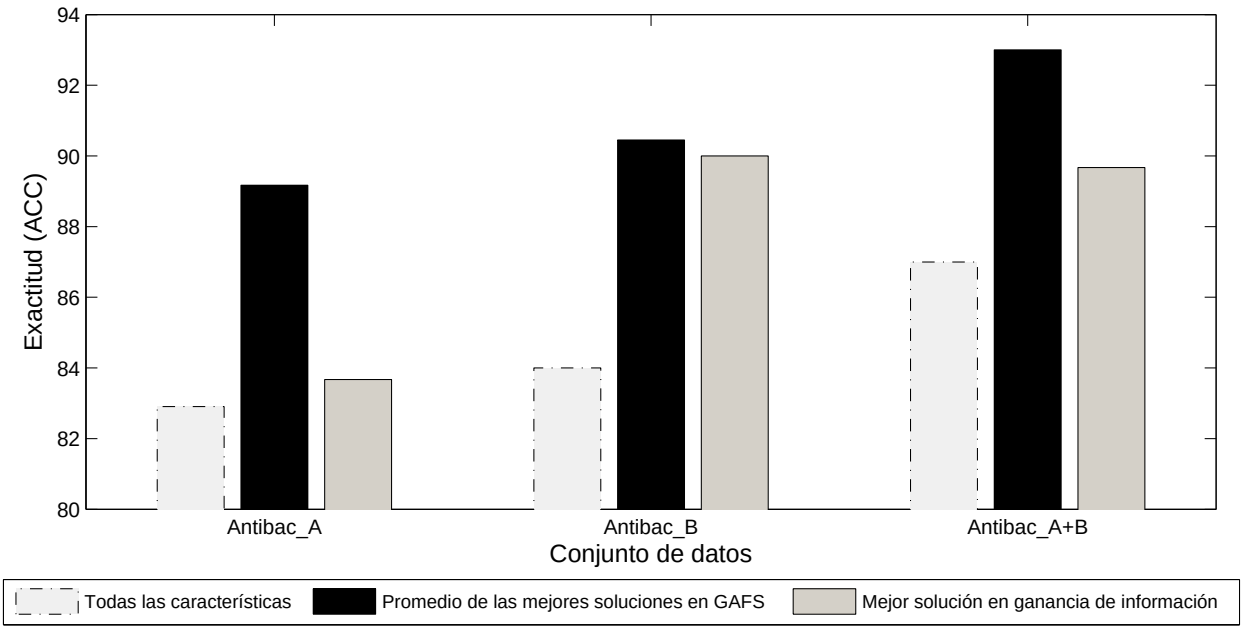


Figura 35: Comparación entre los conjuntos de datos antes y después de aplicar el algoritmo de selección de características GAFS para el conjunto de datos Antibac.

Tabla 20: Lista de las mejores soluciones encontradas por el algoritmo GAFS para el conjunto de datos AMP.

Conjunto de datos	**Nro. de características seleccionadas**	**ACC (Aptitud)**
AMP_A	9	93.7 (94.88)
AMP_B	71	93.3 (94.56)
AMP_A+B	94	96.3 (97.59)

Tabla 21: Lista de las mejores soluciones encontradas por el algoritmo GAFS para el conjunto de datos Antibac.

Conjunto de datos	**Nro. de características seleccionadas**	**ACC (Aptitud)**
Antibac_A	18	92 (93.02)
Antibac_B	217	92 (93.00)
Antibac_A+B	128	93.3 (94.51)

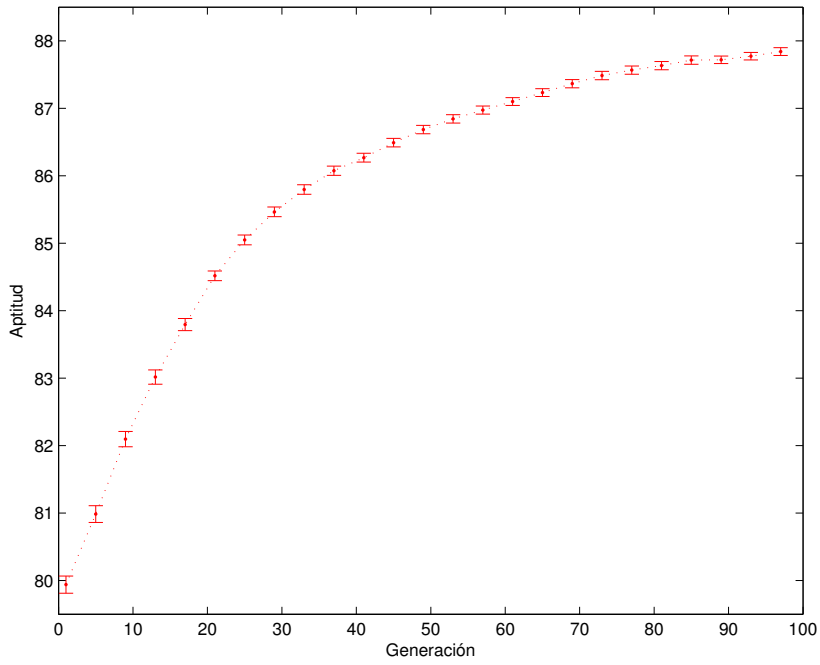


Figura 36: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos AMP_A.

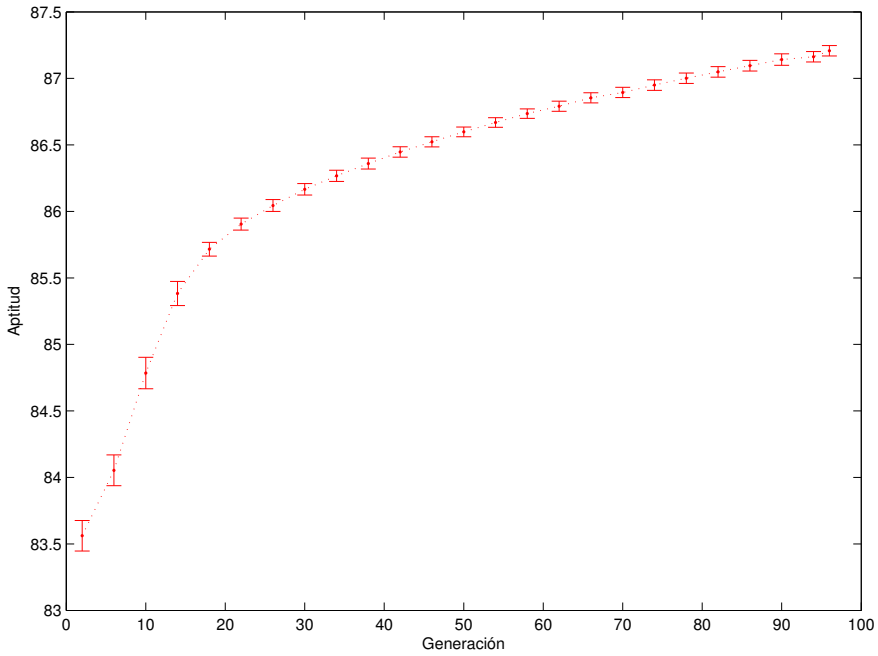


Figura 37: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos AMP_B.

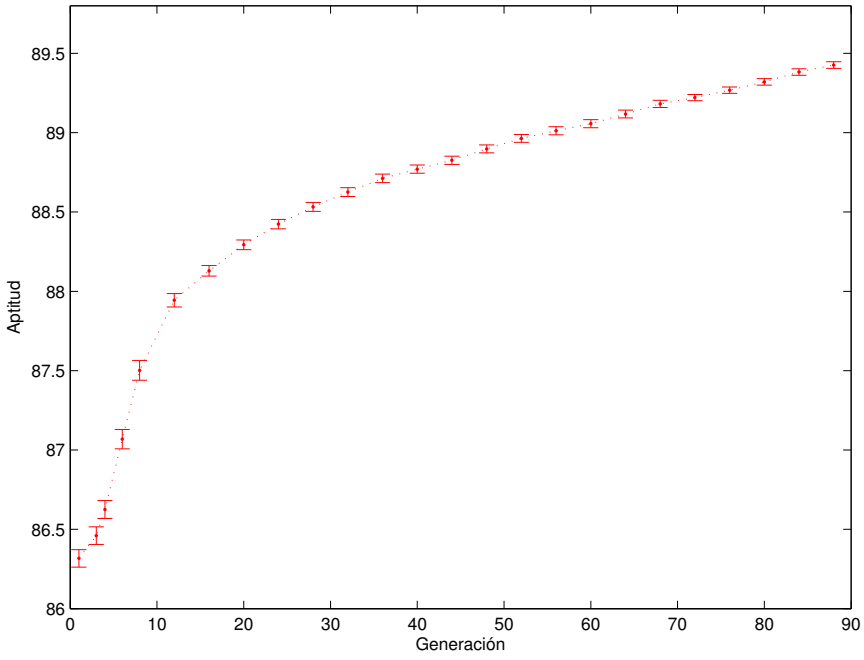


Figura 38: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos AMP_A+B

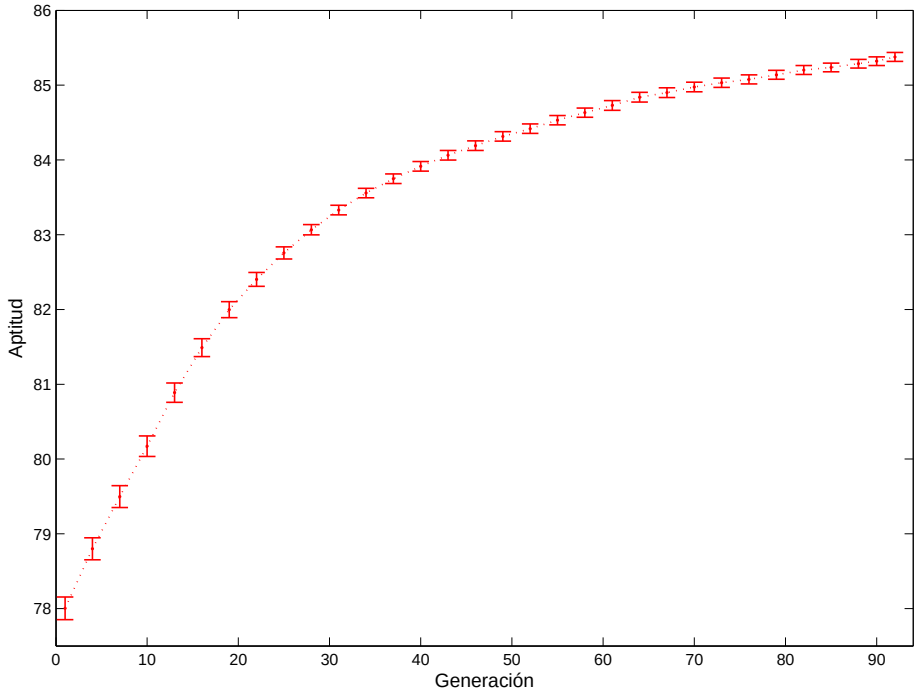


Figura 39: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos Antibac_A.

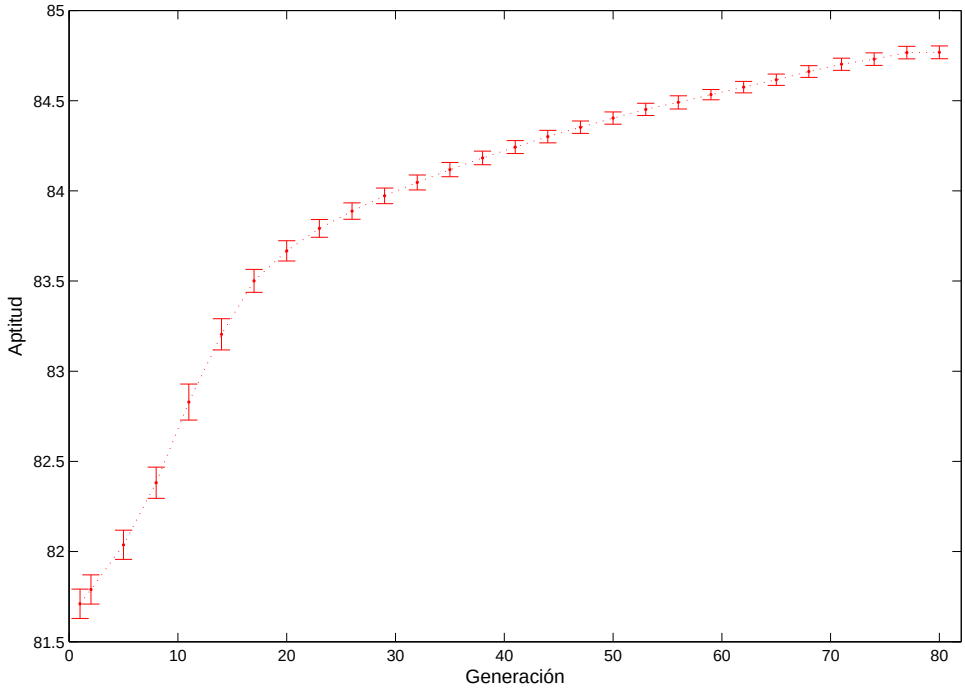


Figura 40: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos Antibac_B.

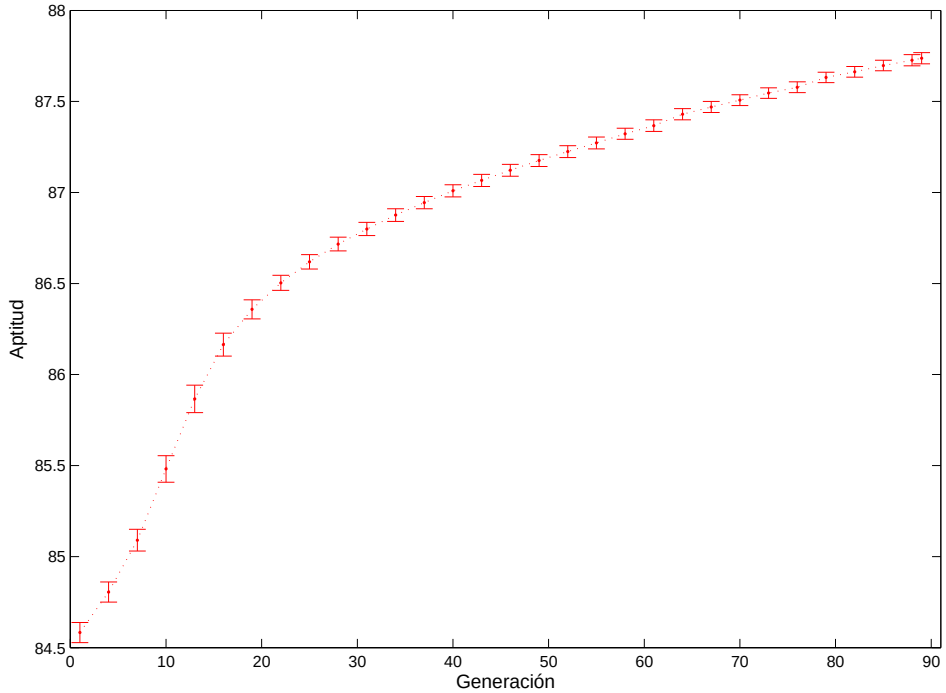


Figura 41: Aptitud promedio de la población con un 95 % de intervalo de confianza para el algoritmo genético utilizando el conjunto de datos Antibac_A+B.

Tabla 22: Tiempo promedio para encontrar el mejor subconjunto de características en el algoritmo GAFS.

Conjunto de datos	**Tiempo GA promedio (min)**	**Tiempo de evaluación promedio por individuo (ms)**
AMP_A	38.033	44.967
AMP_B	310.357	154.5
AMP_A+B	635.497	231.93
Antibac_A	39.096	46.333
Antibac_B	733.005	265.542
Antibac_A+B	733.052	265.577

Tiempo de ejecución

En la Tabla 22 se presenta el tiempo de ejecución promedio de GAFS para los conjuntos de datos AMP y Antibac. Para calcular el tiempo promedio se tomaron las 30 ejecuciones de GAFS. Por cada ejecución se midió el tiempo que tarda el algoritmo desde su inicialización hasta que el criterio de paro se satisface (pasos del 1 al 9, Algoritmo 1), este tiempo se indica en la segunda columna de la tabla 22. Además, dentro del tiempo de GAFS se incluye el tiempo que tarda la evaluación del individuo, que implica construir, entrenar y probar el clasificador dado un subconjunto de características. El tiempo de evaluación del individuo se muestra en la tercera columna de la Tabla 22, este es el resultado de promediar lo siguiente: primero se obtiene el tiempo promedio de un individuo con respecto a su población. Con el resultado de todos los promedios de cada individuo se calcula el promedio por generación, a su vez, con los promedios de cada generación se calcula el promedio obtenido por todas las ejecuciones.

Características relevantes en AMPs

En las figuras 42, 43 y 44 se muestran los índices de las características que más se repiten al aplicar el algoritmo GAFS 30 veces para cada uno de los conjuntos de datos. Los nombres de los índices de los descriptores se describen en la Apéndice B. A continuación se presenta un análisis de la importancia de algunas de las características presentes en las mejores soluciones de GAFS y para las cuáles en la literatura se ha señalado de su importancia en los AMPs.

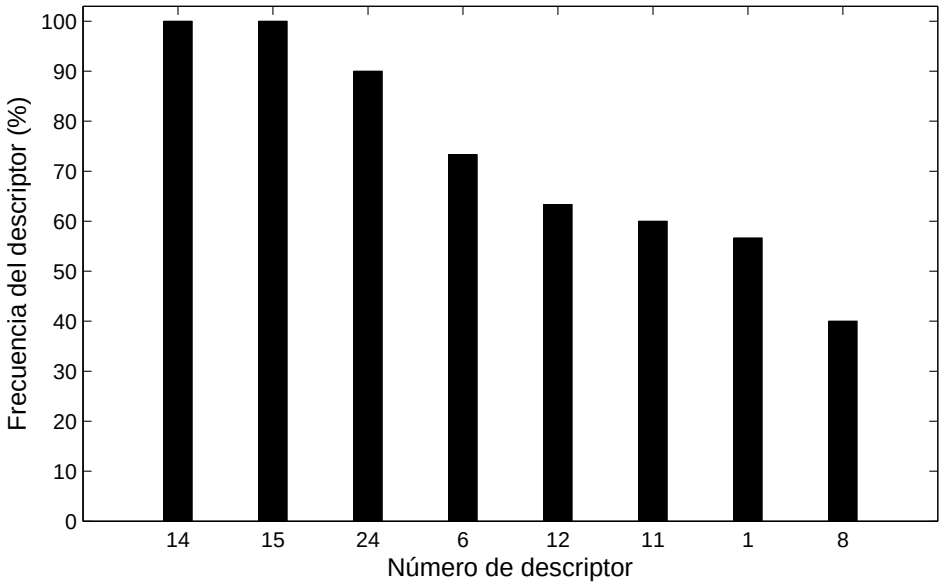


Figura 42: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos AMP_A.

Carga neta (Z): la carga es una propiedad importante en los AMPs, debido a que contribuye en la unión con la membrana mediante interacciones electrostáticas, la mayoría de los AMPs se caracterizan por tener una carga positiva (+2 a +4) (Yeaman y Yount, 2003). En el presente trabajo la carga se calculó a tres diferentes pH, debido a que la carga de algunos péptidos puede variar a diferentes pH, de acuerdo con Piotto *et al.* (2012), este parámetro es decisivo para las simulaciones de péptidos en tejidos específicos. Por otra parte, se puede destacar que la característica de carga a un pH 5 y pH 9 estuvo presente en las 30 mejores soluciones de GAFS para el conjunto de datos AMP_A y AMP_A+B (ver Figura 42, índices 14, 15 y Figura 44 índices 2, 1).

Momento Hidrofóbico (Hm_Hk(θ)): la mayoría de los AMP forman estructuras anfipáticas (*i.e.*, poseen regiones tanto hidrófilas como hidrófobas). Una de las conformaciones más sencillas en la que un AMP es anfipático es la α-hélice (hélice anfipática), ésta tiene una periodicidad de 3 a 4 residuos por hélice, es decir, cada residuo tiene que girar alrededor de 100 grados (Yeaman y Yount, 2003). Para medir de manera cuantitativa la anfipaticidad se utiliza el **momento hidrofóbico**. En el presente trabajo se calculó el momento hidrofóbico a diferentes ángulos debido a la variedad de estructuras secundarias que pueden adquirir los

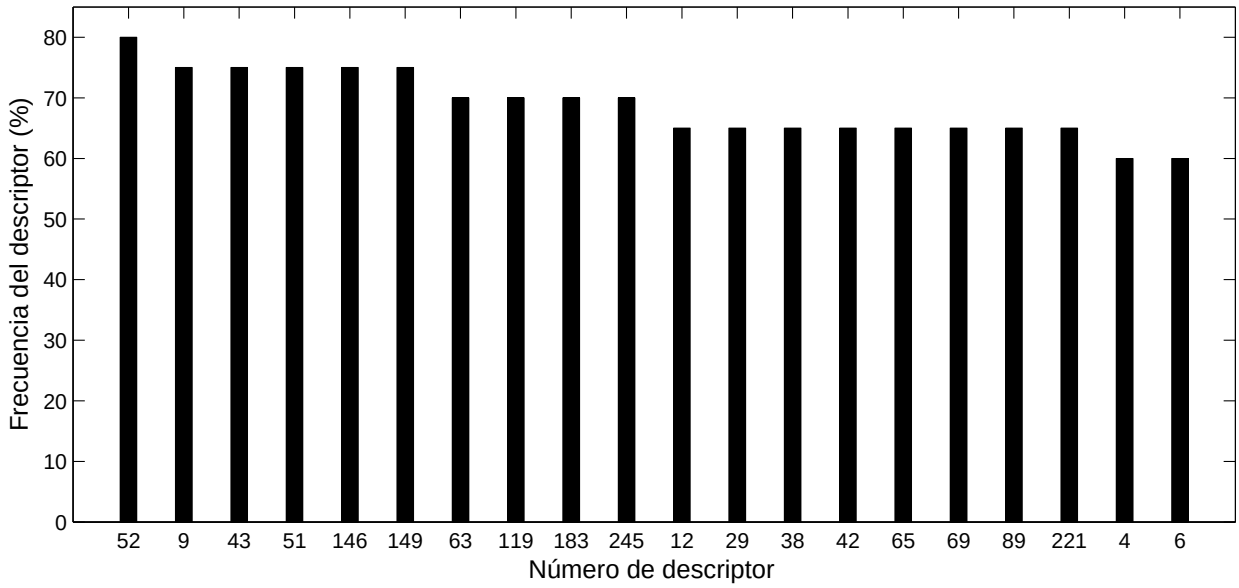


Figura 43: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos AMP_B.

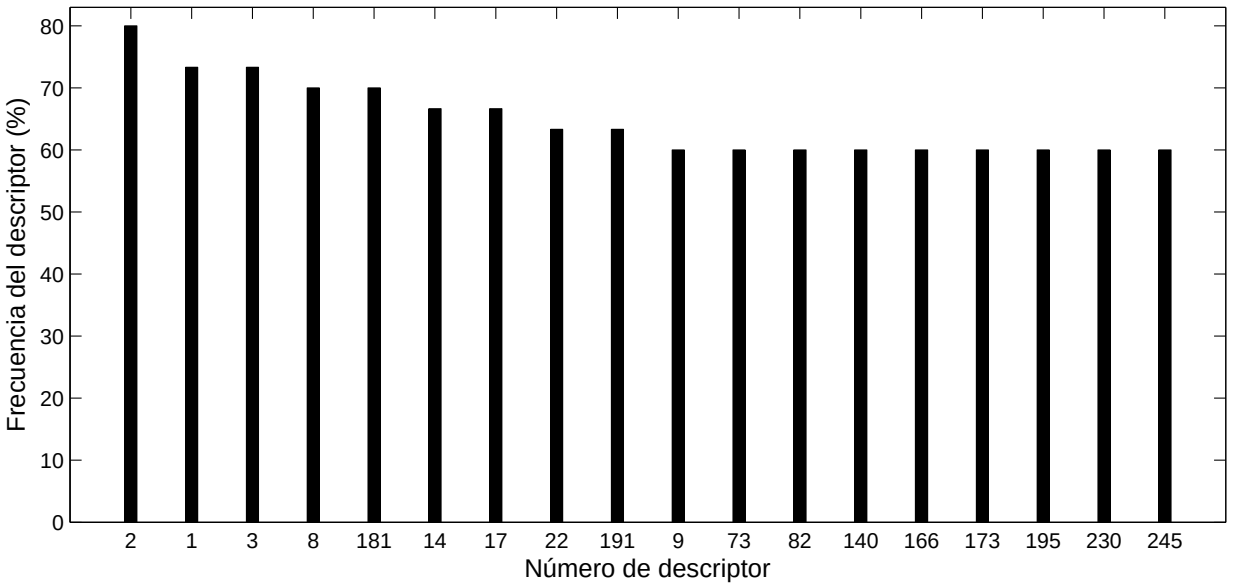


Figura 44: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos AMP_A+B.

