# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

## Maestría y Doctorado en Ciencias Bioquímicas

**Desarrollo de software para predicción y caracterización *in silico* de péptidos antimicrobianos en diferentes organismos**

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias


PRESENTA:
Jose Santiago Sanchez Fragoso

DIRECTOR DE TESIS
Dr. Fidel Alejandro Sánchez Flores
Instituto de Biotecnología

MIEMBROS DEL COMITÉ TUTOR

Dra. Blanca Itzel Taboada Ramírez
Instituto de Biotecnología

Dr. Iván Arenas Sosa
Instituto de Biotecnología



Cuernavaca, Morelos. Diciembre, 2020

El presente trabajo se realizo en la Unidad de Secuenciación Masiva y Bioinformática del Instituto de Biotecnología de la Universidad Nacional Autónoma de México, bajo el programa de Maestría y Doctorado en Ciencias Bioquímicas.

Este trabajo se realizo gracias a la beca de CONACyT para estudios de posgrado nivel maestría, con numero de becario 686410.

## Agradecimientos

A Alejandro Sánchez por brindarme sus conocimientos y guía, pero sobre todo por proporcionarme un espacio ideal para mi desarrollo intelectual.

A los miembros de mi comité tutor, Blanca Taboada y Iván Arenas, por sus excelentes observaciones y recomendaciones, que ayudaron encaminar este proyecto.

A todos los habitantes de la Unidad de Bioinformática por compartir su experticia técnica, así como por crear una atmosfera de confianza y espíritu científico.

A mi familia por proporcionarme y reforzar los valores necesarios para concluir este paso profesional.

## Tabla de Contenido

## Resumen

La utilización indiscriminada y masiva de los antibióticos ha generado una gran presión selectiva en organismos patógenos, favoreciendo la supervivencia de cepas con resistencia a estos compuestos. Ante la inminente crisis que esto representa, existe interés en los péptidos antimicrobianos (AMPs, por sus siglas del inglés "AntiMicrobial Peptides"), que pueden ser una solución al problema de la resistencia. El estudio de los AMPs ayudará a entender sus mecanismos de acción, así como asistir al diseño de AMPs sintéticos. Actualmente existen algunas herramientas de predicción de AMPs pero su poder predictivo disminuye drásticamente al utilizar secuencias que no contemplan el procesamiento post-tradicional de los péptidos.

Utilizando algoritmos de aprendizaje de maquina ("machine learning"), en este proyecto de investigación se ha desarrollado un método para identificar secuencias peptídicas con posible función antimicrobiana en artrópodos y plantas. El método se compone de dos pasos: 1) Se hace una predicción de la secuencia de AMP funcional (péptido maduro), y 2) utilizando 11 características fisicoquímicas e información de estructura primaria se identifica función antimicrobiana en secuencias de aminoácidos.

El método está implementado como un programa llamado proAmps, el cual se encuentra disponible en la página https://github.com/SantiagoSanchezF/proAmps. El poder predictivo de proAmps es evaluado y comparado con las herramientas CAMP, ampscanner y iAMP-2L. Mediante esta evaluación, se determinó que proAmps es el método con mayor sensibilidad y especificidad cuando son utilizadas secuencias de precursores no funcionales. Utilizando proAmps se identificaron 146 AMPs putativos de *Drosophila melanogaster* y 182 de *Phaseolus vulgaris.*

Tras un análisis evolutivo de las secuencias utilizadas en nuestro conjunto de datos, proponemos una clasificación de 4 grupos en los que se pueden subdividir los AMPs de plantas y artrópodos. El análisis sugiere tanto convergencia como divergencia funcional de los AMPs de plantas y artrópodos.

## Abstract

The indiscriminate and massive use of antibiotics has generated great selective pressure on pathogenic organisms, favouring the survival of strains with resistance to these compounds. Given the imminent crisis that this represents, there is interest in antimicrobial peptides (AMPs), which can be a solution to the problem of resistance. The study of AMPs will help to understand their mechanisms of action, as well as assist in the design of synthetic AMPs. Currently, AMP prediction tools are available, but their predictive power decreases dramatically when using sequences that do not contemplate post-traditional processing of peptides.

Using machine learning algorithms, this research project developed a method to identify peptide sequences with possible antimicrobial function in arthropods and plants. The method consists of two steps: 1) A prediction of the functional AMP (mature peptide) sequence, and 2) using 11 physicochemical characteristics and primary structure information, antimicrobial function is identified in amino acid sequences.

The method is implemented as a software called proAmps, which is available at https://github.com/SantiagoSanchezF/proAmps. The predictive power of proAmps is benchmarked with the CAMP, ampscanner and iAMP-2L tools. Through this evaluation, it was determined that proAmps is the method with the highest sensitivity and specificity when non-functional precursor sequences are used. Using proAmps, 146 putative AMPs from Drosophila melanogaster and 182 from Phaseolus vulgaris were identified.

After an evolutionary analysis of the sequences used in our data set, we propose a classification of 4 clusters into which the AMPs of plants and arthropods can be subdivided. The analysis suggests both functional convergence and divergence of plant and arthropod AMPs.

# Introducción

El descubrimiento de la penicilina en 1928 y los esfuerzos posteriores para identificar y caracterizar nuevos antibióticos, incrementó drásticamente la esperanza de vida durante el siglo XX. Sin embargo, desde la administración de la penicilina para uso clínico, se observó la existencia de bacterias resistentes a dicho compuesto. Desde entonces, la utilización indiscriminada y masiva de los antibióticos para uso humano, así como en animales y agricultura, ha generado una gran presión selectiva favoreciendo la supervivencia de cepas resistentes, cuya selección y prevalencia ha provocado que algunas enfermedades consideradas como controladas hasta hace poco (*e.g*. neumonía y pulmonía), resulten en casos de infecciones mortales durante la última década. Ante esta crisis de la resistencia de los antibióticos, se ha generado interés en alternativas para el tratamiento de enfermedades infecciosas. Una alternativa (en algunos casos) a los antibióticos, son los péptidos con actividad antimicrobiana (Bobone, 2014).

Los péptidos antimicrobianos (AMPs, por sus siglas del inglés "AntiMicrobial Peptides") son polipéptidos producidos como sistema de defensa o competencia en una gran variedad de organismos, desde bacterias hasta vertebrados (Nawrot *et al.,* 2014) y se ha reportado un amplio espectro de actividad biocida contra bacterias y hongos. Los AMPs fueron aislados por primera vez a principios de los años 80's, y actualmente se han descrito más de dos mil quinientos AMPs, producidos por organismos de todos los reinos (Wang *et al.*, 2016). Las características comunes, pero no únicas, que comparten gran parte de los AMPs hasta ahora caracterizados son la carga positiva de sus aminoácidos, un tamaño entre 10-40 aminoácidos, una alta proporción de residuos alifáticos y una estructura secundaria de alfa hélice al asociarse con membranas lipídicas

Los mecanismos de acción de los AMPs no se han esclarecido en lo general, identificando familias de péptidos con acción en diferentes blancos. Sin embargo, existe evidencia de que una gran cantidad de AMPs actúan como moléculas multifuncionales, es decir su capacidad de combatir una infección no es exclusiva de un solo mecanismo y pueden funcionar a nivel sistémico. Algunos de los mecanismos descritos son: formadores de poros, inductores de curvatura gaussiana negativa (lo cual facilita la fagocitosis) en la membrana celular, inhibidores de enzimas, captores de precursores de lípidos de membrana, estimulantes del sistema inmune, entre otros (Brogden, 2005; Lai y Gallo, 2009; Schneider *et al.,* 2010; Brown, 2017).

Por otro lado, el desarrollo de la secuenciación masiva de ADN ha producido un incremento dramático en la cantidad y complejidad de datos de sistemas biológicos. Estos datos son una fuente rica para el descubrimiento de nuevos AMPs de la naturaleza, lo cual resulta importante para entender mejor sus mecanismos de acción, así como desarrollar su uso como alternativa a los antibióticos. Sin embargo, muchos AMPs poseen características que los hacen difíciles de identificar mediante técnicas convencionales, por ejemplo, gran variabilidad de estructura primaria, derivarse de precursores no funcionales, y multifuncionalidad. Para aprovechar mejor la información derivada de la secuenciación masiva se requiere un análisis bioinformático de alto nivel. Una tendencia en los últimos años es realizar este tipo de análisis utilizando técnicas de aprendizaje automático asistido por computadora conocido en inglés como "machine learning" (ML) (Camacho *et al.*, 2018). ML es una disciplina de las ciencias de la computación, la cual nos permite desarrollar modelos predictivos. A partir de un conjunto de reglas matemáticas, los

programas de ML reconocen patrones en un conjunto de datos, y estos patrones pueden ser utilizados para hacer predicciones sobre datos no caracterizados.

Por lo tanto, este proyecto implementa un proceso de identificación de AMPs en dos pasos, utilizando técnicas de ML. En el primer paso se realiza una identificación de los sitios de corte proteolíticos en proteínas completas sintetizadas a partir del gen completo para así predecir secuencias que son producto de procesamiento enzimático (posibles proteínas funcionales como AMPs), las cuales en este trabajo consideramos como proteínas maduras. En el segundo paso se realiza la identificación de AMPs con la secuencia madura predicha, utilizando dos tipos de datos: información de la secuencia lineal de aminoácidos e información de 11 propiedades fisicoquímicas de los péptidos. Este método de identificación, que se fundamenta en estos dos parámetros relevantes en la función de un AMP, lo definimos como método mixto. El objetivo de nuestro trabajo es mejorar la sensibilidad y especificidad en la predicción de función de AMPs utilizando ambos enfoques y compararlos contra métodos ya existentes para la predicción o anotación de AMPs.

## Antecedentes

A la fecha se han empleado diversas estrategias computacionales para identificar AMPs a partir de secuencias de aminoácidos. Entre ellos destacan los métodos de clasificación por ML y de homología de secuencias.

ML se le denomina al estudio, desarrollo y aplicación de algoritmos cuyo objetivo es realizar modelos predictivos de una o mas variables, a partir de un conjunto de datos llamados de entrenamiento. Los algoritmos de ML se pueden clasificar como supervisados a aquellos que se entrenan con información explicita de la variable a predecir, y los no supervisados en los que el objetivo es inferir "clusters" implícitos en los datos de entrenamiento. A su vez los algoritmos de ML supervisados se clasifican como métodos de regresión en los cuales la o las variables a predecir se encuentran en un espacio continuo, o métodos de clasificación en los que la o las variables a predecir tiene categorías discretas. Existen diversos algoritmos de clasificación por ML, de los cuales para la identificación de AMPs han destacado:

- Bosque aleatorio (RF, por sus siglas en ingles "Random Forest"). Método que emplea un ensamble de arboles de decisión e introduce parámetros estocásticos en la selección de las variables utilizadas por los arboles. Este método reduce el sobreajuste del modelo a los datos de entrenamiento, típico en arboles de decisión individuales.
- Maquinas de vectores de soporte (SVC, por sus siglas en ingles, "Support Vector machines Classifier"). El proceso de aprendizaje consiste en modelar en un hiperplano un margen que maximice la separación entre clases en un hiperespacio de variables.
- Redes neuronales artificiales (ANN, por sus siglas en ingles "Artificial Neural Networks"). Sistemas que se componen de capas de neuronas artificiales, las cuales reciben una señal, la procesan y de acurdo al resultado del procesamiento, la señal puede ser propagada a otras neuronas artificiales conectadas a esta. El proceso de aprendizaje consiste en ajustar el peso de cada neurona, el cual intensifica o reduce la intensidad de la señal. Se denominan así por sus similitudes a las redes neuronales biológicas.
- Clasificador Bayesiano Ingenuo ("Naive Bayes", por su nombre en ingles). Algoritmo basado en el teorema de Bayes, el cual calcula la probabilidad de que una observación pertenezca a cierta clase dada cierta evidencia (valor de variables). La probabilidad *a posteriori* se actualiza de forma "ingenua" ya que asume que la probabilidad de las distintas combinaciones de valores de variables es la misma. El método se entrena al calcular la probabilidad de cada clase de tener cada valor.
- Análisis de discriminante (DA, por sus siglas en ingles "Discriminat Analysis"). Método el cual busca encontrar una combinación lineal de variables que mejor caracterice las clases en el conjunto de entrenamiento. (Kuhn y Johnson, 2013).

Los métodos de asignación de función por homología de secuencia son actualmente los más confiables e utilizados. La lógica detrás de esta asociación es que secuencias similares producen estructuras secundarias y terciarias con características fisicoquímicas similares, las cuales les confieren función a las proteínas. Sin embargo, la gran cantidad de especies nuevas y sus genomas secuenciados evidencian que las bases de referencia utilizadas actualmente no representan adecuadamente todo el universo de proteínas. Esto genera una limitante importante al querer

asignar función a péptidos que no sean similares a otros ya descritos. Algunos métodos destacables de este enfoque son:

- Alineamiento múltiple de secuencias de proteínas. Técnica para inferir homología, dominios conservados y mutaciones entre secuencias alineadas.
- BLAST (por sus siglas en ingles "Basic Local Alignment Search Tool"). Es un algoritmo y programa para comparar pares de secuencias primarias y sus sub-secuencias (Altschul *et al.,* 1990).
- Modelos ocultos de Markov (HMM, por sus siglas en ingles "Hidden Markov Models"). Modelos probabilísticos de secuencia, los cuales asignan una probabilidad a cada posición de la secuencia de tener un aminoácido, una deleción o una inserción. A partir de ellos se obtiene una probabilidad de que una secuencia tenga relación con el conjunto de secuencias a partir del cual se construyo el modelo.

Otra forma de clasificar los métodos de identificación de AMPs es por el tipo de información que utilizan como entrada:

- Métodos que utilizan "cores" de AMPs. Estos métodos se basan en la observación de que existe información contenida en subcadenas ("cores") de los péptidos maduros, que otorga actividad antimicrobiana. Algunos ejemplos son el "gamma-core*"* propuesto por Yount y Yeaman (2004) y el método de Chang *et al.* (2015) alimentados por información de *"*cores*"* extraídos de AMPs. Ambos trabajos proponen regiones criticas para la función e identificacion de AMPs.

- Métodos que utilizan AMPs maduros. A la fecha, la información de secuencia de los AMPs maduros ha sido la más utilizada para desarrollar métodos de predicción de función antimicrobiana. Estos métodos incluyen la utilización de firmas consenso (Thomas *et al.,* 2010; Sigrist *et al.*, 2012), estrategias basadas en HMM y análisis de homología con BLAST (Altschul *et al.,* 1990; Schutte *et al.,* 2002). Otros métodos realizan análisis del espacio químico de los aminoácidos, ya sea utilizando solo esta información (Wang, *et al.,* 2016), o en combinación con análisis de secuencia, discriminando no AMP de AMPs por DA, ANN, RF y SVC (Lata *et al.,* 2010; Xiao *et al.,* 2013). Otros métodos utilizan ANN y un alfabeto de aminoácidos reducido para la identificación de AMPs con información de secuencia, un ejemplo es la herramienta *ampscaner* de Veltri *et al.* (2018).

- Métodos que utilizan propéptidos conservados. Se ha observado que muchos péptidos antimicrobianos de las mismas familias a pesar de tener gran variabilidad en su secuencia madura, su propéptido es altamente conservado. Sin embargo, utilizar esta característica para su predicción puede generar una gran cantidad de falsos positivos, ya que péptidos con propéptido similar pueden presentar funciones no antimicrobianas (Wang, 2017); un ejemplo es el trabajo de Yang *et al.* (2012), en el cual describieron 662 AMPs nuevos de la piel de *Odorrana schmackeri.*

- Métodos basados en enzimas procesadoras o transportadores. Otro enfoque es hacer identificación de genes biosintéticos de AMPs mediante el escrutinio de secuencias homólogas a enzimas procesadoras o transportadores de AMPs conocidos. En teoría, este

enfoque permite hacer descubrimiento de nuevos tipos de AMPs ya que esa búsqueda no está restringida a la homología con los AMPs de entrenamiento. Algunos trabajos que utilizan este enfoque son: O'Sullivan *et al.* (2011), que identifica exitosamente 9 operones de genes biosintéticos de lantibioticos; y Morton *et al*. (2015), en el cual se presenta el desarrollo de la herramienta BOA, la cual identifica operones y bloques de genes de bacteriocinas a partir de HMMs.

- Métodos basados en contexto genómico. Existe otro tipo de predicciones las cuales se basan en la idea de que muchos AMPs se encuentran en "clusters" de genes biosintéticos, los cuales contienen un gen estructural acompañado de otros genes de modificación postraduccional, transporte, y regulación. Una herramienta desarrollada de propósitos más generales pero con capacidad de identificar "clusters" de genes antimicrobianos es antiSMASH (Weber *et al.,* 2015).

Algunos de estos métodos se encuentran disponibles en la red como herramientas de predicción como AMP.Biosino (Wang *et al.,* 2011), APD3 (Wang *et al.,* 2016), CAMP (Thomas *et al.,* 2010), AMPer (Fjell *el al.,* 2007), AMPscanner (Veltri *et al.,* 2018) y AntiBP2 (Lata *et al.,* 2010).

Además de las estrategias de predicción de AMPs descritas arriba, existen diversas bases de datos de AMPs, dentro de las cuales destacan APD3, PhytAMP (Hammami *et al.,* 2009), BACTIBASE (Hammami *et al.,* 2007), CAMP (Thomas *et al.,* 2010), DBAASP (Pirtskhalava *et al*., 2015) y DRAMPS (Fan *et al.,* 2016).

## Justificación

El descubrimiento de nuevos AMPs ayudará a esclarecer sus mecanismos de acción, así como asistir al diseño de nuevos AMPs sintéticos. Actualmente existen herramientas de predicción disponibles en la red que tienen la ventaja de estar ligadas directamente a una base de datos que se actualiza constantemente (*e.g.*, APD). Además, son implementaciones que cuentan con varias versiones que se han perfeccionado en cada nuevo lanzamiento. Sin embargo, existen algunas desventajas asociadas a estas herramientas: algunas resultan poco versátiles ya que a pesar de estar descrito su algoritmo, su código fuente no es abierto, por lo que no es posible examinar a detalle el proceso de clasificación y tampoco es posible crear un nuevo modelo predictivo a partir de conjuntos de secuencias dadas por el usuario; otra desventaja es que ninguna de estas herramientas de predicción (salvo antiSMASH) se encuentran disponibles para instalación local. Muchas de estas herramientas no permiten el análisis de predicción de múltiples secuencias simultáneamente; otro problema asociado a los métodos disponibles es que tienen mal desempeño al utilizar información de propéptidos, lo cual limita las predicciones a partir de transcriptomas o de genes reportados en genomas. El objetivo de nuestro trabajo es mejorar la sensibilidad y especificidad en la anotación de AMPs utilizando un enfoque mixto y compararlo contra métodos ya existentes para la predicción o anotación de AMPs.

## Hipótesis

El empleo de un método mixto para la identificación de AMPs basado en secuencia y en 11 propiedades fisicoquímicas, mejorará el desempeño general de un modelo predictivo respecto las herramientas CAMP, ampscanner y iAMP-2L al utilizar información de secuenciación masiva.

## Objetivos

General:
Desarrollar un método de identificación de péptidos antimicrobianos basado en un método mixto, capaz de identificarlos en datos derivados de secuenciación masiva, con un enfoque de secuencias y características fisicoquímicas.

Particulares:

- Desarrollar un algoritmo de identificación de AMPs utilizando métodos mixtos (secuencia y propiedades fisicoquímicas).
- Implementar el algoritmo en un "software".
- Realizar un comparativo con los trabajos de identificación de AMPs CAMP, ampscanner y iAMP-2L.
- Identificar *in silico* potenciales AMPs en dos especies de distintos reinos.
- Realizar un análisis evolutivo de los AMPs, detectando posible divergencia o convergencia funcional.

# Metodología

## Desarrollo de un algoritmo utilizando métodos mixtos de identificación de AMPs

Para realizar el objetivo particular uno, el cual incorpora características de secuencia y propiedades fisicoquímicas, se determinó la necesidad de desarrollar un algoritmo de identificación de dos pasos: predicción del péptido maduro y predicción de AMP sobre la secuencia madura predicha. A continuación, se describe el método empleado para la realización de este objetivo.

1)_Conjuntos de datos seleccionados:

- Maduros Positivo: Como conjunto positivo de secuencias maduras se utilizaron secuencias de AMPs de origen *Viridiplantae* o *Arthropoda,* las cuales sólo contuvieran aminoácidos estándares y una longitud en aminoácidos mayor a 6 y menor de 102. Las secuencias se obtuvieron de la base datos de péptidos antimicrobianos APD3 (http://aps.unmc.edu/AP/main.php) la cual cuenta con 838 secuencias de estas características. Ciento setenta y cinco secuencias fueron elegidas aleatoriamente para formar un conjunto de prueba de AMPs maduros.
- Secuencias negativas: Aleatoriamente se eligieron 6,500 secuencias de Uniprot (UniProt Consortium, 2015) que tuvieran anotación curada manualmente, una longitud menor a 102 aminoácidos, y que no tuvieran anotación de "antimicrobial", "toxin" o "toxic". Con el fin de eliminar redundancia en el conjunto, se generaron "clusters" de secuencias con al menos 50% de similitud utilizando el programa CD-HIT (Fu *et al.*, 2012). El valor de 50% de similitud se determinó empíricamente, ya que a este corte se observo agrupamiento por función. De cada uno de estas "clusters" se conservó la secuencia representativa que arroja CD-HIT, la cual forma parte del conjunto de secuencias negativas final que se compone de 2,733 secuencias. Quinientas treinta y nueve secuencias fueron elegidas aleatoriamente para formar un conjunto de prueba de péptidos no-AMPs.
- Sitios de corte N-terminal positivos: Para determinar los precursores y regiones propepídicas de cada AMP maduro se realizó una búsqueda de secuencias que, respecto a alguna secuencia de maduro positivo tuvieran: similitud del 100%, longitud mayor al AMP maduro, su organismo fuente sea del mismo género que el AMP, e iniciaran con metionina. Utilizando este conjunto de precursores, se determinó en qué posición del propéptido se encuentra la región proteolítica N-terminal respecto de la metionina inicial, y cuál es el octámero (secuencia de 8 aminoácidos) que se centra en este sitio de corte. Posteriormente se eliminó redundancia de octámeros. En total se obtuvieron 348 octámeros únicos de 838 secuencias analizadas.
- Sitios de corte N-terminal negativos: Para generar este conjunto de datos se utilizaron octámeros sobrelapados en +/- 5 posiciones respecto a los sitios de corte de región propeptídica N-terminal positivos. Además, se agregaron 1,000 octámeros elegidos aleatoriamente provenientes del conjunto de secuencias negativas, obteniendo un total de 4,480 octámeros únicos.

- Sitios de corte C-terminal positivos: Se utilizó la misma metodología empleada para determinar los sitios de corte de región propeptídica N-terminal positivos, con la diferencia de ubicar el sitio de corte con respecto al extremo C-terminal. En total se determinaron 135 octámeros únicos con estas características de 838 secuencias analizadas.
- Sitios de corte C-terminal negativos: Se utilizó la misma metodología empleada para determinar sitios de corte N-terminal negativos. En total se determinaron 2,350 octámeros únicos en este conjunto.

