UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Synthetic antimicrobial peptides generation using recurrent neural networks

Andrés Vélez [a], Carlos Mera [b], Sergio Orduz [c] & John W. Branch [a]

[a] *Universidad Nacional de Colombia - Sede Medellín, Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión, Medellín, Colombia. anvelezec@unal.edu.co, jwbranch@unal.edu.co*
[b] *Departamento de Sistemas de Información, Instituto Tecnológico Metropolitano - ITM, Medellín, Colombia, carlosmera@itm.edu.co*
[c] *Universidad Nacional de Colombia - Sede Medellín, Facultad de Ciencias, Escuela de Biociencias, Medellín, Colombia, sorduzp@unal.edu.co.*

Received: July 2nd, 2020. Received in revised form: January 2nd, 2021. Accepted: January 18th, 2021.

Abstract
The antimicrobial peptides (AMPs) have taken importance in the development of new antibiotics because of their role as an inhibitor, not only of bacteria but also of viruses, fungi and parasites, among others. Since the discovery of AMPs, thousands have been reported, however, many of them are not suitable for therapeutic applications due to their long amino acid sequences, low antimicrobial potency and high production costs. In this work, we propose to use recurrent neural networks (RNN) with LSTM cells in order to generate more potent and economical peptides. We perform different experiments generating synthetic AMPs between 12 and 20 amino acids. The results show that we can use RNN and improve the generation process compared with the template method.

Keywords: antimicrobial resistance; synthetic peptides; virtual screening; deep learning.

Generación de péptidos antimicrobianos mediante redes neuronales recurrentes

Resumen
Los péptidos antimicrobianos (AMP) han tomado importancia en el desarrollo de nuevos antibióticos debido a su papel como inhibidores, no solo de bacterias sino también de virus, hongos y parásitos, entre otros. Desde el descubrimiento de los AMP, se han reportado miles, sin embargo, muchos de ellos no son adecuados para aplicaciones terapéuticas debido a sus largas secuencias de aminoácidos, baja potencia antimicrobiana y altos costos de producción. En este trabajo, proponemos utilizar redes neuronales recurrentes (RNN) con células LSTM para generar péptidos más potentes y económicos. Realizamos diferentes experimentos generando AMP sintéticos entre 12 y 20 aminoácidos. Los resultados muestran que podemos usar RNN y mejorar el proceso de generación en comparación con el método de plantillas manuales.

Palabras clave: resistencia antimicrobiana; péptidos sintéticos; virtual screening; aprendizaje profundo.

1. Introducción

La resistencia de las bacterias, virus, hongos y parásitos a los medicamentos ha generado una preocupación creciente alrededor del mundo. En la actualidad, se estima la presencia de cerca de 30 millones de casos de sepsis al año a nivel global, de los cuales cerca de 5 millones terminan en muertes como resultado de infecciones que resisten los tratamientos médicos con los antibióticos tradicionales [1]. Este panorama ha llevado a la búsqueda de nuevos antibióticos que permitan combatir las bacterias resistentes.

Ejemplos específicos de resistencia son los que han desarrollado las bacterias *Neisseria gonorrhoeae* FC428 clone2 y *Klebsiella pneumoniae*. Esta última bacteria es una especie que desempeña un papel importante como agente causante de sepsis y enfermedades contagiosas como la neumonía, infecciones del torrente sanguíneo y del tracto urinario; especialmente, en pacientes con diabetes mellitus y neonatos, entre otros [2].

How to cite: Vélez, A., Mera, C., Orduz, S. and Branch, J.W., Generación de péptidos antimicrobianos mediante redes neuronales recurrentes.. DYNA, 88(216), pp. 210-219, January - March, 2021

K. pneumoniae es tratada con antibióticos carbapenémicos, los cuales son el grupo más usado entre los antibióticos disponibles. No obstante, la resistencia desarrollada a este tipo de antibióticos ha llevado a que los tratamientos fracasen en casi la mitad de los pacientes [2]. A pesar de que es clara la necesidad de investigar y desarrollar nuevas moléculas antimicrobianas, existe un desinterés por parte de las compañías farmacéuticas para hacerlo [3]. Esto se debe, principalmente, al escaso retorno de la inversión que representa el desarrollo de nuevos antibióticos, en comparación con los medicamentos utilizados para el tratamiento de enfermedades crónicas como la diabetes y la hipertensión [2]. Como consecuencia, la Organización Mundial de la Salud (OMS) y la Asamblea General de las Naciones Unidas, han alentado a la comunidad científica para aumentar los esfuerzos en la búsqueda de nuevos antibióticos que permitan combatir esta y otras especies de bacterias resistentes a los antibióticos tradicionales [44].