AMPs. Por ejemplo cuando Hm_Hk($\theta = 100$) es cuando el AMP adquiere una estructura α-hélice, Hm_Hk($\theta = 180$) es cuando el AMP adquiere una estructura β-lámina y cuando Hm_Hk($\theta = 160$) es cuando las dos estructuras anteriores están incluidas en el AMP. La característica Hm_Hk($\theta = 100$) estuvo presente en las 30 mejores soluciones para el conjunto de datos AMP_A (índice 24, Figura 42) y para el conjunto de datos AMP_A+B estuvo presente un 66.6 % de las veces (índice 17, Figura 44).

Hidrofobicidad (Hk, He): la hidrofobicidad es una característica esencial para el plegamiento y la interacción del AMP con la membrana celular (Yeaman y Yount, 2003; Piotto *et al.*, 2012). La característica de hidrofobicidad (Hk) utilizando la escala de Kyte y Doolittle (1982) estuvo presente en 63.3 % de las mejores soluciones de GAFS, mientras que la hidrofobicidad (He) utilizando la escala de Eisenberg *et al.* (1984) fue de 60 % para el conjunto de datos AMP_A. En lo que respecta al conjunto de datos AMP_A+B, la hidrofobicidad utilizando la escala de Kyte y Doolittle (1982) y Eisenberg *et al.* (1984) estuvo presente un 70 % y 66.6 % de los casos, respectivamente.

Características relevantes en péptidos antibacterianos

Los descriptores que más se repiten en las mejores soluciones de GAFS para el conjunto Antibac se muestran en las figuras 45, 46 y 47. Los nombres de los índices se describen en el Apéndice B.

Para el conjunto Antibac con la representación de características Antibac_A, el descriptor molecular presente en las 30 ejecuciones del algoritmo GAFS fue la carga neta Z(pH7) (índice 14, Figura 45). La carga neta (ZpH9, índice 15, Figura 45) estuvo presente en el 86.6 % de las mejores soluciones y la hidrofobicidad (Hk, índice 12, Figura 45) utilizando la escala de Kyte y Doolittle (1982) estuvo presente en el 76.6 % de las soluciones. Estos tres descriptores moleculares también estuvieron presentes con una frecuencia similar en las mejores soluciones del algoritmo GAFS para el conjunto AMP (ver Figura 42).

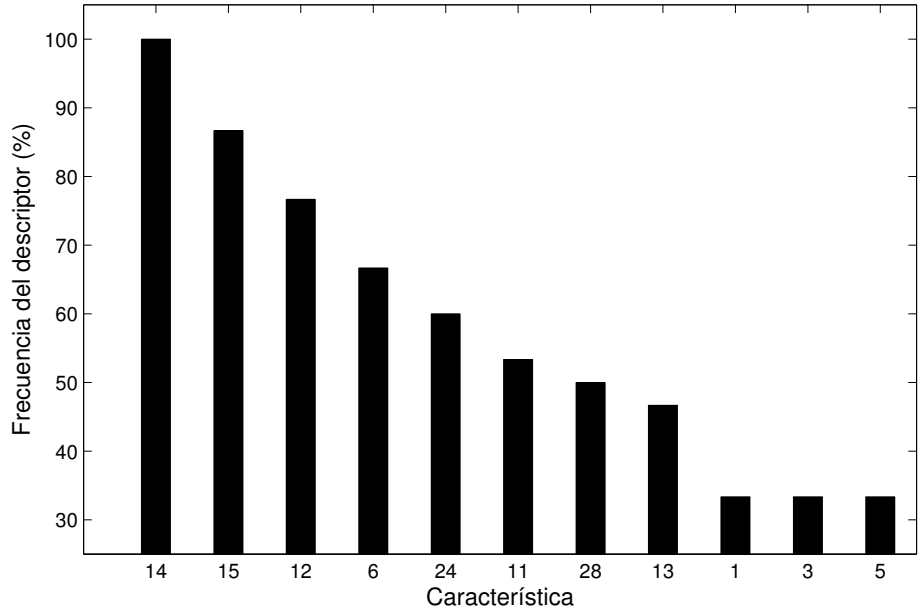


Figura 45: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos Antibac_A.

Para la representación Antibac_B los descriptores moleculares con mayor frecuencia fueron MLFER_BH (índice 92, Figura 46) y ETA_Shape_p (índice 112, Figura 46) con un 76.9 % de aparición en las mejores soluciones para el algoritmo GAFS.

Por último, para la representación Antibac_A+B los descriptores con mayor frecuencia fueron Z(pH7) (índice 2, Figura 47) y el momento hidrofóbico Hm_Hk(100) (índice 17, Figura 47) con un 83.3 % y 75 % de frecuencia de aparición en las mejores soluciones.

5.4.3. Máquina de soporte vectorial (SVM)

En la Sección 5.4.2 se presentaron los mejores subconjuntos de características que encontró el algoritmo GAFS para los conjuntos de datos AMP y Antibac. A partir de los mejores subconjuntos de características se construyeron 6 clasificadores utilizando máquinas de soporte vectorial lineal (SVM) (*i.e.,* 3 subconjuntos para AMPs y 3 subconjutnos para Antibac), el objetivo es validar la calidad de los subconjuntos encontrados por GAFS para cada conjunto de datos (ver tablas 20 y 21). A continuación se describen los resultados.

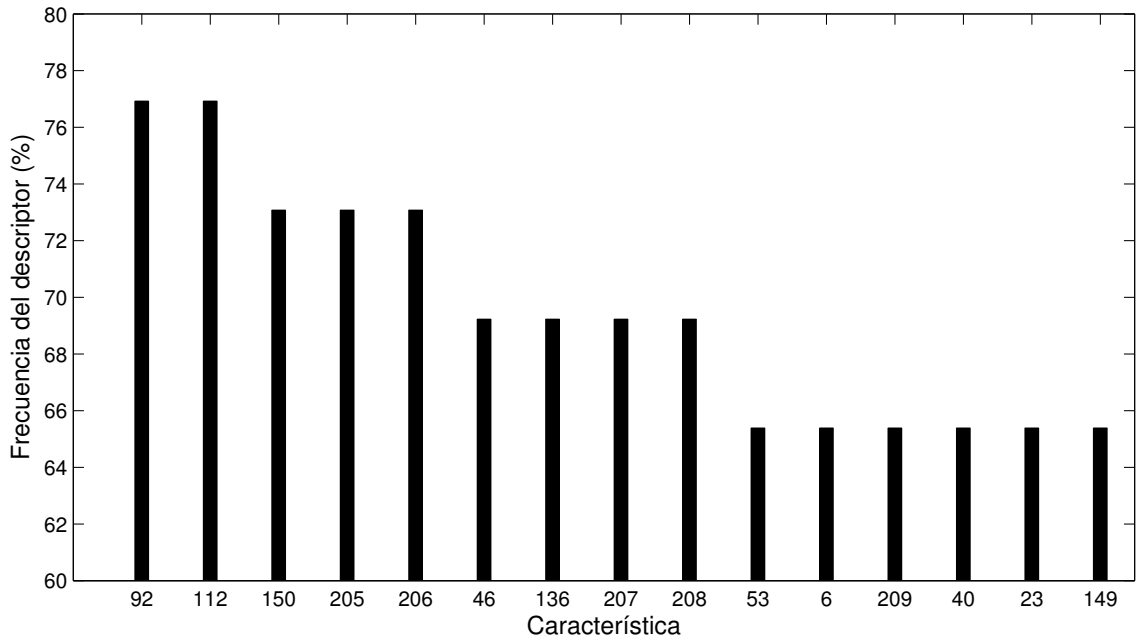


Figura 46: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos Antibac_B.

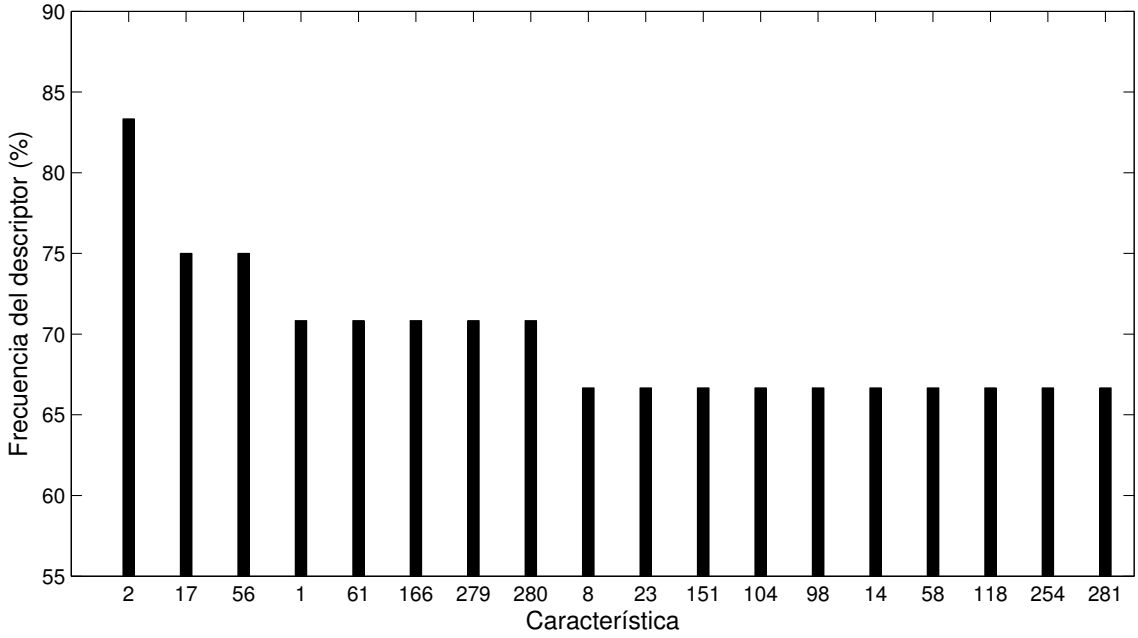


Figura 47: Características con mayor frecuencia de aparición en el algoritmo GAFS en 30 repeticiones para el conjunto de datos Antibac_A+B.

Los índices correspondientes a los 1500 péptidos antimicrobianos y 1500 péptidos sin actividad antimicrobiana conocida para los conjuntos AMP_A, AMP_B y AMP_A+B se muestran en las figuras 48, 49 y 50, respectivamente. Por otra parte, los índices correspondientes a 1500 péptidos con actividad antibacteriana y sin actividad antibacteriana para el conjunto Antibac_A se muestra en la Figura 51. Para el resto de las representaciones del conjunto Antibac (*i.e.*, Antibac_B y Antibac_A+B) no se muestran las gráfica debido al gran número de índices que contienen. Cada gráfica contiene las líneas correspondientes a los valores promedio de los descriptores para cada grupo espectro (Tomás-Vert *et al.*, 2000). En el espectro se muestran zonas con clara diferencia entre los valores que toman los grupos para ciertos índices; por ejemplo, en la Figura 48, el índice número 12 muestra que existe una clara diferencia entre los valores que toma el grupo AMP y no AMP, donde para la característica con índice 12 la clase AMP tiene un valor promedio de 0.6 y para los no AMP tiene un valor de 0.5.

Para la construcción de la SVM, los conjuntos de datos AMP y Antibac se dividieron en dos partes (ver Tabla 12): un conjunto de entrenamiento (90 % de los péptidos) y un conjunto de prueba (10 % de los péptidos). Por convención cada SVM-lineal se denominó de acuerdo con el nombre que toman los conjuntos de datos representados por un conjunto específico de características, es decir: SVM AMP_A, SVM AMP_B, SVM AMP _A+B, SVM Antibac_A, SVM Antibac_B y SVM Antibac_A+B.

Los resultados para los modelos correspondientes al conjunto de datos AMP se muestran en la Tabla 23. La exactitud en general para los modelos son de $ACC \geq 93\,\%$ y un $MCC \geq 0.86$ para el conjunto de prueba, lo que indica que los clasificadores tuvieron un buen desempeño al clasificar tanto las clases positivas (AMP) como las clases negativas (noAMP). Por otra parte, para el conjunto de validación la exactitud en general de los clasificadores disminuyó de 9 % con respecto a la exactitud para el conjunto de prueba. El modelo con mejor desempeño para la predicción de AMPs fue el que se construyó con el conjunto de datos AMP_A+B (SVM AMP_A+B), con una exactitud de 96.33 % y con una especificidad de 0.986, lo que indica que existen pocos falsos positivos en el predictor.

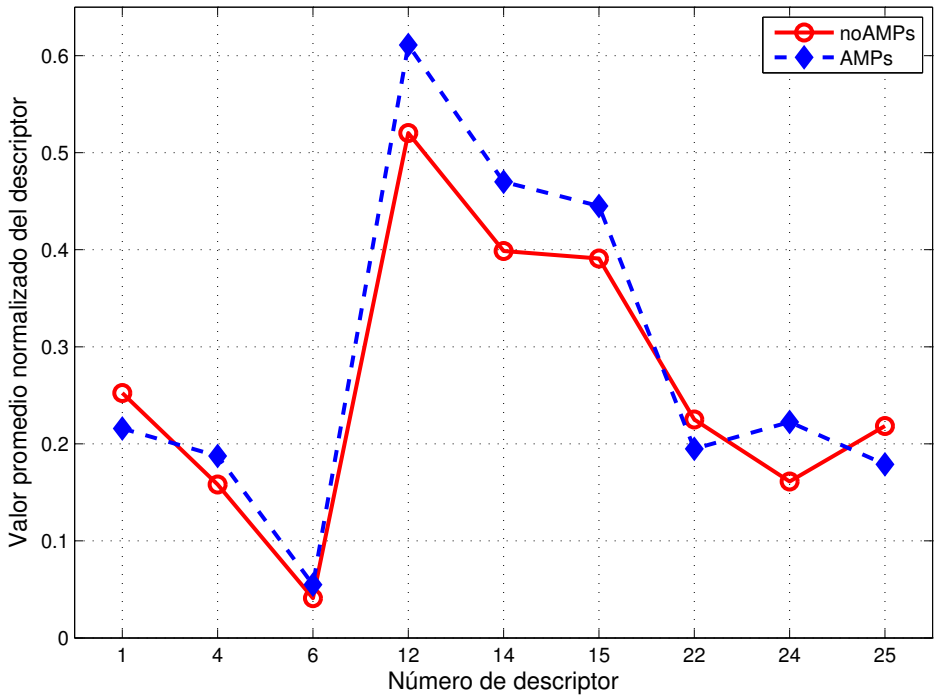


Figura 48: Distribución de los valores promedio de los descriptores moleculares entre los AMPs y noAMPs para el conjunto de datos AMP.

El desempeño de los modelos de clasificación correspondiente al conjunto de datos Antibac se muestran en la Tabla 24. El modelo con mejor desempeño fue el que se construyó con el conjunto de características Antibac_A+B con un valor de 93.33 % para ACC en pruebas y 82.27 % para ACC en validación.

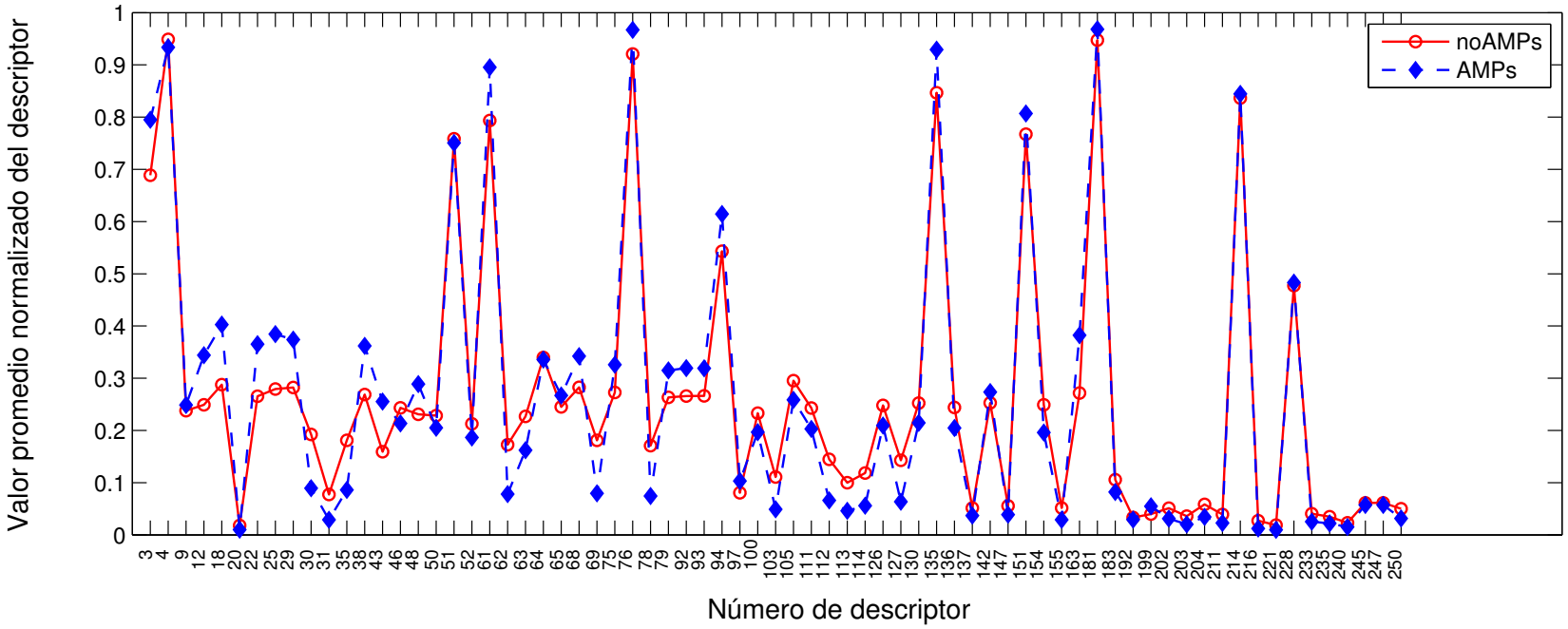


Figura 49: Distribución de los valores promedio de los descriptores moleculares entre los AMPs y noAMPs para el conjunto de datos AMP_B.

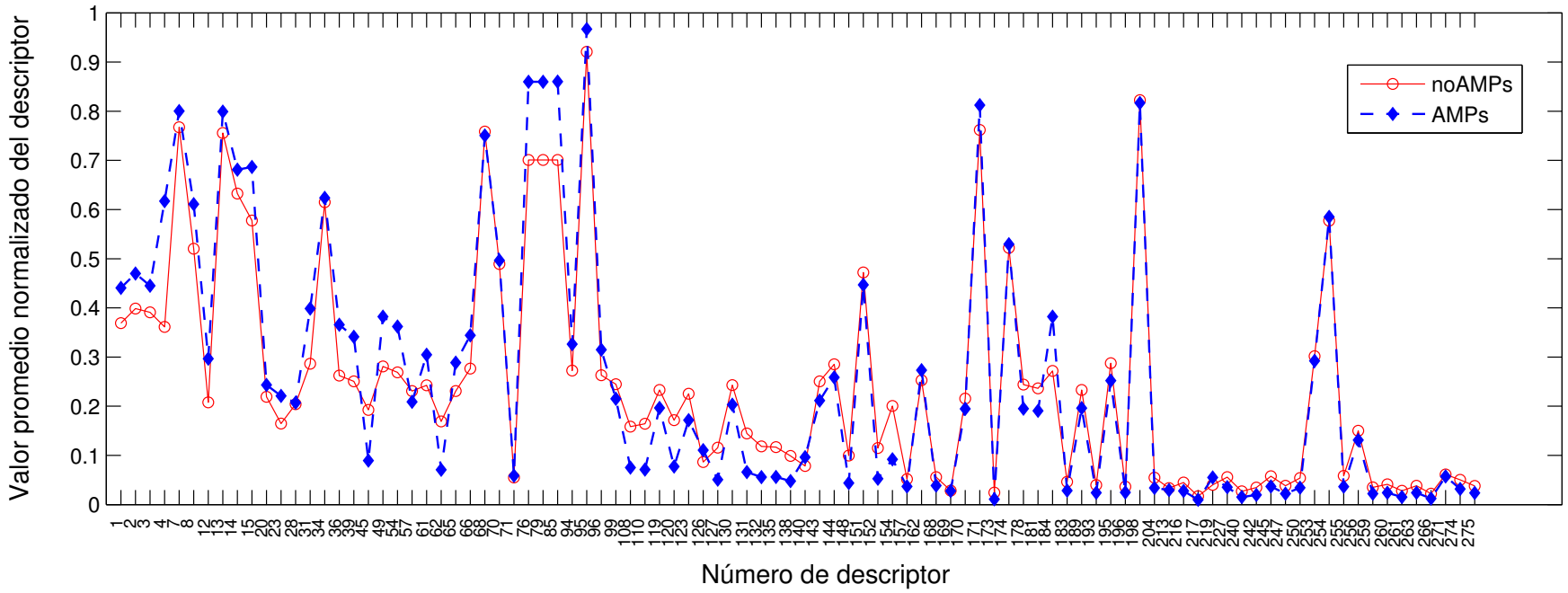


Figura 50: Distribución de los valores promedio de los descriptores moleculares entre los AMPs y noAMPs para el conjunto de datos AMP_A+B.

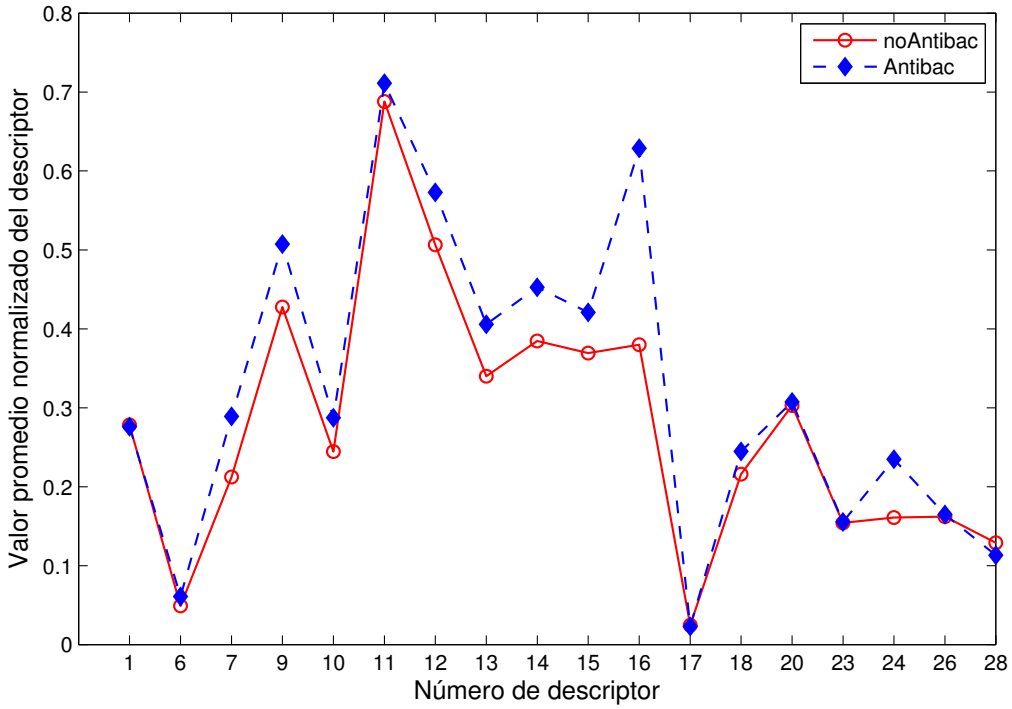


Figura 51: Distribución de los valores promedio de los descriptores moleculares entre los Antibac y noAntibac para el conjunto de datos Antibac_A.