Los conjuntos de datos se encuentran disponibles en el repositorio del "software" derivado de este trabajo (https://github.com/SantiagoSanchezF/proAmps/tree/master/datasets).

2) Entrenamiento del modelo predictivo de sitios de corte proteolíticos y péptido maduro

Con los datos de sitios de corte de región propeptídica N/C-terminal positivos y negativos se realizó una codificación tipo "one hot encoding" (OHE), el cual tiene como representación de cada observación un arreglo de 161 dimensiones, 160 derivadas de 20 aminoácidos*8 posiciones, mas una dimensión extra para representar la posición del octámero respecto de la metionina de inicio.

Partiendo de estos datos, se seleccionaron aleatoriamente el 25% de las observaciones para ser utilizadas como conjunto de prueba después de afinar los modelos. Las secuencias de precursores que contienen los octámeros del conjunto de prueba fueron utilizadas para evaluar el modelo global. El 75% restante de los octámeros se utilizó como conjunto de entrenamiento y validación. Utilizando la librería de Python SciKit-learn v0.20 (Pedregosa *et al.,* 2011) se entrenaron modelos predictivos con los métodos de DA, "Quadratic discriminant analysis", Naive Bayes, Adaboost, "K-nearest neighbors", RF y SVC. La estrategia de "Grid Search" la cual consiste en explorar todos los valores combinados de los parámetros y la validación cruzada de 3 rondas ayudó a determinar los parámetros que optimizan la función AC (*ex infra*) balanceada por clase para estos modelos. También mediante validación cruzada de tres rondas y utilizando Keras v2.2/TensorFlow v1.10 (Chollet *et al.,* 2015; Abadi *et al.,* 2016) se evaluaron diversas arquitecturas de ANN. La función a optimizar utilizada para los modelos ANN fue "binary cross entropy". Se seleccionaron los 3 mejores modelos para ser incorporados en un ensamble de modelos, el cual hace un promedio sopesado de las probabilidades determinadas individualmente por cada modelo, de ser una observación positiva o negativa. Los pesos que maximizan el desempeño del ensamble se estimaron mediante la estrategia "Grid Search".

3) Entrenamiento del modelo predictivo de AMPs

Utilizando los datos de maduros positivos y secuencias negativas se realizaron dos transformaciones independientes que corresponden a los dos tipos de información utilizada en nuestro método. La información de secuencia se codifica por el método de OHE, permitiendo que la máxima longitud de esta codificación sea de 102 posiciones de aminoácidos, así cada observación es representada como un vector de 2,040 dimensiones. En las secuencias con longitud menor a 120 aminoácidos, inicia la codificación en la posición 0 del vector, dejando las posiciones de mayor longitud vacías. La segunda transformación es el cálculo computacional de las variables fisicoquímicas: número de aminoácidos, es decir, correlación con la longitud del péptido; peso

molecular dado que los AMPs tienen un peso molecular bajo; composición de grupos de aminoácideos en diminutos, pequeños, aromáticos, alifáticos, polares, no-polares, cargados, básicos, y ácidos; carga neta generalmente positiva, lo que le permite interactuar con membranas bacterianas; punto isoeléctrico, el cual describe el pH al cual su carga neta es 0 y que en los AMPs se espera que sea de alrededor de 10 para no precipitarse en la membrana; índice alifático que indica la termoestabilidad de los péptidos a partir de los aminoácidos alifáticos ya que los AMPs tienden a ser termoestables; índice de inestabilidad que indica la estabilidad general de los péptidos basados en la composición de aminoácidos y penalizada por la longitud, y que en AMPs se espera que sea baja; índice de Boman que representa el potencial de interacción proteína-proteína, el cual debe ser bajo puesto que se espera que sea mayor la interacción con membranas; índice de hidrofobicidad que es la propiedad principal que permite la interacción con la membrana; índice de momento hidrofóbico, medida que cuantifica la anfipatía del péptido y que en AMPs se espera que sea alta; y posición de membrana que define cada endecámero (secuencia de 11 aminoácidos) del péptido como globular, superficial, o transmembranal.

Todos estos valores se calcularon mediante el paquete estadístico "Peptides" de R (Osorio *et al.*, 2015). Posterior a la estandarización de los datos, se cuantificó la correlación entre variables fisicoquímicas, y de aquellas variables que tuvieran alta correlación (> 0.9) se conservó la que tuviera interpretación biológica más simple. Posteriormente, utilizando las variables fisicoquímicas conservadas se entrenaron modelos predictivos con los algoritmos de DA, "Quadratic discriminant analysis", Naive Bayes, Adaboost, "K-nearest neighbors", RF y SVC, usando la librería Scikit-Learn. Mediante la estrategia de "Grid Search" y validación cruzada de 3 rondas se determinaron los parámetros que optimizan la función de exactitud (AC) balanceada por clase para estos modelos.

Utilizando las librerías Keras v2.2/TensorFlow v1.10 se entrenaron diversas arquitecturas de ANN que permiten tener entradas tanto de los vectores de propiedades fisicoquímicas, como la transformación de las secuencias. Se utilizó como capa de salida una función sigmoide, la cual regresa valores entre 0 y 1 que funcionan como medida de probabilidad. La función a optimizar utilizada para este tipo de modelos predictivos fue "binary cross entropy".

Para generar un mejor modelo predictivo, en este proceso también desarrolló un ensamble de modelos, en el cual se determinó por "Grid Search" cuáles son los pesos de cada modelo que maximizan el desempeño del ensamble.

Cabe destacar que para mitigar los efectos del desbalance de clases en los conjuntos de datos, en todos los entrenamientos se designo un peso especifico a cada clase. Para cada clase este peso fue determinado por la formula n_o / (n_c * n_oc), donde n_o, n_c y n_oc es el numero de total observaciones, numero de clases y el numero de observaciones de la clase, respectivamente.

**Implementación de un algoritmo en un "software"**

Se implementó el algoritmo de identificación como una librería de Python v3.7 (Rossum, 1995), la cual tiene como dependencias: "software" estadístico R v3.2; el paquete de R "Peptides"*;* la librería rpy2 (Gautier, 2008) para comunicarse con R; la librerías de ML SciKit-learn, Keras, y Tensorflow; las librerías estadísticas de Python Numpy (Oliphant, 2006) y Pandas (McKinney, 2010). Para facilitar su instalación, asegurar el correcto funcionamiento de las dependencias, y la compatibilidad de las versiones, se creó un archivo yml para la instalación del "software" en un ambiente conda (Analytics, 2016).

El "software" resultante de este objetivo se hizo público en el repositorio de github https://github.com/SantiagoSanchezF/proAmps.

1) Métricas de desempeño de los modelos predictivos

La evaluación de nuestros modelos predictivos, y las comparaciones entre este trabajo y otros se realizó con matrices de confusión de cada modelo y las métricas derivadas: sensibilidad (Sn) que nos permite conocer la cantidad de AMPs positivos que son correctamente clasificados como tal, especificidad (Sp) que nos permite conocer el número de AMPs negativos que son clasificados así, exactitud (AC) que utilizamos como una medida del desempeño general de un método de clasificación binaria, y el coeficiente de correlación de Matthews (MCC) métrica del desempeño general de un modelo de clasificación binaria que evita los problemas asociados al AC cuando las clases son desbalanceadas (Powers, 2011). La Figura 1 ilustra el procedimiento para generar la matriz de confusión y las métricas derivadas.

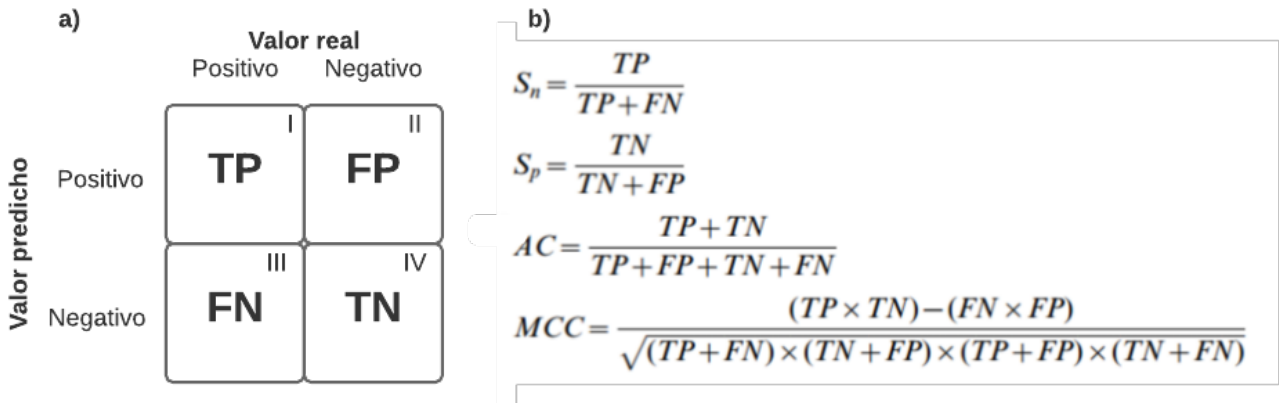


Figura 1. a) Consideraciones para generar la matriz de confusión. Cuadrante I, verdaderos positivos (TP), secuencias positivas correctamente identificadas. Cuadrante II, falsos positivos (FP), secuencias negativas identificadas como positivas incorrectamente. Cuadrante III, falsos negativos (FN), secuencias positivas incorrectamente identificadas como negativas. Cuadrante IV, verdaderos negativos (TN), secuencias negativas identificadas como negativas. b) Fórmulas utilizadas para obtener las métricas derivados de la matriz de confusión. Sensibilidad (Sn), especificidad (Sp), exactitud (AC), coeficiente de correlación de Matthews (MCC).

**Comparación entre los trabajos de identificación de AMPs CAMP, ampscanner y iAMP-2L**

Con el fin de comparar el desempeño de este método respecto a otro trabajos reportados y que están en uso, se realizaron predicciones de AMPs utilizando los conjuntos de datos de este trabajo (714 secuencias en conjunto de prueba de péptidos maduros, y 252 como conjunto de prueba de propéptidos, expuestos en esta misma sección), los datos utilizados en el trabajo de Xiao y col. (2013) compuestos por 1,840 secuencias, los datos utilizados en el trabajo de Thomas *et al.* (2010) (216 secuencias), y los datos del trabajo de Veltri *et al.* (2018) (312 secuencias). Estos conjuntos de datos fueron clasificados como AMPs/no-AMP por las herramientas de predicción del que se derivan, así como por los otros métodos a comparar. Es decir iAmp-2L (Xiao *et al.,* 2013; http://www.jci-bioinfo.cn/iAMP-2L), CAMP (Thomas *et al.,* 2010; http://www.camp3.bicnirrh.res.in/predict/), ampscanner (Veltri *et al.,* 2018; www.ampscanner.com), y proAmps (método derivado de este trabajo). ProAmps aporta dos conjuntos de datos a esta comparación, uno que utiliza los péptidos maduros de prueba, y otro que utiliza los precursores con propéptidos de prueba de nuestro método. Así mismo es necesario destacar que todas las pruebas realizadas con proAmps se ejecutaron utilizando la opción de predicción de AMPs con previa predicción de péptido maduro (*ex infra*).

A partir de los resultados de estas predicciones se generaron matrices de confusión para cada par conjunto de datos - predictor, y se obtuvieron las métricas de desempeño por los métodos expuestos en esta sección. También se realizó análisis de ROC (por sus siglas en ingles “Receiver Operating Characteristic”), que es una gráfica de la sensibilidad contra la especificidad de un clasificador variando el umbral de discriminación para cada uno de los métodos, utilizando el conjunto de datos de precursores de AMPs.

1) Comportamiento de proAmps en transcriptomas sintéticos

En este mismo objetivo se plantea una prueba de concepto y análisis del comportamiento de proAmps en datos de tipo transcriptómico. Para lograrlo se seleccionaron aleatoriamente 201,500 subcadenas de las secuencias de precursores de AMPs, lo cual simula los fragmentos ensamblados obtenidos al realizar secuenciación de RNA utilizando tecnologías de secuenciación masiva. En cada una de estas secuencias sintéticas se registró: fracción del AMP maduro que esta representada en el fragmento, fracción de la región propeptídica que esta representada en el fragmento, longitud de fragmento, y distancia del fragmento al extremo N-terminal. A este conjunto de secuencias se le llamó transcriptoma sintético de AMPs.

A partir de las predicciones obtenidas por proAmps se generaron mapas colorimétricos, con el fin de representar como varía la probabilidad promedio de ser clasificado como AMP respecto a la fracción del AMP y región propeptídica.

También se generó un transcriptoma sintético a partir del ensamble genómico de *Drosophila melanogaster* descrito en el siguiente objetivo. Se generó con una estrategia igual a la del transcriptoma sintético de AMPs, en este caso el transcriptoma sintético cuenta con 4,694,520 secuencias. A este conjunto de secuencias se le llamó transcriptoma sintético de *Drosophila.*

**Identificación *in silico* de potenciales AMPs en dos especies de distintos reinos**

Se realizó la predicción de AMPs utilizando el método proAmps, sobre 5,297 proteínas derivadas del ensamble genómico de *Drosophila melanogaster dm6* (Hoskins *et al., 2015;* GenBank "assembly accession": GCA_000001215.2), filtradas por longitud menor a 200 aminoácidos. También se realizó identificación sobre 7,048 proteínas derivadas del ensamble genómico de *Phaseolus vulgaris* disponible por Phytozome (Goodstein *et al*., 2012; http://phytozome.jgi.doe.gov) las cuales tienen una longitud menor a 200 aminoácidos. Utilizando las secuencias clasificadas positivamente como AMPs se realizó un análisis manual de la anotación funcional de las proteínas del genoma de origen. También se determinó mediante InterPro (Finn *et al.*, 2016) los dominios funcionales con los que cuenta la proteína, y si se ha reportado relación entre estos y actividad antimicrobiana.

En este proceso de identificación se utilizó la opción de predicción de AMPs, con previa predicción de péptido maduro de nuestro método, ya que al utilizar datos derivados de secuenciación masiva los cuales cuentan con toda la región codificante, sin modificaciones postraduccionales, esta es la opción recomendada.

**Análisis evolutivo de los AMPs, detectando posible convergencia o divergencia funcional**

Se realizó el análisis de agrupaciones en las secuencias de péptidos maduros positivos por dos enfoques distintos: 1) Un enfoque dependiente de secuencia, formando agrupaciones por similitud mínima de 30% de secuencia, utilizando CD-HIT. Este valor de corte fue seleccionado ya que por las características de longitud corta y alta variabilidad en secuencia primaria es difícil agrupar a los AMPs con menor similitud. A los grupos derivados de este proceso se les llamo "clusters" de secuencia. 2) Otro enfoque utiliza las propiedades fisicoquímicas calculadas por el paquete de R "Peptides" basado en las mismas variables que se utilizaron para entrenar el modelo de predicción de AMPs. Se realizó un análisis de componentes principales (PCA) con las variables fisicoquímicas para determinar las relaciones entre ellas y si son adecuadas para realizar clasificación en los datos. En este enfoque se determinaron las distancias entre observaciones por el método Manhattan, que es la métrica de distancia preferida en aplicaciones de alta dimencionalidad (Aggarwal *et al.* (2001), y se realizo agrupamiento por el método "Partitions Around Medoids" (PAM) (Kaufman y Rousseeuw, 1987), método de agrupamiento particional, en el que se busca minimizar la distancia entre las observaciones etiquetadas en un grupo y el punto designado como el centro de ese agrupamiento. El número de "clusters" se eligió realizando gráficas de valor de silueta promedio (Rousseeuw, 1987) contra número de "clusters". A los "clusters" derivados de este proceso se les llamo "clusters" *fisicoquímicos.* Posteriormente, mediante análisis del valor de coeficiente de silueta, se determinó qué observaciones son valores atípicos en el "cluster" al que fue asignado. El coeficiente de silueta es una métrica utilizada para evaluar la calidad de un agrupamiento y es utilizado para determinar el numero ideal de "clusters", parámetro de entrada del algoritmo PAM. Las observaciones que tuvieran un promedio de coeficiente de silueta menor a 0.2 se descartaron del "cluster" al que fue asignado. El conjunto de datos derivado de este proceso se llamó conjunto de análisis evolutivo y consta de 520 secuencias.

Posteriormente se realizó el alineamiento pareado del conjunto de análisis evolutivo contra sí mismo utilizando la herramienta BLAST. Con esta información se creó una matriz de valores promedio de identidad, en la cual se promedia la identidad derivada de BLAST de cada par de AMP del conjunto de análisis evolutivo siempre y cuando el par obtenga un “e-value” menor a 1e-5. En caso de no cumplir con el criterio de “e-value”, la identidad del par se fija en 0. Esta matriz fue utilizada para hacer un análisis detallado sobre las relaciones entre las secuencias contenidas en cada ”cluster” fisicoquímico.

El objetivo de realizar un análisis evolutivo se logró al construir y contrastar los ”clusters” fisicoquímicos y los ”clusters” de secuencia, incorporando al análisis la anotación que recibe cada AMP en la base de datos APD.

# Resultados y Discusión

## Desarrollo de un algoritmo utilizando métodos mixtos de identificación de AMPs

Para poder desarrollar un método que utilice entradas mixtas (secuencias y propiedades fisicoquímicas), es necesario que las secuencias de entrada sean péptidos maduros. Si se utiliza secuencias que contengan pre/pro-péptidos el método calculará propiedades fisicoquímicas erróneas que no corresponden al péptido funcional. Es por esto que para utilizar información derivada de ensambles genómicos, nuestro método debe incluir un paso de predicción de péptido maduro. Con nuestro algoritmo es posible la identificación de AMPs a partir de secuencias de péptidos maduros, y a partir de secuencias obtenidas por métodos de secuenciación (*e.g.* proteínas derivadas de ensambles genómicos), ya que tenemos la certeza que el método buscará sitios de corte potenciales de AMPs, y hará una predicción de las secuencias para péptidos maduros.

Es importante resaltar que el primer paso de este algoritmo no pretende ser un predictor de sitios proteolíticos generalista, como lo es el programa ProP (Duckert *et al.,* 2004; http://www.cbs.dtu.dk/services/ProP/), en lugar de eso, este paso busca identificar específicamente sitios de corte en el propéptido de un posible percusor de AMP de plantas o artrópodos.

1) Entrenamiento de modelo predictivo de sitios de corte y péptido maduro

Se entrenaron diversos modelos con el fin de compararlos y elegir los mejores. En los casos de los algoritmos que no utilizan ANN (a los que en su conjunto llamaremos "shallow"*),* se muestran los modelos que utilizan los parámetros que optimizan la función de AC balanceado por clases. En el caso de los métodos de ANN, solo se ejemplifica el modelo que tuvo el mejor desempeño, puesto que mostrar los diferentes modelos construidos con variaciones de hiperpárametros resulta impráctico. Cabe destacar que ninguna de estas variaciones se desempeño peor que los métodos "shallow".

El método de predicción de péptido maduro se basa en dos modelos predictivos, uno para cada extremo del péptido. A continuación, se explica a detalle los resultados obtenidos para ambos tipos de predicción.

Extremo N-terminal

Como se muestra en la Figura 2, los modelos predictivos SVC, RF, y ANN de sitio de corte en el extremo N-terminal, son los de mejor desempeño. Si juzgamos únicamente por el MCC de cada modelo, RF es sin duda el mejor, esto debido a su alta especificidad, y el desbalance de clases en el conjunto de prueba. El modelo de SVC también tuvo alta especificidad, conservando una sensibilidad importante. Sin embargo, el modelo con mayor sensibilidad es de ANN, el cual muestra también tener una buena especificidad. Al analizar estos resultados se decidió unificar estos modelos en un ensamble, en el cual se pudiera tener las ventajas de los modelos altamente específicos (RF y SVC), combinadas con el modelo más sensible (ANN). Empíricamente se determinó que los pesos que generan un modelo que mantiene la sensibilidad y maximiza la

especificidad son 0.5, 0.1, y 0.4 para ANN, RF, y SVC respectivamente. Este ensamble tiene como métricas de desempeño: sensibilidad 0.913, especificidad 0.993, AC 0.993, y MCC 0.369.

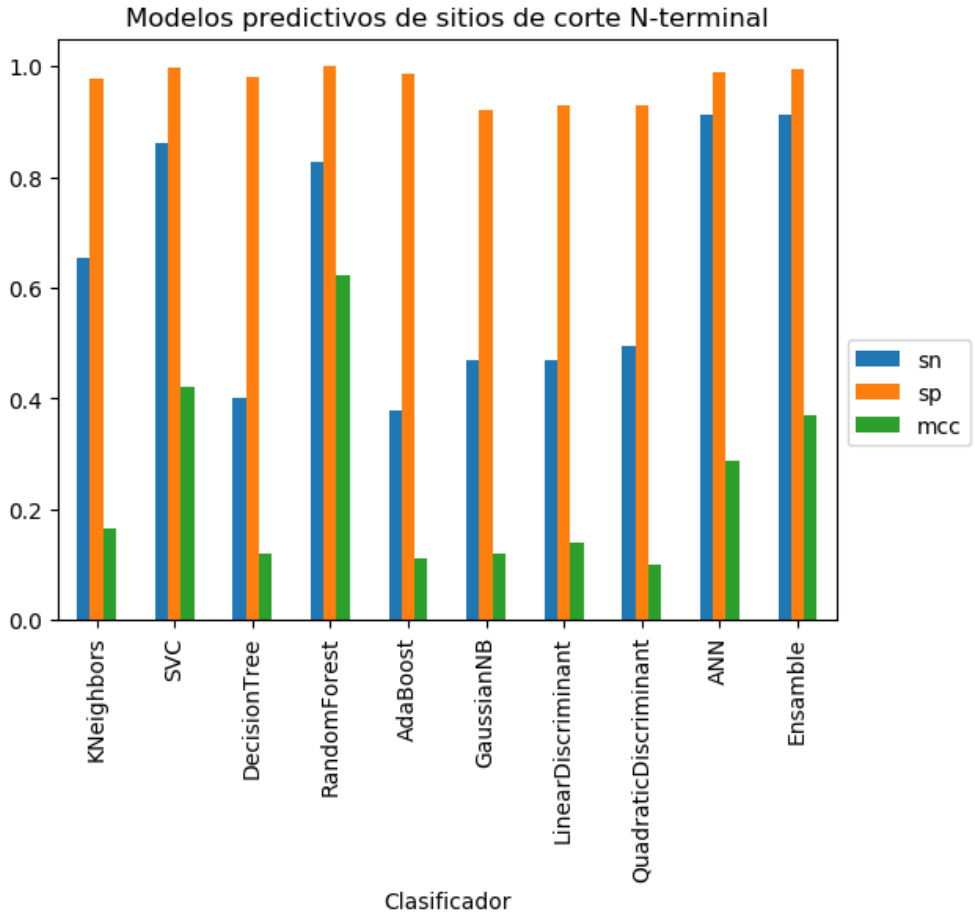


Figura 2. Comparativa de las métricas de desempeño de los modelos entrenados para la predicción del sitio de corte del extremo N-terminal.