Los péptidos antimicrobianos (o AMPs, de su acrónimo en inglés *Antimicrobial Peptides*) han tomado importancia en el desarrollo de nuevos antibióticos por su rol como agente inhibidor, no solo de bacterias sino también de virus, hongos y parásitos, entre otros [4,5]. Los AMPs son parte esencial de todos los organismos vivos y configuran la primera línea de defensa contra bacterias, microbios y parásitos. Este tipo de péptidos causan la muerte de los microbios, bien sea interfiriendo las funcionalidades de la membrana celular o interrumpiendo sus funciones intracelulares [6,7].

En términos generales, los AMPs se caracterizan por estar compuestos por menos de 50 aminoácidos y tener una carga catiónica positiva con rangos entre +1 y +9. Además, son capaces de adoptar estructuras anfipáticas, es decir, con una región hidrofílica que es soluble en agua y otra que es hidrofóbica, la cual rechaza el agua [8]. A pesar de presentar una gran variabilidad estructural, los AMPs se pueden agrupar en cuatro categorías: hélice, hojas beta, estructura extendida lineal y mezclas entre ellas [9,10].

Aparte de la búsqueda tradicional de nuevos medicamentos, que incluye un trabajo exhaustivo de análisis de la biodiversidad, la identificación manual de AMPs implica analizar secuencias de péptidos de diferentes tamaños que pueden ser extraídas de una proteína sintetizada por un organismo vivo. En este análisis se examinan las características fisicoquímicas de diferentes secuencias de aminoácidos a fin de seleccionar un grupo de péptidos candidatos. Estos se sintetizan para determinar las concentraciones inhibitorias mínimas (MIC) contra ciertas bacterias, como *K. pneumoniae*. La actividad antimicrobiana de los péptidos analizados permite, posteriormente, determinar su viabilidad para el desarrollo de nuevos antibióticos.

Desde el descubrimiento de la Maganina por Zasloff en 1987 [11], miles de péptidos antimicrobianos han sido reportados en la literatura, pero muchos de ellos no son adecuados para aplicaciones terapéuticas debido a su tamaño (más de 30 aminoácidos) y baja potencia antimicrobiana. Esto sin contar los altos costos de desarrollo que tienen para las farmacéuticas, los cuales inician desde 100 millones y pueden ir hasta los 1300 millones de dólares, además de los largos periodos de tiempo que toma su producción que se estima que puede ser de 11 años o más [12].

Dada la necesidad de buscar soluciones a la resistencia antimicrobiana, en este trabajo se propone el uso de redes neuronales recurrentes para la generación de péptidos sintéticos que puedan tener una alta probabilidad de ser antimicrobianos. Las redes neuronales recurrentes son un tipo de redes neuronales que pueden aprender adecuadamente los elementos que componen una secuencia. En el caso del lenguaje natural este tipo de redes se han usado para encontrar las relaciones entre letras y palabras que componen una oración, un párrafo y un texto en general. De manera similar, puesto que un péptido es una secuencia de aminoácidos (letras), las redes neuronales recurrentes se pueden usar para que aprendan su estructura de aminoácidos y posteriormente generar nuevos péptidos con base en la estructura de los péptidos de entrenamiento.

Este trabajo se ha organizado de la siguiente manera. En la Sección 2 se presentan los conceptos esenciales y se hace una revisión del estado del arte. En la Sección 3 se describe la metodología usada para la generación de péptidos con redes neuronales recurrentes. En la Sección 4 se presentan los experimentos y se discuten los resultados. Finalmente, se presentan las conclusiones en la Sección 5.