Tabla 23: Los resultados muestran qué tan bien el predictor SVM separa los AMPs de los no AMPs para los conjuntos de prueba y validación.

Método	Conjunto	TN	FP	FN	TP	Sens	Espec	ACC	MCC
SVM AMP_A	Prueba	146	5	14	135	**0.967**	0.906	93.7	0.875
	Validación	329	55	39	163	**0.857**	0.807	83.96	0.653
SVM AMP_B	Prueba	129	10	10	151	0.938	0.928	93.33	0.866
	Validación	312	72	23	179	0.886	0.813	83.78	0.671
SVM AMP_A+B	Prueba	143	2	9	146	0.942	**0.986**	**96.33**	**0.928**
	Validación	349	35	45	157	0.777	**0.909**	**86.348**	**0.695**

Tabla 24: Los resultados muestran qué tan bien el predictor SVM separa los Antibac de los no Antibac para los conjuntos de prueba y validación.

Método	Conjunto	TN	FP	FN	TP	Sens	Espec	ACC	MCC
SVM Antibac_A	Prueba	138	12	12	138	0.920	0.920	92	0.84
	Validación	565	142	127	587	**0.822**	0.799	81.07	0.62
SVM Antibac_B	Prueba	134	18	6	142	**0.959**	0.882	92	0.84
	Validación	565	142	156	558	0.782	0.799	79.03	0.58
SVM Antibac_A+B	Prueba	143	7	13	137	0.913	**0.953**	**93.33**	**0.87**
	Validación	610	97	155	559	0.783	**0.863**	**82.27**	**0.65**

5.5. Comparación de la calidad con los métodos del estado del arte

Nuestro mejor resultado se comparó con los trabajos de la literatura que se muestran en la Tabla 26. Es relevante señalar que existen importantes diferencias entre los trabajos tales como: los conjuntos de datos utilizados para construir el modelo, el tamaño de los conjuntos de validación, el método de clasificación, entre otros.

Se puede observar que nuestro método tiene un mejor desempeño para el conjunto de prueba con un MCC de 0.93. Mientras que para el conjunto de validación el método que destaca es el de ANFIS (Fernandes *et al.*, 2012) que tiene un MCC de 0.94. Sin embargo, dada las diferencias de los trabajos mencionados anteriormente, la comparación entre los métodos no es equitativa.

Para realizar una comparación justa se utilizaron los métodos de DA, RF, ANN y SVM disponibles en el sitio http://www.camp.bicnirrh.res.in/predict/ (Waghu *et al.*, 2014) y el conjunto de validación (ver Tablas 29 y 31, Apéndice B). En la Tabla 25 se presentan los resultados obtenidos por los clasificadores dado el conjunto de validación. Los resultados indican que nuestro método supera a los métodos propuestos por Waghu *et al.* (2014) en los criterios de especificidad (Espec), exactitud (ACC) y coeficiente de correlación de Matthews (MCC). Por otra parte, los métodos que tuvieron mejor desempeño son los que utilizan SVM como clasificador.

Tabla 25: Comparación entre nuestro clasificador y otros algoritmos de la literatura para el conjunto de validación.

Método	**TN**	**FP**	**FN**	**TP**	**Sens**	**Espec**	**ACC**	**MCC**
SVM_CAMP	320	64	33	169	0.837	0.833	83.4	0.65
ANN_CAMP	310	74	51	151	0.748	0.807	78.7	0.54
DA_CAMP	291	93	48	154	0.762	0.758	75.9	0.5
RF_CAMP	297	87	16	186	**0.921**	0.773	82.4	0.66
SVM AMP_A+B	349	35	45	157	0.777	**0.909**	**86.4**	**0.67**

Tabla 26: Resultados comparativos de los métodos para la predicción de AMPs.

Método	**Bases de datos**		**Número de descriptores**	**Desempeño**			**Referencia**
	Conjunto de datos positivos	**Conjunto de datos negativos**		**Entrenamiento**	**Validación**	**Prueba**	
ANN	Generación aleatoria	Generación aleatoria	44		MCC=0.88		(Fjell *et al.*, 2009; Cherkasov *et al.*, 2008)
ANN	CAMP	Uniprot	8	MCC=0.79	MCC=0.797	MCC=0.74	(Torrent *et al.*, 2011)
SVM	CAMP	Uniprot	8		ACC=75 %		(Torrent *et al.*, 2011)
ANFIS	APD2	PDB	8		**MCC=0.94**		(Fernandes *et al.*, 2012)
SVM	APD	SwissProt				MCC=0.84	(Lata *et al.*, 2010)
DA	CAMP	Uniprot	64	MCC=0.75	ACC=87.5	MCC=0.74	(Thomas *et al.*, 2010)
RF	CAMP	Uniprot	64	MCC=0.82	ACC=92.5	MCC=0.84	(Waghu *et al.*, 2014)
ANN	CAMP	Uniprot	64	MCC=0.72	ACC=86.3	MCC=0.72	(Waghu *et al.*, 2014)
SVM	CAMP	Uniprot	64	MCC=0.91	ACC=91.5	MCC=0.83	(Waghu *et al.*, 2014)
SVM AMP_A+B	CAMP	Uniprot	94		ACC=86.4	MCC =0.93 **ACC= 96.3**	Este trabajo

5.6. Discusión

5.6.1. Conjunto de pruebas

Para la selección del conjunto de péptidos negativos (*i.e.*, sin actividad antimicrobiana) (ver subsección 4.1.1) se utilizó la base de datos Uniprot (Apweiler *et al.*, 2004) y se le aplicó un filtro basado en la metodología de Fernandes *et al.* (2012). El motivo por el cual se construyó el conjunto de datos negativos fue principalmente debido a que no existen bases de datos de péptidos con la anotación "sin actividad antimicrobiana" y los casos negativos que utilizan los algoritmos del estado del arte no se encuentran disponibles.

Por otra parte, a pesar que el conjunto de péptidos negativos es disjunto con respecto al conjunto de pétidos antimicrobiamos (Fernandes *et al.*, 2012), no se tiene la certeza de que este conjunto represente de forma adecuada al conjunto negativo de péptidos. Idealmente, se desearía que estos péptidos estuvieran experimentalmente validados tal como el conjunto de péptidos antimicrobianos.

5.6.2. Descriptores moleculares

Respecto a los descriptores moleculares que mejor representan a los péptidos antimicrobianos, es decir, el subconjunto que obtuvo la mayor exactitud en la predicción de AMPs se encuentra dentro de la representación de características AMP_A+B y tiene un tamaño de 94 descriptores.

Por otra parte, se puede observar que el algoritmo GAFS seleccionó un mayor número de descriptores para el caso en el que se tiene que identificar AMPs con actividad específica en contra de bacterias.

Sería interesante aplicar el algoritmo GAFS para identificar AMPs con actividad en contra de hongos o virus, para observar si el número y el tipo de descriptores cambian al variar el pátogeno de interés.

5.6.3. Comparación de métodos

Especificidad

Nuestro método mostró un mejor desempeño en la exactitud de los casos negativos (Espec = 0.909) comparado con los métodos propuestos por Waghu *et al.* (2014), lo cual es un resultado importante debido a que el modelo es específico en la identificación de AMPs. Por lo tanto, el modelo de clasificación disminuye la probabilidad de asignar un péptido no antimicrobiano como antimicrobiano. Esta característica es muy importante para el diseño de péptidos *in silico*, ya que se reduce el número de péptidos a probar en laboratorio, con lo cual se ayuda a disminuir el costo y el tiempo de evaluación de los péptidos en forma experimental (Maccari *et al.*, 2013). Por lo anterior, un modelo de clasificación que tenga un buen desempeño en especificidad cumple con el objetivo.

Tiempo de ejecución

Los métodos del estado del arte no reportan el tiempo de ejecución. En lo que respecta a nuestro método de clasificación el tiempo de ejecución más costoso es en la selección de las características (10.5 hrs para el Algoritmo 1), sin embargo, el esfuerzo se realiza sólo una vez. Después que se seleccionan el subconjunto óptimo de características construir el modelo de clasificación se realiza de manera rápida (*e.g.,* para el modelo SVM AMP_A+B el tiempo es de 23.1 seg.).

Capítulo 6. Conclusiones

En este capítulo se presentan las conclusiones a las que se llegó en este trabajo de tesis así como algunas perspectivas de investigación sobre el problema abordado.

6.1. Sumario

Se analizó el problema de selección de descriptores moleculares para el conjunto de AMPs. El problema se modeló como uno de selección de características y se diseñó un método de selección de envoltura, compuesto principalmente por dos elementos: una estrategia de búsqueda, que en este caso es un algoritmo genético; una función de evaluación para determinar la calidad de los subconjuntos, en este caso se eligió una máquina de soporte vectorial lineal para construir el modelo de clasificación y evaluar su exactitud. Para el algoritmo genético se propusieron dos casos de pruebas: el primero para la identificación de AMPs y el segundo para la identificación de péptidos con actividad antibacteriana. Cada caso de prueba se representó mediante diferentes conjuntos de descriptores moleculares. Por último, se propusieron una serie de experimentos computacionales para estudiar el desempeño del selector propuesto.

A continuación se exponen las conclusiones a las que se llegó con base en los experimentos realizados en el presente trabajo de investigación.

6.2. Conclusiones

Dada la gran cantidad de descriptores que se pueden calcular actualmente en los péptidos, seleccionar aquellos que mejor caracterizan a los AMPs tiene un gran impacto en la eficiencia de clasificación de la actividad biológica del péptido. Con el método de selección de características mostramos que se puede aumentar la eficiencia de clasificación en un 7 %, reduciendo al menos en un 50 % el número de características para el conjunto de AMPs.

Por otra parte, el tamaño del subconjunto de características seleccionadas no es determinante para decidir la calidad de la solución. Debido a que la desviación estándar del tamaño

de los subconjuntos presenta una variación apreciable (*e.g.,* para el conjunto AMP_A+B la desviación fue de 55.37 características, lo que representa un 44.26 % del valor promedio).

El subconjunto de descriptores que tiene una mayor eficiencia de clasificación de péptidos antimicrobianos son aquellos que incluyen descriptores de dimensión cero (0D), dimensión uno (1D) y de dimensión dos (2D). Esto nos habla de la importancia de tomar en consideración tanto la constitución de los aminoácidos como la distribución de los mismos en la secuencia del péptido para la caracterización de los AMPs.

Las medidas utilizadas para evaluar la calidad de los modelos de predicción (*i.e.,* exactitud (ACC), especificidad (Espec), sensibilidad (Sens) y coeficiente de correlación de Matthews (MCC)) muestran que nuestro método tiene un desempeño comparable con los reportados en el estado del arte e incluso, superándolos en algunas medidas de calidad.

Por último, el método propuesto permite el cribado *in silico* de un gran número de secuencias de péptidos con una actividad desconocida de una manera rápida y confiable.

6.3. Propuestas de trabajo futuro

Algoritmo genético

Para el algoritmo genético se propone implementar un método de filtrado que permita realizar una búsqueda local en los individuos de la población para eliminar características irrelevantes. Por lo anterior, se propone un algoritmo genético híbrido que combine los dos métodos de selección de características: el primero, el método de envoltura que servirá para elegir el subconjunto de características con la mayor eficiencia de predicción; por otro lado, el método de filtrado que servirá para disminuir las características del subconjunto sin que impacte negativamente en la calidad de la solución.

Modelo de regresión

Se propone cambiar el modelo de clasificación del método de envoltura por un modelo de regresión, esto con el objetivo de predecir la actividad del péptido en términos de la mínima

concentración inhibitoria (MIC). El MIC es la menor concentración para impedir el crecimiento de un microorganismo después de su incubación. Ahora, este nuevo modelo servirá para identificar cuáles de los péptidos antimicrobianos tienen una alta actividad antimicrobiana (*i.e.,* valor pequeño de MIC).

Implementación de un algoritmo para el diseño de AMPs

Se propone crear un algoritmo que permita diseñar péptidos *in silico* de forma *de novo* o a partir de secuencias de péptidos para las que se conoce su actividad. Para la función de evaluación se propone utilizar el modelo de clasificación de AMPs implementado en el presente trabajo de tesis.

Lista de referencias

Amaldi, E. y Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, **209**(1): 237–260.

Andrews, J. M. (2001). Determination of minimum inhibitory concentrations. *Journal of Antimicrobial Chemotherapy*, **48**(suppl 1): 5–16.

Aoki, W. y Ueda, M. (2013). Characterization of antimicrobial peptides toward the development of novel antibiotics. *Pharmaceuticals*, **6**(8): 1055–1081.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., y Yeh, L.-S. L. (2004). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, **32**(suppl 1): D115–D119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, y T, J. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1): 25–29.

B Hadley, E. y EW Hancock, R. (2010). Strategies for the discovery and advancement of novel cationic antimicrobial peptides. *Current topics in medicinal chemistry*, **10**(18): 1872–1881.

Bollobas, B. (2004). *Extremal Graph Theory*. Dover Publications, Incorporated.

Brodgen, K. A. (2005). Antimicrobial peptides: pore former or metabolic inhibitors in bacteria? *Nature Reviews Microbiology*, **3**(3): 238–250.

Chang, C.-C. y Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**(3): 27:1–27:27.

Cherkasov, A. y Jankovic, B. (2004). Application of 'inductive qsar descriptors for quantification of antibacterial activity of cationic polypeptides. *Molecules*, **9**(12): 1034–1052.

Cherkasov, A., Hilpert, K., Jenssen, H., Fjell, C. D., Waldbrook, M., Mullaly, S. C., Volkmer, R., y Hancock, R. E. (2008). Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS chemical biology*, **4**(1): 65–74.

Clote, P. y Backofen, R. (2000). *Computational Molecular Biology: An Introduction*. Wiley. New York, NY, USA.

Corona de la Fuente, R. I. (2010). *Análisis comparativo de dos heurísticas para el problema de empaquetamiento de la cadena lateral en proteínas*. Tesis de maestría en ciencias, Centro de Investigación Científica y de Educación Superior de Ensenada. 160 p.

Dash, M. y Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, **1**(3): 131 –156.

Del Pozo Menéndez, B., Villamor Martín, R., y A., H. M. (2011). Escarlatina. *Recuperado de: www.guia-abe.es*.

Dondoshansky, I. y Wolf, Y. (2002). Blastclust (NCBI software development toolkit). *NCBI, Bethesda, Md*.

Dosler, S. y Mataraci, E. (2013). In vitro pharmacokinetics of antimicrobial cationic peptides alone and in combination with antibiotics against methicillin resistant staphylococcus aureus biofilms. *Peptides*, **49**: 53–58.

Duda, R. O., Hart, P. E., y Stork, D. G. (2000). *Pattern Classification*. John Wiley & Sons, segunda edición. New York, NY.

Eisenberg, D., Weiss, R. M., Terwilliger, T. C., y Wilcox, W. (1982). Hydrophobic moments and protein structure. En: *Faraday Symposia of the Chemical Society*. Royal Society of Chemistry, Vol. 17, pp. 109–120.

Eisenberg, D., Weiss, R. M., y Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences*, **81**(1): 140–144.

Emmanouilidis, C., Hunter, A., y MacIntyre, J. (2000). A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. En: *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on*. IEEE, Vol. 1, pp. 309–316.

Fernandes, F. C., Ridgen, 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., Fries, P., Aich, P., Griebel, P., Hilpert, K., Hancock, R. E. W., y Cherkasov, A. (2008). Identification of novel host defense peptides and the absence of α-defensins in the bovine genome. *Proteins: Structure, Function, and Bioinformatics*, **73**(2): 420–430.

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. *Journal of Medicinal Chemistry*, **52**(7): 2006–2015.

Fjell, C. D., Jenssen, H., Cheung, W. A., Hancock, R. E. W., y Cherkasov, A. (2011). Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chemical Biology & Drug Design*, **77**(1): 48–56.

Fjell, C. D., Hiss, J. A., Hancock, R. E. W., y Schneider, G. (2012). Designing antimicrobial peptides: form follows function. *Nature reviews Drug discovery*, **11**(1): 37–51.

Goodarzi, M., Dejaegher, B., y Heyden, Y. V. (2012). Feature selection methods in qsar studies. *Journal of AOAC International*, **95**(3): 636–651.

Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**: 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, **11**(1): 10–18.

Hancock, R. E. (1997). Peptide antibiotics. *The Lancet*, **349**(9049): 418–422.

Hancock, R. E. y Sahl, H.-G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature biotechnology*, **24**(12): 1551–1557.

Helguera, A. M., Combes, R. D., González, M. P., y Cordeiro, M. N. D. S. (2008). Applications of 2d descriptors in drug design: a dragon tale. *Current topics in medicinal chemistry*, **8**(18): 1628–55.

Hellberg, S., Sjoestroem, M., Skagerberg, B., y Wold, S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *Journal of medicinal chemistry*, **30**(7): 1126–1135.

Hilpert, K., Fjell, C. D., y Cherkasov, A. (2008). Short linear cationic antimicrobial peptides: screening, optimizing, and prediction. En: *Peptide-Based Drug Design*. Springer, pp. 127–159.

Huang, J., Cai, Y., y Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, **28**(13): 1825 – 1844.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, **14**(1): 55–63.

Jang, J.-S. (1996). Input selection for anfis learning. En: *Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on*, Sep. Vol. 2, pp. 1493–1499.

Japelj, B. (2005). *PEDES Reference Manual*.

Jenssen, H., Hamill, P., y Hancock, R. E. W. (2006). Peptide antimicrobial agents. *Clinical Microbiology Reviews*, **19**(3): 491–511.

Jenssen, H., Lejon, T., Hilpert, K., Fjell, C. D., Cherkasov, A., y Hancock, R. E. (2007). Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward p. aeruginosa. *Chemical biology & drug design*, **70**(2): 134–142.

Joseph, S., Karnik, S., Nilawe, P., Jayaraman, V. K., y Idicula-Thomas, S. (2012). Classamp: A prediction tool for classification of antimicrobial peptides. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**(5): 1535–1538.

Käll, L., Krogh, A., y Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction?the phobius web server. *Nucleic Acids Research*, **35**(suppl 2): W429–W432.

Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *ARTIFICIAL INTELLIGENCE*, **97**(1): 273–324.

Kudo, M. y Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, **33**(1): 25 – 41.

Kyte, J. y Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, **157**(1): 105–132.

Lata, S., Sharma, B. K., y Raghava, G. P. S. (2007). Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, **8**(1): 263.

Lata, S., Mishra, N. K., y Raghava, G. P. (2010). Antibp2: improved version of antibacterial peptide prediction. *BMC bioinformatics*, **11**(1): 263.

Lathrop, R. H., Rogers, R. G., Bienkowska, J., Bryant, B. K., Buturović, L. J., Gaitatzes, C., Nambudripad, R., White, J. V., y Smith, T. F. (1998). *Computational Methods in Molecular Biology*, Vol. 12, capítulo Analysis and Algorithms for Protein Sequence-Structure Alignment, pp. 227–283. Elsevier Press.

Leid, J. G. (2009). Bacterial biofilms resist key host defenses. *Microbe*, **4**(2): 66–70.

Liu, H. y Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, **17**(4): 491–502.

Loose, C., Jensen, K., Rigoutsos, I., y Stephanopoulos, G. (2006). A linguistic model for the rational design of antimicrobial peptides. *Nature*, **443**(7113): 867–869.

Maccari, G., Nifosí, R., y Luca, M. D. (2013). *Microbial pathogens and strategies for combating them: science, technology and education*, capítulo Rational development of antimicrobial peptides for therapeutic use: design and production of highly active compounds, pp. 1265–1277. Formatex Research Center.

MDL (2005). *CT file format*. MDL Information Systems Inc, 14600 Catalina St., San Leandro, CA 94577.

Merrifield, E., Mitchell, S., J., U., Boman, H., Andreu, D., y Merrifield, R. (1995). D-enantiomers of 15-residue cecropin a-melittin hybrids. *International Journal of Peptide and Protein Research*, **46**: 214–220.

Molina, L. C., Belanche, L., y Nebot, À. (2002). Feature selection algorithms: A survey and experimental evaluation. En: *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, pp. 306–313.

Nguyen, L., Schibli, D., y Vogel, H. (2005). Structural studies and model membrane interactions of two peptides derived from bovine lactoferricin. *Journal of Peptide Science*, **11**(7): 379–389.

Pavan, M., Consonni, V., y Todeschini, R. (2005). Partial ranking models by genetic algorithm variable subset selection (gavss) approach for environmental priority settings. *MATCH Commun. Math. Comput. Chem*, **54**: 583–609.

Pavan, M., Netzeva, T. I., y Worth, A. P. (2006). Validation of a qsar model for acute toxicity. *SAR and QSAR in Environmental Research*, **17**(2): 147–171.

Piotto, S. P., Sessa, L., Concilio, S., y Iannelli, P. (2012). Yadamp: yet another database of antimicrobial peptides. *International Journal of Antimicrobial Agents*, **39**(4): 346 – 351.

Quinn, R. W. (1982). Epidemiology of group a streptococcal infections–their changing frequency and severity. *The Yale journal of biology and medicine*, **55**(3-4): 265–270.

Scott, M. G., Dullaghan, E., Mookherjee, N., Glavas, N., Waldbrook, M., Thompson, A., Wang, A., Lee, K., Doria, S., Hamill, P., Yu, J. J., Li, Y., Donini, O., Guarna, M. M., Finalay, B. B., North, J. R., y Hancock, R. E. W. (2007). An anti-infective peptide that selectively modulates the innate immune response. *Nature biotechnology*, **25**(4): 465–472.

Sima, C. y Dougherty, E. R. (2008). The peaking phenomenon in the presence of feature-selection. *Pattern Recognition Letters*, **29**(11): 1667 – 1674.

Taboureau, O. (2010). *Antimicrobial Peptides: Methods and Protocols*, capítulo Methods for Building Quantitative Structure-Activity Relationship (QSAR) Descriptors and Predictive Models for Computer-Aided Design of Antimicrobial Peptides. Human Press.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., y Idicula-Thomas, S. (2010). Camp: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, **38**: D774–D780.

Todeschini, R. y Consonni, V. (2000). *Handbook of molecular descriptors*. Wiley.

Tomás-Vert, F., Perez-Gimenez, F., Salabert-Salvador, M. T., Garcıa-March, F., y Jaen-Oltra, J. (2000). Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *Journal of Molecular Structure: THEOCHEM*, **504**(1): 249–259.

Torrent, M., Andreu, D., Nogués, V. M., y Boix, E. (2011). Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS ONE*, **6**(2): e16968.

Waghu, F. H., Gopi, L., Barai, R. S., Ramteke, P., Nizami, B., y Idicula-Thomas, S. (2014). Camp: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Research*, **42**(D1): D1154–D1158.

Wang, G., Li, X., y Wang, Z. (2009). Apd2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, **37**(suppl 1): D933–D937.

Wang, G., Li, X., y Zasloff, M. (2010). *A Database View of Naturally Occurring Antimicrobial Peptides: Nomenclature, Classification and Amino Acid Sequence Analysis*, pp. 1–21. CABI.

Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., Zheng, W., Li, L., Tan, M., Chen, Z., *et al.* (2011a). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one*, **6**(4): e18476.

Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., Zheng, W., Li, L., Tan, M., Chen, Z., Song, H., Cai, Y.-D., y Chou, K.-C. (2011b). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS ONE*, **6**(4): e18476.