En el caso del extremo N-terminal el modelo de ANN con mejor desempeño es una red "fully-connected" de 4 capas ocultas, con 150, 100, 100, y 100 neuronas artificiales respectivamente. La activación de todas las neuronas de capas ocultas utiliza la función de unidad lineal rectificada (ReLu, del inglés "Rectified Linear Unit"). Durante el entrenamiento de este modelo se realizó "Dropout" del 50% de neuronas en la capa oculta 2 con el fin de regularizar el modelo. La capa de salida del modelo consiste en una función sigmoide. El modelo de RF utiliza 1,200 árboles y una profundidad máxima de 100 nodos por árbol. El modelo SVC tiene un parámetro gama es de 0.01, y utiliza la función "Radial basis" para definir el hiperplano de separación de clases o "kernel".

Extremo C-terminal

También en el caso de los modelos desarrollados para el extremo C-terminal los métodos que muestran mejor desempeño, juzgando por MCC, son RF, SVC y ANN, en ese orden (Figura 3). Como en el caso del extremo N-terminal, los modelos ANN y SVC muestran tener una alta

especificidad. Sin embargo, las mejores métricas de sensibilidad fueron obtenidas por RF y "K neighbours". Se realizó una búsqueda de distintos pesos que pudieran hacer un modelo ensamble con mejor desempeño que los modelos individuales, para esta búsqueda se utilizaron los modelos RF, SVC, "K neighbours" y ANN. El ensamble resultante, a pesar de no tener un mejor MCC que RF en lo individual, logra conservar la sensibilidad de RF con la ventaja de tener una especificidad muy alta. Los pesos asignados a cada modelo para este ensamble son 0.02, 0.89, 0.0, y 0.09 para ANN, RF, "K neighbours" y SVC, por lo tanto, se eliminó "K neighbours". Este modelo tiene como métricas de evaluación sensibilidad 0.441, especificidad 0.999, AC 0.999, y un MCC 0.606.

El modelo predictivo de ANN utilizado en el ensamble del extremo C-terminal consiste en una red de 2 capas ocultas "fully-connected" de 350 y 400 neuronas respectivamente. Utiliza la función de activación ReLu con la cual se propaga de una neurona a otra. La capa de salida del modelo consiste en una función sigmoide. El modelo RF consiste en 200 árboles, con una profundidad máxima de 30 nodos para cada uno. El modelo SVC utiliza un parámetro "cost*"* de 1,000, gama de 0.01 y utiliza como función de kernel "Radial basis".

Los modelos individuales de predicción de sitio de corte se integran en un algoritmo el cual consiste en tomar una secuencia de entrada, dividirla en octámeros consecutivos y sobrelapados, los cuales se obtienen al recorrer en una posición de aminoácidos respecto del octámero anterior. Esto da como resultado N-7 octámeros para cada proteína a predecir su secuencia madura, donde N es la longitud del péptido. Cada uno de estos octámeros es evaluado tanto por el modelo de N-terminal como por el de C-terminal. El octámero con mayor probabilidad de tener el sitio de corte N-terminal es registrado, y si su probabilidad es mayor a 0.5, la unión peptídica entre el cuarto y quinto aminoácido se consideran el sitio de corte de la región propeptídica N-terminal. El mismo procedimiento es aplicado con el modelo del sitio de corte C-terminal, sin embargo para aceptar un sitio de corte C-terminal en los casos que existe una predicción de propéptido N-terminal, es necesario que este sea predicho en una posición posterior al corte N-terminal.

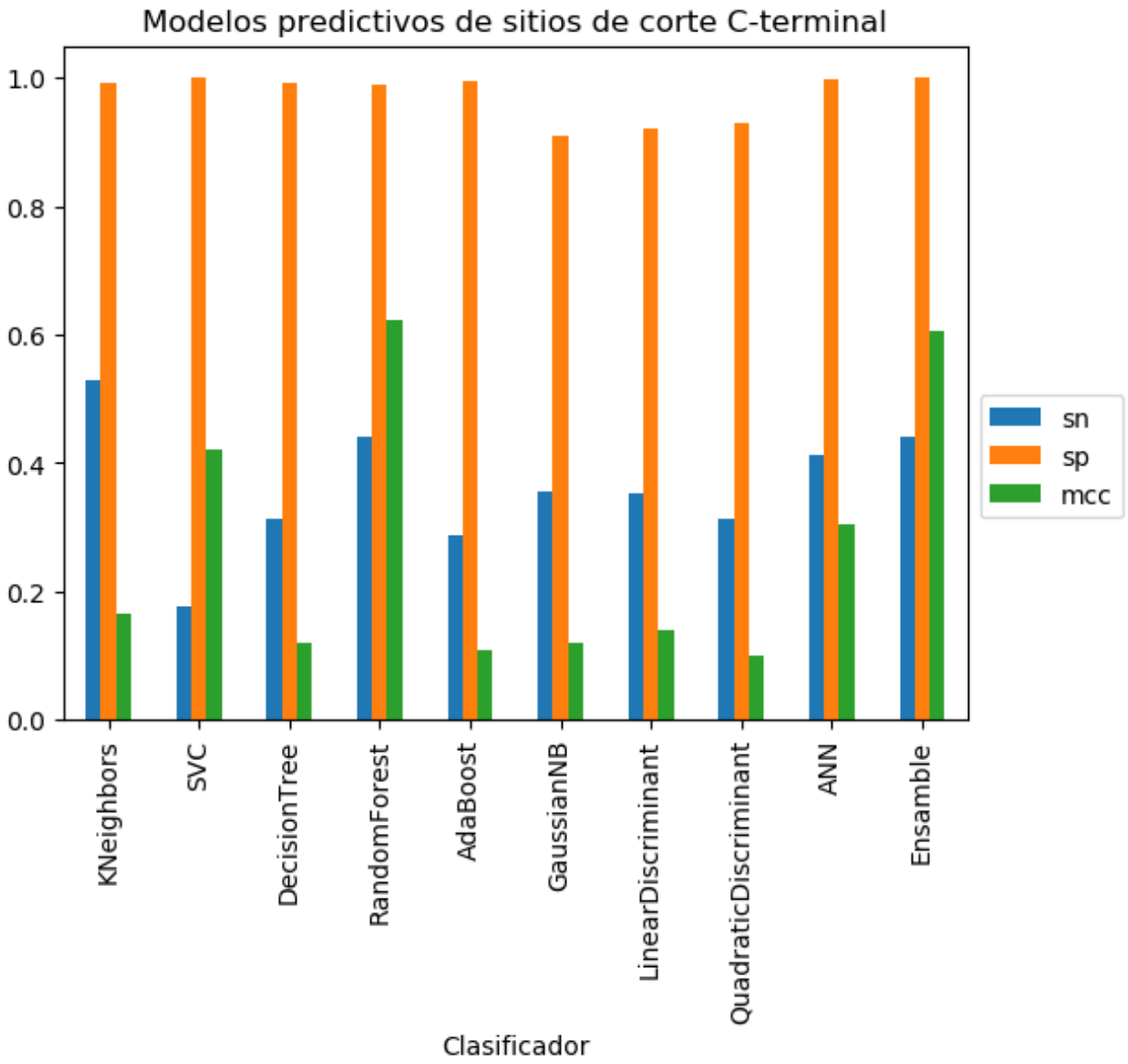


Figura 3. Comparativa de las métricas de desempeño de los modelos entrenados para la predicción del sitio de corte del extremo C-terminal.

Para evaluar nuestro algoritmo de predicción de péptido maduro se utilizaron las secuencias de precursores con propéptidos cuyos sitios de corte, ya sea en N o C-terminal, no se hayan usado para el entrenamiento de los modelos individuales. Bajo este criterio se utilizaron 155 secuencias como conjunto de prueba positivo de propéptidos. El conjunto de prueba negativo de propéptidos se compone de 97 secuencias del conjunto de maduros positivos no derivados de propéptidos. Un verdadero positivo se considera a aquellas secuencias del conjunto de prueba positivo de propéptidos en las que el método identifique los sitios de corte en las posiciones exactas, es decir que prediga un péptido maduro igual al validado experimentalmente. Falsos negativos son aquellas secuencias de este mismo conjunto que el método no prediga la secuencia madura exactamente igual a la validada experimentalmente. Un verdadero negativo es aquella secuencia de conjunto de

prueba negativo de propéptidos a la que no se le identifica ningún sitio de corte. Falsos positivos son aquellas secuencias de conjunto de prueba negativo de propéptidos a las que se le identifique algún sitio de corte.

A nuestro conocimiento sólo existen dos herramientas que predicen péptidos maduros sobre secuencias precursoras de proteínas funcionales, en las cuales sus clases objetivo sobrelapan con proAmps. Una de ellas es proP, la cual utiliza ANN para hacer predicción de sitios proteolíticos por proteínas de la familia proproteína convertasa, enzimas que se caracterizan por activar las funciones de un percusor protéico. Esta herramienta es muy general, ya que su propósito es encontrar sitios proteolíticos de todo el dominio *eukarya*. La otra herramienta disponible es SpiderP (Wong *et al.,* 2013; http://www.arachnoserver.org/spiderP.html) la cual hace predicción de secuencias maduras de toxinas proteicas de origen *Arachnida,* utilizando el método de SVC.

En la Tabla 1 se muestran los resultados de las métricas de desempeño de cada una de estas herramientas utilizando los conjuntos y criterios descritos arriba. En la evaluación general de estos métodos proAmps es el que tiene un mejor desempeño, esto es debido a que tiene una sensibilidad superior y una especificidad que sobrepasa ligeramente a los otros métodos. Sin embargo, es de considerarse que una comparación entre estas herramientas es superficial, ya que el universo potencial de predicción de proP es superconjunto del universo de búsqueda de proAmps. También existe un conjunto de intersección entre el universo de búsqueda de proAmps y SpiderP, pero es de destacar que existe un gran espacio de complemento. No obstante, esta comparación es ilustrativa para fines comparativos de este trabajo.

Tabla 1. Comparativa de desempeño de los métodos disponibles de predicción de péptido maduro.

| Método | Sensibilidad | Especificidad | Coef. correlación de Matthews |
|---|---|---|---|
| proAmps | 0.484 | 0.979 | 0.489 |
| proP 1.0 | 0.290 | 0.959 | 0.306 |
| SpiderP | 0.303 | 0.958 | 0.315 |

Conocer las secuencias maduras de AMPs a partir de secuencias obtenidas de datos de secuenciación masiva tiene relevancia biotecnológica. En primera instancia podemos mencionar el uso que se le da en este método. Se sabe que sustituciones puntuales de aminoácidos puede afectar dramáticamente la potencia y función de los AMPs (Zelezetsky *et al.,* 2005), por ello, consideramos que incluir un paso de predicción de AMPs maduros para después predecir con base a secuencia y características fisicoquímicas, permite una mayor especificidad (*ex infra*). Los métodos que se basan estrictamente en secuencia dejan de lado las sutiles variaciones en secuencia que afectan la función antimicrobiana.

Idealmente debería existir un modelo de predicción de sitio de corte por cada proteasa, sin embargo, actualmente es difícil relacionar directamente una proteína y la proteasa que hace procesamiento postraduccional sobre ella. Es por ello que una aproximación menos rigurosa para

predecir los sitios de corte resulta útil. Una limitante para el perfeccionamiento de esta aproximación es el sesgo a organismos modelo que existe en las bases de datos de proteínas, la cual puede ocultar los propéptidos de un AMP proveniente de un organismo menos estudiado, y así privar al entrenamiento de esa información. A medida que se conozcan nuevos AMPs y se elimine el sesgo mencionado se podrán desarrollar métodos de predicción más precisos.

Gracias al abaratamiento del poder de cómputo, el desarrollo de nuevas técnicas de miniaturización de experimentos y simulación computacional, y a los avances en inteligencia artificial, el descubrimiento automatizado de fármacos comienza a ser una realidad (Baskin *et al.,* 2016). Sin embargo, para poder utilizar estos métodos con proteínas de origen natural es necesario conocer la forma funcional de estas proteínas. La predicción de AMPs maduros a partir de proAmps puede ayudar a la detección sistemática *"*high-throughput*"* de nuevos fármacos al hacer predicciones sobre la estructura primaria funcional de los AMPs.

2) Entrenamiento de modelo predictivo de AMPs

Con el fin de desarrollar este objetivo se entrenaron y compararon diversos modelos, para después elegir los mejores. En los casos de los algoritmos "shallow" se muestran los modelos que utilizan los parámetros que optimizan la función de AC balanceado por clases. En el caso de los métodos de ANN, solo se muestra el modelo que tuvo el mejor desempeño.

Se crearon diversos modelos predictivos. Los modelos "shallow" entrenados utilizan exclusivamente información de las propiedades fisicoquímicas. Para el caso de las ANN se crearon diversas arquitecturas, algunas dependientes de secuencia, otras dependientes de las propiedades fisicoquímicas, y las mixtas las cuales resultaron tener mejor desempeño. En la Figura 4 se distingue que el desempeño de los modelos que dependen exclusivamente de las propiedades fisicoquímicas es bajo en comparación con los de entrada mixta, siendo RF el mejor de ellos. La ANN desarrollada supera con creces cualquiera de los modelos "shallow", sin embargo, al crear un ensamble del mejor de los modelos "shallow" con la mejor arquitectura de ANN se obtiene un método con mejor desempeño que sus componentes individuales. En este caso se observa que el ensamble mejora su especificidad con un sacrificio pequeño de sensibilidad. La estructura del modelo ensamble se observa en la Figura 5, el cual consiste en un componente de capas densas "fully-connected" para la entrada del vector de propiedades fisicoquímicas, y una sucesión de capas convolucionales de una dimensión con capas de redes recurrentes para la entrada de secuencia. Posteriormente, la salida de estos componentes se concatena y se conecta a una función sigmoidal. Paralelamente con el vector de propiedades fisicoquímicas se realiza predicción con el modelo de RF. Se realiza un promedio sopesado de las salidas del componente ANN y el "shallow", con pesos de 0.61 y 0.39 respectivamente.

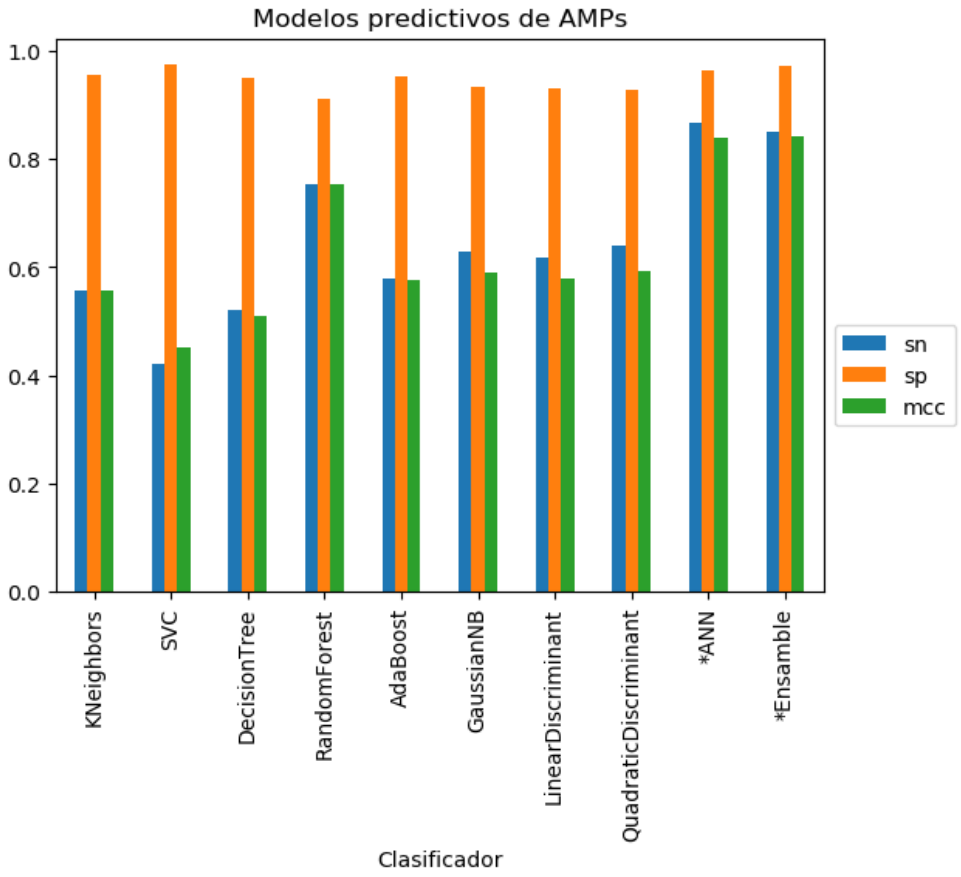


Figura 4. Comparativa de las métricas de desempeño de los modelos entrenados para la predicción de AMPs a partir de péptidos maduros. * modelos con entradas mixtas (secuencia y variables fisicoquímicas).

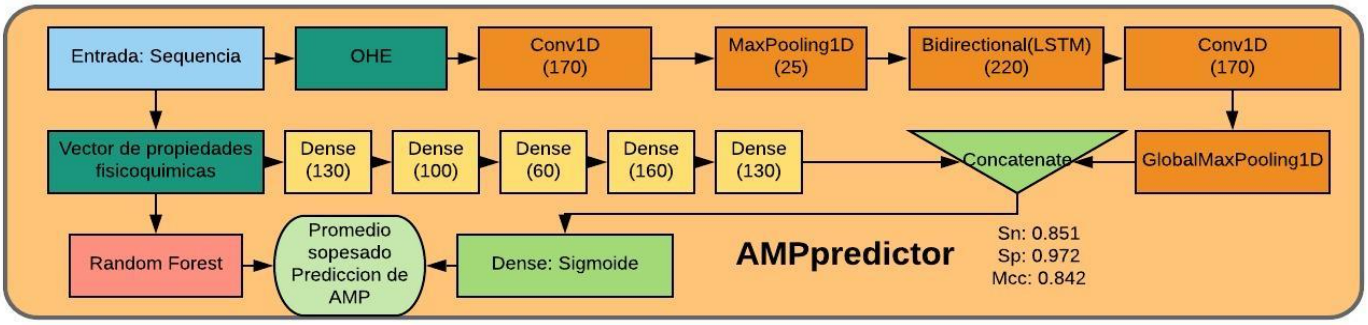


Figura 5. Esquema del algoritmo de identificación de AMPs maduros (AMPpredictor) a partir de péptidos maduros. Entre paréntesis se muestra el número de elementos utilizados en cada capa. En el caso de las capas convolucionales se muestran los filtros.

Los modelos de redes convolucionales combinadas con redes recurrentes son muy populares para el procesamiento de lenguaje natural y de secuencias biológicas (Lai *et al.,* 2015; Veltri *et al.,* 2018; Armenteros *et al.,* 2019) por su capacidad de abstraer patrones en datos de secuencia. En este trabajo utilizamos las ventajas de estas representaciones en la obtención de modelos de alto poder predictivo, sin embargo, la supremacía de desempeño de este tipo de arquitecturas en

modelado de datos de secuencia ha sido recientemente cuestionada. La incorporación del mecanismo de 'atención' para modelar este tipo de problemas se ha propuesto como una alternativa con mejores resultados generalmente (Vaswani *et al.,* 2017). Por lo tanto, es de considerarse que en el futuro muchos de los modelos predictivos desarrollados sean entrenados utilizando las nuevas arquitecturas de redes. Actualmente, a pesar de estar descritos los mecanismos de 'atención', las implementaciones de estos métodos son pocas y deficientemente documentadas. Esto dificulta la accesibilidad generalizada a estos métodos, razón por la que este tipo de arquitecturas no fueron consideradas en este trabajo.

**Implementación de un algoritmo en un "software"**

Producto de este objetivo se obtuvo un "software" de predicción de péptidos antimicrobianos en plantas y artrópodos a partir de estructura primaria de proteínas llamado proAmps. Este programa toma como entrada un archivo en formato FASTA, el cual puede ser procesado por una de las tres opciones de proAmps:

- Predicción de AMPs sin predicción de péptido maduro. Esta opción permite identificar AMPs a partir de proteínas a las cuales se les conoce su secuencia funcional (*e.g.* información de secuenciación Edman). Al utilizar las secuencias directamente del archivo de entrada, esta opción es más rápida que la opción que requiere el paso de identificación de región propeptídica.
- Predicción de AMPs, con previa predicción de péptido maduro. Esta opción realiza predicción de regiones propeptídicas e identifica AMPs a partir de las secuencias maduras predichas. Esta es la opción más recomendable cuando se utiliza información derivada de secuenciación masiva, ya que generalmente estos datos no contienen información sobre el procesamiento postraduccional de las proteínas. En comparativa con la opción anterior, ésta es más lenta puesto que requiere de predicción de sitios de corte en cada octámero sobrelapado.
- Predicción de péptido maduro. Esta opción no realiza predicción de AMPs, solo hace la predicción de las regiones propeptídicas.

El resultado de procesar un archivo de secuencias con alguna de estas opciones es un archivo FASTA, el cual en los "headers" de secuencia contiene la siguiente información:

```
>"identificador de péptido" | Posición sitio de corte y probabilidad de región propeptídica N-ter | Posición sitio de corte y probabilidad de región propeptídica C-ter | probabilidad de AMP y predicción de ”cluster” fisicoquímico.
```

Las secuencias en el archivo de salida son idénticas a las de entrada en el caso de la predicción de AMP sin predicción de péptido maduro. Con las opciones restantes la salida son secuencias a las cuales se le eliminó la región propeptídica predicha.

Este "software" está disponible para instalación local (https://github.com/SantiagoSanchezF/proAmps) en sistemas operativos tipo UNIX, y se ejecuta directamente desde la línea de comandos. El repositorio incluye un archivo README, con la documentación sobre instalación y uso de proAmps.

La mayoría de los "software" de identificación de AMPs (e.g. Veltri *et al.,*2018; Xiao *et al.,* 2013; Thomas *et al.,* 2010) se encuentran implementados en un sitio de internet. Este tipo de implementaciones tiene la ventaja de ser accesible para el público con bajo perfil informático, ya que no requieren de proceso de instalación, y no utiliza la línea de comandos. Otra ventaja es que el usuario no realiza el proceso de manera local, sino que utiliza los recursos computacionales del servidor. Sin embargo, las implementaciones en servidores "web" imponen límites a los usuarios en cuanto el volumen de datos y/o el tiempo de procesamiento ya que en algunos casos la predicción de AMP se debe hacer secuencia por secuencia (Wang *et al.*, 2016). El objetivo de proAmps de realizar predicción masiva de AMPs dificulta la implementación en servidores web, ya que los recursos computacionales deben ser muy grandes para poder garantizar el uso de varios usuarios del servidor. Es por ello que se decidió implementar nuestro método para instalación local.