2. Péptidos y péptidos antimicrobianos

2.1 Péptidos

Los péptidos son moléculas biológicas naturales que se encuentran en todos los organismos vivos y juegan un papel clave en todo tipo de actividad biológica, al igual que las proteínas. Hay 20 aminoácidos naturales y, como las letras de una palabra, se pueden combinar para generar una inmensa variedad de moléculas diferentes. Generalmente, cuando una molécula tiene entre 2 y 50 aminoácidos se llama péptido, mientras que a las moléculas más largas se les denomina proteínas [13]. La función que desempeña un péptido depende de los aminoácidos que lo componen. Así mismo, la distribución de esos aminoácidos determina el tipo de estructura y de las propiedades fisicoquímicas del péptido. Estas a su vez permiten o no al péptido interactuar con la membrana celular, bien sea como transmisor, enzima, hormona o antibiótico [13].

2.2 Péptidos antimicrobianos

Los AMPs son un grupo específico de péptidos y el espectro de su actividad es amplio y abarca antivirales, antifúngicos, antitumorales y antibacterianos. Así, los AMPs son una familia de sustancias polifacéticas que contemplan complejos mecanismos de acción relacionados con la interacción con la membrana celular de un patógeno, desestabilizándola o penetrándola a fin de afectar blancos internos, como la replicación del ADN, la síntesis de proteínas o interactuando con el huésped, regulando el proceso inflamatorio y de cicatrización [14]. El mecanismo de acción de un AMP está relacionado con su carácter catiónico y anfipático. Esto facilita su interacción e inserción en las paredes celulares aniónicas y membranas de fosfolípidos de los microorganismos para luego ejercer una

acción detergente sobre la membrana celular [8].

A pesar de que los AMPs se caracterizan por su variedad estructural, se han reportado ciertas características fisicoquímicas comunes en la mayoría de los péptidos con actividad antimicrobiana experimentalmente validada. En general, se habla de AMPs con propiedad de carga positiva la cual oscila entre +1 y +9, un porcentaje de hidrofobicidad que oscila entre el 40% y el 60%, un momento hidrofóbico relativo mayor a 50% y un punto isoeléctrico mayor a nueve [45]. Todas estas características, en conjunto, permiten incrementar la probabilidad de interacción de un péptido con una membrana celular y crear una disrupción en esta.

2.3 Generación de péptidos antimicrobianos

En la literatura se han propuesto diferentes aproximaciones que llevan a obtener péptidos antimicrobianos. De acuerdo con Fjell et. al [15], existen tres corrientes de investigación que agrupan los trabajos en este campo: el análisis manual de los péptidos para la creación de plantillas, la modelación biofísica y el uso de algoritmos de inteligencia artificial, denominado en inglés *virtual screening*. En esta última corriente existen dos líneas de estudio: la generación sintética de AMPs y la predicción de la actividad antimicrobiana, más conocida como modelos QSAR por su acrónimo en inglés *Quantitative Structure Activity Relationship* [9].

El análisis de la actividad de los péptidos para el diseño de plantillas busca identificar, de manera manual, aquellos péptidos que tienen el mayor grado de actividad antimicrobiana [15]. Los péptidos usados en el análisis se obtienen de secuencias en las que se van modificando, sistemáticamente, cada uno de los aminoácidos a fin de observar el comportamiento que dicho cambio produce en la estructura y las propiedades del péptido.

Esto ayuda a identificar el orden y la posición de los aminoácidos importantes para poder generar plantillas con estructuras que permitan el diseño de péptidos con actividad antimicrobiana [16]. A pesar de los resultados prometedores de esta corriente de investigación, su limitación es que se concentra en enfoques locales; es decir, al realizar el análisis de la actividad de los péptidos no se tiene en cuenta las interacciones entre los aminoácidos que influyen en la conformación tridimensional del mismo [17].