Whelan, C., Roark, B., y Sonmez, K. (2010). Designing antimicrobial peptides with weighted finite-state transducers. En: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, pp. 764–767.

Whittaker, R. (1969). New concepts of kingdoms of organisms: Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *SCIENCE*, **163**(3863): 150–160.

WHO (2014). Antimicrobial resistance. *Recuperado de: www.who.int*.

Wimley, W. C. (2010). *5 Discovery of Novel Antimicrobial Peptides Using Combinatorial Chemistry and High Throughput Screening*, capítulo 5, pp. 87–99.

Yap, C. W. (2011). Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, **32**(7): 1466–1474.

Yasri, A. y Hartsough, D. (2001). Toward an optimal procedure for variable selection and qsar model building. *Journal of chemical information and computer sciences*, **41**(5): 1218–1227.

Yeaman, M. R. y Yount, N. Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacological Reviews*, **55**(1): 27–55.

Yount, N. Y. y Yeaman, M. R. (2004). Multidimensional signatures in antimicrobial peptides. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(19): 7363–7368.

Zhao, H. (2003). *Mode of Action of Antimicrobial Peptides*. Tesis de maestría en ciencias, Helsinki Biophysics and Biomembrane Group. 184 p.

Apéndice A. Clasificación de los aminoácidos

En este apéndice se muestra una clasificación de los aminoácidos según sus propiedades químicas, estas propiedades están relacionadas con la estructura que adquiere la proteína o péptido. Los aminoácidos de acuerdo con la propiedad química que tengan puede generar cierta afinidad con su entorno, por ejemplo los aminoácidos que rechazan el agua (*i.e.* hidrofóbicos) tienden a estar enterrados en la estructura del péptido.

A.1. Aminoácidos hidrofóbicos (no-polares)

Los aminoácidos hidrofóbicos se encuentran en las partes internas de las proteínas debido que se ocultan del medio acuoso. La lista de estos se presenta a continuación:

- Alanina (Ala, A)
- Isoleucina (Ile, I)
- Leucina (Leu, L)
- Metionina (Met, M)
- Fenilalanina (Phe, F)
- Prolina (Pro, P)
- Triptófano (Thr, T)
- Valina (Val, V)

Valores de hidrofobicidad en los aminoácidos

En un intento por describir aspectos cuantitativos del plegado de la proteína en términos de un carácter hidrofóbico o hidrofílico, se han propuesto varias escalas para asignar de manera numérica la hidrofobicidad a cada tipo de aminoácido. Los valores de hidrofobicidad corresponden a la energía libre, resultado de transferir la cadena lateral de un aminoácido desde un medio acuoso a uno polar (Eisenberg *et al.*, 1982). Diferentes escalas han sido

Tabla 27: Valores de hidrofobicidad por cada aminoácido (representado en código de una letra).

Aminoácido	He[a]	Hk[b]
A	0.25	1.8
C	0.04	2.5
D	-0.72	-3.5
E	-0.62	-0.35
F	0.61	2.8
G	0.16	-0.4
H	-0.4	-3.2
I	0.73	4.5
K	-1.1	-3.9
L	0.53	3.8
M	0.26	1.9
N	-0.64	-3.5
P	-0.07	-1.6
Q	-0.69	-3.5
R	-1.8	-4.5
S	-0.26	-0.8
T	-0.18	-0.7
V	0.54	4.2
W	0.37	-0.9
Y	0.02	-1.3

[a]Hidrofobicidad según la escala de Eisenberg
[b]Hidrofobicidad según la escala de Kyte-Doolittle

propuestas, sin embargo las más usadas son la escala de Eisenberg *et al.* (1982) y la de Kyte y Doolittle (1982) (ver Tabla 27).

A.2. Aminoácidos hidrófilo (polares)

Los aminoácidos hidrófilos son aquellos que tienen afinidad por el medio acuoso, por lo general estos aminoácidos se encuentran en las partes externas de la proteína. Los aminoácidos hidrófilos son:

- Asparagina (Asn, N)
- Cisteina (Cys, C)
- Glutamina (Glu, Q)
- Glicina (Gly, G)

- Serina (Ser, S)
- Treonina (Thr, T)
- Tirosina (Tyr, Y)

A.3. Aminoácidos cargados

De acuerdo con la carga que presentan los aminoácidos se clasifican en: básicos o con carga positiva; ácidos o con carga negativa. Los aminoácidos con carga positiva son los siguientes:

- Arginina (Arg, R)
- Histidina (His, H)
- Lisina (Lys, K)

Los de carga negativa son:

- Ácido aspártico (Asp, D)
- Ácido glutámico (Glu, E)

La carga media neta de los péptidos está determinada por la frecuencia del número de aminoácidos positivos y la frecuencia del número de aminoácidos negativos.

Apéndice B. Selección del conjunto de datos

En este apéndice se muestran los péptidos que se obtuvieron al aplicar la metodología descrita en la Sección 4.1. Los péptidos que se enlistan son representados con un identificador y son agrupados en péptidos en AMPs y noAMPs. En las tablas 28 y 30 se enlistan los AMPs y no AMPs que fueron utilizados para las pruebas y entrenamiento. Por otra parte, en las tablas 29 y 31 se enlistan los péptidos que fueron utilizados para la validación.

Tabla 28: Casos positivos: Péptidos antimicrobianos. Conjunto de prueba y entrenamiento, compuesto por 1500 péptidos recuperados de la base de datos CAMP.

CAMPSQ791	CAMPSQ318	CAMPSQ325	CAMPSQ1109	CAMPSQ1109	CAMPSQ1104	CAMPSQ3519	CAMPSQ3087
CAMPSQ792	CAMPSQ3541	CAMPSQ568	CAMPSQ3529	CAMPSQ3529	CAMPSQ341	CAMPSQ3518	CAMPSQ4176
CAMPSQ550	CAMPSQ1116	CAMPSQ327	CAMPSQ568	CAMPSQ571	CAMPSQ583	CAMPSQ570	CAMPSQ4175
CAMPSQ551	CAMPSQ3537	CAMPSQ569	CAMPSQ327	CAMPSQ330	CAMPSQ584	CAMPSQ3988	CAMPSQ5018
CAMPSQ793	CAMPSQ1117	CAMPSQ1110	CAMPSQ569	CAMPSQ331	CAMPSQ101	CAMPSQ3987	CAMPSQ3270
CAMPSQ552	CAMPSQ3536	CAMPSQ3773	CAMPSQ1110	CAMPSQ573	CAMPSQ585	CAMPSQ2658	CAMPSQ1092
CAMPSQ310	CAMPSQ1118	CAMPSQ1111	CAMPSQ3773	CAMPSQ574	CAMPSQ344	CAMPSQ3742	CAMPSQ941
CAMPSQ794	CAMPSQ3539	CAMPSQ329	CAMPSQ1111	CAMPSQ332	CAMPSQ586	CAMPSQ3984	CAMPSQ3034
CAMPSQ311	CAMPSQ1119	CAMPSQ1105	CAMPSQ329	CAMPSQ575	CAMPSQ345	CAMPSQ3741	CAMPSQ700
CAMPSQ795	CAMPSQ3538	CAMPSQ3768	CAMPSQ1105	CAMPSQ576	CAMPSQ587	CAMPSQ3983	CAMPSQ1097
CAMPSQ312	CAMPSQ1112	CAMPSQ3525	CAMPSQ3768	CAMPSQ577	CAMPSQ346	CAMPSQ2894	CAMPSQ1098
CAMPSQ796	CAMPSQ1113	CAMPSQ3767	CAMPSQ3525	CAMPSQ578	CAMPSQ588	CAMPSQ3743	CAMPSQ701
CAMPSQ554	CAMPSQ3532	CAMPSQ1106	CAMPSQ3767	CAMPSQ579	CAMPSQ104	CAMPSQ580	CAMPSQ3033
CAMPSQ314	CAMPSQ1114	CAMPSQ2438	CAMPSQ1106	CAMPSQ3762	CAMPSQ105	CAMPSQ581	CAMPSQ702
CAMPSQ798	CAMPSQ3535	CAMPSQ3527	CAMPSQ2438	CAMPSQ3761	CAMPSQ589	CAMPSQ990	CAMPSQ3036
CAMPSQ3540	CAMPSQ3777	CAMPSQ3769	CAMPSQ3527	CAMPSQ3999	CAMPSQ347	CAMPSQ991	CAMPSQ3277
CAMPSQ315	CAMPSQ3534	CAMPSQ3522	CAMPSQ3769	CAMPSQ3998	CAMPSQ106	CAMPSQ994	CAMPSQ3035
CAMPSQ558	CAMPSQ1115	CAMPSQ3764	CAMPSQ3522	CAMPSQ3759	CAMPSQ3991	CAMPSQ510	CAMPSQ703
CAMPSQ316	CAMPSQ563	CAMPSQ1101	CAMPSQ3764	CAMPSQ3753	CAMPSQ349	CAMPSQ753	CAMPSQ3272
CAMPSQ3781	CAMPSQ322	CAMPSQ3521	CAMPSQ1101	CAMPSQ3995	CAMPSQ3990	CAMPSQ996	CAMPSQ704
CAMPSQ1120	CAMPSQ564	CAMPSQ1102	CAMPSQ3521	CAMPSQ2422	CAMPSQ107	CAMPSQ512	CAMPSQ3271
CAMPSQ3542	CAMPSQ565	CAMPSQ3763	CAMPSQ1102	CAMPSQ3997	CAMPSQ3993	CAMPSQ754	CAMPSQ3032
CAMPSQ3300	CAMPSQ566	CAMPSQ1103	CAMPSQ3763	CAMPSQ3754	CAMPSQ3992	CAMPSQ755	CAMPSQ1095
CAMPSQ3784	CAMPSQ567	CAMPSQ1104	CAMPSQ1103	CAMPSQ3996	CAMPSQ109	CAMPSQ3342	CAMPSQ707
CAMPSQ1163	CAMPSQ3334	CAMPSQ3325	CAMPSQ4407	CAMPSQ545	CAMPSQ985	CAMPSQ974	CAMPSQ3031
CAMPSQ3100	CAMPSQ4668	CAMPSQ1148	CAMPSQ3318	CAMPSQ546	CAMPSQ723	CAMPSQ1084	CAMPSQ1096
CAMPSQ998	CAMPSQ1158	CAMPSQ775	CAMPSQ3313	CAMPSQ304	CAMPSQ4103	CAMPSQ717	CAMPSQ4119
CAMPSQ756	CAMPSQ3337	CAMPSQ534	CAMPSQ4402	CAMPSQ305	CAMPSQ3013	CAMPSQ4110	CAMPSQ3027
CAMPSQ3570	CAMPSQ3579	CAMPSQ3562	CAMPSQ3555	CAMPSQ1131	CAMPSQ1078	CAMPSQ718	CAMPSQ4115
CAMPSQ3569	CAMPSQ3336	CAMPSQ3320	CAMPSQ3797	CAMPSQ3550	CAMPSQ725	CAMPSQ3020	CAMPSQ714
CAMPSQ3329	CAMPSQ4667	CAMPSQ3561	CAMPSQ1134	CAMPSQ548	CAMPSQ3492	CAMPSQ1085	CAMPSQ715
CAMPSQ3324	CAMPSQ1159	CAMPSQ778	CAMPSQ3312	CAMPSQ3553	CAMPSQ726	CAMPSQ3019	CAMPSQ1082
CAMPSQ4897	CAMPSQ3578	CAMPSQ779	CAMPSQ3315	CAMPSQ3311	CAMPSQ1071	CAMPSQ4108	CAMPSQ716
CAMPSQ3566	CAMPSQ762	CAMPSQ3564	CAMPSQ4888	CAMPSQ3795	CAMPSQ1072	CAMPSQ709	CAMPSQ3293
CAMPSQ4896	CAMPSQ520	CAMPSQ537	CAMPSQ1136	CAMPSQ3310	CAMPSQ727	CAMPSQ3016	CAMPSQ4382
CAMPSQ1147	CAMPSQ521	CAMPSQ1144	CAMPSQ1137	CAMPSQ307	CAMPSQ3491	CAMPSQ1079	CAMPSQ3054
CAMPSQ3568	CAMPSQ522	CAMPSQ3563	CAMPSQ3314	CAMPSQ549	CAMPSQ3010	CAMPSQ4104	CAMPSQ3296
CAMPSQ3567	CAMPSQ764	CAMPSQ539	CAMPSQ4403	CAMPSQ309	CAMPSQ1073	CAMPSQ4106	CAMPSQ4385
CAMPSQ3341	CAMPSQ523	CAMPSQ4891	CAMPSQ3556	CAMPSQ3306	CAMPSQ728	CAMPSQ3490	CAMPSQ3053
CAMPSQ3102	CAMPSQ766	CAMPSQ3560	CAMPSQ781	CAMPSQ1127	CAMPSQ1074	CAMPSQ1070	CAMPSQ3295
CAMPSQ3586	CAMPSQ3331	CAMPSQ1140	CAMPSQ782	CAMPSQ3547	CAMPSQ719	CAMPSQ961	CAMPSQ4384
CAMPSQ3344	CAMPSQ1153	CAMPSQ3317	CAMPSQ783	CAMPSQ3305	CAMPSQ3008	CAMPSQ720	CAMPSQ3048
CAMPSQ3585	CAMPSQ525	CAMPSQ3559	CAMPSQ541	CAMPSQ3308	CAMPSQ3489	CAMPSQ721	CAMPSQ3287
CAMPSQ3101	CAMPSQ3572	CAMPSQ1139	CAMPSQ784	CAMPSQ3549	CAMPSQ1069	CAMPSQ964	CAMPSQ3286

Continúa en la siguiente página

Tabla 28 Casos positivos: Péptidos antimicrobianos – *Continuación*

CAMPSQ516	CAMPSQ3333	CAMPSQ4889	CAMPSQ300	CAMPSQ3307	CAMPSQ3488	CAMPSQ722	CAMPSQ932
CAMPSQ3580	CAMPSQ527	CAMPSQ3316	CAMPSQ785	CAMPSQ1123	CAMPSQ3007	CAMPSQ1075	CAMPSQ4133
CAMPSQ518	CAMPSQ529	CAMPSQ4405	CAMPSQ301	CAMPSQ3544	CAMPSQ3006	CAMPSQ3011	CAMPSQ3288
CAMPSQ1160	CAMPSQ1150	CAMPSQ3319	CAMPSQ544	CAMPSQ9	CAMPSQ972	CAMPSQ1076	CAMPSQ935
CAMPSQ3582	CAMPSQ3571	CAMPSQ4408	CAMPSQ302	CAMPSQ3301	CAMPSQ731	CAMPSQ965	CAMPSQ4130
CAMPSQ3704	CAMPSQ3957	CAMPSQ124	CAMPSQ115	CAMPSQ3917	CAMPSQ3543	CAMPSQ166	CAMPSQ936
CAMPSQ3946	CAMPSQ3715	CAMPSQ366	CAMPSQ357	CAMPSQ173	CAMPSQ3785	CAMPSQ167	CAMPSQ3285
CAMPSQ3703	CAMPSQ3956	CAMPSQ367	CAMPSQ358	CAMPSQ189	CAMPSQ3546	CAMPSQ169	CAMPSQ3043
CAMPSQ3945	CAMPSQ3714	CAMPSQ3726	CAMPSQ359	CAMPSQ3900	CAMPSQ1125	CAMPSQ3922	CAMPSQ937
CAMPSQ3940	CAMPSQ3951	CAMPSQ3968	CAMPSQ118	CAMPSQ2811	CAMPSQ3303	CAMPSQ3920	CAMPSQ3042
CAMPSQ3941	CAMPSQ3950	CAMPSQ3720	CAMPSQ119	CAMPSQ180	CAMPSQ3545	CAMPSQ3929	CAMPSQ3038
CAMPSQ3709	CAMPSQ2861	CAMPSQ3961	CAMPSQ3982	CAMPSQ181	CAMPSQ5	CAMPSQ160	CAMPSQ3037
CAMPSQ3948	CAMPSQ3711	CAMPSQ3722	CAMPSQ3977	CAMPSQ3904	CAMPSQ2	CAMPSQ161	CAMPSQ3039
CAMPSQ3947	CAMPSQ3953	CAMPSQ3721	CAMPSQ3976	CAMPSQ182	CAMPSQ1	CAMPSQ3928	CAMPSQ1091
CAMPSQ3708	CAMPSQ3952	CAMPSQ360	CAMPSQ3734	CAMPSQ3903	CAMPSQ3309	CAMPSQ3927	CAMPSQ3839
CAMPSQ140	CAMPSQ3959	CAMPSQ3729	CAMPSQ3978	CAMPSQ185	CAMPSQ153	CAMPSQ163	CAMPSQ4928
CAMPSQ382	CAMPSQ3958	CAMPSQ373	CAMPSQ3736	CAMPSQ3905	CAMPSQ154	CAMPSQ175	CAMPSQ493
CAMPSQ3949	CAMPSQ3716	CAMPSQ374	CAMPSQ3731	CAMPSQ110	CAMPSQ155	CAMPSQ176	CAMPSQ263
CAMPSQ383	CAMPSQ3719	CAMPSQ132	CAMPSQ3973	CAMPSQ594	CAMPSQ397	CAMPSQ177	CAMPSQ501
CAMPSQ3707	CAMPSQ372	CAMPSQ375	CAMPSQ3730	CAMPSQ595	CAMPSQ156	CAMPSQ178	CAMPSQ502
CAMPSQ4071	CAMPSQ384	CAMPSQ133	CAMPSQ3975	CAMPSQ353	CAMPSQ399	CAMPSQ179	CAMPSQ3473
CAMPSQ4070	CAMPSQ142	CAMPSQ134	CAMPSQ3739	CAMPSQ111	CAMPSQ159	CAMPSQ3913	CAMPSQ505
CAMPSQ4073	CAMPSQ143	CAMPSQ136	CAMPSQ350	CAMPSQ112	CAMPSQ3935	CAMPSQ3912	CAMPSQ506
CAMPSQ4072	CAMPSQ386	CAMPSQ379	CAMPSQ120	CAMPSQ596	CAMPSQ391	CAMPSQ3919	CAMPSQ3470
CAMPSQ4075	CAMPSQ144	CAMPSQ137	CAMPSQ362	CAMPSQ354	CAMPSQ392	CAMPSQ3918	CAMPSQ748
CAMPSQ4074	CAMPSQ145	CAMPSQ138	CAMPSQ121	CAMPSQ597	CAMPSQ3939	CAMPSQ170	CAMPSQ507
CAMPSQ4077	CAMPSQ146	CAMPSQ3955	CAMPSQ363	CAMPSQ113	CAMPSQ393	CAMPSQ171	CAMPSQ508
CAMPSQ4060	CAMPSQ389	CAMPSQ3713	CAMPSQ122	CAMPSQ355	CAMPSQ394	CAMPSQ3915	CAMPSQ3471
CAMPSQ4061	CAMPSQ147	CAMPSQ2866	CAMPSQ364	CAMPSQ598	CAMPSQ164	CAMPSQ172	CAMPSQ509
CAMPSQ4067	CAMPSQ3701	CAMPSQ3712	CAMPSQ365	CAMPSQ356	CAMPSQ165	CAMPSQ3914	CAMPSQ1046
CAMPSQ4088	CAMPSQ618	CAMPSQ867	CAMPSQ3133	CAMPSQ3129	CAMPSQ2928	CAMPSQ4064	CAMPSQ3467
CAMPSQ831	CAMPSQ3120	CAMPSQ1176	CAMPSQ4222	CAMPSQ3126	CAMPSQ190	CAMPSQ4063	CAMPSQ1047
CAMPSQ832	CAMPSQ1184	CAMPSQ3113	CAMPSQ3375	CAMPSQ3368	CAMPSQ191	CAMPSQ4051	CAMPSQ3466
CAMPSQ3144	CAMPSQ609	CAMPSQ3597	CAMPSQ3132	CAMPSQ3367	CAMPSQ2925	CAMPSQ4057	CAMPSQ3469
CAMPSQ4232	CAMPSQ3119	CAMPSQ626	CAMPSQ4221	CAMPSQ3125	CAMPSQ2924	CAMPSQ4059	CAMPSQ3468
CAMPSQ3143	CAMPSQ3599	CAMPSQ868	CAMPSQ3374	CAMPSQ3127	CAMPSQ2926	CAMPSQ4058	CAMPSQ4315
CAMPSQ3145	CAMPSQ1178	CAMPSQ1177	CAMPSQ844	CAMPSQ1180	CAMPSQ196	CAMPSQ4053	CAMPSQ4150
CAMPSQ3140	CAMPSQ3357	CAMPSQ869	CAMPSQ3377	CAMPSQ850	CAMPSQ2914	CAMPSQ4055	CAMPSQ3067
CAMPSQ3381	CAMPSQ4204	CAMPSQ627	CAMPSQ845	CAMPSQ852	CAMPSQ2913	CAMPSQ4049	CAMPSQ4398
CAMPSQ3384	CAMPSQ3598	CAMPSQ3591	CAMPSQ3376	CAMPSQ610	CAMPSQ4095	CAMPSQ4280	CAMPSQ4156
CAMPSQ838	CAMPSQ3356	CAMPSQ628	CAMPSQ846	CAMPSQ853	CAMPSQ4094	CAMPSQ3193	CAMPSQ4155
CAMPSQ3383	CAMPSQ1179	CAMPSQ3590	CAMPSQ605	CAMPSQ612	CAMPSQ4090	CAMPSQ3192	CAMPSQ4158
CAMPSQ839	CAMPSQ4205	CAMPSQ3351	CAMPSQ1192	CAMPSQ1185	CAMPSQ4099	CAMPSQ4046	CAMPSQ4152
CAMPSQ4230	CAMPSQ860	CAMPSQ629	CAMPSQ3371	CAMPSQ855	CAMPSQ2907	CAMPSQ4045	CAMPSQ4180
CAMPSQ4229	CAMPSQ861	CAMPSQ3593	CAMPSQ3370	CAMPSQ3121	CAMPSQ2906	CAMPSQ4048	CAMPSQ5030
CAMPSQ4226	CAMPSQ862	CAMPSQ3350	CAMPSQ848	CAMPSQ613	CAMPSQ2903	CAMPSQ4047	CAMPSQ4182
CAMPSQ3137	CAMPSQ621	CAMPSQ3108	CAMPSQ1193	CAMPSQ614	CAMPSQ2902	CAMPSQ3195	CAMPSQ3090
CAMPSQ4468	CAMPSQ622	CAMPSQ3349	CAMPSQ3373	CAMPSQ615	CAMPSQ2905	CAMPSQ4286	CAMPSQ3028
CAMPSQ3136	CAMPSQ623	CAMPSQ3107	CAMPSQ849	CAMPSQ1188	CAMPSQ2904	CAMPSQ4044	CAMPSQ3062
CAMPSQ3139	CAMPSQ3595	CAMPSQ3109	CAMPSQ1194	CAMPSQ3123	CAMPSQ4082	CAMPSQ4039	CAMPSQ914
CAMPSQ4227	CAMPSQ3111	CAMPSQ1167	CAMPSQ3131	CAMPSQ857	CAMPSQ4081	CAMPSQ197	CAMPSQ4396
CAMPSQ3138	CAMPSQ4200	CAMPSQ3588	CAMPSQ3130	CAMPSQ1181	CAMPSQ4084	CAMPSQ199	CAMPSQ3064
CAMPSQ1190	CAMPSQ3594	CAMPSQ3104	CAMPSQ3372	CAMPSQ858	CAMPSQ4083	CAMPSQ2923	CAMPSQ4395
CAMPSQ840	CAMPSQ866	CAMPSQ4679	CAMPSQ1195	CAMPSQ616	CAMPSQ4086	CAMPSQ2922	CAMPSQ907
CAMPSQ842	CAMPSQ3352	CAMPSQ3348	CAMPSQ4218	CAMPSQ1182	CAMPSQ4085	CAMPSQ2929	CAMPSQ909
CAMPSQ3680	CAMPSQ410	CAMPSQ1034	CAMPSQ1038	CAMPSQ632	CAMPSQ3872	CAMPSQ3877	CAMPSQ4148
CAMPSQ1017	CAMPSQ411	CAMPSQ3695	CAMPSQ3699	CAMPSQ874	CAMPSQ4714	CAMPSQ3638	CAMPSQ3059
CAMPSQ1018	CAMPSQ654	CAMPSQ406	CAMPSQ880	CAMPSQ3463	CAMPSQ3625	CAMPSQ3637	CAMPSQ3290
CAMPSQ3437	CAMPSQ412	CAMPSQ3453	CAMPSQ640	CAMPSQ1042	CAMPSQ3867	CAMPSQ3874	CAMPSQ3292
CAMPSQ3679	CAMPSQ655	CAMPSQ3690	CAMPSQ641	CAMPSQ876	CAMPSQ3866	CAMPSQ3632	CAMPSQ4381
CAMPSQ1019	CAMPSQ897	CAMPSQ649	CAMPSQ400	CAMPSQ3462	CAMPSQ3869	CAMPSQ4720	CAMPSQ3291
CAMPSQ3439	CAMPSQ414	CAMPSQ3450	CAMPSQ642	CAMPSQ877	CAMPSQ4716	CAMPSQ3876	CAMPSQ4380
CAMPSQ1013	CAMPSQ1020	CAMPSQ3692	CAMPSQ885	CAMPSQ635	CAMPSQ3626	CAMPSQ4723	CAMPSQ3056