La instalación de dependencias de un "software" se facilita mediante el uso de manejadores de paquetes como conda, por lo que se decidió crear un ambiente conda para la instalación de proAmps*,* volviéndolo atractivo para usuarios con bajo perfil informático. Este tipo de implementaciones son cada vez más populares, como se observa con el número creciente de paquetes en el canal de paquetes especializado en bioinformática *BIOCONDA* (Dale *et al.,* 2017).

ProAmps es un "software" de fácil instalación y uso, sin embargo, aún es optimizable. La incorporación de la opción de cómputo en paralelo y un análisis del proceso permitirá mejorar la eficiencia de proAmps, objetivo que está fuera del alcance de este trabajo.

**Comparación de los trabajos de identificación de AMPs CAMP, ampscanner y iAMP-2L**

En la Tabla 2 se observan las métricas de desempeño de los distintos métodos comparados, utilizando los distintos conjuntos de datos.

Tabla 2. Comparativa del desempeño de distintas herramientas de identificación de AMPs.

| | **Conjunto de datos*** | | | | |
|---|---|---|---|---|---|
| **Metodo** | **CAMP** | **ampscanner** | **iAMP-l2** | **proAmps M** | **proAmps P** |
| | Sn,Sp,MCC, | Sn,Sp,MCC, | Sn,Sp,MCC, | Sn,Sp,MCC, | Sn,Sp,MCC, |
| **C-SVC** | **0.92,0.57,0.54** | **0.89,0.82,0.71** | **0.87,0.48,0.29** | **0.86,0.50,0.32** | **0.74,0.50,0.21** |
| **C-RF** | **0.97,0.73,0.75** | **0.92,0.84,0.77** | **0.94,0.43,0.31** | **0.91,0.51,0.37** | **0.70,0.50,0.18** |
| **C-ANN** | **0.89,0.57,0.50** | **0.83,0.84,0.68** | **0.80,0.71,0.43** | **0.78,0.64,0.36** | **0.62,0.65,0.23** |
| **C-DA** | **0.89,0.42,0.37** | **0.87,0.82,0.69** | **0.83,0.53,0.30** | **0.87,0.54,0.62** | **0.64,0.54,0.15** |
| **ampscanner** | **0.97,0.89,0.87** | **0.98,0.99,0.97** | **0.96,0.69,0.54** | **0.94,0.68,0.55** | **0.81,0.69,0.42** |
| **iAMP-2L** | **0.91,0.98,0.88** | **0.86,0.85,0.71** | **0.87,0.87,0.70** | **0.91,0.82,0.66** | **0.20,0.82,0.03** |
| **proAmps** | **0.85,1.00,0.83** | **0.71,0.92,0.65** | **0.75,0.95,0.72** | **0.85,0.97,0.84** | **0.73,0.97,0.80** |

*Utilizando los conjuntos de prueba. Los resultado se presentan en formato Sn,Sp,MCC, con el MCC en rojo para facilitar su ubicación. En las columnas se encuentra el conjunto de prueba de cada método donde proAmps M y P representan el conjunto de prueba de péptidos maduros y propéptidos, respectivamente. En cada fila se muestra el método con el que fue identificado. Los métodos que comienzan con "C-" son los derivados de CAMP.

Al utilizar el conjunto de datos de CAMP, se observa que el método con mejor desempeño es ampscaner, el cual tiene una sensibilidad similar a las del método de RF utilizado en CAMP, pero con la ventaja de tener una especificidad superior. ProAmps es el método con mayor especificidad sobre este conjunto de datos. Es destacable que los métodos utilizados por CAMP, son los que obtienen un MCC más bajo, siendo RF su método más robusto. Por otro lado, en el conjunto de datos de ampscanner se obtienen métricas de desempeño más homogéneas entre los distintos métodos. Aquí el método de ampscanner es el que mejor desempeño obtiene en todas las métricas de evaluación. En este conjunto, proAmps es el segundo en la métrica de especificidad. En los datos utilizados para desarrollar iAMP-l2, proAmps tiene el mejor desempeño general, seguido de iAMP-l2.

Se realizaron dos comparaciones utilizando los conjuntos de datos de ese trabajo. Una utiliza el conjunto de prueba de péptidos maduros y la otra utiliza el conjunto de propéptidos. Se observa que para el conjunto de proteínas maduras ampscanner es el que mayor sensibilidad alcanza, seguido de iAMP-l2 y CAMP. Sin embargo, en la métrica general es proAmps el que mejor evaluación obtiene debido a su alta especificidad.

Cuando se realiza el comparativo utilizando el conjunto de datos de propéptido es evidente que proAmps es el método con mejor desempeño en las métricas de evaluación, excepto la sensibilidad. Ampscanner es el segundo mejor método en este tipo de datos, el cual alcanza una sensibilidad más alta a la de nuestro método.

Estos resultados nos arrojan que los mejores métodos de identificación de AMP maduros son ampscanner y proAmps. Cabe resaltar que estos dos métodos son los más nuevos, y han sido entrenados con metodologías e innovaciones de inteligencia artificial que muy recientemente se han popularizado (LeCun *et al.,* 2015) debido a su gran poder predictivo e implementaciones fáciles de usar para un público generalizado (*e.g.* redes convolucionales y recurrentes, activación ReLu, regularización de redes), por lo tanto, los métodos previamente desarrollados no son capaces de aprovechar el potencial hasta ahora conocido de las ANN.

El desempeño de sensibilidad de proAmps frente a los otros métodos es similar, pero siempre más bajo. Este resultado es de esperar, ya que proAmps fue entrenado con el propósito de identificar AMPs de plantas y de artrópodos, y no en una visión generalista. Sin embargo, el hecho de que la sensibilidad sea aceptable, nos sugiere que muchas de las características (tanto de secuencia como fisicoquímicas) de los AMPs de un clado, pueden ser extrapolados para la identificación de AMPs en otros clados.

Otro aspecto interesante de este resultado es la superioridad en la métrica de especificidad de proAmps en todos los conjuntos excepto en el de apscanner. El otro método que supera con creces a los demás en esta métrica es el de iAMP-l2, método de identificación por valores umbral de características fisicoquímicas. Esto nos sugiere que la alta especificidad de estos métodos puede estar relacionada con la utilización de parámetros fisicoquímicos. Una explicación posible es que los métodos que se basan estrictamente en homología pueden perder sensibilidad a las sutiles variaciones en secuencia que afectan las características fisicoquímicas del péptido, lo cual puede llevar a pérdida de su función antimicrobiana. Es por ello, que a pesar de que proAmps no es el método con mayor sensibilidad, el paso de identificación de regiones propeptídicas es provechoso para el método, ya que sin predicción de péptido maduro no se podría realizar predicción de AMPs utilizando propiedades fisicoquímicas, y nuestro método perdería la especificidad que lo destaca sobre los otros.

En un análisis más detallado sobre el comportamiento de los distintos métodos en el conjunto de datos de propéptidos se observa que proAmps es el método con mejor desempeño (Figura 6) sin importar el valor del umbral de decisión entre AMP/no-AMP que se utilice, por lo tanto, es el método con mayor área bajo la curva. Este método muestra una gran especificidad y solo decae drásticamente cuando se intenta obtener una sensibilidad muy alta en cualquier conjunto de datos evaluado.

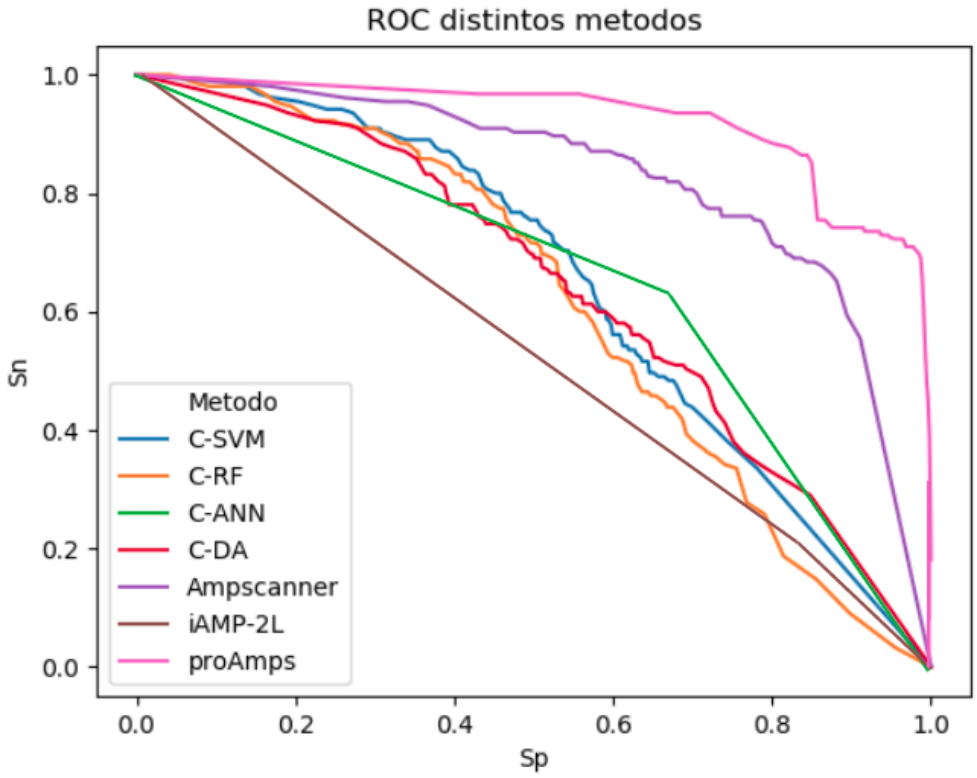


Figura 6. Curvas ROC al variar la característica probabilidad umbral de decisión. Los métodos que comienzan con "C-" son los derivados de CAMP.

Los métodos de CAMP y iAMP-l2 son evidentemente los que tienen peor desempeño al utilizar información con secuencias de propéptidos, su área bajo la curva es similar a la recta que obtendría un modelo predictivo al azar en un conjunto balanceado de casos positivos y negativos.

Por otro lado, ampscanner resulta ser un método muy bueno para la identificación de AMPs a partir de propéptidos. Incluso en el valor umbral de decisión predeterminado de proAmps y ampscanner la sensibilidad del segundo es más alta comparado con el primero. Una explicación posible del buen desempeño de este método es que, al ser dependiente de secuencia, cuando es alimentado con péptidos no maduros estos contienen la información sobre el péptido funcional, y así en el proceso de deconvolución los patrones de los AMP maduros pueden ser detectados por el método. Sin embargo, este método también padece de una baja especificidad cuando se busca tener sensibilidad superior a 0.7.

La ventaja de proAmps sobre los otros métodos en este aspecto, emerge por la introducción del método de identificación de secuencias maduras, las cuales permiten hacer una clasificación que contempla aspectos fisicoquímicos. Este enfoque no ayuda a detectar más AMPs (sensibilidad), sino a discriminar correctamente proteínas que no sean AMPs. Los métodos que se basan estrictamente en secuencia no son sensibles a las sutiles variaciones en secuencia que afectan la función antimicrobiana.

De acuerdo con nuestra evaluación, proAmps es el método de mejor desempeño cuando se utilizan secuencias precursoras de AMPs de plantas y artrópodos. Nuestra herramienta proAmps ocupa un nicho hasta ahora solo explorado por SpiderP (Wong *et al.,* 2013), en la identificación de toxinas partiendo de la predicción de un precursor del péptido maduro. Sin embargo, tienen diferencias fundamentales, SpiderP es especializado en toxinas de arácnidos, clase objetivo que sobrelapa con la de proAmps sin ser ninguna subconjunto de la otra. Además, la implementación de SpiderP no es en un "pipeline" automatizado y está restringida a la predicción de propéptidos en el extremo N-terminal

El poder predictivo de las herramientas disponibles para identificar AMPs incrementará en el futuro. El desarrollo de nuevas metodologías de ML, el incremento de la evidencia experimental sobre AMPs, los avances en la dilucidación de sus distintos mecanismos de acción, y la incorporación de metadatos bibliográficos a las bases de datos especializadas de AMPs, ayudarán a los modelos "state-of-art*"* del futuro a su identificación. Por lo tanto, es de esperar que las herramientas con mejor desempeño hoy, sean reemplazadas en el futuro.

1) Comportamiento de proAmps en transcriptomas sintéticos

Una de las potenciales aplicaciones de proAmps es en la predicción de AMPs en datos transcriptómicos, sin embargo, la naturaleza molecular del RNA y el proceso de secuenciación deriva en alta fragmentación de las secuencias. Para resolver este problema y poder determinar la capacidad de proAmps para identificar AMPs en este tipo de datos se realizó el siguiente experimento.

Utilizando el conjunto transcriptoma sintético de AMPs, fue examinado el valor tipo probabilidad de ser AMP en un espacio de dimensiones: Longitud de fragmento, fracción de AMP representada en el fragmento, y fracción de la región propeptídica representada en el fragmento. La probabilidad de ser AMP muestra una baja correlación (cor=-0.029, "test p-value" < 2.2e-16) con la longitud de los fragmentos con longitud en el intervalo 1-102. En las figuras 7 y 8 se observa la probabilidad promedio de ser AMP contra las variables fracción de AMP representada en el fragmento, y fracción de la región propeptídica representada en el fragmento. En estas figuras se observan los resultados obtenidos en ambas opciones analizadas, AMPs sin predicción de péptido maduro y predicción de AMPs con previa predicción de péptido maduro, respectivamente.

Los fragmentos que contiene mayor fracción del AMP maduro son aquellos que se clasifican como AMP con mayor probabilidad. En el caso de la opción que no realiza predicción previa del péptido maduro (Figura 7) es evidente que al incorporar información de secuencia que no está directamente relacionada con la actividad antimicrobiana, la probabilidad de ser clasificado como AMP disminuye, siendo las proteínas maduras la información óptima para que este método funcione con mayor confiabilidad. Es importante hacer notar que si el método realiza predicción de péptido maduro (Figura 8) puede identificar AMP en un espacio de mucho mayor dimensiones: fracción AMP, fracción propéptido, y con mayor confianza. Notaese la diferencia de la escala color en cada figura. En esta figura también es evidente que este método realiza una predicción óptima cuando en las secuencias existe entre el 20% y 40% de la región propeptídica. Esto puede deberse a que cuando se realiza el paso de predicción de péptido maduro, si el algoritmo es alimentado por una secuencia madura, existe cierta probabilidad de que el método realice predicción de sitios de corte

incorrectas. Sin embargo, debe notarse que este método no tiene un mal comportamiento cuando se alimenta con secuencias maduras. En la figura se muestra que incluso sin contar con ninguna fracción de propéptido, las fracciones se clasifican como positivas aún cuando existe una baja representación del péptido maduro.

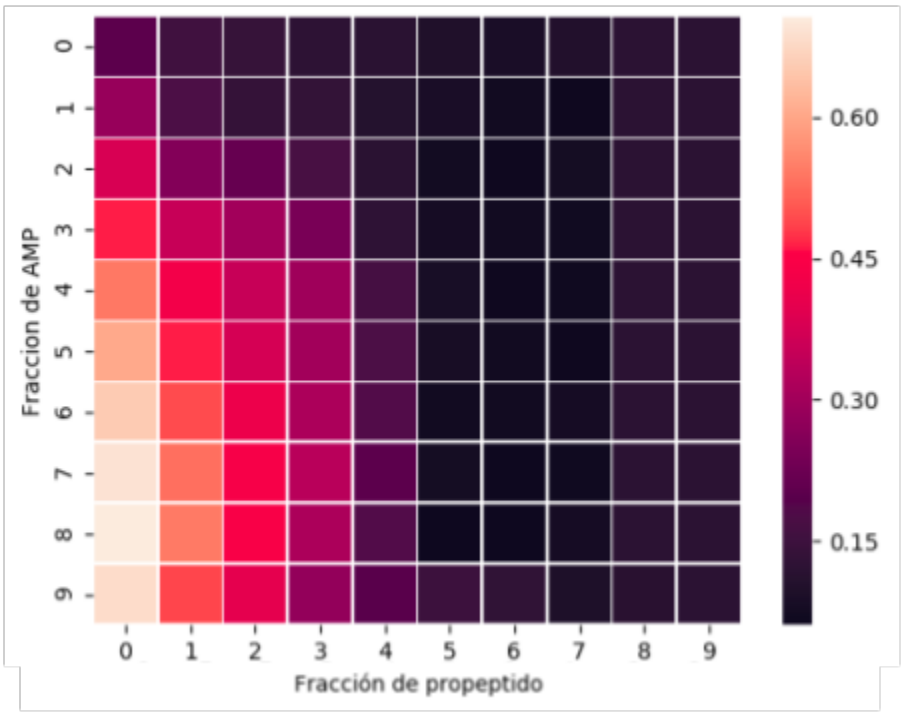


Figura 7. Mapa calorimétrico de la probabilidad promedio de ser AMP de las secuencias del transcriptoma sintético de AMPs utilizando **Predicción de AMP, sin previa identificación de proteína madura.** Los valores del eje vertical deben ser multiplicados por 0.1, y representan la fracción del AMP maduro contenida en la secuencia del transcriptoma. Los valores del eje horizontal deben ser multiplicados por 0.1, y representan la fracción de la región propeptídica contenida en la secuencia.

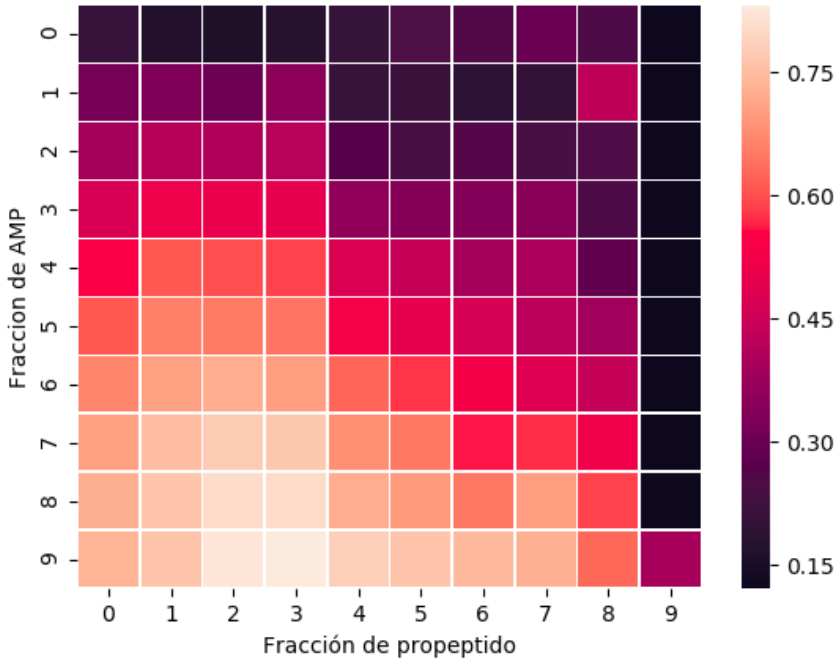


Figura 8. Mapa calorimétrico de la probabilidad promedio de ser AMP de las secuencias del transcriptoma sintético de AMPs utilizando **Predicción de AMP, con previa identificación de proteína madura.** Los valores del eje vertical deben ser multiplicados por 0.1, y representan la fracción del AMP maduro contenida en la secuencia del transcriptoma. Los valores del eje horizontal deben ser multiplicados por 0.1, y representan la fracción de la región propeptídica contenida en la secuencia.

De las secuencias clasificadas en el transcriptoma sintético de *Drosophila*, se encontró que de las 4,694,520 secuencias en este conjunto, 196,320 fueron clasificadas como AMPs, es decir el 4.18% de las secuencias. También es destacable que el 75.4% de los fragmentos que provienen de AMPs validados experimentalmente se identificaron con una probabilidad mayor a 0.5. Las variables de la longitud del fragmento y su posición respecto al N-terminal, no mostró ninguna correlación significativa con la probabilidad de ser clasificado como AMP. Este resultado nos indica que proAmps es capaz de trabajar con grandes volúmenes de datos, generando pocos FPs, así como mantener su capacidad predictiva en secuencias proteicas fragmentadas.

A partir del análisis de estos resultados podemos concluir que proAmps es un "software" de identificación de AMPs efectivo en secuencias de proteínas maduras, así como utilizando datos provenientes de NGS, siempre y cuando se elija la opción que existe para cada caso. Comparado con las demás herramientas evaluadas, proAmps destaca por su alta especificidad, además de tener un desempeño superior al utilizar datos de péptidos no maduros.

**Identificación *in silico* de potenciales AMPs en dos especies de distintos reinos**

A pesar de que proAmps es un método altamente específico, la probabilidad de obtener un verdadero AMP dado una proteína clasificada como tal se ve disminuida cuando la ocurrencia del fenómeno es baja. Por ello es que elegimos reportar las proteínas con una probabilidad superior a 0.8, y así evitar FPs.

*Drosophila melanogaster*

De las 30,493 proteínas derivadas del ensamble genómico de *Drosophila melanogaster dm,* 397 fueron identificadas como AMPs por proAmps. En este proteoma se identificaron 145 proteínas con una probabilidad mayor a 0.8 de ser AMPs, las cuales se encuentran en la Tabla 1 de la sección de Anexos. De estas secuencias, a 138 se les predijo región propeptídica. Entre estas secuencias se encuentran 6 de los 10 AMP de la base de datos APD que cumplen nuestros criterios de tamaño y tipo de aminoácidos.

Estás 145 proteínas fueron anotadas por InterPro, para detectar dominios funcionales. Posteriormente se analizó manualmente la anotación funcional, y se realizó una búsqueda bibliográfica sobre evidencia de estos dominios con función antimicrobiana. Este análisis muestra que 98 de estas proteínas no tiene ningún dominio funcional que coincida con los caracterizados en las distintas bases de datos de InterPro. Veintiocho de éstas tienen anotación distinta a la función de AMP, y 19 tienen una anotación directamente relacionada con función biocida. En una exploración más minuciosa de las proteínas anotadas con función distinta a la de actividad biocida, se buscó si existía reporte de alguna proteína con estos dominios que tuviera relación con la defensa de los organismos que la producen. En estos casos se comparó la proteína reportada con nuestro fragmento predicho para comprobar homología significativa de estos péptidos. Se encontró que 6 de los péptidos con anotación distinta a la de AMP cumplen con estas condiciones.