Por su parte, las investigaciones basadas en la modelación biofísica buscan comprender la actividad de los AMPs y diseñar variantes mejoradas, bien sea mediante el análisis de un péptido en ambientes hidrofóbicos o usando modelación a nivel atómico [15]. Una dificultad en este tipo de investigaciones es la complejidad de la cantidad de variaciones y componentes que interactúan entre sí y que conlleva a simplificaciones impuestas por restricciones para realizar la modelación. De esta forma, las predicciones que provienen de tales modelos no pueden ser automáticamente transferidas a la configuración en vivo, lo que conlleva a un aumento en el número de pruebas en laboratorio que se deben realizar, generando sobrecostos al proceso [15].

Por último, las investigaciones basadas en *virtual screening* utilizan las propiedades fisicoquímicas de los péptidos y su estructura, en conjunto con diversos algoritmos de aprendizaje de máquina o algoritmos evolutivos para relacionar esas propiedades con la actividad antimicrobiana [18]. Adicionalmente, han empezado a surgir trabajos en los que se utilizan las redes neuronales de aprendizaje profundo, no para clasificar un péptido como antimicrobiano o no, sino para generar péptidos sintéticos antimicrobianos [19] y anticancerígenos [20].

Tabla 1.
Evolutivo de trabajos representativos referentes a la generación sintética y clasificación de AMPs.

Año	Algoritmo	Estrategia de diseño	Descripción	Ref.
1995	RN	Algoritmo evolutivo	Los péptidos son generados como un proceso evolutivo de simulación maximizando la clasificación de la RN.	[21]
2006	Reglas gramaticales	Modelo lingüístico	Los AMPs son generados con base en reglas gramaticales extraídas de péptidos naturales.	[16]
2009	RN		Clasificación de péptidos con QSAR y *machine learning*.	[22]
2010	RN		Presentan AntiBP2, un algoritmo basado en redes neuronales para clasificar la actividad antimicrobiana de un péptido.	[23]
2011	RN	Algoritmo genético	Se usa un algoritmo genético para el diseño de nuevos péptidos.	[18]
2013	Fuzzy *k*-means		Clasificador multi-etiqueta basado en las propiedades fisicoquímicas de un AMP. El sistema desarrollado clasifica los AMPs en diez categorías.	[24]
2014	SVM		Uso del kernel espectro-p de un péptido para la clasificación de AMPS.	[25]
2016	Random forest, RN, SVM		CAMP-R3 permite clasificar la actividad antimicrobiana de un péptido usando redes neuronales, máquinas de soporte vectorial y random forest.	[26]
2018	AG	Algoritmo genético	Mejorar la actividad antimicrobiana de péptidos contra *E. coli*.	[5]
2018	Random forest		Presenta un algoritmo basado en random forest para clasificar la actividad antimicrobiana.	[27]
2018	RN profundas		Usa redes convoluciones y recurrentes para determinar si un péptidos es antimicrobiano o no	[28]
2018		RN profundas	Se basan en la representación de secuencias internas de alta dimensión para la generación de nuevos AMPs tomando características aprendidas por las bases suministradas.	[19]
2018		RN profundas	Usa redes neuronales con una estructura *auto-encoder* normalizado para el diseño de AMPs	[29]
2018		RN profundas	Uso de *fine tuning* para focalizar el conocimiento de un modelo previamente entrenado.	[20]
2019	Random forest, SVM		Uso de modelos de *machine learning* para la predicción de múltiples categorías.	[30]
2019	RN		Usa una red neuronal convolucional multi-escala para la identificación de AMPs.	[47]
2020	RN		Usa una rede neuronal convolucional para la predicción de AMPs cortos	[48]
2020	Random forest		N-grams + Random forest	[49]

Fuente: Los Autores

Entre las tres aproximaciones, *virtual screening* toma importancia dado que optimiza el proceso de búsqueda de un AMP haciendo de esta aproximación la más viable en términos económicos y de tiempo. Lo anterior porque se disminuye el número de secuencias a sintetizar y el número de pruebas de laboratorio que se requieren, comparada con las otras dos aproximaciones [18]. A modo de resumen, la Tabla 1 muestra algunos de los trabajos representativos de esta aproximación.