Continúa en la siguiente página

Tabla 28 Casos positivos: Péptidos antimicrobianos – *Continuación*

CAMPSQ3676	CAMPSQ2351	CAMPSQ1030	CAMPSQ643	CAMPSQ1044	CAMPSQ3868	CAMPSQ4722	CAMPSQ920
CAMPSQ1014	CAMPSQ656	CAMPSQ3691	CAMPSQ401	CAMPSQ3465	CAMPSQ4715	CAMPSQ3875	CAMPSQ4387
CAMPSQ3675	CAMPSQ3441	CAMPSQ3449	CAMPSQ644	CAMPSQ636	CAMPSQ3863	CAMPSQ3639	CAMPSQ3297
CAMPSQ3436	CAMPSQ3440	CAMPSQ1028	CAMPSQ645	CAMPSQ3464	CAMPSQ3621	CAMPSQ4728	CAMPSQ3058
CAMPSQ3678	CAMPSQ1021	CAMPSQ3207	CAMPSQ3452	CAMPSQ1045	CAMPSQ3862	CAMPSQ691	CAMPSQ3057
CAMPSQ1015	CAMPSQ3682	CAMPSQ3448	CAMPSQ11	CAMPSQ638	CAMPSQ3623	CAMPSQ461	CAMPSQ4388
CAMPSQ1016	CAMPSQ2352	CAMPSQ1029	CAMPSQ403	CAMPSQ2370	CAMPSQ3865	CAMPSQ220	CAMPSQ923
CAMPSQ3677	CAMPSQ415	CAMPSQ1024	CAMPSQ3210	CAMPSQ639	CAMPSQ3864	CAMPSQ221	CAMPSQ3052
CAMPSQ3435	CAMPSQ3443	CAMPSQ3445	CAMPSQ1031	CAMPSQ3461	CAMPSQ3622	CAMPSQ222	CAMPSQ3294
CAMPSQ660	CAMPSQ2353	CAMPSQ3444	CAMPSQ888	CAMPSQ3460	CAMPSQ4717	CAMPSQ223	CAMPSQ3821
CAMPSQ661	CAMPSQ658	CAMPSQ1026	CAMPSQ10	CAMPSQ3698	CAMPSQ3628	CAMPSQ225	CAMPSQ2974
CAMPSQ420	CAMPSQ3685	CAMPSQ3447	CAMPSQ3693	CAMPSQ1035	CAMPSQ4719	CAMPSQ226	CAMPSQ3820
CAMPSQ662	CAMPSQ659	CAMPSQ3688	CAMPSQ1032	CAMPSQ3455	CAMPSQ870	CAMPSQ468	CAMPSQ2973
CAMPSQ422	CAMPSQ1023	CAMPSQ3446	CAMPSQ404	CAMPSQ3213	CAMPSQ871	CAMPSQ469	CAMPSQ4394
CAMPSQ423	CAMPSQ3442	CAMPSQ890	CAMPSQ3454	CAMPSQ1036	CAMPSQ630	CAMPSQ3870	CAMPSQ732
CAMPSQ424	CAMPSQ419	CAMPSQ650	CAMPSQ1033	CAMPSQ3697	CAMPSQ631	CAMPSQ228	CAMPSQ3001
CAMPSQ3672	CAMPSQ3681	CAMPSQ651	CAMPSQ3696	CAMPSQ1037	CAMPSQ873	CAMPSQ229	CAMPSQ3485
CAMPSQ3854	CAMPSQ2992	CAMPSQ2933	CAMPSQ22	CAMPSQ3818	CAMPSQ40	CAMPSQ1010	CAMPSQ733
CAMPSQ3612	CAMPSQ2530	CAMPSQ2930	CAMPSQ21	CAMPSQ2728	CAMPSQ3812	CAMPSQ668	CAMPSQ3484
CAMPSQ2522	CAMPSQ86	CAMPSQ24	CAMPSQ2943	CAMPSQ4906	CAMPSQ2965	CAMPSQ669	CAMPSQ734
CAMPSQ3611	CAMPSQ3861	CAMPSQ23	CAMPSQ19	CAMPSQ273	CAMPSQ2723	CAMPSQ427	CAMPSQ3003
CAMPSQ3618	CAMPSQ3860	CAMPSQ2939	CAMPSQ18	CAMPSQ3817	CAMPSQ2722	CAMPSQ3432	CAMPSQ735
CAMPSQ3617	CAMPSQ85	CAMPSQ2936	CAMPSQ16	CAMPSQ50	CAMPSQ3811	CAMPSQ1011	CAMPSQ3487
CAMPSQ470	CAMPSQ84	CAMPSQ2935	CAMPSQ15	CAMPSQ2954	CAMPSQ2964	CAMPSQ1012	CAMPSQ977
CAMPSQ3619	CAMPSQ83	CAMPSQ2938	CAMPSQ14	CAMPSQ2953	CAMPSQ3814	CAMPSQ429	CAMPSQ978
CAMPSQ4708	CAMPSQ82	CAMPSQ2937	CAMPSQ2941	CAMPSQ2955	CAMPSQ39	CAMPSQ3427	CAMPSQ736
CAMPSQ483	CAMPSQ81	CAMPSQ472	CAMPSQ13	CAMPSQ48	CAMPSQ4903	CAMPSQ3669	CAMPSQ737
CAMPSQ241	CAMPSQ3614	CAMPSQ80	CAMPSQ2940	CAMPSQ47	CAMPSQ2725	CAMPSQ2338	CAMPSQ738
CAMPSQ242	CAMPSQ3613	CAMPSQ230	CAMPSQ12	CAMPSQ46	CAMPSQ38	CAMPSQ1007	CAMPSQ3483
CAMPSQ90	CAMPSQ2527	CAMPSQ231	CAMPSQ290	CAMPSQ45	CAMPSQ2966	CAMPSQ1008	CAMPSQ2393
CAMPSQ243	CAMPSQ3616	CAMPSQ232	CAMPSQ293	CAMPSQ2951	CAMPSQ2724	CAMPSQ1009	CAMPSQ739
CAMPSQ485	CAMPSQ2526	CAMPSQ233	CAMPSQ294	CAMPSQ3809	CAMPSQ2961	CAMPSQ1003	CAMPSQ3478
CAMPSQ486	CAMPSQ3615	CAMPSQ234	CAMPSQ295	CAMPSQ280	CAMPSQ36	CAMPSQ3664	CAMPSQ3477
CAMPSQ244	CAMPSQ4704	CAMPSQ476	CAMPSQ33	CAMPSQ281	CAMPSQ3810	CAMPSQ3667	CAMPSQ3479
CAMPSQ487	CAMPSQ3857	CAMPSQ235	CAMPSQ32	CAMPSQ3804	CAMPSQ34	CAMPSQ3425	CAMPSQ980
CAMPSQ246	CAMPSQ3852	CAMPSQ477	CAMPSQ31	CAMPSQ282	CAMPSQ3819	CAMPSQ3424	CAMPSQ981
CAMPSQ489	CAMPSQ3610	CAMPSQ479	CAMPSQ30	CAMPSQ2957	CAMPSQ3816	CAMPSQ274	CAMPSQ741
CAMPSQ247	CAMPSQ2520	CAMPSQ237	CAMPSQ2932	CAMPSQ283	CAMPSQ4905	CAMPSQ277	CAMPSQ500
CAMPSQ99	CAMPSQ4940	CAMPSQ88	CAMPSQ29	CAMPSQ284	CAMPSQ270	CAMPSQ278	CAMPSQ499
CAMPSQ3850	CAMPSQ3851	CAMPSQ238	CAMPSQ2931	CAMPSQ296	CAMPSQ271	CAMPSQ279	CAMPSQ257
CAMPSQ95	CAMPSQ79	CAMPSQ87	CAMPSQ28	CAMPSQ297	CAMPSQ3815	CAMPSQ43	CAMPSQ259
CAMPSQ94	CAMPSQ4701	CAMPSQ239	CAMPSQ27	CAMPSQ299	CAMPSQ2726	CAMPSQ2970	CAMPSQ2990
CAMPSQ4160	CAMPSQ2523	CAMPSQ4161	CAMPSQ4189	CAMPSQ264	CAMPSQ2995	CAMPSQ93	CAMPSQ3340
CAMPSQ5016	CAMPSQ3099	CAMPSQ2731	CAMPSQ2750	CAMPSQ265	CAMPSQ89	CAMPSQ92	CAMPSQ1162
CAMPSQ3079	CAMPSQ4188	CAMPSQ3827	CAMPSQ62	CAMPSQ266	CAMPSQ4931	CAMPSQ2998	CAMPSQ3339
CAMPSQ902	CAMPSQ5038	CAMPSQ2737	CAMPSQ61	CAMPSQ267	CAMPSQ3842	CAMPSQ3603	CAMPSQ3338
CAMPSQ905	CAMPSQ5033	CAMPSQ260	CAMPSQ60	CAMPSQ269	CAMPSQ4938	CAMPSQ3845	CAMPSQ3335
CAMPSQ3075	CAMPSQ4185	CAMPSQ2979	CAMPSQ2745	CAMPSQ76	CAMPSQ3849	CAMPSQ3844	CAMPSQ3577
CAMPSQ4159	CAMPSQ5034	CAMPSQ261	CAMPSQ3834	CAMPSQ2980	CAMPSQ3848	CAMPSQ2755	CAMPSQ3840
CAMPSQ5007	CAMPSQ4184	CAMPSQ3829	CAMPSQ2987	CAMPSQ74	CAMPSQ4937	CAMPSQ2997	CAMPSQ4932
CAMPSQ1080	CAMPSQ3095	CAMPSQ262	CAMPSQ2986	CAMPSQ73	CAMPSQ3609	CAMPSQ2513	CAMPSQ3843
CAMPSQ950	CAMPSQ4187	CAMPSQ3828	CAMPSQ4922	CAMPSQ72	CAMPSQ481	CAMPSQ3602	CAMPSQ2996
CAMPSQ710	CAMPSQ3097	CAMPSQ5040	CAMPSQ3833	CAMPSQ70	CAMPSQ2519	CAMPSQ4933	CAMPSQ4184
CAMPSQ4112	CAMPSQ4186	CAMPSQ4194	CAMPSQ3836	CAMPSQ2734	CAMPSQ240	CAMPSQ4936	CAMPSQ3095
CAMPSQ1086	CAMPSQ5029	CAMPSQ4190	CAMPSQ2746	CAMPSQ2975	CAMPSQ482	CAMPSQ3847	CAMPSQ4187
CAMPSQ4111	CAMPSQ3081	CAMPSQ4196	CAMPSQ3830	CAMPSQ2733	CAMPSQ4939	CAMPSQ2516	CAMPSQ3097
CAMPSQ1087	CAMPSQ3080	CAMPSQ5044	CAMPSQ3832	CAMPSQ2978	CAMPSQ494	CAMPSQ3605	CAMPSQ4186
CAMPSQ712	CAMPSQ3082	CAMPSQ5045	CAMPSQ2985	CAMPSQ2977	CAMPSQ495	CAMPSQ2999	CAMPSQ5029
CAMPSQ3025	CAMPSQ3089	CAMPSQ4195	CAMPSQ2743	CAMPSQ2972	CAMPSQ253	CAMPSQ3604	CAMPSQ3081
CAMPSQ713	CAMPSQ3088	CAMPSQ4198	CAMPSQ2984	CAMPSQ2730	CAMPSQ254	CAMPSQ3846	CAMPSQ3080
CAMPSQ4114	CAMPSQ4174	CAMPSQ3092	CAMPSQ490	CAMPSQ69	CAMPSQ496	CAMPSQ2994	CAMPSQ3082
CAMPSQ955	CAMPSQ3084	CAMPSQ4181	CAMPSQ491	CAMPSQ2971	CAMPSQ497	CAMPSQ3841	CAMPSQ3089
CAMPSQ956	CAMPSQ4173	CAMPSQ3091	CAMPSQ492	CAMPSQ68	CAMPSQ256	CAMPSQ2993	CAMPSQ3088

Continúa en la siguiente página

Tabla 28 Casos positivos: Péptidos antimicrobianos – *Continuación*

CAMPSQ714	CAMPSQ3087	CAMPSQ4180	CAMPSQ3839	CAMPSQ3821	CAMPSQ499	CAMPSQ3840	CAMPSQ4174
CAMPSQ715	CAMPSQ4176	CAMPSQ5030	CAMPSQ4928	CAMPSQ2974	CAMPSQ257	CAMPSQ4932	CAMPSQ3084
CAMPSQ1082	CAMPSQ4175	CAMPSQ4182	CAMPSQ493	CAMPSQ3820	CAMPSQ259	CAMPSQ3843	CAMPSQ4173
CAMPSQ716	CAMPSQ5018	CAMPSQ3090	CAMPSQ263	CAMPSQ2973	CAMPSQ2990	CAMPSQ2996	CAMPSQ5007
CAMPSQ3293	CAMPSQ3270	CAMPSQ3028	CAMPSQ501	CAMPSQ4394	CAMPSQ985	CAMPSQ974	CAMPSQ1080
CAMPSQ4382	CAMPSQ1092	CAMPSQ3062	CAMPSQ502	CAMPSQ732	CAMPSQ723	CAMPSQ1084	CAMPSQ950
CAMPSQ3054	CAMPSQ941	CAMPSQ914	CAMPSQ3473	CAMPSQ3001	CAMPSQ4103	CAMPSQ717	CAMPSQ710
CAMPSQ3296	CAMPSQ3034	CAMPSQ4396	CAMPSQ505	CAMPSQ3485	CAMPSQ3013	CAMPSQ4110	CAMPSQ4112
CAMPSQ4385	CAMPSQ700	CAMPSQ3064	CAMPSQ506	CAMPSQ733	CAMPSQ1078	CAMPSQ718	CAMPSQ1086
CAMPSQ3053	CAMPSQ1097	CAMPSQ4395	CAMPSQ3470	CAMPSQ3484	CAMPSQ725	CAMPSQ3020	CAMPSQ4111
CAMPSQ3295	CAMPSQ1098	CAMPSQ907	CAMPSQ748	CAMPSQ734	CAMPSQ3492	CAMPSQ1085	CAMPSQ1087
CAMPSQ4384	CAMPSQ701	CAMPSQ909	CAMPSQ507	CAMPSQ3003	CAMPSQ726	CAMPSQ3019	CAMPSQ712
CAMPSQ3048	CAMPSQ3033	CAMPSQ4148	CAMPSQ508	CAMPSQ735	CAMPSQ1071	CAMPSQ4108	CAMPSQ3025
CAMPSQ3287	CAMPSQ702	CAMPSQ3059	CAMPSQ3471	CAMPSQ3487	CAMPSQ1072	CAMPSQ709	CAMPSQ713
CAMPSQ3286	CAMPSQ3036	CAMPSQ3290	CAMPSQ509	CAMPSQ977	CAMPSQ727	CAMPSQ3016	CAMPSQ4114
CAMPSQ932	CAMPSQ3277	CAMPSQ3292	CAMPSQ1046	CAMPSQ978	CAMPSQ3491	CAMPSQ1079	CAMPSQ955
CAMPSQ4133	CAMPSQ3035	CAMPSQ4381	CAMPSQ3467	CAMPSQ736	CAMPSQ3010	CAMPSQ4104	CAMPSQ956
CAMPSQ3288	CAMPSQ703	CAMPSQ3291	CAMPSQ1047	CAMPSQ737	CAMPSQ1073	CAMPSQ4106	CAMPSQ3829
CAMPSQ935	CAMPSQ3272	CAMPSQ4380	CAMPSQ3466	CAMPSQ738	CAMPSQ728	CAMPSQ3490	CAMPSQ262
CAMPSQ4130	CAMPSQ704	CAMPSQ3056	CAMPSQ3469	CAMPSQ3483	CAMPSQ1074	CAMPSQ1070	CAMPSQ3828
CAMPSQ936	CAMPSQ3271	CAMPSQ920	CAMPSQ3468	CAMPSQ2393	CAMPSQ719	CAMPSQ961	CAMPSQ5040
CAMPSQ3285	CAMPSQ3032	CAMPSQ4387	CAMPSQ4315	CAMPSQ739	CAMPSQ3008	CAMPSQ720	CAMPSQ4194
CAMPSQ3043	CAMPSQ1095	CAMPSQ3297	CAMPSQ4150	CAMPSQ3478	CAMPSQ3489	CAMPSQ721	CAMPSQ4190
CAMPSQ937	CAMPSQ707	CAMPSQ3058	CAMPSQ3067	CAMPSQ3477	CAMPSQ1069	CAMPSQ964	CAMPSQ4196
CAMPSQ3042	CAMPSQ3031	CAMPSQ3057	CAMPSQ4398	CAMPSQ3479	CAMPSQ3488	CAMPSQ722	CAMPSQ5044
CAMPSQ3038	CAMPSQ1096	CAMPSQ4388	CAMPSQ4156	CAMPSQ980	CAMPSQ3007	CAMPSQ1075	CAMPSQ5045
CAMPSQ3037	CAMPSQ4119	CAMPSQ923	CAMPSQ4155	CAMPSQ981	CAMPSQ3006	CAMPSQ3011	CAMPSQ4195
CAMPSQ3039	CAMPSQ3027	CAMPSQ3052	CAMPSQ4158	CAMPSQ741	CAMPSQ972	CAMPSQ1076	CAMPSQ4198
CAMPSQ1091	CAMPSQ4115	CAMPSQ3294	CAMPSQ4152	CAMPSQ500	CAMPSQ731	CAMPSQ965	CAMPSQ3092
CAMPSQ3886	CAMPSQ206	CAMPSQ3897	CAMPSQ826	CAMPSQ3164	CAMPSQ801	CAMPSQ3589	CAMPSQ4181
CAMPSQ3408	CAMPSQ3652	CAMPSQ3413	CAMPSQ3395	CAMPSQ4011	CAMPSQ4025	CAMPSQ3105	CAMPSQ3091
CAMPSQ680	CAMPSQ207	CAMPSQ681	CAMPSQ827	CAMPSQ817	CAMPSQ802	CAMPSQ3180	CAMPSQ2987
CAMPSQ210	CAMPSQ449	CAMPSQ682	CAMPSQ2063	CAMPSQ807	CAMPSQ4262	CAMPSQ4270	CAMPSQ2986
CAMPSQ695	CAMPSQ3894	CAMPSQ440	CAMPSQ828	CAMPSQ4009	CAMPSQ803	CAMPSQ4276	CAMPSQ4922
CAMPSQ211	CAMPSQ3893	CAMPSQ683	CAMPSQ818	CAMPSQ4006	CAMPSQ4020	CAMPSQ3187	CAMPSQ3833
CAMPSQ453	CAMPSQ208	CAMPSQ441	CAMPSQ4238	CAMPSQ4005	CAMPSQ3172	CAMPSQ4037	CAMPSQ3836
CAMPSQ696	CAMPSQ209	CAMPSQ684	CAMPSQ672	CAMPSQ4008	CAMPSQ3175	CAMPSQ4279	CAMPSQ2746
CAMPSQ454	CAMPSQ3890	CAMPSQ200	CAMPSQ430	CAMPSQ3391	CAMPSQ4264	CAMPSQ4278	CAMPSQ3830
CAMPSQ213	CAMPSQ3889	CAMPSQ442	CAMPSQ673	CAMPSQ820	CAMPSQ4263	CAMPSQ4031	CAMPSQ3832
CAMPSQ697	CAMPSQ4735	CAMPSQ443	CAMPSQ431	CAMPSQ3397	CAMPSQ4021	CAMPSQ4273	CAMPSQ2985
CAMPSQ214	CAMPSQ3888	CAMPSQ201	CAMPSQ432	CAMPSQ3155	CAMPSQ806	CAMPSQ3183	CAMPSQ2743
CAMPSQ699	CAMPSQ4738	CAMPSQ685	CAMPSQ676	CAMPSQ4002	CAMPSQ4259	CAMPSQ4275	CAMPSQ2984
CAMPSQ215	CAMPSQ3406	CAMPSQ444	CAMPSQ677	CAMPSQ3399	CAMPSQ4258	CAMPSQ3186	CAMPSQ490
CAMPSQ457	CAMPSQ4737	CAMPSQ686	CAMPSQ436	CAMPSQ4004	CAMPSQ4019	CAMPSQ4033	CAMPSQ491
CAMPSQ3881	CAMPSQ3643	CAMPSQ202	CAMPSQ679	CAMPSQ3157	CAMPSQ4013	CAMPSQ3185	CAMPSQ492
CAMPSQ458	CAMPSQ3400	CAMPSQ203	CAMPSQ437	CAMPSQ3398	CAMPSQ3166	CAMPSQ4032	CAMPSQ4937
CAMPSQ459	CAMPSQ3642	CAMPSQ445	CAMPSQ438	CAMPSQ824	CAMPSQ4254	CAMPSQ4274	CAMPSQ3609
CAMPSQ3880	CAMPSQ4731	CAMPSQ446	CAMPSQ1001	CAMPSQ3156	CAMPSQ4012	CAMPSQ4028	CAMPSQ481
CAMPSQ3883	CAMPSQ3403	CAMPSQ204	CAMPSQ3420	CAMPSQ4003	CAMPSQ3165	CAMPSQ4269	CAMPSQ2519
CAMPSQ3641	CAMPSQ3645	CAMPSQ688	CAMPSQ3415	CAMPSQ3151	CAMPSQ4015	CAMPSQ4260	CAMPSQ240
CAMPSQ218	CAMPSQ2798	CAMPSQ447	CAMPSQ3417	CAMPSQ3393	CAMPSQ813	CAMPSQ4266	CAMPSQ482
CAMPSQ4730	CAMPSQ3887	CAMPSQ3892	CAMPSQ3896	CAMPSQ825	CAMPSQ814	CAMPSQ4023	CAMPSQ4939
CAMPSQ3882	CAMPSQ3644	CAMPSQ3891	CAMPSQ3656	CAMPSQ3392	CAMPSQ815	CAMPSQ4265	CAMPSQ494
CAMPSQ3635	CAMPSQ3402	CAMPSQ448	CAMPSQ3898	CAMPSQ3150	CAMPSQ3161	CAMPSQ800	CAMPSQ495
CAMPSQ3680	CAMPSQ410	CAMPSQ1034	CAMPSQ1038	CAMPSQ632	CAMPSQ3872	CAMPSQ3877	CAMPSQ253
CAMPSQ1017	CAMPSQ411	CAMPSQ3695	CAMPSQ3699	CAMPSQ874	CAMPSQ4714	CAMPSQ3638	CAMPSQ254
CAMPSQ1018	CAMPSQ654	CAMPSQ406	CAMPSQ880	CAMPSQ3463	CAMPSQ3625	CAMPSQ3637	CAMPSQ496
CAMPSQ3437	CAMPSQ412	CAMPSQ3453	CAMPSQ640	CAMPSQ1042	CAMPSQ3867	CAMPSQ3874	CAMPSQ497
CAMPSQ3679	CAMPSQ655	CAMPSQ3690	CAMPSQ641	CAMPSQ876	CAMPSQ3866	CAMPSQ3632	CAMPSQ256
CAMPSQ1019	CAMPSQ897	CAMPSQ649	CAMPSQ400	CAMPSQ3462	CAMPSQ3869	CAMPSQ4720	CAMPSQ74
CAMPSQ3439	CAMPSQ414	CAMPSQ3450	CAMPSQ642	CAMPSQ877	CAMPSQ4716	CAMPSQ3876	CAMPSQ73
CAMPSQ1013	CAMPSQ1020	CAMPSQ3692	CAMPSQ885	CAMPSQ635	CAMPSQ3626	CAMPSQ4723	CAMPSQ72
CAMPSQ3676	CAMPSQ2351	CAMPSQ1030	CAMPSQ643	CAMPSQ1044	CAMPSQ3868	CAMPSQ4722	CAMPSQ70