El gran número de proteínas predichas como AMPs que no cuentan con anotación representan una potencial fuente de exploración experimental para el descubrimiento de nuevos AMPs. Las proteínas que cuentan con dominios distintos a la actividad biocida deberán ser investigadas sobre posible multifuncionalidad, ya que como muestra nuestro análisis, algunas de ellas pueden estar relacionadas en funciones de defensa de los organismos. Casos ejemplares de este tipo de proteínas son NP_001286570.1 y NP_611322.1 (identificadores del ensamble dm), las cuales se anotan como Glutatión S-transferasas, proteínas que catalizan la conjugación de glutatión reducido a compuestos xenobióticos. A primera vista esta actividad podría parecer relacionada a la defensa contra agentes patógenos y no una relación directa con actividad biocida, sin embargo, un estudio reciente reporta la existencia de un cripto péptido con función de AMP en una proteína con esta actividad (Horam *et al.,* 2018). El caso de las dos proteínas identificadas como AMPs en este trabajo y anotadas con esta función deben ser consideradas seriamente para validarse experimentalmente. Otro caso interesante es la proteína NP_651220.3, la cual se anota como “elafin-like”, un inhibidor de proteasas. Para este tipo de proteínas existen reportes sobre su capacidad para inhibir la proliferación de *Pseudomonas aeruginosa* (Ballemare *et al.,*2008).

Las proteínas identificadas por nuestro método también fueron sometidas a calificación por las distintas herramientas comparadas en este trabajo. En la Tabla 3 se muestra el número de proteínas

identificadas por cada método como AMPs, y entre paréntesis se muestra el número de proteínas con anotación distinta a la antimicrobiana que son detectadas por estos métodos.

Tabla 3. Conteo de proteínas con probabilidad mayor a 0.8 de ser AMPs y la evidencia que arroja la anotación por InterPro

| proAMPs | S/A | O/A | AMP | defensa | ampscan | iamp | c-svc | c-rf | c-ann | c-da |
|---|---|---|---|---|---|---|---|---|---|---|
| 145 | 98 | 28 | 19 | 6 | 88(8) | 29(2) | 116(26) | 113(26) | 50(6) | 109(26) |

S/A, sin anotación. O/A, otra anotación. AMP, anotación de AMP. Defensa, péptidos con O/A que la literatura sugiere relación a la defensa antimicrobiana. En naranja se muestra el número de AMPs identificados por proAmps y que fueron así clasificados por otros métodos. En paréntesis el número de péptidos del conjunto O/A, que también son identificados por otros métodos.

Las proteínas predichas por nuestro método como AMPs y que tienen una anotación distinta a las de AMP, son predichas casi en su mayoría por los métodos de SVC y RF de la herramienta CAMP. Esto refuerza la idea de que estas proteínas deben ser analizadas para caracterizar su potencial función antimicrobiana.

*Phaseolus vulgaris*

De las 31,638 proteínas del ensamble genómico de *P. vulgaris*, 766 fueron identificadas como AMPs por proAmps. Tomando en cuenta las proteínas predichas como AMPs con una probabilidad superior a 0.8, se obtienen 181 AMP putativos (ver Tabla 2 en la sección de Anexos). Ciento ochenta y una de estas secuencias cuentan con predicción de región propeptídica. Las dos proteínas de *P. vulgaris* con evidencia experimental de actividad antimicrobiana (identificador APD 00563 y 01677) fueron detectadas como AMPs por proAmps con probabilidades superiores a 0.95.

Estas proteínas fueron analizadas con la herramienta InterPro para hacer una anotación funcional con base a sus dominios. Se encontró que a 105 de las proteínas no se les puede asignar una función. A 49 de ellas se les identifica dominios funcionales distintos a los de actividad antimicrobiana, y 18 de estas proteínas tienen dominios de AMPs. Se realizó una búsqueda bibliográfica para encontrar reportes sobre proteínas con los dominios no correspondientes a AMPs buscando relación con la defensa de los organismos que las producen. Se encontró que este es el caso para 19 de las 49 proteínas.

Un caso ejemplar de este tipo de proteínas es Phvul.003G269100.1 (identificador del ensamble), la cual se contiene un dominio de fosfoserina aminotransferasa. La búsqueda bibliográfica muestra una proteína con el mismo dominio, del género *Eucalyptus*, la cual se sobreexpresa al ser infectado por *Calonectria pseudoreteaudii* (Chen *et al.,* 2015). En un análisis más detallado, encontramos que la proteína identificada en *P. vulgaris* y la proveniente de eucalipto comparten un 70% de identidad con un "e-value" significativo. A pesar que el reporte no muestra una evidencia directa de esta proteína con actividad antimicrobiana, los análisis de expresión muestran que su producción aumenta como respuesta a una infección.

Otro caso ejemplar de estas proteínas es Phvul.009G030400.1, la cual se anota con un dominio Pumilio. Estos dominios se caracterizan por unirse con mRNA, y así reprimir la expresión de

proteínas específicas. Sin embargo, también existen reportes en *D. melanogaster* que estas proteínas están relacionadas con la respuesta inmune antimicrobiana (Sim *et al.,* 2017). En el análisis detallado de la proteína se encontró que comparte una identidad del 60% con la identificada para *P. vulgaris.*

Tabla 4. Conteo de proteínas con probabilidad mayor a 0.8 de ser AMPs y la evidencia que arroja la anotación por InterPro.

| proAMPs | S/A | O/A | AMP | defensa | ampscan | Iamp | c-svc | c-rf | c-ann | c-da |
|---|---|---|---|---|---|---|---|---|---|---|
| 172 | 105 | 49 | 18 | 19 | 101(28) | 10(4) | 148(43) | 144(41) | 63(18) | 136(41) |

S/A, sin anotación. O/A, otra anotación. AMP, anotación de AMP. Defensa, Péptidos con O/A que la literatura sugiere relación a la defensa antimicrobiana. En naranja se muestran el número de AMPs identificados por proAmps y que fueron así clasificados por otros métodos. En paréntesis el número de péptidos del conjunto O/A, que también son identificados por otros métodos.

En este conjunto de proteínas predichas como AMPs también se realizó la identificación de esta función, usando las mismas herramientas con las que se comparo nuestro metodo, proAMPs. En el caso de los distintos métodos de CAMP se observó, en *P. vulgaris* que la mayoría de las proteínas que se anotaron con una función distinta a las antimicrobiana también fueron predichas como AMPs por estos métodos (Tabla 4).

La comparativa entre los resultados de proAmps y la evidencia obtenida por InterPro, muestra las desventajas generales de los métodos de identificación clásicos. El primer problema es el sesgo asociado a la base de datos empleada y los métodos que se derivan de ellos, por ejemplo, Uniprot y su filiación a organismos modelos. Otro problema es que el universo de proteínas aún no está representado en su totalidad en estas bases de datos, por lo tanto es de esperar que muchas proteínas no cuenten con homólogos conocidos y sea difícil caracterizar patrones de los dominios funcionales. Un problema adicional es la falta de asociación de la información disponible en la literatura con los elementos proteicos de las bases de datos. La información contenida en la literatura podría ayudar enormemente a la identificación de AMPs (Danchin *et al.,* 2018), sin embargo, los esfuerzos para realizar este objetivo aún son aislados; además gran parte de las anotaciones disponibles son muy genéricas y no apuntan directamente a una función. Un ejemplo de esto es el caso de los dos AMPs de *P. vulgaris* validados experimentalmente (Wong *et al.,*2006; Games *et al.,* 2008), de los cuales ninguno se encuentra en Swissprot, la anotación automatizada con la que cuenta es sobre su contenido de cisteínas y no sobre su función biocida. Otro problema es que la pequeña longitud de los péptidos antimicrobianos pueda causar coincidencias en secuencia con proteínas mucho más grandes y a pesar de contar con una alta identidad, funcionalmente son proteínas muy distantes. Adicionalmente, se sabe que gran parte de los péptidos antimicrobianos actúan como moléculas multifuncionales lo que complica el análisis (Lai y Gallo, 2009). Todos estos problemas contribuyen en conjunto a una caracterización o búsqueda de AMPs por métodos bioinformáticos que aún esta lejos de explorar el universo real de proteínas y que es propenso a sobreestimar los FPs.

En este objetivo se reportaron 145 proteínas como AMPs proveniente de *D. melanogaster* y 181 de origen *P. vulgaris*, sin embargo, los resultados aquí expuestos deben ser aún validados experimentalmente.

**Análisis evolutivo de los AMPs, detectando posible convergencia o divergencia funcional**

Como resultado de agrupar el conjunto de maduros positivo por homología se determinaron 152 agrupaciones o ”clusters” que comparten similitud de secuencia de al menos 30%. En estas agrupaciones destacan 2 por su gran número de miembros: 87 y 43 secuencias. Ambos ”clusters” se componen exclusivamente de péptidos de la familia cyclotide de plantas. Otros 2 de los ”clusters” con 20 y 18 miembros, están compuestos de defensinas de plantas. De los ”clusters” compuestos de secuencias de artrópodos, resaltan dos compuestos por 15 y 13 secuencias, en los cuales están representados miembros de las familias de AMPs de artrópodos cecropina, y defensinas respectivamente. Las agrupaciones generadas por CD-HIT son consistentes con la matriz de identidad generada a partir del análisis de BLAST.

Para la clasificación de estos datos se construyo un PCA utilizando las variables fisicoquímicas con el fin de eliminar variables redundantes del objetivo uno. En la Figura 9 se observa que los primeros dos componentes principales (PC) explican alrededor del 60% de la varianza de los datos. Por ello se conservaron estas variables para los análisis posteriores.

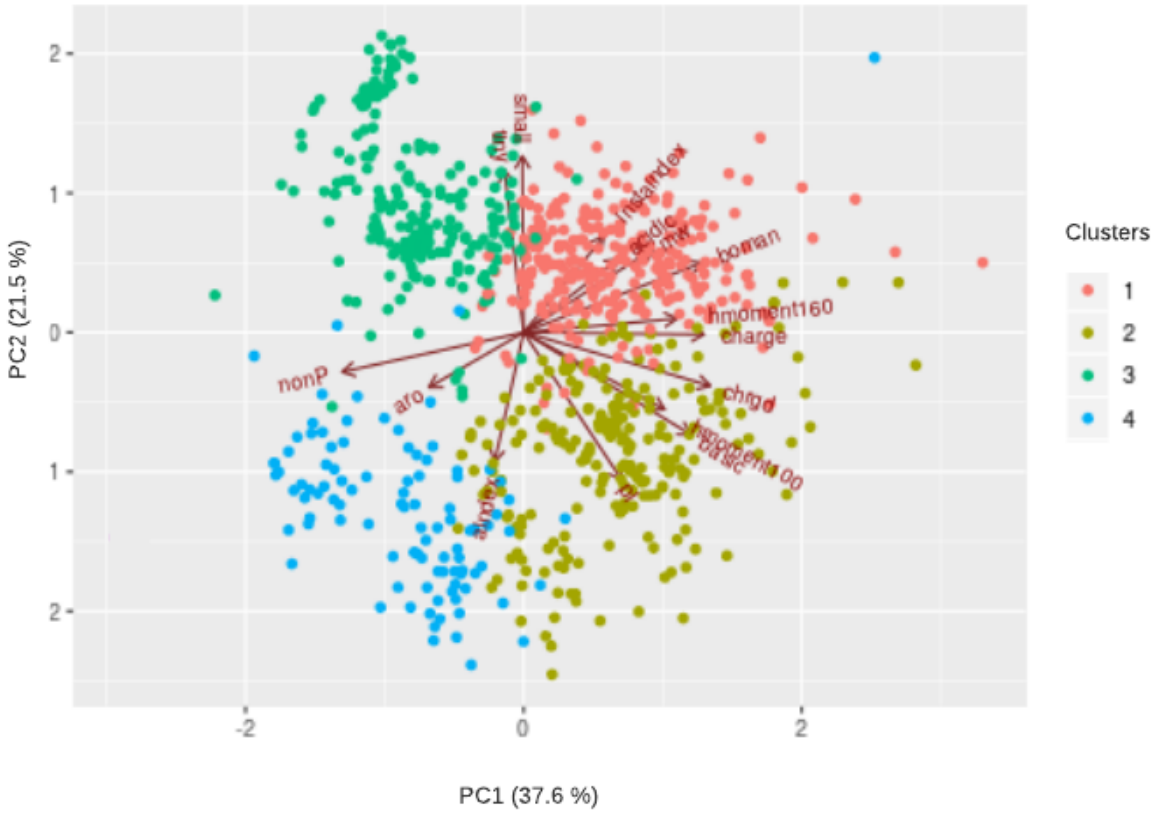


Figura 9. Gráfico de dispersión de las observaciones sobre los primeros dos componentes principales (PC) derivados de un PCA. Entre paréntesis se indica el porcentaje de la varianza explicada por cada componente principal. Las flechas indican los vectores de peso para las variables fisicoquímicas. mw, Peso molecular. charge, Carga neta. pI, Punto isoeléctrico. aindex, Índice alifático. instaindex, Índice de inestabilidad. Boman, Índice de Boman. hmoment100 y hmoment160, Momento hidrofóbico (100°, 160° rotación de hélice). Composición de grupos: “tiny”, Diminutos. “small”, Pequeños. aro, Aromáticos. ali, Alifáticos. nonP, No polares. chrgd, Cargados. basic, Basicos. acidic, Ácidos.

Derivado del proceso descrito en la metodología se obtuvieron 4 ”clusters” fisicoquímicos. En la Figura 10 se observa un mapa calorimétrico de la distancia entre las observaciones agrupadas por los ”clusters” fisicoquímicos determinados. En esta figura podemos observar que la distancia entre las observaciones intra-”cluster” es menor a la distancia entre puntos inter-”cluster”. Algunas observaciones de estas agrupaciones tienen distancia corta con observaciones de otros ”cluster”. A continuación se hace una descripción de cada uno de estos ”clusters”.

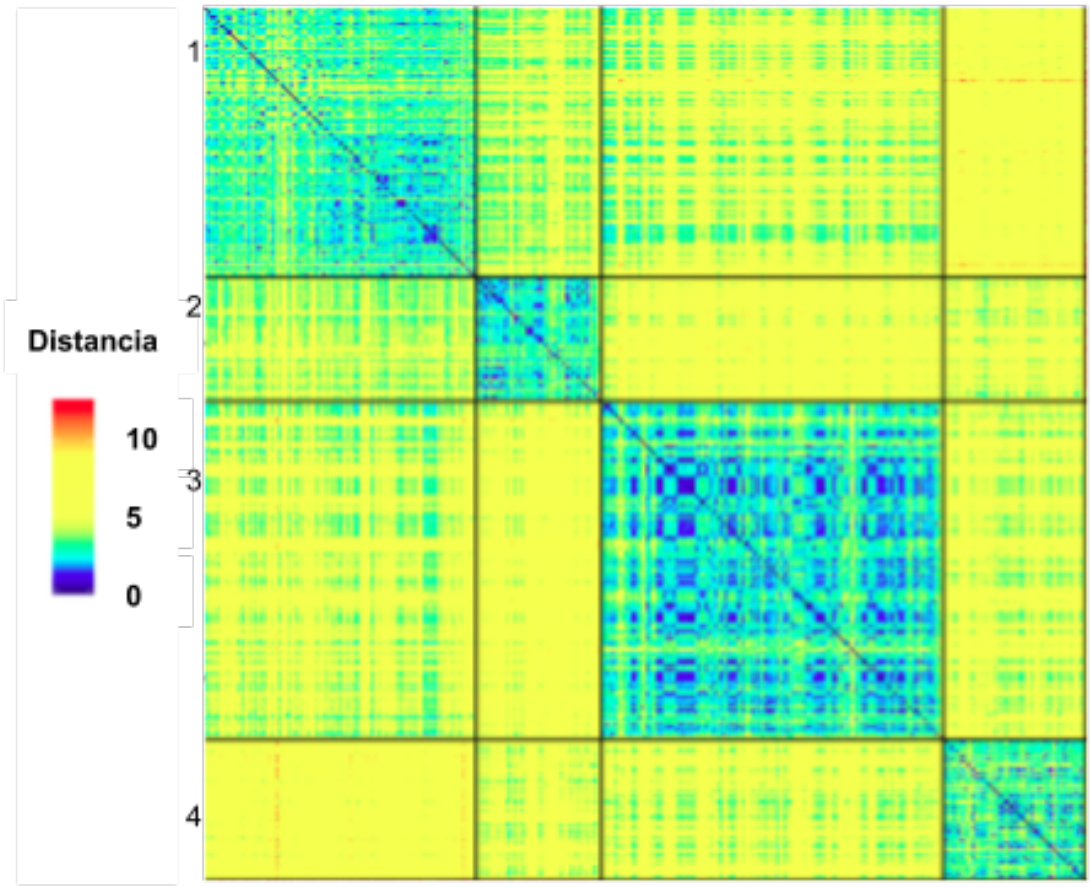


Figura 10. Mapa calorimétrico de la distancia entre observaciones del conjunto de análisis evolutivo agrupado por los ”clusters” fisicoquímicos. En el borde izquierdo se marca con número el ”cluster” al que corresponde cada cuadrante.

**”Cluster” 1**. Se compone por 160 AMPs de origen de planta (86) y artrópodos (74). De acuerdo con los ”clusters” de secuencia, 57 grupos se formaron con una similitud mínima de 30% dentro de esta agrupación. El ”cluster” fisicoquímico 1 contiene gran parte de AMP clasificados en APD como defensinas, tanto de plantas como de artrópodos.

Fisicoquímicamente este ”cluster” tiene un bajo índice alifático e índice de Boman alto, respecto de las demás agrupaciones. En la Tabla 3 de la sección de Anexos, se muestra las estadísticas principales de los ”clusters” fisicoquímicos. También es destacable que este grupo es el de mayor peso molecular y el más catiónico. Estas características fisicoquímicas corresponden a los mecanismos de acción reportados para un gran número de observaciones del conjunto: las defensinas. Se sabe que la permeabilización a través de las membranas bacterianas es un paso crucial en la función antimicrobiana de las defensinas (Ganz, 2003). Dicha permeabilización se logra por medio de fuerzas electrostáticas entre las cabezas de los fosfolípidos cargadas negativamente y los aminoácidos catiónicos de los AMPs.

Es importante notar que la agrupación por variables fisicoquímicas logró recuperar gran parte de las secuencias caracterizadas como defensinas en un mismo conjunto, contrastando con los

métodos de homología que las subdividen. Este grupo de defensinas comparte características fisicoquímicas además de tener homología a nivel de estructura primaria.

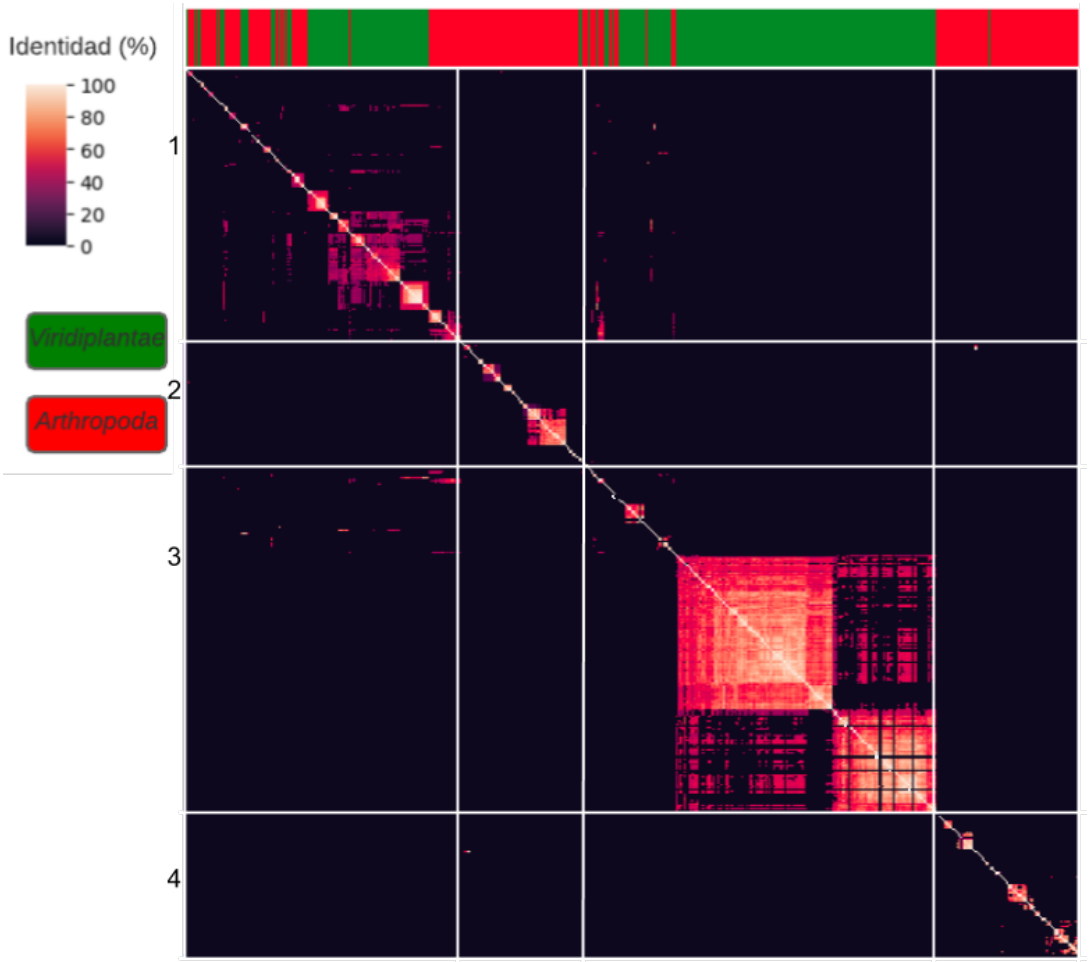


Figura 11. Mapa calorimétrico de la identidad promedio entre pares de AMPs divididos por ”clusters” fisicoquímicos. En el borde superior del mapa se muestra con código de color indicando el clado al que pertenece la observación. En el borde izquierdo se marca con número el ”cluster” al que corresponde cada cuadrante.

En la Figura 11 se observa un mapa calorimétrico de la identidad que comparten las secuencias del conjunto de análisis evolutivo. En la parte inferior derecha del cuadrante correspondiente al ”cluster” 1, se observa un conjunto de secuencias que tienen una identidad alta, estas secuencias son defensinas, pertenecientes a los dos clados estudiados. Esta observación apoya la conclusión derivada por Shafee *et al.* (2017), de que las defensinas de origen de plantas y artrópodos, llamadas *cis*-defensinas, tienen un mismo origen evolutivo, a diferencia de las defensinas de vertebrados, las cuales deben su similitud a convergencia de estructura terciaria. Sin embargo, aseverar esto resulta muy difícil, ya que con la gran variabilidad de secuencia, puentes disulfuro, estructura y función que presentan las defensinas, comprobar si son secuencias homólogas, análogas, o tienen un ancestro común requiere de un estudio de mayor profundidad.

En este mismo cuadrante (1,1) de la Figura 11, los AMPs que se observan en la parte superior derecha son de origen *arthropoda*. Sin embargo, no se conoce evidencia sólida de que estos AMPs esten relacionados por secuencia con el grupo de las defensinas. Esto parece indicar que estos

péptidos han convergido fisicoquímicamente con las defensinas. Sin embargo, la falta de estandarización de los ensayos de actividad antimicrobiana, y la falta de esclarecimiento de los mecanismos de acción de estos, dificulta poder asegurar una convergencia funcional.