3. Metodología para la generación de AMPs usando RNN

La metodología para la generación de péptidos sintéticos usando redes neuronales recurrentes abarca los siguientes pasos:

i. Recolección de AMPs: con el fin de aprender la estructura de un péptido antimicrobiano, es necesario recopilar un conjunto de péptidos de los cuales se tenga certeza que tienen capacidad antimicrobiana.
ii. Preprocesamiento de los péptidos: el paso siguiente consiste en reprocesar los péptidos a fin de transformar las secuencias de aminoácidos en una representación dispersa que pueda ser utilizada por la red neuronal recurrente.
iii. Entrenamiento del modelo: para entrenar los modelos de aprendizaje profundo es necesario seleccionar y configurar aspectos como la función de pérdida, el tamaño del lote de ejecución, el algoritmo de optimización y las métricas de rendimiento del proceso de entrenamiento. Estos elementos influyen en la capacidad predictiva de los modelos y por tanto, de la generación de los péptidos sintéticos.
iv. Generación de péptidos: una vez se ha entrenado el modelo de aprendizaje profundo, éste se utiliza para generar péptidos sintéticos. En nuestro caso, dicha generación considera secuencias que sean económicamente viables de sintetizar y producir. Finalmente, los péptidos generados se filtran y se clasifican para obtener aquellos que tengan mayor probabilidad de ser antimicrobianos y puedan ser llevados al laboratorio para iniciar las pruebas de inhibición con los mismos. A continuación, se describen en detalle cada uno de estos pasos.

3.1 Recolección de AMPs

En la literatura se encuentran diferentes trabajos en los que se han presentado y usado diversas bases de datos de péptidos, normalmente, para fines de clasificación. Entre esos trabajos, las bases de datos más comunes son:

- ADAM la cual contiene 7007 secuencias de AMPs de diferentes tipos de organismos [31].
- APD3 que es una base de datos que se centra en péptidos antimicrobianos naturales, con secuencia y actividad definidas, e incluye un total de 2619 secuencias [32].
- MilkAMP que se caracteriza por tener 371 secuencias derivadas de la leche; sin embargo, solo 23 han sido validadas como AMPs [33].
- LAMP la cual cuenta con 5547 secuencias de las cuales 3904 son péptidos antimicrobianos naturales y los 1643 restantes son sintéticos [34].

Una vez agrupados los péptidos antimicrobianos de las bases de datos descritas, se procedió a realizar el siguiente proceso de filtrado:

i. Eliminación de duplicados: con el fin de evitar secuencias duplicadas, se analizaron las cuatro bases de datos y se eliminaron aquellas secuencias que se repiten dejando una sola ocurrencia.
ii. Eliminación de péptidos con aminoácidos X: debido a que este aminoácido representa un amidado en el extremo C-terminal.
iii. Eliminación de péptidos con cisteína (aminoácido C): la distribución de los aminoácidos de un péptido determina su estructura. Entre las cuatro estructuras posibles, los péptidos con una conformación helicoidal tienden a tener altos momentos hidrofóbicos [8]. Esta característica los hace interesantes porque mejora la capacidad de contacto del péptido y la adherencia a la membrana celular de las bacterias, razón por la cual tienen mayor capacidad de ser antimicrobiano. Así, los péptidos que contienen cisteína se eliminan porque este aminoácido puede generar puentes de disulfuro [35] que evitan que los péptidos adopten una estructura helicoidal. Después del proceso de filtrado, se obtuvo una lista de 2668 péptidos antimicrobianos, cuyas propiedades fisicoquímicas se resumen en la Tabla 2.

3.2 Preprocesamiento de los péptidos

Para el preprocesamiento de las secuencias de aminoácidos que conforman los AMPs, se tomó como referencia el trabajo de Müller et al. [19], en el cual se proponen los siguientes pasos:

i. Se adicionan los caracteres "B" y "_" al inicio y al final de cada péptido, respectivamente. Esto es necesario para establecer el punto de partida y el punto de finalización de una secuencia de aminoácidos.
ii. Se determina la longitud ($lmax$) del péptido con el mayor número de aminoácidos en las secuencias recolectadas. Esto se hace con el fin de homogeneizar el tamaño de los péptidos. Así, aquellos péptidos que tienen un tamaño menor a $lmax$ se rellenan con el carácter "_" hasta alcanzar la longitud de $lmax$.