Continúa en la siguiente página

Tabla 28 Casos positivos: Péptidos antimicrobianos – *Continuación*

CAMPSQ1014	CAMPSQ656	CAMPSQ3691	CAMPSQ401	CAMPSQ3465	CAMPSQ4715	CAMPSQ3875	CAMPSQ2734
CAMPSQ3675	CAMPSQ3441	CAMPSQ3449	CAMPSQ644	CAMPSQ636	CAMPSQ3863	CAMPSQ3639	CAMPSQ2975
CAMPSQ3436	CAMPSQ3440	CAMPSQ1028	CAMPSQ645	CAMPSQ3464	CAMPSQ3621	CAMPSQ4728	CAMPSQ2733
CAMPSQ3678	CAMPSQ1021	CAMPSQ3207	CAMPSQ3452	CAMPSQ1045	CAMPSQ3862	CAMPSQ691	CAMPSQ2978
CAMPSQ1015	CAMPSQ3682	CAMPSQ3448	CAMPSQ11	CAMPSQ638	CAMPSQ3623	CAMPSQ461	CAMPSQ2977
CAMPSQ1016	CAMPSQ2352	CAMPSQ1029	CAMPSQ403	CAMPSQ2370	CAMPSQ3865	CAMPSQ220	CAMPSQ2972
CAMPSQ3677	CAMPSQ415	CAMPSQ1024	CAMPSQ3210	CAMPSQ639	CAMPSQ3864	CAMPSQ221	CAMPSQ2730
CAMPSQ3435	CAMPSQ3443	CAMPSQ3445	CAMPSQ1031	CAMPSQ3461	CAMPSQ3622	CAMPSQ222	CAMPSQ69
CAMPSQ660	CAMPSQ2353	CAMPSQ3444	CAMPSQ888	CAMPSQ3460	CAMPSQ4717	CAMPSQ223	CAMPSQ2971
CAMPSQ661	CAMPSQ658	CAMPSQ1026	CAMPSQ10	CAMPSQ3698	CAMPSQ3628	CAMPSQ225	CAMPSQ68
CAMPSQ420	CAMPSQ3685	CAMPSQ3447	CAMPSQ3693	CAMPSQ1035	CAMPSQ4719	CAMPSQ226	CAMPSQ2997
CAMPSQ662	CAMPSQ659	CAMPSQ3688	CAMPSQ1032	CAMPSQ3455	CAMPSQ870	CAMPSQ468	CAMPSQ2513
CAMPSQ422	CAMPSQ1023	CAMPSQ3446	CAMPSQ404	CAMPSQ3213	CAMPSQ871	CAMPSQ469	CAMPSQ3602
CAMPSQ423	CAMPSQ3442	CAMPSQ890	CAMPSQ3454	CAMPSQ1036	CAMPSQ630	CAMPSQ3870	CAMPSQ4933
CAMPSQ424	CAMPSQ419	CAMPSQ650	CAMPSQ1033	CAMPSQ3697	CAMPSQ631	CAMPSQ228	CAMPSQ4936
CAMPSQ3672	CAMPSQ3681	CAMPSQ651	CAMPSQ3696	CAMPSQ1037	CAMPSQ873	CAMPSQ229	CAMPSQ3847
CAMPSQ3854	CAMPSQ2992	CAMPSQ2933	CAMPSQ22	CAMPSQ3818	CAMPSQ40	CAMPSQ1010	CAMPSQ2516
CAMPSQ3612	CAMPSQ2530	CAMPSQ2930	CAMPSQ21	CAMPSQ2728	CAMPSQ3812	CAMPSQ668	CAMPSQ3605
CAMPSQ2522	CAMPSQ86	CAMPSQ24	CAMPSQ2943	CAMPSQ4906	CAMPSQ2965	CAMPSQ669	CAMPSQ2999
CAMPSQ3611	CAMPSQ3861	CAMPSQ23	CAMPSQ19	CAMPSQ273	CAMPSQ2723	CAMPSQ427	CAMPSQ3604
CAMPSQ3618	CAMPSQ3860	CAMPSQ2939	CAMPSQ18	CAMPSQ3817	CAMPSQ2722	CAMPSQ3432	CAMPSQ3846
CAMPSQ3617	CAMPSQ85	CAMPSQ2936	CAMPSQ16	CAMPSQ50	CAMPSQ3811	CAMPSQ1011	CAMPSQ2994
CAMPSQ470	CAMPSQ84	CAMPSQ2935	CAMPSQ15	CAMPSQ2954	CAMPSQ2964	CAMPSQ1012	CAMPSQ3841
CAMPSQ3619	CAMPSQ83	CAMPSQ2938	CAMPSQ14	CAMPSQ2953	CAMPSQ3814	CAMPSQ429	CAMPSQ2993
CAMPSQ4708	CAMPSQ82	CAMPSQ2937	CAMPSQ2941	CAMPSQ2955	CAMPSQ39	CAMPSQ3427	CAMPSQ3075
CAMPSQ483	CAMPSQ81	CAMPSQ472	CAMPSQ13	CAMPSQ48	CAMPSQ4903	CAMPSQ3669	CAMPSQ4159
CAMPSQ241	CAMPSQ3614	CAMPSQ80	CAMPSQ2940	CAMPSQ47	CAMPSQ2725	CAMPSQ2338	CAMPSQ4185
CAMPSQ242	CAMPSQ3613	CAMPSQ230	CAMPSQ12	CAMPSQ46	CAMPSQ38	CAMPSQ1007	CAMPSQ5034
CAMPSQ90	CAMPSQ2527	CAMPSQ231	CAMPSQ290	CAMPSQ45	CAMPSQ2966	CAMPSQ1008	CAMPSQ2979
CAMPSQ243	CAMPSQ3616	CAMPSQ232	CAMPSQ293	CAMPSQ2951	CAMPSQ2724	CAMPSQ1009	CAMPSQ261
CAMPSQ485	CAMPSQ2526	CAMPSQ233	CAMPSQ294	CAMPSQ3809	CAMPSQ2961	CAMPSQ1003	CAMPSQ2745
CAMPSQ486	CAMPSQ3615	CAMPSQ234	CAMPSQ295	CAMPSQ280	CAMPSQ36	CAMPSQ3664	CAMPSQ3834
CAMPSQ244	CAMPSQ4704	CAMPSQ476	CAMPSQ33	CAMPSQ281	CAMPSQ3810	CAMPSQ3667	CAMPSQ76
CAMPSQ487	CAMPSQ3857	CAMPSQ235	CAMPSQ32	CAMPSQ3804	CAMPSQ34	CAMPSQ3425	CAMPSQ2980
CAMPSQ246	CAMPSQ3852	CAMPSQ477	CAMPSQ31	CAMPSQ282	CAMPSQ3819	CAMPSQ3424	CAMPSQ3849
CAMPSQ489	CAMPSQ3610	CAMPSQ479	CAMPSQ30	CAMPSQ2957	CAMPSQ3816	CAMPSQ274	CAMPSQ3848
CAMPSQ247	CAMPSQ2520	CAMPSQ237	CAMPSQ2932	CAMPSQ283	CAMPSQ4905	CAMPSQ277	CAMPSQ3844
CAMPSQ99	CAMPSQ4940	CAMPSQ88	CAMPSQ29	CAMPSQ284	CAMPSQ270	CAMPSQ278	CAMPSQ2755
CAMPSQ3850	CAMPSQ3851	CAMPSQ238	CAMPSQ2931	CAMPSQ296	CAMPSQ271	CAMPSQ279	CAMPSQ902
CAMPSQ95	CAMPSQ79	CAMPSQ87	CAMPSQ28	CAMPSQ297	CAMPSQ3815	CAMPSQ43	CAMPSQ905
CAMPSQ94	CAMPSQ4701	CAMPSQ239	CAMPSQ27	CAMPSQ299	CAMPSQ2726	CAMPSQ2970	CAMPSQ5038
CAMPSQ4160	CAMPSQ2523	CAMPSQ4161	CAMPSQ4189	CAMPSQ264	CAMPSQ2995	CAMPSQ93	CAMPSQ5033
CAMPSQ5016	CAMPSQ3099	CAMPSQ2731	CAMPSQ2750	CAMPSQ265	CAMPSQ89	CAMPSQ92	CAMPSQ2737
CAMPSQ3079	CAMPSQ4188	CAMPSQ3827	CAMPSQ62	CAMPSQ266	CAMPSQ4931	CAMPSQ2998	CAMPSQ260
CAMPSQ61	CAMPSQ267	CAMPSQ3842	CAMPSQ3603				
CAMPSQ60	CAMPSQ269	CAMPSQ4938	CAMPSQ3845				

Tabla 29: Casos positivos: Péptidos antimicrobianos. Conjunto de validación compuesto por 202 péptidos recuperados de la base de datos CAMP.

CAMPSQ313	CAMPSQ3299	CAMPSQ219	CAMPSQ3004	CAMPSQ3548	CAMPSQ408	CAMPSQ4127	CAMPSQ1081
CAMPSQ2982	CAMPSQ3071	CAMPSQ3923	CAMPSQ1121	CAMPSQ17	CAMPSQ409	CAMPSQ467	CAMPSQ3094
CAMPSQ3433	CAMPSQ123	CAMPSQ405	CAMPSQ114	CAMPSQ224	CAMPSQ533	CAMPSQ4087	CAMPSQ2533
CAMPSQ933	CAMPSQ1170	CAMPSQ317	CAMPSQ3456	CAMPSQ1187	CAMPSQ450	CAMPSQ4892	CAMPSQ2540
CAMPSQ3979	CAMPSQ1175	CAMPSQ592	CAMPSQ637	CAMPSQ433	CAMPSQ603	CAMPSQ3765	CAMPSQ4191
CAMPSQ7	CAMPSQ747	CAMPSQ2899	CAMPSQ916	CAMPSQ3380	CAMPSQ4052	CAMPSQ328	CAMPSQ708
CAMPSQ4272	CAMPSQ3419	CAMPSQ3921	CAMPSQ3994	CAMPSQ865	CAMPSQ428	CAMPSQ41	CAMPSQ3954
CAMPSQ513	CAMPSQ4117	CAMPSQ20	CAMPSQ4721	CAMPSQ361	CAMPSQ3740	CAMPSQ4267	CAMPSQ3146
CAMPSQ5019	CAMPSQ847	CAMPSQ2518	CAMPSQ3835	CAMPSQ3009	CAMPSQ4209	CAMPSQ1165	CAMPSQ4733
CAMPSQ4312	CAMPSQ3482	CAMPSQ245	CAMPSQ174	CAMPSQ407	CAMPSQ3096	CAMPSQ2515	CAMPSQ4710
CAMPSQ3565	CAMPSQ3884	CAMPSQ413	CAMPSQ1052	CAMPSQ25	CAMPSQ1006	CAMPSQ2514	CAMPSQ3278
CAMPSQ887	CAMPSQ1191	CAMPSQ4386	CAMPSQ3747	CAMPSQ3093	CAMPSQ711	CAMPSQ4092	CAMPSQ1135

Continúa en la siguiente página

Tabla 29 Casos positivos: Péptidos antimicrobianos – *Continuación*

CAMPSQ3055	CAMPSQ292	CAMPSQ2962	CAMPSQ3674	CAMPSQ750	CAMPSQ670	CAMPSQ2901	CAMPSQ4193
CAMPSQ663	CAMPSQ3159	CAMPSQ3916	CAMPSQ141	CAMPSQ687	CAMPSQ3624	CAMPSQ4054	CAMPSQ3024
CAMPSQ217	CAMPSQ3506	CAMPSQ3813	CAMPSQ3504	CAMPSQ381	CAMPSQ931	CAMPSQ3160	CAMPSQ3737
CAMPSQ3974	CAMPSQ3735	CAMPSQ2988	CAMPSQ3655	CAMPSQ724	CAMPSQ692	CAMPSQ3194	CAMPSQ3189
CAMPSQ4179	CAMPSQ463	CAMPSQ3666	CAMPSQ3901	CAMPSQ186	CAMPSQ769	CAMPSQ3631	CAMPSQ3040
CAMPSQ3188	CAMPSQ843	CAMPSQ2720	CAMPSQ3627	CAMPSQ3379	CAMPSQ2981	CAMPSQ3526	CAMPSQ3630
CAMPSQ3486	CAMPSQ553	CAMPSQ511	CAMPSQ3533	CAMPSQ3760	CAMPSQ3766	CAMPSQ3636	CAMPSQ5006
CAMPSQ917	CAMPSQ3871	CAMPSQ4256	CAMPSQ611	CAMPSQ872	CAMPSQ3330	CAMPSQ4383	CAMPSQ1022
CAMPSQ3378	CAMPSQ4220	CAMPSQ3124	CAMPSQ4893	CAMPSQ4131	CAMPSQ439	CAMPSQ3936	CAMPSQ343
CAMPSQ3885	CAMPSQ3873	CAMPSQ953	CAMPSQ3825	CAMPSQ3044	CAMPSQ3895	CAMPSQ255	CAMPSQ1077
CAMPSQ5039	CAMPSQ4215	CAMPSQ4207	CAMPSQ653	CAMPSQ634	CAMPSQ3924	CAMPSQ582	CAMPSQ987
CAMPSQ4423	CAMPSQ3826	CAMPSQ216	CAMPSQ4038	CAMPSQ883	CAMPSQ3581	CAMPSQ3128	CAMPSQ3665
CAMPSQ805	CAMPSQ4122	CAMPSQ4043	CAMPSQ3700	CAMPSQ3738	CAMPSQ227	CAMPSQ3937	CAMPSQ1166
CAMPSQ524							
CAMPSQ2969							

Tabla 30: Casos Negativos : Péptidos no antimicrobianos. Conjunto de prueba y entrenamiento compuesto por 1500 péptidos recuperados de la base de datos Uniprot.

Q9PSP2	Q9R583	Q7M038	Q9R4I4	P83890	UniRef90_B3A0I4
Q9R5K3	P82006	1XH5_P61925	1JMX_P0A182	P84819	Q9TWN5
1D5S_P01009	1EJY_P05221	1MV9_Q15596	P19864	P14487	1K1F_P11274
P84917	1UKL_Q12772	1JBU_1JBU	Q7M1T2	P80625	1EMU_P25054
Q9R5Q2	Q7M0Z8	Q9PRX8	P36987	P80453	Q9UWM6
P81363	Q7M062	1SM3_P15941	P30879	P15871	P0CAP8
Q9R4I5	Q9TRF9	1XH6_P61925	P85404	A8I8G4	2BFX_O13024
UniRef90_P86284	1LKY_P41212	1XXA_P0A6D0	P85267	Q9TRL1	P82005
UniRef90_P0C1X3	P84535	1FXR_P00210	Q7LZ26	Q9R4N8	1KZZ_Q92844
1GFW_P42574	Q9TRB9	P85337	Q7M3G2	P83145	1K74_Q15788
Q9R5S3	Q9TWG3	2C1B_P61925	Q9C6P0	P0DMB7	Q68K21
1PEG_P61831	UniRef90_P86466	Q9S8A2	P85323	Q7M0W5	1X7E_1X7E
1S9V_1S9V	P0CAT0	P80606	P81136	Q7M0J1	1LUJ_Q9NSA3
P0DMB9	Q0PGA5	Q9TQQ8	P81666	P82952	P58689
P10094	UniRef90_P86465	Q7M4E1	Q9TSA4	Q53128	Q7M2Q9
P80792	Q9R5T5	4CC7_O00401	Q9TRF1	Q9TRD8	P0DJL1
Q9TRR9	P0DJN8	Q9UWG9	P20141	P58690	Q9R4A1
P0DJF7	1P0T_Q96RJ3	Q9TWF6	O11822	P06884	Q9S8C2
UniRef90_P42987	Q7M3P2	UniRef90_G3E7W2	P28525	Q9T2R7	1DLH_P04664
Q7M0Y8	Q9TR12	Q7M1P6	P82999	Q9S9H9	Q7M1L5
UniRef90_P84189	P84986	1F47_P0A9A6	O36236	P81747	P12666
UniRef90_P17684	P80701	P20416	Q9PRV5	5R1R_P69924	Q7M0C8
1U3R_1U3R	P80349	P19917	UniRef90_P62328	P80462	Q9TRA4
P81247	P81638	Q9LUX3	P81716	P14467	Q9S8N4
Q9S8X7	P81778	P34153	Q9TR69	P84477	F6Z4Q6
1W5C_Q8DIP0	Q7M2P5	Q9PSR6	P82681	Q9TR44	P14485
1B0X_Q03137	UniRef90_P08611	P84565	P83048	P16351	2GSI_P00644
Q9R4C8	P83012	P21227	Q9TWI3	2UZW_Q3SX13	Q9PRW2
Q9S8Y1	UniRef90_P82092	P85006	P80974	Q7M1H1	P86207
Q7M0A6	P85218	P39092	1TAF_Q27272	UniRef90_I3AXJ7	Q9TRP1
1W5C_Q8DIN9	UniRef90_P82098	1W8X_P27391	P14476	Q9TWF2	P80610
1L0H_P0A6A8	P19122	2OH0_2OH0	Q9R5U5	P85394	P83507
P83598	UniRef90_P82097	1PTQ_P28867	1OV3_4557505	Q9S8B0	P84541
P36984	B1AXT0	P86248	P85960	Q9R4Y0	Q7M2F8
Q9TR29	P16093	P83540	Q1I165	Q94I97	1BAZ_P03050
1H64_Q9V0Y8	1KRL_P02881	Q9S8P2	Q9Q3N3	P85910	Q9R4M9
Q9R4T8	A8IZA8	UniRef90_I3B8J6	Q7LZT6	P85353	Q9PS26
Q7M4A8	Q7LZ35	UniRef90_B3A090	P20412	1EBD_P11961	P0DJJ7
1CFS_1CFS	1KRL_P02882	Q9UWM1	P80347	P58602	Q9TRF7
Q9TRN6	Q7M0M1	Q9S8J5	P68121	P80348	P04576
1BR8_1BR8	Q9R4Z1	1RTF_1RTF	P13858	P86876	P63181
UniRef90_K8BDW9	P83918	P81543	P84468	Q7M435	Q1HVA2
1SVE_P61926	1D0D_P17726	1MZW_O43172	P85062	P85321	2G01_Q9UQF2

Continúa en la siguiente página

Tabla 30 Casos Negativos: Péptidos no antimicrobianos – *Continuación*

P81246	P84000	Q7M0I9	Q7M0Q1	Q9ST33	1UB4_P0AE72
1KG0_1KG0	O10481	1TAF_P49847	P86001	Q9R5F7	UniRef90_B3DSG7
P80770	1C94_1C94	P84532	P85985	Q5Y971	P14539
Q9R5E6	Q9TRT3	1M45_P19524	P83246	2F7X_Q71U53	1TMT_P28504
Q9PRU0	Q7M3E6	Q9PRU8	Q9T2K8	UniRef90_E3SZP6	1TET_P32890
1P22_P35222	Q14SA3	P27063	P0DJN0	P09684	Q7M0K8
1BE3_P07552	Q7M198	Q9R4A7	P80368	UniRef90_B0M2U1	1MIK_NOR00033
P86712	P81882	P80487	UniRef90_P0C7S3	P80760	D3Z0A5
Q9R5Q0	P86069	UniRef90_P84719	Q7M1K4	Q9S922	P82248
P0C7B1	Q7M0N2	6R1R_P69924	O34199	P84351	Q9S9H1
4CPA_P01075	P0DJF4	2H96_Q9UQF2	P0C8D3	Q86H22	Q7M046
Q9TWE2	Q9TRP6	Q9R4D8	Q7M103	Q9R5S2	Q9TRM6
Q9S8I7	UniRef90_I3B5B6	1QGW_Q00433	Q7M3X9	Q10985	2CJC_Q9HPW4
P85112	Q7M088	P0C1H2	Q38782	2PUY_Q96BD5	P81072
P83728	Q9R4R9	2CK0_2CK0	Q9PRL7	Q9PRQ2	P83444
Q7LZE6	1G6G_1G6G	P81949	P85428	C0HJG6	Q9TWM8
P83092	1PGX_P06654	1YCP_P02671	P82332	Q69H22	P84619
P86467	P86787	UniRef90_Q9R582	P19094	P22775	1MZN_Q15596
Q7M2W1	Q9R5C1	1LDD_Q12440	Q7M236	P85338	P0DJB0
Q7M381	Q9R4Z7	P81987	P85935	Q8VYX9	Q7M483
1CWJ_NOR00036	Q7M1A7	1QOJ_P0A8F8	2EBO_Q05320	P0C8C7	Q7M4J9
P84784	Q9R4D3	Q7M1D6	P14477	P85401	Q7M0N9
P84554	Q41503	Q9R4C7	2RMA_NOR00033	P80629	P83899
Q9T2U1	P0C2W4	P23032	Q9TWG7	Q09053	UniRef90_P80578
Q7M0J4	Q7M3Q4	P80787	P85430	P84542	B3EWP2
Q9S8T0	Q00M74	Q29431	Q9PRR0	Q8LJU4	UniRef90_Q7M0L2
P81622	Q9R4A0	P84545	P85926	1BOG_1BOG	P00728
1FM0_P30748	UniRef90_P08609	Q9T2Q9	P13642	Q9TRP3	P81802
1RV1_Q00987	P80442	B7FFP2	P81869	2H1P_2H1P	Q9R4J0
Q9TWR0	P84462	C5HA90	P82335	1AJJ_P01130	Q7LZH2
A8JK20	UniRef90_P82089	P86420	P0DL20	P81874	1DM0_Q7BQ98
Q9UWN0	Q9S8M5	Q9TWJ4	B9HKE5	P60262	Q9TWR2
P82989	Q9R5N5	P56651	P81883	1TME_P13899	P80582
B2BGU9	Q7M1A4	Q9R551	P55823	1CWK_NOR00036	Q7M2X3
Q9TRG2	Q9R5D7	Q70AA1	P86534	1BH0_P01275	P68123
1VCC_P68698	2DVQ_P02309	B7FG99	P0C5S7	Q9TWP2	1MFT_1MFT
UniRef90_B0M3D5	Q9S8N0	P83477	Q7M3E7	Q4YDA8	A8HVY0
Q9TR14	P85842	P30947	1GNG_Q92837	1ET1_P01270	Q9TXF8
Q9TRB6	Q7M1Q2	Q9R5I3	B3EWS0	P55749	Q9R4U0
Q7M165	Q6LDN4	P82297	P01182	P14536	P85905
UniRef90_P0DJC3	P18657	P86074	1JBP_P61926	Q9S878	1CYN_NOR00033
P84877	P86365	Q9TR03	Q9TWJ9	P10846	P0C577
P0DKT7	Q7M1R6	Q9R5P6	Q9UWL1	Q9R4W6	Q9TS81
Q7M264	1LUZ_P18378	Q9S8W8	Q9R4N2	1LJO_O29885	Q9PRG2
Q9R4C6	Q7M170	1JYI_1JYI	P86343	1SVH_P63249	1I3Q_P40422
1GL2_Q9Z2Q7	1MHM_Q04694	2E3K_P02309	Q42271	P80444	2CPK_P63248
Q9R5E1	Q7M3S1	Q7M3Z5	Q9S8Z9	Q9R4I7	A6N1T7
A3QP71	P82796	2EIL_P13183	P85444	P35756	P80612
P85325	Q7M355	1AS4_P01011	Q7M3V2	C0HJD0	P91948
1VJW_P46797	P81875	Q9R4A5	UniRef90_B3A0K4	B3EWS6	1CIQ_P01053
1FXD_P00209	Q9S8U3	D3Z606	Q9S8H1	Q9TRE8	P84721
1OXG_P00766	P68354	D3YZ25	Q9S8Q6	Q9S8T9	UniRef90_P83181
1ICF_P07711	P24927	1FI8_P23827	Q7M1B4	Q7M0Z0	P80626
Q7M0W2	Q7M166	1A3B_P28504	Q9TR80	1FPT_P03300	Q7LZ24
P20739	1CI6_P28033	P83657	P85366	P46380	1RDT_Q92793
Q9TRP4	P0DMG8	1SYQ_Q9Y490	Q7M2R9	Q7DM06	P22028
Q9PSQ3	P84109	UniRef90_P84726	B7PXR4	P17698	P82301
Q9TWL5	Q9S8R1	1YCR_P04637	Q7M3M1	P82327	2PRG_Q15788
P02875	P0DJN6	Q7M0G3	P86366	P84027	P20242
P58909	P83347	Q9TRS9	P86080	P68124	1CWL_NOR00033
Q41065	Q7M0E1	P37300	Q7LZ40	P0CH73	Q9TQZ6
Q7M2Y3	Q9PRR4	1BH8_Q15543	B3EWN4	Q9S908	Q7M262
P59682	Q9TWU8	P85932	P80537	Q7M544	Q26181