En el cuadrante (1,3) de la Figura 11, observamos como algunos AMPs clasificados en el ”cluster” 3 tienen identidad significativa con AMPs del ”cluster” 1, en específico con defensinas. Esta observación resulta muy interesante ya que existe evidencia de que las defensinas tienen una gran diversidad de funciones, en la cuales el motivo estructural de hélices alfa-hoja beta estabilizados por puentes disulfuro (CSαβ) sirve como plantilla que otorga estabilidad estructural a estas proteínas, y adiciones a este motivo confieren diferencias funcionales (de Oliveira Diaz y Franco, 2015; van der Weerden y Anderson, 2019). Las defensinas que se agruparon en el ”cluster” 3 divergen fisicoquímicamente del grueso de las defensinas contenidas en el ”cluster” 1, conservando el “core” a nivel de secuencia.

**”Cluster” 2**. Este grupo está compuesto por 72 AMPs de origen de artrópodo y dos de origen de plantas. Realizando agrupaciones con CD-HIT este ”cluster” se pueden clasificar con 32 conjuntos que compartan al menos un 30% de similitud. Los AMPs contenidos en este ”cluster” no comparten una clasificación o anotación común en la base de datos APD, sin embargo observando la Figura 11, cuadrante (2,2) se observan claras agrupaciones por homología. En el grupo que comparte identidad, ubicado en la parte inferior derecha del cuadrante, se encuentran péptidos clasificados en la literatura como cecropinas.

Fisicoquímicamente este grupo contiene la mayor frecuencia de aminoácidos básicos, así como una carga positiva alta (9.8), cercana a la del ”cluster” 1. Otra característica de este conjunto de datos es que tiene el promedio de punto isoeléctrico más alto (promedio 11.5), esta característica señala que este grupo de AMPs funciona en un intervalo grande de pH, ya que se requeriría de un medio muy básico (~11.5) para que estos péptidos pierdan su cationicidad, y por lo tanto su capacidad de interactuar con la bicapa lipídica. Este punto es apoyado por el estudio de Lee *et al.,* (1997), en el cual se demuestra que la acción antimicrobiana de las cecropinas (AMPs agrupados en este ”cluster”) no son dependientes de pH. Como se observa en este “cluster” fisicoquímico, muchas de las secuencias no comparten homología, lo cual nos sugiere que estos AMPs han convergido en un espacio fisicoquímico que les permite tener actividad biocida en un amplio espectro de pH.

En este ”cluster” se encuentran dos AMPs descritos como *meucinas*, los cuales comparten una identidad significativa con un AMP perteneciente al ”cluster” fisicoquímico 4. Estos tres AMPs comparten un “core” común (FFGHLFKLATKIIPS), los AMPs del ”cluster” 2 tienen 7 aminoácidos extra en el extremo C-terminal, mientras que el AMP del ”cluster” 4 tiene 3 aminoácidos extra. A pesar de ser pequeñas las diferencias en secuencia entre estos AMPs, se puede observar que las diferencias fisicoquímicas son dramáticas puesto que son agrupados por separado. En un estudio de Gao *et al.* (2009) se demostró que las meucinas son genes ortólogos en diversas especies de vertebrados, y en los cuales existen diferencias dramáticas en la potencia antimicrobiana al remover pequeños fragmentos de estas proteínas. Las diferencias fisicoquímicas de estos AMPs, y su identidad a nivel de secuencia, nos sugiere una divergencia funcional de estos AMPs.

**”Cluster” 3.** Compuesto por 16 secuencias de origen artrópodo y 186 secuencias de origen de plantas, se pueden distinguir a hasta 36 ”clusters” que comparten por lo menos 30% de similitud. El mapa calorimétrico de la identidad de estos AMPs se encuentra en la Figura 11, cuadrante (3,3).

Fisicoquímicamente, este ”cluster” es el de menor frecuencia de aminoácidos básicos, así como con mayor frecuencia de aminoácidos pequeños.

En esta agrupación se encuentra gran parte de las secuencias clasificadas por la literatura como “cyclotides” lineales, circulares, y anudados de plantas. Cabe resaltar que existen AMPs de artrópodos con anotación de péptidos anudados que también se clasificaron fisicoquímicamente en este grupo. Es importante notar que esta agrupación fisicoquímica logra recuperar la mayoría de las secuencias “cyclotides” en el mismo conjunto, mientras que al agrupar por estructura primaria estas secuencias se dispersan en diversos grupos, mismo fenómeno que se observa al comparar la identidad entre pares de muchos de lo “cyclotides” (regiones "difusas" en el sub”cluster” ubicado en la parte inferior derecha del cuadrante (3,3) de la Figura 11). Estas observaciones nos sugieren que los “cyclotides” son AMPs que han convergido evolutivamente. Estas mismas conclusiones son derivadas del trabajo de Porto *et al.* (2016) en su estudio con “cyclotides” de la familia *Poaceae*. A diferencia de este trabajo, ese estudio se realizó exclusivamente con información sobre secuencia.

Esta agrupación también contiene AMP clasificados como defensinas de ambos clados, los cuales comparten identidad con AMPs clasificados en el ”cluster” 1 (*ex supra*).

**”Cluster” 4**. En este ”cluster” se encuentra solo un AMP de plantas y 83 AMPs de artrópodos. Dentro de él, se pueden generar 35 grupos con al menos un 30% de similitud de secuencia. Los miembros de esta agrupación muestran baja homología intra-”cluster”, solo en algunos subgrupos pequeños (Figura 11, cuadrante (4,4)) se observa identidad significativa. Sin embargo, al revisar la anotación de los péptidos homólogos, se observa que no cuentan con una clasificación común en APD.

Fisicoquímicamente este ”cluster” tiene un índice alifático significativamente mayor al del resto de los ”clusters”. También el índice de Boman es más bajo respecto a los otros ”clusters”. La magnitud de estas variables es correspondiente con la actividad descrita de algunos de sus miembros, los cuales son conocidos en su mayoría como mastoparan. Existen diversos reportes en la literatura de que estos AMPs actúan en membranas plasmáticas, así como en blancos de membranas intracelulares, permeabilizando la membrana mitocondrial o inhibiendo $K_{atp}$ en organelos lo cual deriva en apoptosis (Eddlestone *et al.*, 1995; Pfeiffer *et al.*, 1995; Moreno y Giralt, 2015). Para lograr este mecanismo de acción es necesario que estos péptidos tengan un alto índice alifático, lo cual les permite penetrar las bicapas lipídicas. También es necesario tener un bajo potencial de interacción proteína-proteína (medido con el índice de Boman), ya que al pasar por el citosol hacia sus objetivos intracelulares tienen que sortear un ambiente rico en proteínas.

Dentro de este ”cluster” se encuentra un AMP de plantas, el cual no tiene identidad a nivel de secuencia con los demás miembros del ”cluster”. Desafortunadamente no existen reportes del mecanismo de acción de este péptido, por lo tanto no se puede comparar directamente con aquellos reportados para los otros miembros del ”cluster”.

A pesar de que en la Figura 11 se observa que los miembros de este ”cluster” fisicoquímico comparten baja homología, es difícil sacar conclusiones sobre la convergencia funcional de estos AMPs, ya que al ser péptidos cortos (promedio 14.8 aminoácidos), pequeñas divergencias en la secuencia afectan en gran medida la identidad entre ellos.

En la Tabla 3 de Anexos se muestra la clasificación resultante de los dos enfoques de este objetivo, así como sus propiedades fisicoquímicas calculadas.

En este objetivo hemos identificado 4 ”clusters” en los cuales los AMPs de plantas y artrópodos contenidos en la base de datos APD, con longitud menor a 100 aminoácidos, se pueden agrupar. Por medio de observaciones en las propiedades fisicoquímicas y la homología de los AMPs hemos hipotetizado sobre la convergencia y divergencia de algunos AMPs, y en el caso de algunos ”clusters” hemos corroborado lo concluido en otros trabajos. Sin embargo, para poder hacer un mapa de la historia evolutiva de los AMPs más completo, se debe tener en cuenta los siguientes puntos:

- Integración de nuevos AMPs a las bases de datos a medida que sean descubiertos y caracterizados. Las relaciones evolutivas de los AMPs serán esclarecidas a medida que los sesgos de investigación en organismos modelo y de descubrimiento de nuevos AMPs redundantes en secuencia desaparezcan. El método proAmps es una herramienta que acelerará potencialmente este proceso. Eliminar estos sesgos también esclarecerá si existen clasificaciones discretas, como la aquí propuesta, entre los AMPs, o un *continuum* de actividad antimicrobiana, como el propuesto por Lee *et al.* (2018). Ese trabajo propone que las propiedades fisicoquímicas de los AMPs no están correlacionadas con un mecanismo específico de acción biocida, o con la potencia antimicrobiana, si no con la capacidad de inducir curvatura gaussiana negativa.
- Integración de metadatos de la literatura a las bases de datos. Contar con información fácilmente accesible y computable de datos que no estén estrictamente relacionados con la actividad antimicrobiana (*e.g.* condiciones de los ensayos, tipos de cepas utilizadas, etc.), ayudará enormemente a hacer una mejor clasificación de los AMPs. Así los mecanismos se podrán dilucidar más fácilmente, permitiendo discernir entre propiedades intrínsecas de los AMPs y el ruido de variables de confusión.

Clasificar correctamente los AMP de acuerdo a su mecanismo de acción permitirá hacer diseño racional de AMPs de forma más eficaz, mitigando efectos no deseados de los AMPs, y mejorando aquellos que confieren utilidad. Este objetivo pretende ayudar este tipo de procesos.

## Conclusiones

En este trabajo proponemos un método novedoso de identificación de AMPs a partir de datos procesados de NGS. El método de dos pasos (predicción de AMP maduro e identificación por métodos mixtos) optimiza la precisión con la que se detectan AMPs de plantas y artrópodos, ya sea utilizando proteomas derivados de ensambles genómicos o en traducciones de transcriptomas secuenciados.

Este método fue implementado en una herramienta, que nombramos proAmps, de fácil acceso, uso, e interpretabilidad de los resultados. Comparado con otros trabajos publicados anteriormente de detección de AMPs, nuestra herramienta es la de mejor desempeño cuando las secuencias a identificar no han sido procesadas postraduccionalmente. Se realizó una prueba del concepto generando sintéticamente un transcriptoma, la cual mostró la efectividad del método.

Utilizando proAmps se identificaron 145 AMPs putativos de *Drosophila melanogaster* y 181 de *Phaseolus vulgaris*. Seis proteínas de las proteínas identificadas como AMPs de *D. melanogaster* y 2 de *P. vulgaris* han demostrado tener actividad antimicrobiana en laboratorio, el resto deberán ser validadas por métodos experimentales.

Tras un análisis de agrupamiento, utilizando propiedades fisicoquímicas y homología de secuencia, proponemos una clasificación de 4 grupos en los que se pueden subdividir los AMPs de plantas y artrópodos. El análisis nos sugiere tanto convergencia como divergencia funcional de los AMPs de plantas y artrópodos. Nuestras conclusiones sobre la historia evolutiva de algunos AMPs son apoyadas por trabajos previos. Esta clasificación potencialmente ayudará al diseño racional de AMPs.

Nuestra herramienta desarrollada, proAmps, a pesar de ser una herramienta con margen de optimización, potencialmente asistirá al descubrimiento de nuevos AMPs.

## Referencias


Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. y Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation (16) (pp. 265-283).

Aggarwal, C. C., Hinneburg, A., y Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420-434). Springer, Berlin, Heidelberg.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Analytics, C. (2016). Anaconda software distribution. Computer software Vers, 2-2.

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S. y Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature biotechnology, 1.

Baskin, I. I., Winkler, D., y Tetko, I. V. (2016). A renaissance of neural networks in drug discovery. Expert opinion on drug discovery, 11(8), 785-795.

Bellemare, A., Vernoux, N., Morisset, D., y Bourbonnais, Y. (2008). Human pre-elafin inhibits a Pseudomonas aeruginosa-secreted peptidase and prevents its proliferation in complex media. Antimicrobial agents and chemotherapy, 52(2), 483-490.

Brogden, K. A. (2005). Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?. Nature reviews microbiology, 3(3), 238-250.

Brown, M. F. (2017). Soft matter in lipid–protein interactions. Annual review of biophysics, 46, 379-410.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., y Collins, J. J. (2018). Next-generation machine learning for biological networks. Cell, 173(7), 1581-1592.

Chang, K. Y., Lin, T. P., Shih, L. Y., y Wang, C. K. (2015). Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields. *PLoS One*, *10*(3), e0119490.

Chen, Q., Guo, W., Feng, L., Ye, X., Xie, W., Huang, X., y Liu, J. (2015). Transcriptome and proteome analysis of Eucalyptus infected with Calonectria pseudoreteaudii. Journal of proteomics, 115, 117-131.

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

Danchin, A., Ouzounis, C., Tokuyasu, T., y Zucker, J. D. (2018). No wisdom in the crowd: genome annotation in the era of big data–current status and future prospects. Microbial biotechnology, 11(4), 588-605.

de Oliveira Dias, R., y Franco, O. L. (2015). Cysteine-stabilized αβ defensins: from a common fold to antibacterial activity. Peptides, 72, 64-72.

Duckert, P., Brunak, S., y Blom, N. (2004). Prediction of proprotein convertase cleavage sites. Protein Engineering Design and Selection, 17(1), 107-112.

Eddlestone, G. T., Komatsu, M., Shen, L., y Sharp, G. W. (1995). Mastoparan increases the intracellular free calcium concentration in two insulin-secreting cell lines by inhibition of ATP-sensitive potassium channels. Molecular pharmacology, 47(4), 787-797.

Fan, L., Sun, J., Zhou, M., Zhou, J., Lao, X., Zheng, H., y Xu, H. (2016). DRAMP: a comprehensive data repository of antimicrobial peptides. Scientific reports, 6, 24482.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztányi, Z., El-Gebali, S., Fraser, M. y Gough, J. (2016). InterPro in 2017—beyond protein family and domain annotations. Nucleic acids research, 45(D1), D190-D199.

Fjell, C. D., Hancock, R. E., y Cherkasov, A. (2007). AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics, 23(9), 1148-1155.

Fu, L., Niu, B., Zhu, Z., Wu, S., y Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150-3152.

Games, P.D., dos Santos, I.S., Mello, É.O., Diz, M.S., Carvalho, A.O., de Souza-Filho, G.A., Da Cunha, M., Vasconcelos, I.M., Ferreira, B.D.S. y Gomes, V. M. (2008). Isolation, characterization and cloning of a cDNA encoding a new antifungal defensin from Phaseolus vulgaris L. seeds. Peptides, 29(12), 2090-2100.

Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. Nature reviews immunology, 3(9), 710.

Gao, B., Sherman, P., Luo, L., Bowie, J., y Zhu, S. (2009). Structural and functional characterization of two genetically related meucin peptides highlights evolutionary divergence and convergence in antimicrobial peptides. The FASEB Journal, 23(4), 1230-1245.

Gautier, L. (2008). rpy2: A Simple and Efficient Access to R from Python. URL http://rpy. sourceforge. net/rpy2. html.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. y Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic acids research, 40(D1), D1178-D1186.

Hammami, R., Ben Hamida, J., Vergoten, G., y Fliss, I. (2009). PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic acids research, 37(suppl_1), D963-D968.

Hammami, R., Zouhir, A., Hamida, J. B., y Fliss, I. (2007). BACTIBASE: a new web-accessible database for bacteriocin characterization. Bmc Microbiology, 7(1), 89.

Horam, S., Raj, S., Tripathi, V.C., Pant, G., Kalyan, M., Reddy, T.J., Arockiaraj, J. y Pasupuleti, M. (2018). Xenobiotic Binding Domain of Glutathione S-Transferase Has Cryptic Antimicrobial Peptides. International Journal of Peptide Research and Therapeutics, 1-13.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R. y Krzywinski, M. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome research*, *25*(3), 445-458.

Kaufman, L., Rousseeuw, P. J., y Dodge, Y. (1987). Clustering by Means of Medoids. Data Analysis based on the L1-Norm and Related Methods: 405-416.

Kuhn, M., y Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

Lai, S., Xu, L., Liu, K., y Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence.

Lai, Y., y Gallo, R. L. (2009). AMPed up immunity: how antimicrobial peptides have multiple roles in immune defense. Trends in immunology, 30(3), 131-141.

Lata, S., Mishra, N. K., y Raghava, G. P. (2010). AntiBP2: improved version of antibacterial peptide prediction. BMC bioinformatics, 11(1), 1-7.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. Nature, 521(7553), 436.

Lee, E. Y., Wong, G. C., y Ferguson, A. L. (2018). Machine learning-enabled discovery and design of membrane-active peptides. Bioorganic y medicinal chemistry, 26(10), 2708-2718.

Lee, I. H., Cho, Y., y Lehrer, R. I. (1997). Effects of pH and salinity on the antimicrobial properties of clavanins. Infection and immunity, 65(7), 2898-2903.

McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).

Moreno, M., y Giralt, E. (2015). Three valuable peptides from bee and wasp venoms for therapeutic and biotechnological use: melittin, apamin and mastoparan. Toxins, 7(4), 1126-1150.

Morton, J. T., Freed, S. D., Lee, S. W., y Friedberg, I. (2015). A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. BMC bioinformatics, 16(1), 1-9.

Nawrot, R., Barylski, J., Nowicki, G., Broniarczyk, J., Buchwald, W., y Goździcka-Józefiak, A. (2014). Plant antimicrobial peptides. Folia microbiologica, 59(3), 181-196.

Oliphant, T. E. (2006). A guide to NumPy (Vol. 1, p. 85). USA: Trelgol Publishing.

Osorio, D., Rondón-Villarrea, P., y Torres, R. (2015). Peptides: a package for data mining of antimicrobial peptides. R Journal, 7(1).

O'Sullivan, O., Begley, M., Ross, R. P., Cotter, P. D., y Hill, C. (2011). Further Identification of novel lantibiotic operons using LanM-based genome mining. Probiotics and Antimicrobial Proteins, 3(1), 27-40.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. y Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pfeiffer, D. R., Gudz, T. I., Novgorodov, S. A., y Erdahl, W. L. (1995). The peptide mastoparan is a potent facilitator of the mitochondrial permeability transition. Journal of Biological Chemistry, 270(9), 4923-4932.

Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H.L., Squires, R.B., Hurt, D.E., Grigolava, M., Chubinidze, M., Gogoladze, G., Vishnepolsky, B. y Alekseev, V. (2015). DBAASP v. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucleic acids research, 44(D1), D1104-D1112.

Porto, W. F., Miranda, V. J., Pinto, M. F., Dohms, S. M., y Franco, O. L. (2016). High-performance computational analysis and peptide screening from databases of cyclotides from poaceae. Peptide Science, 106(1), 109-118.

Powers, D. (2007). Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation (Tech. Rep.). Adelaide, Australia.

Rossum, G. V. (1995). Python tutorial, technical report CS-R9526. Centrum voor Wiskunde en Informatica (CWI), Amsterdam.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Schneider, T., Kruse, T., Wimmer, R., Wiedemann, I., Sass, V., Pag, U., Jansen, A., Nielsen, A.K., Mygind, P.H., Raventós, D.S. y Neve, S. (2010). Plectasin, a fungal defensin, targets the bacterial cell wall precursor Lipid II. Science, 328(5982), 1168-1172.

Schutte, B.C., Mitros, J.P., Bartlett, J.A., Walters, J.D., Jia, H.P., Welsh, M.J., Casavant, T.L. y McCray, P.B. (2002). Discovery of five conserved β-defensin gene clusters using a computational search strategy. Proceedings of the National Academy of Sciences, 99(4), 2129-2133.

Shafee, T. M., Lay, F. T., Hulett, M. D., y Anderson, M. A. (2016). The defensins consist of two independent, convergent protein superfamilies. Molecular biology and evolution, 33(9), 2345-2356.

Shafee, T. M., Lay, F. T., Phan, T. K., Anderson, M. A., y Hulett, M. D. (2017). Convergent evolution of defensin sequence, structure and function. Cellular and molecular life sciences, 74(4), 663-682.

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. y Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. Briefings in bioinformatics, 3(3), 265-274.

Sim, S., Wang, P., Beyer, B. N., Cutrona, K. J., Radhakrishnan, M. L., y Elmore, D. E. (2017). Investigating the nucleic acid interactions of histone-derived antimicrobial peptides. FEBS Letters, 591(5), 706-717.

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(suppl_1), D774-D780.

UniProt Consortium. (2015). UniProt: a hub for protein information. Nucleic acids research, 43(D1), D204-D212.

van der Weerden, N., y Anderson, M. A. (2019). U.S. Patent Application No. 10/174,339.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Veltri, D., Kamath, U., y Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. Bioinformatics, 34(16), 2740-2747.

Wang, G., Li, X., & Wang, Z. (2016). APD3: the antimicrobial peptide database as a tool for research and education. Nucleic acids research, 44(D1), D1087-D1093.

Wang, G. (Ed.). (2017). Antimicrobial peptides: discovery, design and novel therapeutic strategies. Cabi.

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

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Müller, R., Wohlleben, W. y Breitling, R. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic acids research, 43(W1), W237-W243.

Wong, E. S., Hardy, M. C., Wood, D., Bailey, T., y King, G. F. (2013). SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLoS One*, *8*(7), e66279.

Wong, J. H., Zhang, X. Q., Wang, H. X., y Ng, T. B. (2006). A mitogenic defensin from white cloud beans (*Phaseolus vulgaris*). Peptides, 27(9), 2075-2081.

Xiao, X., Wang, P., Lin, W. Z., Jia, J. H., y Chou, K. C. (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical biochemistry, 436(2), 168-177.

Yang, X., Lee, W. H., & Zhang, Y. (2012). Extremely abundant antimicrobial peptides existed in the skins of nine kinds of Chinese odorous frogs. Journal of proteome research, 11(1), 306-319.

Yount, N. Y., y Yeaman, M. R. (2004). Multidimensional signatures in antimicrobial peptides. Proceedings of the National Academy of Sciences, 101(19), 7363-7368.

Zelezetsky, I., Pag, U., Sahl, H. G., y Tossi, A. (2005). Tuning the biological properties of amphipathic α-helical antimicrobial peptides: rational use of minimal amino acid substitutions. Peptides, 26(12), 2368-2376.

Tabla 1. Proteínas de Drosophila melanogaster identificadas como AMPs con una probabilidad mayor a 0.8. **id**, identificador de la proteína. **prob**, probabilidad con la que se predijo cómo AMP por proAmps. **cluster**, subconjunto de AMPs que fue clasificada la proteína. **N-pos**, Posición de sitio de corte del extremo N-terminal. **N-prob**, Probabilidad de que la posición sea un sitio de corte N-terminal. **C-pos**, Posición de sitio de corte del extremo C-terminal. **C-prob**, Probabilidad de que la posición sea un sitio de corte C-terminal.

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 0 | NP_524588.1 | 0.998 | 2 | 23 | 1.0 | 0 | 0 |
| 1 | NP_524589.1 | 0.998 | 2 | 23 | 1.0 | 0 | 0 |
| 2 | NP_523672.1 | 0.995 | 1 | 52 | 0.968 | 0 | 0 |
| 3 | NP_523744.1 | 0.979 | 2 | 21 | 0.602 | 40 | 0.891 |
| 4 | NP_524591.1 | 0.979 | 2 | 23 | 1.0 | 0 | 0 |
| 5 | NP_001246324.1 | 0.979 | 2 | 21 | 0.602 | 40 | 0.891 |
| 6 | NP_524590.1 | 0.969 | 2 | 23 | 0.982 | 0 | 0 |
| 7 | NP_728860.2 | 0.964 | 1 | 26 | 0.907 | 0 | 0 |
| 8 | NP_608810.1 | 0.957 | 1 | 944 | 0.986 | 0 | 0 |
| 9 | NP_001097613.1 | 0.952 | 4 | 355 | 0.841 | 0 | 0 |
| 10 | NP_001163407.1 | 0.948 | 2 | 100 | 0.851 | 0 | 0 |
| 11 | NP_569851.2 | 0.946 | 2 | 379 | 0.713 | 0 | 0 |
| 12 | NP_523729.3 | 0.941 | 3 | 23 | 0.9 | 46 | 0.931 |
| 13 | NP_001163379.1 | 0.94 | 1 | 950 | 0.955 | 0 | 0 |
| 14 | NP_001259302.1 | 0.939 | 2 | 387 | 0.834 | 0 | 0 |
| 15 | NP_788873.1 | 0.939 | 2 | 387 | 0.834 | 0 | 0 |
| 16 | NP_001245991.1 | 0.931 | 2 | 311 | 0.986 | 0 | 0 |
| 17 | NP_602306.1 | 0.931 | 2 | 338 | 0.987 | 0 | 0 |
| 18 | NP_477465.1 | 0.931 | 2 | 338 | 0.987 | 0 | 0 |
| 19 | NP_651242.1 | 0.922 | 2 | 256 | 0.86 | 0 | 0 |
| 20 | NP_001285107.1 | 0.918 | 2 | 325 | 0.753 | 0 | 0 |
| 21 | NP_572681.1 | 0.918 | 2 | 325 | 0.753 | 0 | 0 |
| 22 | NP_651382.2 | 0.91 | 2 | 384 | 0.863 | 0 | 0 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 23 | NP_001287110.1 | 0.91 | 1 | 0 | 0 | 0 | 0 |
| 24 | NP_001246832.1 | 0.91 | 1 | 0 | 0 | 0 | 0 |
| 25 | NP_788859.1 | 0.899 | 2 | 690 | 0.814 | 0 | 0 |
| 26 | NP_001284904.1 | 0.899 | 2 | 809 | 0.857 | 0 | 0 |
| 27 | NP_001284905.1 | 0.899 | 2 | 690 | 0.814 | 0 | 0 |
| 28 | NP_788858.1 | 0.899 | 2 | 890 | 0.881 | 0 | 0 |
| 29 | NP_524587.1 | 0.898 | 2 | 23 | 0.904 | 0 | 0 |
| 30 | NP_001287605.1 | 0.898 | 2 | 23 | 0.904 | 0 | 0 |
| 31 | NP_001261385.1 | 0.89 | 2 | 757 | 0.955 | 0 | 0 |
| 32 | NP_647841.1 | 0.89 | 2 | 757 | 0.955 | 0 | 0 |
| 33 | NP_608791.2 | 0.889 | 2 | 766 | 0.882 | 0 | 0 |
| 34 | NP_788577.1 | 0.884 | 2 | 828 | 0.926 | 0 | 0 |
| 35 | NP_001247137.1 | 0.883 | 2 | 499 | 0.913 | 0 | 0 |
| 36 | NP_001247138.1 | 0.883 | 2 | 474 | 0.912 | 0 | 0 |
| 37 | NP_001163628.1 | 0.883 | 2 | 452 | 0.911 | 0 | 0 |
| 38 | NP_649651.2 | 0.882 | 2 | 458 | 0.725 | 0 | 0 |
| 39 | NP_651755.2 | 0.882 | 1 | 344 | 0.848 | 0 | 0 |
| 40 | NP_001263093.1 | 0.882 | 1 | 342 | 0.846 | 0 | 0 |
| 41 | NP_648679.1 | 0.881 | 2 | 312 | 0.782 | 0 | 0 |
| 42 | NP_652363.1 | 0.88 | 1 | 0 | 0 | 0 | 0 |
| 43 | NP_001014453.1 | 0.88 | 1 | 1072 | 0.896 | 0 | 0 |
| 44 | NP_001259850.1 | 0.88 | 1 | 1072 | 0.896 | 0 | 0 |
| 45 | NP_995981.1 | 0.878 | 1 | 348 | 0.701 | 0 | 0 |
| 46 | NP_647905.1 | 0.875 | 2 | 282 | 0.72 | 0 | 0 |
| 47 | NP_648149.1 | 0.871 | 3 | 185 | 0.884 | 202 | 0.961 |
| 48 | NP_524345.2 | 0.87 | 2 | 480 | 0.863 | 0 | 0 |
| 49 | NP_728833.1 | 0.87 | 2 | 211 | 0.723 | 0 | 0 |
| 50 | NP_726307.1 | 0.869 | 2 | 244 | 0.696 | 0 | 0 |
| 51 | NP_001285492.1 | 0.868 | 2 | 479 | 0.705 | 0 | 0 |
| 52 | NP_608393.1 | 0.868 | 2 | 479 | 0.705 | 0 | 0 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 53 | NP_723144.1 | 0.867 | 2 | 252 | 0.638 | 0 | 0 |
| 54 | NP_608981.1 | 0.867 | 2 | 252 | 0.638 | 0 | 0 |
| 55 | NP_652638.2 | 0.863 | 1 | 950 | 0.9 | 0 | 0 |
| 56 | NP_001246463.1 | 0.862 | 3 | 106 | 0.723 | 0 | 0 |
| 57 | NP_610313.1 | 0.862 | 2 | 108 | 0.694 | 0 | 0 |
| 58 | NP_610964.1 | 0.862 | 1 | 809 | 0.966 | 0 | 0 |
| 59 | NP_001286419.1 | 0.862 | 1 | 809 | 0.966 | 0 | 0 |
| 60 | NP_001246464.1 | 0.862 | 3 | 106 | 0.723 | 0 | 0 |
| 61 | NP_728862.1 | 0.861 | 1 | 26 | 0.636 | 0 | 0 |
| 62 | NP_523673.1 | 0.858 | 2 | 91 | 0.696 | 0 | 0 |
| 63 | NP_001188899.1 | 0.858 | 2 | 91 | 0.696 | 0 | 0 |
| 64 | NP_722889.1 | 0.855 | 2 | 601 | 0.922 | 0 | 0 |
| 65 | NP_001245853.1 | 0.855 | 2 | 600 | 0.921 | 0 | 0 |
| 66 | NP_722890.1 | 0.855 | 2 | 601 | 0.922 | 0 | 0 |
| 67 | NP_651811.1 | 0.855 | 2 | 391 | 0.822 | 0 | 0 |
| 68 | NP_001286356.1 | 0.855 | 2 | 0 | 0 | 0 | 0 |
| 69 | NP_524551.2 | 0.855 | 2 | 313 | 0.68 | 0 | 0 |
| 70 | NP_001286357.1 | 0.855 | 2 | 0 | 0 | 0 | 0 |
| 71 | NP_729846.3 | 0.854 | 2 | 921 | 0.951 | 0 | 0 |
| 72 | NP_523917.2 | 0.854 | 2 | 156 | 0.657 | 0 | 0 |
| 73 | NP_001261086.1 | 0.853 | 4 | 811 | 0.908 | 0 | 0 |
| 74 | NP_542440.1 | 0.845 | 2 | 891 | 0.977 | 0 | 0 |
| 75 | NP_611664.1 | 0.843 | 2 | 139 | 0.617 | 0 | 0 |
| 76 | NP_610369.1 | 0.842 | 1 | 79 | 0.671 | 0 | 0 |
| 77 | NP_524504.2 | 0.841 | 4 | 200 | 0.713 | 0 | 0 |
| 78 | NP_652658.1 | 0.841 | 2 | 91 | 0.695 | 0 | 0 |
| 79 | NP_723803.2 | 0.838 | 2 | 149 | 0.944 | 156 | 0.824 |
| 80 | NP_649165.1 | 0.837 | 1 | 348 | 0.603 | 0 | 0 |
| 81 | NP_573320.1 | 0.837 | 2 | 477 | 0.971 | 0 | 0 |
| 82 | NP_611452.2 | 0.837 | 2 | 373 | 0.925 | 0 | 0 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 83 | NP_610797.1 | 0.836 | 2 | 139 | 0.903 | 0 | 0 |
| 84 | NP_611322.1 | 0.836 | 2 | 205 | 0.784 | 0 | 0 |
| 85 | NP_726260.1 | 0.836 | 2 | 638 | 0.963 | 0 | 0 |
| 86 | NP_001286570.1 | 0.836 | 2 | 205 | 0.784 | 0 | 0 |
| 87 | NP_001246470.1 | 0.836 | 2 | 650 | 0.962 | 0 | 0 |
| 88 | NP_001188835.1 | 0.833 | 1 | 0 | 0 | 0 | 0 |
| 89 | NP_523665.2 | 0.832 | 2 | 143 | 0.763 | 0 | 0 |
| 90 | NP_650773.1 | 0.832 | 2 | 288 | 0.837 | 0 | 0 |
| 91 | NP_649503.1 | 0.832 | 2 | 72 | 0.671 | 0 | 0 |
| 92 | NP_648244.1 | 0.831 | 2 | 518 | 0.985 | 0 | 0 |
| 93 | NP_001188667.1 | 0.83 | 1 | 104 | 0.624 | 0 | 0 |
| 94 | NP_652623.2 | 0.83 | 2 | 485 | 0.757 | 0 | 0 |
| 95 | NP_573250.1 | 0.83 | 1 | 114 | 0.642 | 0 | 0 |
| 96 | NP_651220.3 | 0.83 | 2 | 500 | 0.824 | 0 | 0 |
| 97 | NP_611154.1 | 0.828 | 4 | 171 | 0.657 | 0 | 0 |
| 98 | NP_001246001.1 | 0.827 | 2 | 320 | 0.608 | 0 | 0 |
| 99 | NP_610942.2 | 0.825 | 2 | 815 | 0.972 | 0 | 0 |
| 100 | NP_725371.1 | 0.825 | 2 | 815 | 0.972 | 0 | 0 |
| 101 | NP_725370.1 | 0.825 | 2 | 815 | 0.972 | 0 | 0 |
| 102 | NP_477035.1 | 0.822 | 2 | 573 | 0.977 | 0 | 0 |
| 103 | NP_001260167.1 | 0.822 | 2 | 573 | 0.977 | 0 | 0 |
| 104 | NP_731887.1 | 0.821 | 2 | 815 | 0.914 | 0 | 0 |
| 105 | NP_649357.1 | 0.819 | 1 | 158 | 0.926 | 0 | 0 |
| 106 | NP_609284.3 | 0.818 | 4 | 827 | 0.978 | 0 | 0 |
| 107 | NP_001245834.1 | 0.818 | 1 | 0 | 0 | 0 | 0 |
| 108 | NP_001259862.1 | 0.818 | 1 | 0 | 0 | 0 | 0 |
| 109 | NP_001097131.2 | 0.818 | 4 | 929 | 0.976 | 0 | 0 |
| 110 | NP_001246844.1 | 0.814 | 1 | 370 | 0.959 | 0 | 0 |
| 111 | NP_649213.2 | 0.814 | 1 | 351 | 0.954 | 0 | 0 |
| 112 | NP_572301.2 | 0.814 | 2 | 301 | 0.743 | 0 | 0 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 113 | NP_001096890.1 | 0.814 | 2 | 281 | 0.731 | 0 | 0 |
| 114 | NP_001247004.1 | 0.814 | 1 | 187 | 0.696 | 0 | 0 |
| 115 | NP_525055.1 | 0.811 | 1 | 321 | 0.815 | 0 | 0 |
| 116 | NP_608817.1 | 0.81 | 2 | 533 | 0.913 | 0 | 0 |
| 117 | NP_723135.1 | 0.81 | 1 | 172 | 0.621 | 0 | 0 |
| 118 | NP_477133.1 | 0.808 | 2 | 938 | 0.839 | 0 | 0 |
| 119 | NP_996409.1 | 0.807 | 4 | 438 | 0.839 | 0 | 0 |
| 120 | NP_001036425.1 | 0.807 | 2 | 346 | 0.876 | 0 | 0 |
| 121 | NP_001036426.1 | 0.807 | 2 | 378 | 0.896 | 0 | 0 |
| 122 | NP_001036424.1 | 0.807 | 2 | 338 | 0.869 | 0 | 0 |
| 123 | NP_523663.1 | 0.806 | 4 | 547 | 0.9 | 0 | 0 |
| 124 | NP_724723.1 | 0.806 | 4 | 547 | 0.9 | 0 | 0 |
| 125 | NP_724724.1 | 0.806 | 4 | 547 | 0.9 | 0 | 0 |
| 126 | NP_001262518.1 | 0.805 | 2 | 33 | 0.602 | 0 | 0 |
| 127 | NP_001189083.1 | 0.805 | 2 | 1216 | 0.835 | 0 | 0 |
| 128 | NP_001246715.1 | 0.805 | 2 | 1216 | 0.835 | 0 | 0 |
| 129 | NP_001246716.1 | 0.805 | 2 | 1216 | 0.835 | 0 | 0 |
| 130 | NP_729690.1 | 0.805 | 2 | 1216 | 0.835 | 0 | 0 |
| 131 | NP_001287028.1 | 0.805 | 2 | 1217 | 0.834 | 0 | 0 |
| 132 | NP_001287308.1 | 0.805 | 2 | 33 | 0.602 | 0 | 0 |
| 133 | NP_611198.1 | 0.805 | 2 | 174 | 0.753 | 0 | 0 |
| 134 | NP_648454.1 | 0.805 | 2 | 1216 | 0.835 | 0 | 0 |
| 135 | NP_609749.2 | 0.804 | 1 | 194 | 0.675 | 0 | 0 |
| 136 | NP_001286196.1 | 0.802 | 2 | 516 | 0.924 | 0 | 0 |
| 137 | NP_001262975.1 | 0.802 | 1 | 226 | 0.889 | 0 | 0 |
| 138 | NP_651427.2 | 0.802 | 2 | 256 | 0.891 | 0 | 0 |
| 139 | NP_477117.2 | 0.802 | 2 | 516 | 0.924 | 0 | 0 |
| 140 | NP_788539.1 | 0.802 | 1 | 85 | 0.765 | 0 | 0 |
| 141 | NP_001189130.1 | 0.802 | 1 | 85 | 0.765 | 0 | 0 |
| 142 | NP_001163735.1 | 0.802 | 1 | 224 | 0.889 | 0 | 0 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 143 | NP_651426.1 | 0.802 | 1 | 249 | 0.891 | 0 | 0 |
| 144 | NP_001261789.2 | 0.801 | 2 | 921 | 0.951 | 959 | 0.63 |
| 145 | NP_651439.2 | 0.801 | 1 | 391 | 0.68 | 0 | 0 |

Tabla 2. Proteínas de Phaseolus vulgaris identificadas como AMPs con una probabilidad mayor a 0.8. **id**, identificador de la proteína. **prob**, probabilidad con la que se predijo cómo AMP por proAmps. **cluster**, subconjunto de AMPs que fue clasificada la proteína. **N-pos**, Posición de sitio de corte del extremo N-terminal. **N-prob**, Probabilidad de que la posición sea un sitio de corte N-terminal. **C-pos**, Posición de sitio de corte del extremo C-terminal. **C-prob**, Probabilidad de que la posición sea un sitio de corte C-terminal.