Continúa en la siguiente página

Tabla 30 Casos Negativos: Péptidos no antimicrobianos – *Continuación*

P49820	2IYB_Q9UGI8	1HPI_P38524	P86254	P84538	1SMH_P61926
Q63337	Q7M0L1	P25937	P80805	Q9TSC4	Q7M049
1GL2_O70439	Q7M399	Q9TWQ0	Q7LZL0	Q9R4V9	Q9TS60
4SRN_P61823	Q7M0N7	P80706	P11918	Q9TRS2	Q9T2Q5
Q7M173	P80647	Q7M2N6	P80501	P14466	2IGF_P02247
Q9TS54	1QSN_P61830	1UN0_P32499	C7E2T8	Q9S8D7	Q9PSQ8
1KU3_Q9EZJ8	P85896	1CI6_P18848	1LGB_P12307	P30806	Q7M0P2
1ETR_P00735	P85336	Q8I6R5	P86217	P82325	Q9PRR6
Q9S8L0	P61094	Q9S8Q4	P83961	1XH9_P61925	A8JJF6
Q7M1R8	D8SRQ0	P0DJM8	Q7LZS7	Q9PRM1	P45661
P83357	P86290	1NHG_Q9BH77	Q9PRZ2	P12802	Q9R496
P83445	2HGT_P28504	1FBM_P35444	Q9S9G2	P0C1M8	P86348
Q7M0L4	1C8O_P07385	1I27_P35269	Q42222	Q9TRI6	A8JHV1
1JLU_P61925	Q9R5P4	P85354	P85914	Q7M2L3	Q7M110
Q9S8B1	Q9S8N1	Q9TRG7	Q9NFI4	Q9R4N1	P68119
UniRef90_Q7LZE0	Q7M0S3	P0DJN7	Q7M2F9	Q43281	P86071
P0C8X1	P20732	UniRef90_P0C2A2	P14538	UniRef90_F6LNL7	Q9TWE8
1CWM_NOR00033	P83001	Q7M0X6	1STC_P61926	UniRef90_T2I2F9	Q9R4I9
Q9R5L9	Q0PHC9	Q9S8T6	1WYX_P56945	P86690	Q9T2V5
P60261	P27067	UniRef90_P0C2A1	P84909	Q7M0K9	P0C8J0
Q9UWK4	Q7M059	Q9TWW9	Q9TXH0	Q9S8Q5	Q9R592
UniRef90_T2I2Y8	P85390	Q9PS01	2EIL_P00430	P0CAR5	2EIL_P07470
P80824	P0DJQ1	UniRef90_P0C2A3	Q9R4R4	Q7M383	P29134
P82296	P82816	P83009	Q7M4I4	UniRef90_F6LNM4	P0C2K3
Q7M384	Q7M3Q1	P17877	Q7M486	P86103	Q9TWE3
P80774	P86689	Q8W502	1GL2_Q9WUF4	Q7M403	Q9TRJ5
P82024	Q7M125	P83148	Q9S932	1XH4_P61925	P42706
P23473	P83146	1M5N_P12272	P80736	Q7M1A0	P68396
P20900	P14474	Q7M163	P14447	Q42781	Q7M484
P84564	Q7M2P2	1BBT_O90754	P84821	Q9R546	Q9S8U2
P21843	P83572	1X7R_1X7R	P13062	UniRef90_A7KZR0	1RKC_P54939
P85413	D3YYG4	P82441	P85343	Q9R5Q8	P81107
P83688	P82947	Q7M1M9	P86739	B5AH74	C0HJA7
P82309	Q9PRX6	Q9S909	P80514	P02677	Q9T2R0
1UDI_P14739	Q9PSP3	UniRef90_G3E7N6	Q7M1X7	Q0NZX5	P82895
1MOF_P03385	1FE6_Q54436	1CWA_NOR00033	Q9R5E9	P83352	UniRef90_B3DSA5
2F8E_Q9DS05	Q9UWJ7	Q9S8J2	Q9S8I0	P24475	1THR_P26631
Q9PRS9	Q9R5I2	P80529	Q9S8A6	Q9TXG6	1PTF_P07515
P84552	P80526	P26888	Q7LZG2	Q7M1X4	1XH8_P61925
2C1A_P61925	P83956	Q7M0Y7	Q9S8V2	P45668	P85096
1C26_P04637	P20733	P85843	P85250	Q7M150	O04243
Q9PS38	Q9R5F8	P02243	P84708	1LM8_Q15369	1HLE_P05619
1QGK_P52292	P80772	Q0H636	P85491	Q9R5T3	Q9R4A3
2GMX_Q9UQF2	Q9S8H6	P81083	UniRef90_P85252	1ZEI_P01315	1VIE_P00383
Q9TRM5	Q9TWX0	Q38768	Q9R5C7	P85445	P0C2S2
1YYE_1YYE	Q9R568	UniRef90_B3A0A9	1L0A_Q92844	P18647	P59072
P85493	1KO6_P52948	Q9R4H5	Q9S940	P85261	P80981
1G1S_Q14242	P86468	UniRef90_I3AUC4	Q9R4Y5	Q9TWF7	P09688
1I5K_P49054	1MJ4_P51687	Q9TRX7	Q7M218	1G0Y_1G0Y	P81532
2RMC_NOR00033	2IBF_P18010	P14453	P81164	Q9S8M8	P83061
P32954	Q7M0K0	P84478	P84846	Q9T2V4	P82304
1IAK_Q29431	P81087	P81670	P80742	P82311	P86175
Q9T2N9	B3A0N1	D3YV76	P86979	Q9R4Y6	Q9TR30
P05582	UniRef90_S7K9T3	P68357	P85270	Q9S8T8	1ISU_P33678
Q7LZJ1	P81173	P85269	A6PWK9	Q7M0Q7	Q9M677
Q0H425	Q9R4A8	Q9S8H0	P82439	Q7M276	P86058
Q5SVY1	Q9R5S6	Q9TRH0	Q9T2S6	Q9TR06	Q7M1D1
1X78_1X78	P04362	2DVR_P02309	P82256	P80636	P81110
Q9TR11	Q7M0A0	Q9S8P3	P69502	Q9T2G9	Q9R529
P27687	P0CH59	P68118	1QZM_P0A9M0	P81080	P59760
2VB2_P77214	Q09097	1HT9_P02633	Q79428	P68218	Q9R4Y8
1GL2_O88384	P84731	P83607	Q7M493	P0C6S1	Q7M3G8
O88687	P83364	Q9R5D2	P85122	P85956	P19977

Continúa en la siguiente página

Tabla 30 Casos Negativos: Péptidos no antimicrobianos – *Continuación*

Q7M261	Q5YBB7	Q9TWV4	Q7M417	Q9S938	1QCR_P13272
P82302	Q9R4I1	P85156	P81669	P0DL27	Q9S8J4
2BZW_Q61337	Q7M362	P36215	A1Z199	P82536	Q9UWI4
Q7M3D6	P0CV89	Q9R4L6	Q9PS35	P00168	Q7M3F1
P68219	Q10721	D3Z5A6	Q9R558	F7BP55	Q9R4W0
P30370	Q9R4W4	Q9T2K5	Q9TWR3	P81496	1A3E_P28504
Q6LAD3	P86681	1L6L_P02652	P13570	Q9TWD1	1G2C_P11209
P84470	1K7L_Q15788	P34177	P80472	P0C603	Q7M3G0
Q9PRT6	Q7M2Q0	P0C1G1	Q7M1U7	1POH_P0AA04	P85948
P07855	Q7M433	Q7M2X4	Q7M199	Q7M0T7	2EIL_P10175
Q7M068	Q7LZT3	P0C8X7	P56623	P84558	P84983
1A81_P07766	P21597	Q9R5R7	UniRef90_P85260	P19865	P01496
P52964	Q7M401	Q7M2Z1	Q9R4V3	P68353	P56281
1VPP_1VPP	P84576	Q9TS95	1AAP_P05067	P81720	P81421
P81078	D0SG48	1CKS_P33552	P84812	Q0PEL9	P55739
1X76_Q15788	P06297	2UZT_Q3SX13	P84587	Q9T2H4	P83090
Q7M389	Q7M1C6	D1NCZ2	1M46_P19524	P80532	P0DJF1
P81800	Q1X8P0	Q9T2Q3	Q9S8Q7	Q9PRP2	P84579
UniRef90_P84985	P82452	P80486	A9LIT6	Q7M552	P14465
P19628	Q9S8R7	Q7M3E5	P84795	1HLT_P07204	Q10584
Q9S8P6	P82150	1RPO_P03051	P14449	P82207	Q9TR71
1B13_P00268	Q7M241	Q9S8K2	1SFI_Q4GWU5	Q9TXG4	P22949
1B4B_O31408	A8HVD7	P83729	1JMT_P26368	P83889	P02260
D3YWC8	P80757	P04099	Q7M3B2	P83330	Q9R5E3
1DS5_P67870	P81163	Q7M3H1	Q9S9I3	C0HJG8	Q9T2N5
1UM2_P17255	1PPE_P01074	1A64_P08921	1LJ2_Q04637	2BPT_P20676	Q9TWJ3
P68216	1TUC_P07751	D3Z7I7	P85963	Q9S8M3	P34175
1UTG_P02779	Q7M1L6	P56624	P84557	Q9TR05	1G3J_P70062
Q9R4Q5	Q10987	A2A6G7	Q7M1Y9	P80413	Q9TS61
Q7M3Q8	C7U2Y3	1CC7_P38636	P99505	Q7M1W8	Q7M1L9
P82340	Q7M154	P69658	1FYN_P06241	P27459	Q98Y46
P85111	Q7M1A2	P81593	Q7M272	P85961	Q9TRK2
P84851	B5KP01	1UCR_Q46582	P85903	1HTA_P48781	1YY4_Q15788
Q9PS48	C0HJD1	P85329	Q9T2U3	Q9PSQ1	Q7M4E6
1JEN_P17707	Q9R4P7	1LEW_	Q9R4L5	O09893	P85996
P22950	Q7M140	Q16938	Q9TWI5	1X7J_1X7J	Q7M3F5
Q9TR54	P13571	1SSB_P61823	Q9PRQ4	D3YUN8	A8JJ58
Q7LZF8	P80753	P81151	Q9R4L3	Q9TWK6	2CA3_Q9LA15
P00881	H3BL84	Q7M414	P84727	P01081	P19371
2BVR_2BVR	P18927	P99507	P86529	Q9R542	Q7M186
P80998	P80729	1N64_P29846	Q7M1X3	P05393	Q6B519
P85383	2F7E_P61925	Q9TRN5	Q7M0J2	Q9TWU0	1BJP_Q01468
A2KLM6	Q9UWM3	1VEB_P61925	P86357	Q9T2Q2	1U4L_P13501
Q7M1J4	Q7M2M6	Q9S8K0	P0C2C3	A9SY68	1GTD_O26271
Q9TXG1	Q9PS09	1SIZ_P29603	1JWG_P11717	2OJF_2OJF	P81663
3SRN_P61823	Q7M1V1	1GFF_P03652	P82312	Q7M547	P58570
P83366	A8CYC8	2NX5_P03206	C0SJS5	Q9R4H7	Q7M3H7
Q9TWW0	P86081	1F3J_P00698	P80145	P81106	Q7LZ32
P59681	Q7M3B4	P82329	Q9TQB0	Q9TWJ5	Q7M4K8
Q7M2V8	Q7LZW6	Q9T2P0	Q7M0V7	P84548	Q9S892
Q7M378	P80082	Q9S8F0	P83466	P80633	P99501
1XHA_P61925	UniRef90_Q9TWN4	Q9R4U9	Q9TRG1	Q6LBV5	Q9R5A1
1XH7_P61925	P20016	1SSC_P61823	Q7M2G7	Q9R5M6	P86532
1BGS_P11540	P0DJM7	P11149	P83127	P21983	1BSX_1BSX
P85952	P86073	D2IX31	P80342	Q7M045	2GJ2_Q91LD0
P0A0E0	2HIP_P04168	1AVS_P02588	Q9TWS3	P0DJB7	Q9S8D5
C0HJG5	1DEV_O95405	P86070	1GG2_P63212	E9PZ55	UniRef90_P85198
P80008	Q7M182	P80264	Q9R4P1	1CMK_P61925	Q9TXG2
P20005	Q7M2P9	Q9R4R2	Q9R5J3	A8JJG8	P21791
Q7M072	Q9R4Q9	Q9R5G0	P84976	Q7M310	1C8M_Q82122
Q9T2I6	Q9R4R3	1SVG_P61926	Q84LP2	1SQK_O97428	Q7M369
Q9T2S3	P84540	P85931	Q9TWG2	Q7M2P3	P82902
P86328	Q9R545	Q7M187	Q7M489	Q7M313	C6TMS5

Continúa en la siguiente página

Tabla 30 Casos Negativos: Péptidos no antimicrobianos – *Continuación*

1ENH_P02836	P59683	D9J164	Q9T2R9	UniRef90_P81754	Q7M322
UniRef90_B3A052	P49328	1KKQ_Q9Y618	Q9R5K0	1BCK_NOR00035	Q9S8C1
1QC6_1QC6	Q7DLL2	V9GXY2	Q7M0C0	Q7M3Q6	Q7M1D9
2DX8_Q9D0P5	P50983	P31859	1CSK_P41240	Q7M1J1	P82322
P33556	Q7M0E5	P84520	P13064	P84282	B2CS62
UniRef90_G3E7U1	Q7M283	1LM8_Q16665	Q7M0N6	P83115	Q7LZP6
Q9S8F1	P21988	Q7M101	P80250	P59851	P59891
Q9R505	Q9R4B8	Q7LZ45	P31082	Q9UWI5	Q9TWH5
P22660	P80236	P35430	P10625	Q7M061	Q7M0J0
UniRef90_P85555	P82657	UniRef90_P62567	1KJY_O08773	Q9R507	Q9TWR5
2EIJ_P07471	P85371	UniRef90_P84411	P82338	P20140	P60265
Q7M3I0	P80704	P83162	P84981	P85986	B3F8X7
P01016	Q7M3I5	P22582	Q9T2I7	P86043	Q9S9I5
O48611	Q7M1Z6	1MOX_P01135	3B95_P68431	Q9TWK0	P0C8X0
Q9PSQ4	1IRQ_Q57468	P59073	Q7M371	1CWF_NOR00036	Q7M284
1JYC_1JYC	2NO3_2NO3	Q9R501	P86195	Q9R5P3	Q9TRP2
1HQJ_1HQJ	Q9PRV3	UniRef90_B3A0I5	P0DMB6	Q9S8G5	Q9S915
Q7M3Z9	P84890	1H59_P24593	P80635	Q9S8S4	Q10583

Tabla 31: Casos Negativos : Péptidos no antimicrobianos. Conjunto de validación compuesto por 384 péptidos recuperados de la base de datos Uniprot.

Q7M0I8	P20728	P0DKU0	Q9TWP9	P85955	1CDK_P61926
Q9TXF3	P81242	1T3L_P07293	Q0IXH9	1BX2_P02686	Q9S8V0
P81002	P83629	Q9TRH8	1CWH_NOR00033	Q9PS16	P82945
Q9PRQ0	P80786	Q9UR66	P80806	C0HJH8	P86067
P59066	UniRef90_P85110	1H98_P03942	Q0PKT1	1U3S_Q15788	2OQ1_P20963
1RF3_P36941	Q9PRR7	H3BJU3	P85235	P30833	Q9T2Q1
P83149	Q9R2F6	Q9UWJ6	P21596	P55236	P83508
P55967	P84729	Q9R4T0	P85310	P84037	P86066
P0DL21	1MVC_Q15596	P16117	P86117	Q7LZ33	Q7M2N4
P81124	Q7M2A6	UniRef90_G3E7U4	Q9T2J9	Q9PRZ6	1BE3_P13272
P86525	P85949	1XTC_P01555	1TY4_O61667	Q9TRX2	Q7M1D3
Q9R573	P0CAR0	Q9PRM4	Q7M097	P81647	P83725
P55932	P80899	2DZN_P33298	1IVO_P01133	Q7M1P8	Q7DLK7
P80762	Q7M3B5	Q9R4R7	1CWI_NOR00036	1FMO_P61926	P11530
1X7B_1X7B	P14803	Q7M269	P0DJC8	P28524	P23078
Q1I169	Q7M057	D3YUT0	Q9S8N5	C6K8E9	Q7M2P4
Q7M2S7	P80744	Q9TWF0	1F9F_P06790	UniRef90_F8KLY5	P86497
P86351	P28270	1PEF_1PEF	Q7M404	P81495	P83722
P85305	Q9PRV9	A7LB48	Q9S870	P14456	Q7M2M3
Q9R497	Q7M2N5	Q9UWJ2	P84990	1CN3_P12908	P68110
P14461	P80852	Q9S8J1	Q9PS65	A9XNZ7	P80845
Q7M188	P83079	P84982	Q7M1T9	C0HJC8	P85937
1ICF_P04233	1CSE_P01051	Q9T2Q0	P0CH89	1O9Y_O85094	P0C8C2
P83289	Q7M1L1	1NRO_P25116	P86367	Q7LZ11	2EIL_P04038
Q7M1U8	A5Z1X9	P14459	P81440	Q9R4I0	P21792
Q9R4X8	Q9R4B7	Q9TQX5	UniRef90_C2GUK3	Q7M134	P82108
P27205	P13066	P81671	Q9S8J3	Q9TRI1	Q7M3I6
UniRef90_P86295	P85994	P82648	P83628	P81285	P21794
1CWC_NOR00033	P0C8I7	P0CV91	P14460	P85433	1DTD_P81511
Q7M4E0	P84342	P85326	P85938	Q9TWJ7	P83447
Q7M0P3	Q570I3	Q7M3Y2	P0DJQ0	P81786	Q7M263
P85334	Q9S8H8	P80530	1C4U_P28504	Q7M1S2	P80821
Q7LZ49	Q7M040	Q9S906	Q7M0J7	P81086	P82453
P0C8B7	1ABO_P00520	P80828	P86821	Q7M1U3	P62789
1Q2H_O14492	4CC7_Q6XZF7	Q7M3T0	Q9S8H9	P13283	P32441
Q9TRL5	Q9R528	Q9R5J6	P80733	P0C1H0	1NAY_1NAY
1YDI_O43707	2UZU_Q3SX13	2H4M_P09651	UniRef90_P84724	P18997	Q9UWL9
A2NAI0	Q7M3Z4	Q7M0Z5	2CPG_P13920	P80755	P80527
1N2D_P19524	Q9TWI4	Q9TS71	1SRN_P61823	Q7LZ68	P20903
P84533	P68214	Q7M2N3	P84549	Q9R509	P68359

Continúa en la siguiente página

Tabla 31 Casos Negativos: Péptidos no antimicrobianos – *Continuación*

UniRef90_P84977	Q9R5R5	Q9QUY8	2UZV_Q3SX13	Q9R4V1	P01304
1FAP_P42345	P09691	1YCQ_P04637	1E0B_P40381	Q9TRB1	P83054
Q9TRP5	P86450	B6CHX1	Q9R4T2	P80818	Q9S8L6
Q7M3H8	UniRef90_C5E952	P0CAR8	Q7M0B8	P19979	1DD4_P29396
P81411	P83150	Q9R4H4	P33036	Q9R5P7	P20304
Q9UWL8	P33588	2SEM_P29355	E9Q9J0	P0DJQ2	Q7M1U5
P84725	Q7M128	P84472	Q7M1V0	Q9XQE1	Q9TWC0
UniRef90_T2I3E8	P0DMH2	P05542	P05866	Q7M285	P86463
P68117	Q9R596	P85904	Q9S8Y7	Q7M1H8	P83402
P02343	Q7M2L4	P85929	Q7M0Y1	Q7M127	Q7M3H9
Q7M1U1	P14486	Q90ZX1	Q9TRF4	P83973	1FR3_Q7SIF7
Q9BAC4	Q9TRG9	Q7M2T4	1EQ7_P69776	P85510	P86807
2BFI_2BFI	Q7M3Z2	P83761	1B0N_P23308	P02744	Q9T2I9
P84553	1PPT_P68249	Q9R5A6	1ZAF_Q15788	Q56Z41	P09876
Q9SB20	UniRef90_P86511	1T29_Q9BX63	2F7Z_P61926	1E0F_Q25163	Q9R4U6
1FE0_O00244	P84563	1WAP_P19466	B2L571	P82431	P83215
1CL7_P01869	Q9J188	1SSA_P61823	Q41183	Q9TRI3	Q9S8Q8
UniRef90_Q3SAF2	Q3LA80	P85498	P14445	P86062	P85368
1C4V_1C4V	P81340	Q9S8Y0	P86106	F6XVM5	Q7M0I5
Q7M2T2	P84001	P80972	Q09123	Q9UWL0	P82328
1FSE_P11470	7R1R_P69924	P86498	P85922	P85906	Q9PSN9
P83834	Q9TWU4	D3Z0M2	A2A6Q3	Q7M0L9	2HPZ_2HPZ
P83075	Q9R5D6	Q9R4L1	Q9T2Q6	Q6RJU6	1B7I_P19614
1FS1_Q13309	Q7M1P7	P0CH74	P85984	C0HJG7	P0C2K1

Apéndice C. Lista de los descriptores moleculares

C.1. PaDEL-Descriptor

La lista de los descriptores usando el programa PaDEL-Descriptor (Yap, 2011) se describen en el siguiente vínculo:

http://padel.nus.edu.sg/software/padeldescriptor/Descriptors.xls

Los índices que toma cada descriptor de los utilizados en la presente investigación para el conjunto de datos AMP_B se muestra en la Tabla 32.

Tabla 32: Lista de descriptores para el conjunto de datos AMP_B.