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 0 | Phvul.009G158000.1 | 0.997 | 1 | 26 | 0.957 | 0 | 0.000 |
| 1 | Phvul.002G278400.1 | 0.996 | 1 | 27 | 0.930 | 0 | 0.000 |
| 2 | Phvul.004G100700.1 | 0.971 | 2 | 476 | 0.953 | 0 | 0.000 |
| 3 | Phvul.003G064000.1 | 0.971 | 2 | 198 | 0.630 | 0 | 0.000 |
| 4 | Phvul.001G225900.1 | 0.967 | 2 | 76 | 0.756 | 0 | 0.000 |
| 5 | Phvul.005G071300.1 | 0.964 | 1 | 28 | 0.976 | 0 | 0.000 |
| 6 | Phvul.002G278600.1 | 0.962 | 1 | 28 | 0.959 | 0 | 0.000 |
| 7 | Phvul.004G136200.1 | 0.962 | 4 | 455 | 0.795 | 0 | 0.000 |
| 8 | Phvul.007G171400.1 | 0.962 | 4 | 308 | 0.633 | 0 | 0.000 |
| 9 | Phvul.005G151300.1 | 0.959 | 2 | 510 | 0.958 | 0 | 0.000 |
| 10 | Phvul.009G176100.1 | 0.957 | 2 | 313 | 0.639 | 0 | 0.000 |
| 11 | Phvul.003G282500.1 | 0.954 | 1 | 31 | 0.910 | 0 | 0.000 |
| 12 | Phvul.005G071400.1 | 0.953 | 1 | 28 | 0.976 | 0 | 0.000 |
| 13 | Phvul.006G172700.1 | 0.952 | 2 | 666 | 0.826 | 0 | 0.000 |
| 14 | Phvul.009G008200.1 | 0.951 | 2 | 311 | 0.708 | 0 | 0.000 |
| 15 | Phvul.002G313200.1 | 0.950 | 2 | 374 | 0.926 | 0 | 0.000 |
| 16 | Phvul.008G102800.1 | 0.948 | 2 | 80 | 0.661 | 0 | 0.000 |
| 17 | Phvul.008G120600.1 | 0.941 | 1 | 367 | 0.891 | 0 | 0.000 |
| 18 | Phvul.003G282400.1 | 0.934 | 1 | 31 | 0.803 | 0 | 0.000 |
| 19 | Phvul.003G202200.1 | 0.934 | 2 | 315 | 0.771 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 20 | Phvul.006G104600.1 | 0.931 | 2 | 209 | 0.645 | 0 | 0.000 |
| 21 | Phvul.011G056200.1 | 0.927 | 2 | 497 | 0.916 | 0 | 0.000 |
| 22 | Phvul.003G211800.1 | 0.927 | 1 | 407 | 0.989 | 0 | 0.000 |
| 23 | Phvul.005G026800.1 | 0.926 | 2 | 26 | 0.841 | 0 | 0.000 |
| 24 | Phvul.010G026800.1 | 0.924 | 4 | 234 | 0.756 | 0 | 0.000 |
| 25 | Phvul.003G133100.1 | 0.923 | 1 | 193 | 0.695 | 0 | 0.000 |
| 26 | Phvul.001G023800.1 | 0.922 | 1 | 513 | 0.741 | 0 | 0.000 |
| 27 | Phvul.002G064500.1 | 0.920 | 1 | 318 | 0.809 | 0 | 0.000 |
| 28 | Phvul.001G121000.1 | 0.919 | 2 | 363 | 0.758 | 0 | 0.000 |
| 29 | Phvul.008G058200.1 | 0.918 | 2 | 504 | 0.624 | 0 | 0.000 |
| 30 | Phvul.008G184300.1 | 0.916 | 2 | 716 | 0.905 | 0 | 0.000 |
| 31 | Phvul.005G151400.1 | 0.913 | 2 | 459 | 0.950 | 0 | 0.000 |
| 32 | Phvul.001G255700.1 | 0.913 | 1 | 654 | 0.719 | 0 | 0.000 |
| 33 | Phvul.011G062200.1 | 0.912 | 2 | 1016 | 0.896 | 0 | 0.000 |
| 34 | Phvul.008G043600.1 | 0.905 | 1 | 396 | 0.893 | 0 | 0.000 |
| 35 | Phvul.009G183000.1 | 0.905 | 2 | 575 | 0.614 | 0 | 0.000 |
| 36 | Phvul.011G172200.1 | 0.902 | 2 | 283 | 0.820 | 0 | 0.000 |
| 37 | Phvul.L011400.1 | 0.901 | 1 | 132 | 0.604 | 0 | 0.000 |
| 38 | Phvul.005G082100.1 | 0.897 | 1 | 27 | 0.783 | 0 | 0.000 |
| 39 | Phvul.002G278700.1 | 0.896 | 1 | 27 | 0.870 | 0 | 0.000 |
| 40 | Phvul.009G080400.1 | 0.894 | 1 | 212 | 0.720 | 0 | 0.000 |
| 41 | Phvul.011G056100.1 | 0.893 | 2 | 510 | 0.958 | 0 | 0.000 |
| 42 | Phvul.002G187700.1 | 0.893 | 2 | 475 | 0.919 | 0 | 0.000 |
| 43 | Phvul.003G025600.1 | 0.893 | 1 | 551 | 0.910 | 0 | 0.000 |
| 44 | Phvul.002G232000.1 | 0.892 | 2 | 115 | 0.810 | 0 | 0.000 |
| 45 | Phvul.004G125800.1 | 0.891 | 1 | 76 | 0.629 | 0 | 0.000 |
| 46 | Phvul.003G295500.1 | 0.889 | 2 | 381 | 0.831 | 0 | 0.000 |
| 47 | Phvul.008G093200.1 | 0.889 | 2 | 1053 | 0.894 | 0 | 0.000 |
| 48 | Phvul.008G093200.3 | 0.889 | 2 | 1013 | 0.901 | 0 | 0.000 |
| 49 | Phvul.008G093200.2 | 0.889 | 2 | 1013 | 0.901 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 50 | Phvul.009G051600.1 | 0.888 | 2 | 237 | 0.745 | 0 | 0.000 |
| 51 | Phvul.011G137000.1 | 0.885 | 2 | 70 | 0.782 | 0 | 0.000 |
| 52 | Phvul.003G011600.1 | 0.883 | 1 | 389 | 0.738 | 0 | 0.000 |
| 53 | Phvul.008G173200.1 | 0.883 | 2 | 363 | 0.950 | 0 | 0.000 |
| 54 | Phvul.003G015000.1 | 0.882 | 2 | 398 | 0.785 | 0 | 0.000 |
| 55 | Phvul.004G015400.1 | 0.882 | 1 | 311 | 0.804 | 0 | 0.000 |
| 56 | Phvul.007G041500.1 | 0.881 | 1 | 71 | 0.748 | 0 | 0.000 |
| 57 | Phvul.001G256800.1 | 0.878 | 1 | 356 | 0.870 | 0 | 0.000 |
| 58 | Phvul.008G115400.1 | 0.875 | 2 | 90 | 0.612 | 0 | 0.000 |
| 59 | Phvul.004G055100.1 | 0.875 | 2 | 192 | 0.660 | 0 | 0.000 |
| 60 | Phvul.008G174200.1 | 0.874 | 1 | 184 | 0.785 | 0 | 0.000 |
| 61 | Phvul.003G003400.1 | 0.874 | 2 | 482 | 0.967 | 0 | 0.000 |
| 62 | Phvul.009G106600.1 | 0.873 | 2 | 443 | 0.746 | 0 | 0.000 |
| 63 | Phvul.010G059800.1 | 0.872 | 2 | 369 | 0.761 | 0 | 0.000 |
| 64 | Phvul.008G081300.1 | 0.871 | 1 | 451 | 0.967 | 0 | 0.000 |
| 65 | Phvul.008G081300.2 | 0.871 | 1 | 451 | 0.967 | 0 | 0.000 |
| 66 | Phvul.008G000800.1 | 0.868 | 3 | 346 | 0.765 | 0 | 0.000 |
| 67 | Phvul.001G250600.1 | 0.868 | 2 | 450 | 0.724 | 0 | 0.000 |
| 68 | Phvul.001G243500.1 | 0.865 | 2 | 296 | 0.977 | 0 | 0.000 |
| 69 | Phvul.008G161100.1 | 0.864 | 1 | 825 | 0.906 | 0 | 0.000 |
| 70 | Phvul.008G161100.2 | 0.864 | 1 | 825 | 0.906 | 0 | 0.000 |
| 71 | Phvul.002G279900.1 | 0.864 | 4 | 104 | 0.659 | 0 | 0.000 |
| 72 | Phvul.002G084800.1 | 0.863 | 2 | 218 | 0.656 | 0 | 0.000 |
| 73 | Phvul.008G077000.1 | 0.863 | 1 | 795 | 0.985 | 0 | 0.000 |
| 74 | Phvul.010G137300.1 | 0.861 | 1 | 460 | 0.948 | 0 | 0.000 |
| 75 | Phvul.001G219700.1 | 0.861 | 1 | 117 | 0.754 | 158 | 0.602 |
| 76 | Phvul.003G034100.1 | 0.861 | 4 | 1014 | 0.984 | 0 | 0.000 |
| 77 | Phvul.009G030400.1 | 0.861 | 2 | 663 | 0.833 | 0 | 0.000 |
| 78 | Phvul.002G113600.1 | 0.857 | 2 | 268 | 0.720 | 0 | 0.000 |
| 79 | Phvul.009G113700.3 | 0.857 | 1 | 543 | 0.934 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 80 | Phvul.009G113700.2 | 0.857 | 1 | 543 | 0.934 | 0 | 0.000 |
| 81 | Phvul.011G055900.1 | 0.857 | 2 | 498 | 0.878 | 0 | 0.000 |
| 82 | Phvul.003G280400.1 | 0.855 | 1 | 1012 | 0.952 | 0 | 0.000 |
| 83 | Phvul.007G002700.1 | 0.854 | 1 | 258 | 0.793 | 0 | 0.000 |
| 84 | Phvul.005G064200.1 | 0.852 | 4 | 417 | 0.624 | 0 | 0.000 |
| 85 | Phvul.002G146900.1 | 0.850 | 1 | 1016 | 0.976 | 0 | 0.000 |
| 86 | Phvul.011G154300.1 | 0.849 | 4 | 258 | 0.644 | 0 | 0.000 |
| 87 | Phvul.008G252100.1 | 0.848 | 2 | 610 | 0.994 | 0 | 0.000 |
| 88 | Phvul.004G087800.1 | 0.847 | 1 | 79 | 0.908 | 0 | 0.000 |
| 89 | Phvul.011G014400.1 | 0.846 | 2 | 590 | 0.968 | 0 | 0.000 |
| 90 | Phvul.002G131600.1 | 0.846 | 2 | 193 | 0.799 | 0 | 0.000 |
| 91 | Phvul.002G084700.1 | 0.844 | 2 | 565 | 0.845 | 0 | 0.000 |
| 92 | Phvul.011G145700.1 | 0.843 | 1 | 116 | 0.611 | 0 | 0.000 |
| 93 | Phvul.002G113700.1 | 0.843 | 2 | 268 | 0.699 | 0 | 0.000 |
| 94 | Phvul.007G173200.1 | 0.843 | 1 | 90 | 0.782 | 0 | 0.000 |
| 95 | Phvul.009G143700.1 | 0.842 | 1 | 884 | 0.917 | 0 | 0.000 |
| 96 | Phvul.003G064200.1 | 0.841 | 4 | 749 | 0.982 | 0 | 0.000 |
| 97 | Phvul.002G072200.1 | 0.838 | 1 | 304 | 0.612 | 0 | 0.000 |
| 98 | Phvul.008G100900.1 | 0.838 | 4 | 499 | 0.767 | 0 | 0.000 |
| 99 | Phvul.002G328000.1 | 0.838 | 2 | 656 | 0.891 | 0 | 0.000 |
| 100 | Phvul.011G081400.1 | 0.837 | 2 | 47 | 0.665 | 0 | 0.000 |
| 101 | Phvul.003G151600.1 | 0.836 | 2 | 538 | 0.945 | 0 | 0.000 |
| 102 | Phvul.009G133400.1 | 0.836 | 4 | 214 | 0.986 | 0 | 0.000 |
| 103 | Phvul.009G043300.1 | 0.836 | 1 | 37 | 0.650 | 0 | 0.000 |
| 104 | Phvul.007G130300.1 | 0.835 | 2 | 985 | 0.977 | 0 | 0.000 |
| 105 | Phvul.001G256900.1 | 0.834 | 1 | 405 | 0.887 | 0 | 0.000 |
| 106 | Phvul.010G106600.1 | 0.834 | 1 | 377 | 0.733 | 0 | 0.000 |
| 107 | Phvul.011G056000.1 | 0.834 | 2 | 495 | 0.902 | 0 | 0.000 |
| 108 | Phvul.003G115400.1 | 0.833 | 2 | 296 | 0.978 | 0 | 0.000 |
| 109 | Phvul.008G043200.1 | 0.833 | 2 | 234 | 0.825 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 110 | Phvul.001G242200.1 | 0.832 | 2 | 451 | 0.715 | 0 | 0.000 |
| 111 | Phvul.003G036100.1 | 0.831 | 1 | 287 | 0.667 | 0 | 0.000 |
| 112 | Phvul.003G036100.2 | 0.831 | 1 | 213 | 0.702 | 0 | 0.000 |
| 113 | Phvul.006G077600.1 | 0.830 | 2 | 113 | 0.658 | 0 | 0.000 |
| 114 | Phvul.007G248900.1 | 0.830 | 1 | 80 | 0.771 | 0 | 0.000 |
| 115 | Phvul.010G000600.1 | 0.830 | 4 | 329 | 0.710 | 0 | 0.000 |
| 116 | Phvul.003G022800.1 | 0.829 | 2 | 395 | 0.715 | 0 | 0.000 |
| 117 | Phvul.008G085400.1 | 0.828 | 1 | 295 | 0.858 | 0 | 0.000 |
| 118 | Phvul.009G210700.1 | 0.828 | 2 | 267 | 0.670 | 0 | 0.000 |
| 119 | Phvul.005G067800.1 | 0.827 | 2 | 418 | 0.877 | 0 | 0.000 |
| 120 | Phvul.001G208400.1 | 0.827 | 2 | 267 | 0.909 | 0 | 0.000 |
| 121 | Phvul.003G235300.1 | 0.825 | 4 | 301 | 0.903 | 0 | 0.000 |
| 122 | Phvul.008G224600.1 | 0.825 | 2 | 471 | 0.707 | 0 | 0.000 |
| 123 | Phvul.010G021200.1 | 0.824 | 1 | 195 | 0.773 | 0 | 0.000 |
| 124 | Phvul.002G259000.1 | 0.824 | 2 | 410 | 0.949 | 0 | 0.000 |
| 125 | Phvul.001G008000.1 | 0.823 | 1 | 1020 | 0.890 | 0 | 0.000 |
| 126 | Phvul.011G126700.1 | 0.822 | 1 | 297 | 0.652 | 0 | 0.000 |
| 127 | Phvul.010G163200.3 | 0.822 | 1 | 459 | 0.903 | 0 | 0.000 |
| 128 | Phvul.002G249700.1 | 0.820 | 1 | 415 | 0.722 | 0 | 0.000 |
| 129 | Phvul.007G052700.1 | 0.820 | 1 | 65 | 0.839 | 0 | 0.000 |
| 130 | Phvul.003G244400.1 | 0.820 | 1 | 451 | 0.906 | 0 | 0.000 |
| 131 | Phvul.007G271400.1 | 0.819 | 2 | 980 | 0.977 | 0 | 0.000 |
| 132 | Phvul.003G184700.1 | 0.819 | 2 | 292 | 0.657 | 0 | 0.000 |
| 133 | Phvul.005G153600.2 | 0.819 | 1 | 266 | 0.748 | 0 | 0.000 |
| 134 | Phvul.003G063500.1 | 0.818 | 2 | 992 | 0.992 | 0 | 0.000 |
| 135 | Phvul.005G001800.2 | 0.818 | 3 | 101 | 0.900 | 0 | 0.000 |
| 136 | Phvul.011G214700.1 | 0.817 | 2 | 435 | 0.916 | 0 | 0.000 |
| 137 | Phvul.005G042800.1 | 0.817 | 2 | 518 | 0.905 | 0 | 0.000 |
| 138 | Phvul.001G122400.1 | 0.816 | 1 | 313 | 0.918 | 0 | 0.000 |
| 139 | Phvul.001G179000.1 | 0.816 | 2 | 104 | 0.730 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 140 | Phvul.010G033600.1 | 0.815 | 2 | 87 | 0.737 | 0 | 0.000 |
| 141 | Phvul.009G080500.1 | 0.815 | 1 | 211 | 0.690 | 0 | 0.000 |
| 142 | Phvul.002G018500.1 | 0.815 | 2 | 342 | 0.617 | 0 | 0.000 |
| 143 | Phvul.005G036400.1 | 0.815 | 1 | 252 | 0.929 | 0 | 0.000 |
| 144 | Phvul.002G252600.1 | 0.813 | 1 | 279 | 0.958 | 0 | 0.000 |
| 145 | Phvul.001G151200.1 | 0.813 | 2 | 322 | 0.972 | 0 | 0.000 |
| 146 | Phvul.003G269100.1 | 0.813 | 2 | 393 | 0.744 | 0 | 0.000 |
| 147 | Phvul.002G246200.1 | 0.812 | 2 | 163 | 0.841 | 0 | 0.000 |
| 148 | Phvul.010G033200.1 | 0.812 | 2 | 128 | 0.785 | 0 | 0.000 |
| 149 | Phvul.005G084500.1 | 0.812 | 2 | 334 | 0.843 | 0 | 0.000 |
| 150 | Phvul.010G033300.1 | 0.812 | 2 | 128 | 0.785 | 0 | 0.000 |
| 151 | Phvul.001G057600.1 | 0.811 | 2 | 97 | 0.682 | 0 | 0.000 |
| 152 | Phvul.007G151500.1 | 0.811 | 2 | 280 | 0.745 | 0 | 0.000 |
| 153 | Phvul.007G151500.2 | 0.811 | 2 | 265 | 0.751 | 0 | 0.000 |
| 154 | Phvul.009G178600.1 | 0.808 | 2 | 276 | 0.665 | 0 | 0.000 |
| 155 | Phvul.007G007700.1 | 0.808 | 2 | 122 | 0.719 | 0 | 0.000 |
| 156 | Phvul.010G009700.1 | 0.807 | 2 | 59 | 0.690 | 0 | 0.000 |
| 157 | Phvul.008G055500.1 | 0.807 | 4 | 340 | 0.626 | 0 | 0.000 |
| 158 | Phvul.007G118800.1 | 0.807 | 2 | 301 | 0.848 | 0 | 0.000 |
| 159 | Phvul.009G215200.1 | 0.807 | 1 | 331 | 0.697 | 0 | 0.000 |
| 160 | Phvul.002G158400.1 | 0.806 | 1 | 0 | 0.000 | 0 | 0.000 |
| 161 | Phvul.001G219500.1 | 0.806 | 2 | 879 | 0.885 | 0 | 0.000 |
| 162 | Phvul.007G199800.2 | 0.806 | 2 | 296 | 0.701 | 0 | 0.000 |
| 163 | Phvul.007G199800.1 | 0.806 | 2 | 296 | 0.701 | 0 | 0.000 |
| 164 | Phvul.009G101600.1 | 0.805 | 1 | 38 | 0.686 | 0 | 0.000 |
| 165 | Phvul.004G156100.1 | 0.805 | 2 | 614 | 0.876 | 0 | 0.000 |
| 166 | Phvul.009G249300.1 | 0.804 | 2 | 508 | 0.890 | 0 | 0.000 |
| 167 | Phvul.001G264100.1 | 0.804 | 2 | 874 | 0.975 | 0 | 0.000 |
| 168 | Phvul.004G005400.1 | 0.804 | 2 | 496 | 0.875 | 0 | 0.000 |
| 169 | Phvul.001G211100.1 | 0.803 | 1 | 296 | 0.941 | 0 | 0.000 |

| | id | prob | cluster | N-pos | N-prob | C-pos | C-prob |
|---|---|---|---|---|---|---|---|
| 170 | Phvul.006G181500.2 | 0.803 | 2 | 290 | 0.857 | 0 | 0.000 |
| 171 | Phvul.006G181500.1 | 0.803 | 2 | 290 | 0.857 | 0 | 0.000 |
| 172 | Phvul.005G157800.3 | 0.802 | 1 | 432 | 0.942 | 0 | 0.000 |
| 173 | Phvul.005G157800.2 | 0.802 | 1 | 432 | 0.942 | 0 | 0.000 |
| 174 | Phvul.005G157800.4 | 0.802 | 1 | 432 | 0.942 | 0 | 0.000 |
| 175 | Phvul.005G157800.1 | 0.802 | 1 | 432 | 0.942 | 0 | 0.000 |
| 176 | Phvul.011G128000.1 | 0.802 | 2 | 398 | 0.697 | 0 | 0.000 |
| 177 | Phvul.001G136800.1 | 0.801 | 2 | 105 | 0.883 | 0 | 0.000 |
| 178 | Phvul.006G090900.1 | 0.800 | 1 | 535 | 0.850 | 0 | 0.000 |
| 179 | Phvul.010G144300.1 | 0.800 | 2 | 239 | 0.740 | 0 | 0.000 |
| 180 | Phvul.007G024200.1 | 0.800 | 3 | 48 | 0.654 | 0 | 0.000 |
| 181 | Phvul.004G121600.1 | 0.800 | 2 | 262 | 0.707 | 0 | 0.000 |

Cuadro 3: Tabla 3. Estadísticas principales de los clusters fisicoquímicos.mw, Peso molecular. charge, Carga neta. pI, Punto isoeléctrico. aindex, Índice alifático. instaindex, Índice de inestabilidad. boman, Índice de Boman, hmoment100, hmoment160. Momento hidrofóbico (100, 160 rotación de hélice). Composición de grupos: tiny, Diminutos. small, Pequeños. aro, Aromáticos. ali, Alifáticos. nonP, No polares. chrgd, Cargados. basic, Basicos. acidic, Ácidos.

Cluster 1

| | acidic | aindex | aro | basic | boman | charge | chrgd | hmoment100 | hmoment160 | instaindex | mw | nonP | pI | small | tiny |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 6.688931 | 44.718268 | 11.784769 | 19.215306 | 2.192395 | 11.080610 | 25.904231 | 0.569700 | 0.553082 | 44.768911 | 5969.464681 | 54.276631 | 8.667368 | 57.424737 | 40.716112 |
| std | 3.968757 | 16.064009 | 4.389159 | 3.723140 | 0.590860 | 2.649090 | 5.306881 | 0.109330 | 0.116775 | 18.307272 | 1621.330220 | 6.444767 | 0.878226 | 5.231986 | 6.062100 |
| min | 0.000000 | 10.425532 | 3.571000 | 10.204000 | 1.254396 | 5.999292 | 13.636000 | 0.322565 | 0.284362 | 5.507895 | 3545.166640 | 37.838000 | 4.215889 | 44.068000 | 24.074000 |
| 25 % | 4.000000 | 34.493976 | 8.696000 | 16.817000 | 1.723812 | 8.999657 | 21.510750 | 0.492479 | 0.481055 | 31.522265 | 5004.921015 | 50.893500 | 8.387231 | 54.450500 | 37.255000 |
| 50 % | 6.383000 | 44.810081 | 11.803500 | 19.098500 | 2.097192 | 10.998944 | 25.658500 | 0.566318 | 0.536560 | 44.085710 | 5468.226740 | 54.702000 | 8.743631 | 57.447000 | 40.909000 |
| 75 % | 8.696000 | 52.982955 | 14.642250 | 21.894000 | 2.526283 | 12.249164 | 29.787000 | 0.644554 | 0.621241 | 57.052894 | 6235.738640 | 58.000000 | 9.098601 | 60.887000 | 44.948000 |
| max | 21.505000 | 90.707071 | 24.324000 | 29.787000 | 4.057143 | 20.998959 | 38.298000 | 0.904380 | 1.009224 | 92.238462 | 10962.564240 | 72.917000 | 10.818685 | 69.697000 | 53.030000 |

Cluster 2

| | acidic | aindex | aro | basic | boman | charge | chrgd | hmoment100 | hmoment160 | instaindex | mw | nonP | pI | small | tiny |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 4.575162 | 93.379515 | 8.886284 | 27.762824 | 1.377697 | 9.823920 | 32.338068 | 0.766739 | 0.489350 | 15.388123 | 3603.960609 | 55.691122 | 11.469358 | 41.135338 | 28.062081 |
| std | 2.998510 | 20.206666 | 5.381705 | 5.154714 | 0.555613 | 1.674599 | 5.022867 | 0.141348 | 0.122411 | 18.040464 | 669.189584 | 6.106814 | 0.526351 | 9.813776 | 6.447042 |
| min | 0.000000 | 49.000000 | 0.000000 | 20.000000 | 0.423462 | 6.999590 | 25.000000 | 0.405701 | 0.301640 | -36.220000 | 2209.791240 | 40.000000 | 10.085761 | 20.000000 | 8.000000 |
| 25 % | 2.487250 | 78.918919 | 5.128000 | 24.324000 | 0.946864 | 8.999543 | 28.571000 | 0.672605 | 0.405381 | 2.975682 | 3000.737165 | 51.476250 | 11.116366 | 34.711250 | 24.000000 |
| 50 % | 5.128000 | 96.401312 | 7.947500 | 27.027000 | 1.358995 | 9.999464 | 31.643000 | 0.749665 | 0.477708 | 16.423273 | 3809.590440 | 56.950000 | 11.449492 | 43.243000 | 28.056000 |
| 75 % | 5.714000 | 105.703704 | 11.813250 | 30.635000 | 1.644038 | 10.999453 | 35.005000 | 0.904490 | 0.554866 | 24.724196 | 4086.625865 | 60.000000 | 11.822451 | 48.700750 | 33.333000 |
| max | 13.514000 | 163.600000 | 27.273000 | 41.176000 | 2.845200 | 13.999622 | 45.833000 | 0.987856 | 0.930652 | 67.265714 | 5221.332040 | 68.571000 | 12.546071 | 58.824000 | 41.176000 |

Cluster 3

| | acidic | aindex | aro | basic | boman | charge | chrgd | hmoment100 | hmoment160 | instaindex | mw | nonP | pI | small | tiny |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 4.740287 | 56.143711 | 9.154376 | 7.750248 | 0.606070 | 3.336249 | 12.490604 | 0.318065 | 0.348135 | 31.883024 | 3124.381597 | 66.489262 | 6.760473 | 70.521495 | 50.925812 |
| std | 2.804957 | 24.475495 | 5.120498 | 4.416769 | 0.545369 | 1.454295 | 4.612094 | 0.091714 | 0.086285 | 17.011213 | 629.138342 | 5.531794 | 1.651082 | 7.399427 | 7.968848 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -1.175000 | 0.999544 | 0.000000 | 0.044051 | 0.126368 | -16.758333 | 780.838540 | 45.000000 | 3.550010 | 52.381000 | 22.857000 |
| 25 % | 3.333000 | 36.896552 | 6.667000 | 3.448000 | 0.224707 | 1.999669 | 9.756000 | 0.260777 | 0.283867 | 21.519814 | 2984.663540 | 63.636000 | 5.940668 | 66.667000 | 46.667000 |
| 50 % | 3.448000 | 56.451613 | 6.897000 | 7.407000 | 0.582652 | 2.999669 | 12.903000 | 0.304950 | 0.360156 | 30.424194 | 3145.730740 | 66.667000 | 7.728705 | 69.331500 | 50.000000 |
| 75 % | 6.667000 | 76.305804 | 10.526000 | 10.000000 | 0.906234 | 3.999669 | 14.815000 | 0.365941 | 0.400961 | 41.816845 | 3296.547465 | 70.000000 | 8.111233 | 75.000000 | 55.889000 |
| max | 12.500000 | 114.137931 | 32.143000 | 28.571000 | 2.202500 | 8.999746 | 28.571000 | 0.640814 | 0.603788 | 82.984211 | 5057.838140 | 90.000000 | 10.549999 | 89.655000 | 70.000000 |

Cluster 4

| | acidic | aindex | aro | basic | boman | charge | chrgd | hmoment100 | hmoment160 | instaindex | mw | nonP | pI | small | tiny |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 1.646190 | 159.742592 | 10.829274 | 14.484738 | -0.994013 | 3.106868 | 16.130905 | 0.533816 | 0.280358 | 9.145493 | 1611.182341 | 72.099381 | 10.190283 | 37.788119 | 27.814952 |
| std | 3.407431 | 29.728628 | 7.558282 | 6.529581 | 0.598360 | 1.006216 | 7.532955 | 0.089723 | 0.134255 | 17.467819 | 310.243929 | 7.116149 | 0.965640 | 8.010962 | 7.493612 |
| min | 0.000000 | 83.571429 | 0.000000 | 5.263000 | -2.428333 | 1.999749 | 5.263000 | 0.279061 | 0.088978 | -23.326667 | 804.043340 | 54.545000 | 6.447271 | 23.077000 | 7.143000 |
| 25 % | 0.000000 | 143.529412 | 6.470750 | 7.692000 | -1.307105 | 1.999749 | 7.692000 | 0.479224 | 0.176159 | -3.830769 | 1451.002115 | 69.231000 | 9.700016 | 33.333000 | 22.600750 |
| 50 % | 0.000000 | 157.692308 | 10.526000 | 14.286000 | -0.933344 | 2.999749 | 19.091000 | 0.556836 | 0.241131 | 5.947059 | 1535.419440 | 71.429000 | 9.702002 | 37.652000 | 28.571000 |
| 75 % | 0.000000 | 178.750000 | 15.385000 | 21.429000 | -0.681488 | 3.999749 | 21.429000 | 0.589558 | 0.365443 | 18.025315 | 1770.669740 | 76.923000 | 11.103081 | 41.596250 | 33.333000 |
| max | 15.385000 | 227.500000 | 27.778000 | 29.412000 | 0.536154 | 5.999749 | 30.769000 | 0.710773 | 0.619960 | 76.830769 | 3040.813640 | 88.235000 | 11.651769 | 65.625000 | 40.000000 |