1	nAcid	52	ATSc4	103	ETA_Eta_B_RC	205	MAXDN
2	XLogP	53	ECCEN	104	nHBAcc3	206	SsOH
3	BCUTp-1l	54	WPATH	105	nRotB	207	SsssCH
4	globalTopoChargeIndex	55	BCUTc-1l	106	Kier1	208	SdssC
5	CrippenLogP	56	ETA_Epsilon_2	107	VCH-6	209	SHBint8
6	BCUTw-1h	57	ETA_EtaP_L	108	ETA_Eta_F_L	210	MAXDN2
7	MDEN-11	58	MDEC-11	109	WPOL	211	SHBd
8	MLogP	59	topoRadius	110	ETA_Eta_B	212	minsssCH
9	C1SP3	60	topoDiameter	111	Zagreb	213	gmin
10	VP-0	61	BCUTw-1l	112	ETA_Shape_P	214	RotBtFrac
11	VP-1	62	ETA_Epsilon_5	113	ETA_Eta_R	215	nsOH
12	VPC-5	63	nO	114	ETA_Alpha	216	nHsOH
13	SP-0	64	CrippenMR	115	HybRatio	217	nHBint8
14	C3SP3	65	MLFER_A	116	ETA_Eta_L	218	MDEN-23
15	VP-2	66	topoShape	117	MW	219	SHBint9
16	VP-4	67	nAtomLC	118	ETA_Eta	220	DELS2
17	SPC-6	68	ATSp1	119	VCH-7	221	SHBint10
18	VP-3	69	ETA_Shape_Y	120	SCH-7	222	maxHCsats
19	nAtomP	70	PetitjeanNumber	121	ETA_Eta_R_L	223	maxHCsatu
20	MDEN-13	71	MDEC-12	122	ETA_Beta_s	224	SHBint2
21	SP-1	72	ETA_Epsilon_1	123	ATSm2	225	maxHBint10
22	VPC-4	73	ATSp2	124	Kier3	226	maxsOH
23	SP-2	74	ETA_dEpsilon_A	125	ETA_dBeta	227	maxHBint8
24	VPC-6	75	ATSp4	126	nBonds	228	SssCH2
25	SP-4	76	ETA_dEpsilon_C	127	ETA_BetaP_ns	229	nBondsS2
26	SP-3	77	ETA_Epsilon_4	128	ETA_Eta_F	230	nBondsS
27	VP-5	78	ETA_EtaP_B_RC	129	VCH-5	231	SwHBa
28	SPC-5	79	ATSp3	130	nHeavyAtom	232	maxHBint2
29	SPC-4	80	apol	131	VAdjMat	233	nHBd
30	ETA_AlphaP	81	WTPT-5	132	ETA_Beta	234	maxHssNH
31	MDEN-12	82	ATSp5	133	MDEO-11	235	SHssNH
32	VP-7	83	ETA_BetaP_s	134	ETA_Epsilon_3	236	ndssC
33	SP-5	84	McGowan_Volume	135	ETA_dBetaP	237	nHBint5
34	VP-6	85	Kier2	136	ATSm5	238	maxHBint6
35	ETA_Psi_1	86	ETA_EtaP_F_L	137	maxsCH3	239	SHBint3
36	WTPT-4	87	ETA_dEpsilon_D	138	ATSm4	240	SHBint4
37	VC-3	88	VABC	139	WTPT-3	241	ndO
38	SP-7	89	ETA_EtaP	140	ATSm1	242	nHBint10

Continúa en la siguiente página

Tabla 32 Lista de descriptores para el conjunto de datos AMP_B – *Continuación*

39	ETA_dAlpha_B	90	AMR	141	SC-5	243	C2SP2
40	nS	91	ETA_EtaP_F	142	VC-5	244	n5HeteroRing
41	nH	92	MLFER_BH	143	SC-3	245	nT5Ring
42	SP-6	93	MLFER_BO	144	nBase	246	nHeteroRing
43	WTPT-1	94	BCUTc-1h	145	C2SP3	247	nT5HeteroRing
44	MLFER_S	95	ETA_EtaP_B	146	LipinskiFailures	248	n5Ring
45	ETA_dPsi_A	96	BCUTp-1h	147	maxsNH2	249	mindsCH
46	nC	97	SCH-5	148	minsssN	250	minHCsatu
47	nAtom	98	nRotBt	149	nN	251	nHBint2
48	MLFER_E	99	SCH-6	150	ATSc2	252	nHBint6
49	MLFER_L	100	nBondsS3	151	ATSc5	253	MDEC-23
50	bpol	101	ETA_BetaP	152	ETA_dEpsilon_B	203	SHBint6
51	ATSc3	102	ETA_Beta_ns	153	LipoaffinityIndex	204	nHBa
205	minssCH2	206	ndsCH	207	n6Ring	208	SHdsCH
209	ETA_dEpsilon_A						

C.2. JPEDES (*Java Peptide Descriptors*)

Tabla 33: Lista de descriptores para el conjunto de datos AMP_A.

Índice	Descriptor	Descripción
1	MW	Peso molecular del péptido
2	seq_length	Número de residuos en el péptido
3	no_basic	Número de residuos básicos
4	no_aromatic	Número de residuos aromáticos
5	no_hydrophobic	Número de residuos hidrófobos
6	no_W	Número de residuos de triptófano (W)
7	no_basic/length	Número de residuos básicos entre la longitud del péptido
8	%basic_res	Porcentaje de residuos básicos
9	%hydrophobic +basic_res	Suma del porcentaje de residuos hidrófobos y residuos básicos
10	Sum_aromatic +basic	Suma de residuos aromáticos y básicos
11	He	Hidrofobicidad de acuerdo con la escala de Eisenberg *et al.* (1984)
12	Hk	Hidrofobicidad de acuerdo con la escala de Kyte y Doolittle (1982)
13	Z(pH5)	Carga neta a un pH de 5

Continúa en la siguiente página

Tabla 33 Lista de descriptores para el conjunto de datos AMP_A – *Continuación*

14	Z(pH7)	Carga neta a un pH de 7
15	Z(pH9)	Carga neta a un pH de 9
16	IP	Punto isoeléctrico
17	max_dist_W	Máxima distancia entre residuos de triptófano en la secuencia
18	max_dist_basic	Máxima distancia entre residuos básicos en la secuencia
19	max_dist_aromatic	Máxima distancia entre residuos aromáticos en la secuencia
20	max_dist_hydrophobic	Máxima distancia entre residuos hidrófobos
21	Hm(100)	Momento hidrofóbico a un angulo de 100° utilizando la escala de (Eisenberg *et al.*, 1984)
22	Hm(160)	Momento hidrofóbico a un angulo de 160° utilizando la escala de (Eisenberg *et al.*, 1984)
23	Hm(180)	Momento hidrofóbico a un angulo de 180° utilizando la escala de (Eisenberg *et al.*, 1984)
24	Hm_Hk(100)	Momento hidrofóbico a un angulo de 100° utilizando la escala de (Kyte y Doolittle, 1982)
25	Hm_Hk(160)	Momento hidrofóbico a un angulo de 160° utilizando la escala de (Kyte y Doolittle, 1982)
26	Hm_Hk(180)	Momento hidrofóbico a un angulo de 180° utilizando la escala de (Kyte y Doolittle, 1982)
27	avg_dist_b_basicRes	Distancia promedio de los residuos básicos
28	avg_dist_b _hydrophobicRes	Distancia promedio de los residuos hidrófobos

Tabla 34: Lista de descriptores para el conjunto de datos AMP_A+B.

1	Z(pH5)	71	WPATH	140	VCH-7	210	hmax
2	Z(pH7)	72	Max_dist_hydro	141	ETA_Eta_R_L	211	maxHsOH
3	Z(pH9)	73	BCUTc-1l	142	ETA_Beta_s	212	SHsOH
4	IP	74	Max_dist_basic	143	ATSm2	213	maxdsCH
5	'%hydroBasRes'	75	ETA_Epsilon_2	144	Kier3	214	maxwHBa
6	nAcid	76	ETA_EtaP_L	145	ETA_dBeta	215	SHsNH2
7	XLogP	77	MDEC-11	146	nBonds	216	SHBa
8	Hk	78	topoRadius	147	ETA_BetaP_ns	217	SHBint7
9	BCUTp-1l	79	topoDiameter	148	ETA_Eta_F	218	DELS
10	globalTopoChargeIndex	80	ETA_Epsilon_5	149	VCH-5	219	nF9HeteroRing
11	'%basRes'	81	AVG_dist_b_hydrophobic	150	nHeavyAtom	220	nT9Ring
12	no_basic/length	82	nO	151	VAdjMat	221	nF9Ring
13	CrippenLogP	83	CrippenMR	152	ETA_Beta	222	nT9HeteroRing
14	He	84	MLFER_A	153	MDEO-11	223	nFRing
15	BCUTw-1h	85	topoShape	154	ETA_Epsilon_3	224	maxHsNH2

Continúa en la siguiente página

Tabla 34 Lista de descriptores para el conjunto de datos AMP_A+B – *Continuación*

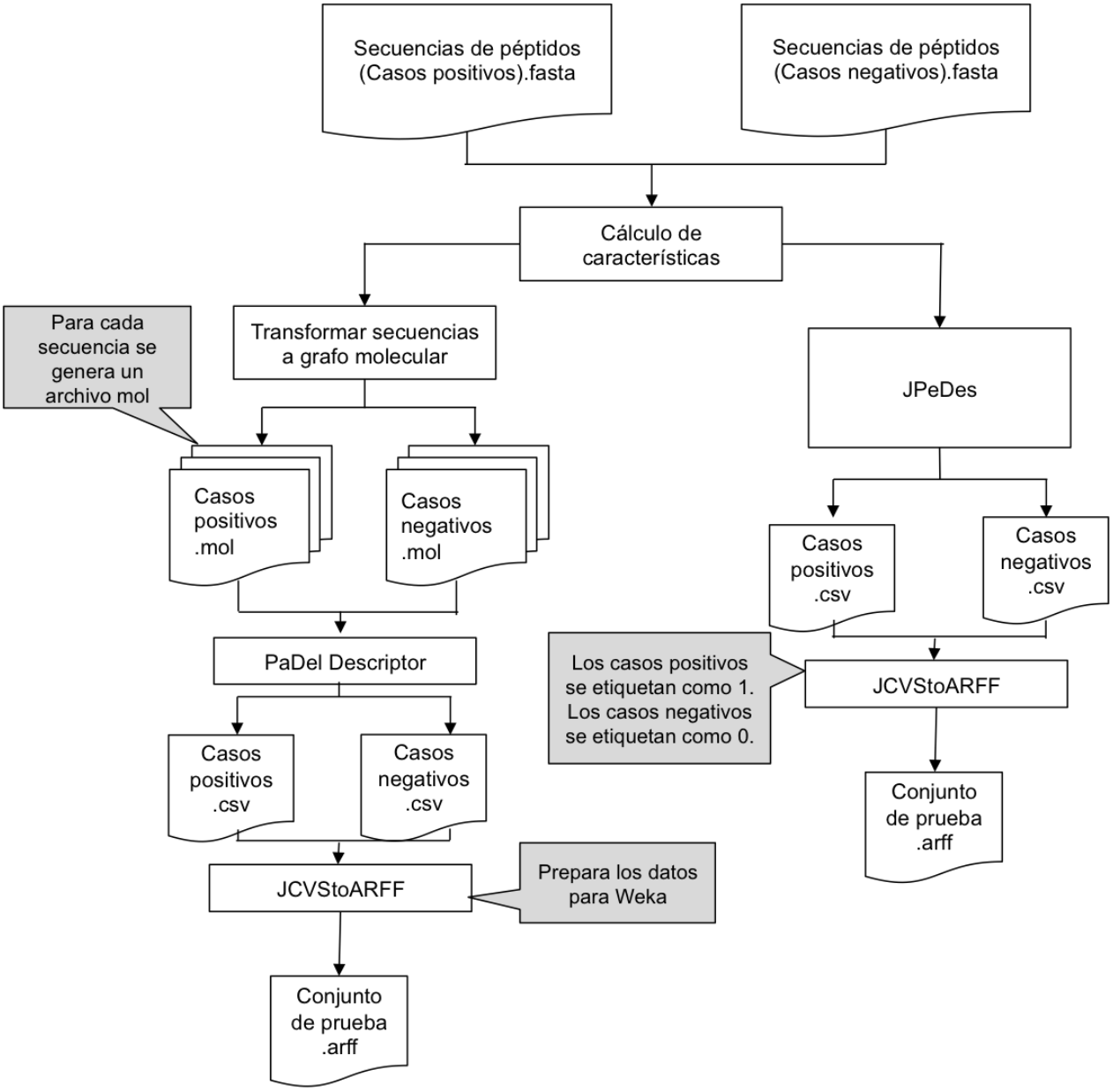
16	MDEN-11	86	nAtomLC	155	ETA_dBetaP	225	SHBint6
17	Hm_Hk(100)	87	ATSp1	156	ATSm5	226	nHBa
18	MLogP	88	ETA_Shape_Y	157	maxsCH3	227	MAXDN
19	no_basic	89	PetitjeanNumber	158	ATSm4	228	no_W
20	'sum Aromatic+Basic Res'	90	MDEC-12	159	WTPT-3	229	SsOH
21	no_hydrophobic	91	ETA_Epsilon_1	160	ATSm1	230	SsssCH
22	C1SP3	92	ATSp2	161	SC-5	231	SdssC
23	Hm(100)	93	ETA_dEpsilon_A	162	VC-5	232	SHBint8
24	SP-0	94	ATSp4	163	SC-3	233	MAXDN2
25	VP-0	95	ETA_dEpsilon_C	164	Hm_Hk(160)	234	SHBd
26	VP-1	96	ETA_Epsilon_4	165	nBase	235	gmin
27	VPC-5	97	ETA_EtaP_B_RC	166	C2SP3	236	minsssCH
28	C3SP3	98	ATSp3	167	LipinskiFailures	237	RotBtFrac
29	seq_length	99	apol	168	maxsNH2	238	nHsOH
30	VP-2	100	WTPT-5	169	minsssN	239	nsOH
31	VP-4	101	ATSp5	170	nN	240	nHBint8
32	SPC-6	102	ETA_BetaP_s	171	ATSc2	241	MDEN-23
33	VP-3	103	McGowan_Volume	172	ATSc5	242	SHBint9
34	nAtomP	104	Kier2	173	ETA_dEpsilon_B	243	DELS2
35	MDEN-13	105	ETA_EtaP_F_L	174	LipoaffinityIndex	244	SHBint10
36	SP-1	106	ETA_dEpsilon_D	175	ATSc1	245	maxHCsatu
37	VPC-4	107	VABC	176	maxHBint7	246	maxHCsats
38	SP-2	108	ETA_EtaP	177	ATSm3	247	SHBint2
39	VPC-6	109	AMR	178	nHBAcc_Lipinski	248	maxHBint10
40	SP-4	110	ETA_EtaP_F	179	nHBAcc	249	maxsOH
41	SP-3	111	MLFER_BH	180	nHBAcc2	250	maxHBint8
42	VP-5	112	MLFER_BO	181	nHBDon	251	SssCH2
43	SPC-5	113	BCUTc-1h	182	nHBDon_Lipinski	252	nBondsS
44	SPC-4	114	ETA_EtaP_B	183	maxHBint3	253	nBondsS2
45	ETA_AlphaP	115	BCUTp-1h	184	WTPT-2	254	SwHBa
46	MDEN-12	116	SCH-5	185	minHssNH	255	maxHBint2
47	VP-7	117	nRotBt	186	C1SP2	256	Max_dist_aromatic
48	AVG_dist_b_basicRes	118	SCH-6	187	nBondsM	257	nHBd
49	SP-5	119	nBondsS3	188	nBondsD	258	maxHssNH
50	VP-6	120	ETA_BetaP	189	nBondsD2	259	SHssNH
51	ETA_Psi_1	121	ETA_Beta_ns	190	minHsNH2	260	ndssC
52	WTPT-4	122	ETA_Eta_B_RC	191	MDEC-13	261	nHBint5
53	VC-3	123	nHBAcc3	192	MDEC-22	262	maxHBint6
54	SP-7	124	nRotB	193	maxHBint4	263	SHBint3
55	ETA_dAlpha_B	125	Kier1	194	minsNH2	264	SHBint4
56	nS	126	VCH-6	195	TopoPSA	265	ndO
57	nH	127	ETA_Eta_F_L	196	minHBd	266	nHBint10
58	SP-6	128	WPOL	197	minHCsats	267	C2SP2
59	BCUTw-1l	129	ETA_Eta_B	198	RotBFrac	268	n5HeteroRing
60	WTPT-1	130	Zagreb	199	nHBint7	269	nT5HeteroRing
61	MLFER_S	131	ETA_Shape_P	200	maxsssN	270	nHeteroRing
62	ETA_dPsi_A	132	ETA_Alpha	201	Hm(160)	271	n5Ring
63	nC	133	ETA_Eta_R	202	minwHBa	272	nT5Ring
64	nAtom	134	HybRatio	203	mindssC	273	mindsCH
65	MLFER_E	135	ETA_Eta_L	204	maxHBd	274	minHCsatu
66	MLFER_L	136	MW	205	ALogp2	275	nHBint2
67	bpol	137	MW2	206	ALogP	276	nHBint6
68	ATSc3	138	ETA_Eta	207	nBonds2	277	no_aromatic
69	ATSc4	139	SCH-7	208	maxdssC	278	MDEC-23
70	ECCEN			209	nwHBa	279	nHdsCH
280	SsCH3	281	nssCH2	282	nssS		

Apéndice D. Implementación de los Algoritmos

D.1. Configuración del algoritmo genético

Tabla 35: Parámetros de configuración para el algoritmo de selección de características.

Conjunto de datos	Número generaciones	Número de individuos	Número de padres	p_c	p_m	Aptitud promedio de los mejores individuos	Aptitud promedio
AMP_A	500	200	200	0.7	0.3	92.48	85.43
	500	150	150	0.7	0.3	92.89	88.27
	500	100	100	0.7	0.5	92.12	87.51
	500	100	100	0.7	0.3	92.23	87.08
	500	100	100	0.7	0.1	92.98	88.08
	500	100	100	0.8	0.5	93.14	87.35
	500	100	100	0.8	0.5	92.64	88.99
	500	100	100	0.8	0.3	92.10	86.78
	500	100	100	0.8	0.1	91.55	87.43
	500	50	50	0.8	0.5	92.49	88.90
AMP_B	600	500	500	0.8	0.5	93.02	86.59
	600	200	200	0.8	0.8	93.02	87.10
	600	300	300	0.8	0.8	93.47	86.91
	600	200	200	0.8	0.8	93.15	87.19
AMP_A+B	550	350	350	0.7	0.3	95.90	89.61
	500	300	300	0.7	0.3	95.51	89.68
	500	200	200	0.8	0.5	94.95	89.35
	550	300	300	0.8	0.5	95.99	89.41


Secuencias de péptidos (Casos positivos).fasta
Secuencias de péptidos (Casos negativos).fasta
Cálculo de características
Transformar secuencias a grafo molecular
Para cada secuencia se genera un archivo mol
JPeDes
Casos positivos .mol
Casos negativos .mol
Casos positivos .csv
Casos negativos .csv
PaDel Descriptor
Los casos positivos se etiquetan como 1. Los casos negativos se etiquetan como 0.
JCVStoARFF
Casos positivos .csv
Casos negativos .csv
Conjunto de prueba .arff
JCVStoARFF
Prepara los datos para Weka
Conjunto de prueba .arff

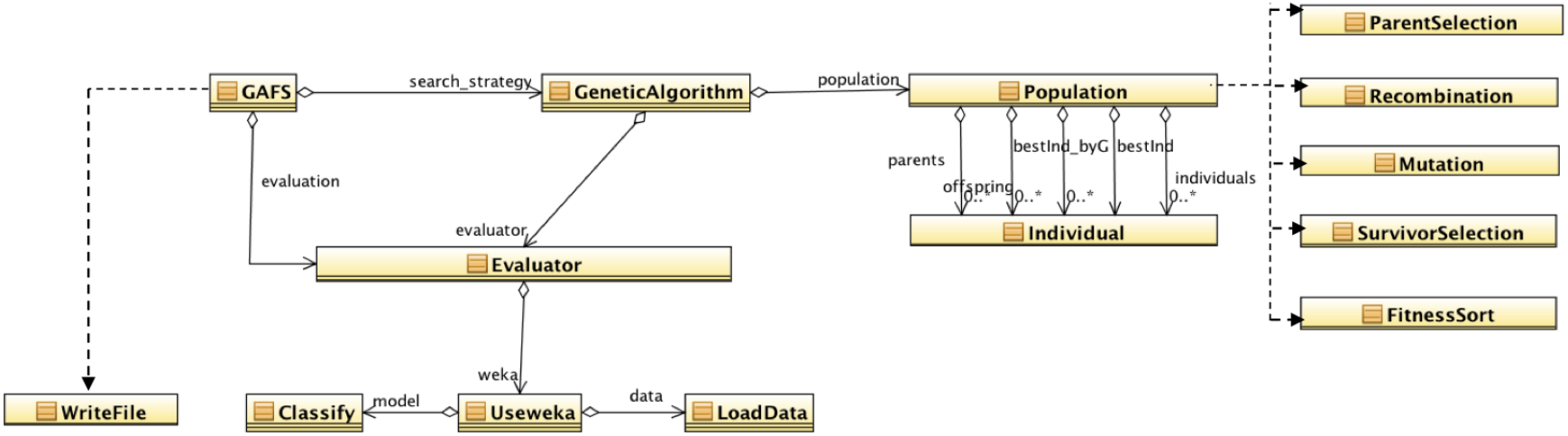


Figura 52: Diagrama de objetos del algoritmo genético para la selección de características.

Apéndice E. Ganancia de información

En la Sección 5.4.1 se presentaron los resultados de ganancia de información tras establecer un umbral θ. A continuación se describe cómo calcular la ganacia de información y el procedimiento llevado acabo para realizar el experimento de la Sección 5.3.

Procedimiento

Se consideran cada uno de los conjuntos de datos a utilizar (ver Sección 4.1) como un conjunto $\mathcal{D} = \cup_{i=1}^{p}\{(\mathbf{x_i}, y_i)|\mathbf{x_i} \in \mathbb{R}^n,\ y_i \in \{0,1\}\}$. Los descriptores moleculares se jerarquizan (*Ranking*) de acuerdo a la función $S(X_k)$, donde valores altos implican que el descriptor es importante. La función $S(X_k)$ está dada por la ganancia de información $I(Y, X_k)$ del descriptor molecular X_k con respecto a la clase Y. A continuación mostramos la fórmula para calcular la ganancia de información.

$$S(X_k) = I(Y, X_k) \tag{19}$$

$$I(Y, X_k) = H(Y) - H(Y|X_k) \tag{20}$$

$$H(Y) = -\sum_{j=1}^{|Y|} p(Y = y_j) \log_2 p(Y = y_j) \tag{21}$$

$$H(Y|X_k) = \sum_{i=1}^{|X_k|} p(X_k = x_i) H(Y|X_k = xi) \tag{22}$$

$$H(Y|X_k = xi) = -\sum_{j=1}^{|Y|} p(Y = y_j|X_k = x_i) \log_2 p(Y = y_j|X_k = x_i) \tag{23}$$

Donde $p(Y = y_i)$ es la probabilidad de que la variable Y tome el valor de y_i, $p(X_k = x_i)$ es la probabilidad de que la característica X_k tome el valor de x_i, $p(Y = y_j|X_k = x_i)$ es la probabilidad de que Y tome el valor de y_j dado que X tomá el valor de x_i.

Pasos a seguir en el experimento

Para cada conjunto de datos de entrada se hace lo siguiente:

- **Paso 1:** Calcular ganancia de información con el programa para cada uno de los descriptores moleculares.
- **Paso 2:** Inicializar $\theta = 0$.
- **Paso 3:** Si $\theta < 0.9$ ir al paso 4, en otro caso ir al paso 8.
- **Paso 4:** Seleccionar sólo los descriptores moleculares con ganancia de información mayores a θ.
- **Paso 5:** Construir un clasificador con la configuración de la Sección 5.2.2 y la base de datos con los atributos seleccionados en el paso 3.
- **Paso 6:** Obtener la exactitud (ACC) del clasificador.
- **Paso 7:** Actualizar $\theta = \theta + 0.001$ y volver al paso 3.
- **Paso8**: Terminar procedimiento.