El modelo de generación de AMPs propuesto considera el uso de modelos de aprendizaje profundo en el que se utiliza una función de activación tipo *softmax*, por ello las secuencias de aminoácidos se codifican usando una estrategia de transformación *one hot encoding*. En este tipo de transformación cada aminoácido de un péptido es representado como un vector de longitud 21 (19 aminoácidos más los caracteres de inicio "B" y fin "_") en el que todas sus posiciones, excepto la que representa al aminoácido, tienen valor cero. Por ejemplo, considere el siguiente péptido: GHKAA. Con base en las reglas de preprocesamiento iniciales y suponiendo que $lmax = 7$ el péptido se transforma a: BGHKAA. Al usar una codificación *one hot encoding* el péptido se representa por una matriz de 7x21 como la que se ilustra en la Fig. 1.

Tabla 2.
Resumen de las propiedades fisicoquímicas de los AMPs recolectados

Propiedades Fisicoquímicas	Límite Inferior	Mediana	Límite Superior
Momento hidrofóbico	0.09	0.38	0.96
Punto isoeléctrico	3.75	10.81	12.83
Carga	-1.00	2.00	8.09
% de aminoácidos hidrofóbicos	15.38	56.25	91.67
Longitud	8.00	21.00	30.00

Fuente: Los Autores

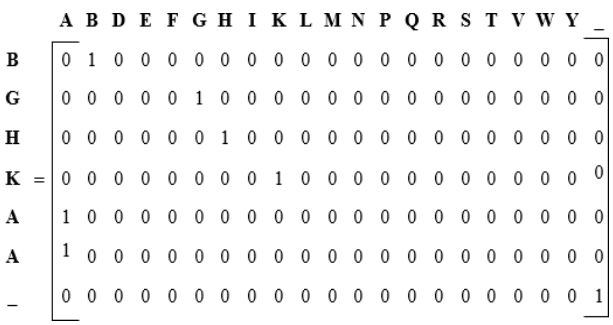

Figura 1. Representación del péptido BHKAA usando *one hot encoding*. Fuente: Los Autores

3.3 Entrenamiento del modelo de aprendizaje

Dado que un AMP puede verse como una palabra que es formada usando un alfabeto determinado y un conjunto de reglas de composición, los métodos que se aplican a la generación de lenguaje natural pueden ser usados para la creación de péptidos con capacidad antimicrobiana. Específicamente, Müller et al. [19] han empleado los modelos de aprendizaje profundo para la generación de AMPs [19]. No obstante, en ese trabajo los autores se centraron en el uso de redes recurrentes con una arquitectura LSTM-Decoder. A diferencia de la propuesta en [19], en este trabajo se explora el uso de una arquitectura *autoencoder* con celdas LSTM, denominada LSTM-LSTM. Este tipo de arquitectura fue propuesta por Sutskever et al. [36] y tiene la capacidad de crear una representación vectorial fija, también conocida como el contexto del texto procesado, el cual ayuda a mejorar la capacidad de memoria del modelo [37]. Este tipo de redes recurrentes ha sido utilizado para la traducción de textos [36, 38] y, de manera similar, puede ser usada para la generación de péptidos sintéticos. Al usar una arquitectura LSTM-LSTM se dota el proceso de generación con memoria a fin de capturar la estructura de la secuencia de los aminoácidos que conforman un AMP.

Por su parte, para entrenar correctamente los modelos de aprendizaje profundo se deben establecer la función de pérdida, el algoritmo de optimización, el tamaño del lote de ejecución, la selección de la métrica de rendimiento y la estrategia de validación a utilizar.

3.3.1 Definición de la función de pérdida

Como el objetivo es predecir un aminoácido, teniendo en cuenta la composición de la cadena de aminoácidos ya generada, la función de pérdida seleccionada es la entropía cruzada, la cual se describe en (1).

$$\boldsymbol{L}(\boldsymbol{X}) = \sum_{j=1}^{M} \sum_{i=1}^{lmax} y_{i,j} * log\left(p\left(y_{i,j}\right)\right)\ i = 1, 2, 3, \ldots, lmax \qquad (1)$$

donde M es el número de péptidos de entrenamiento y equivale a 2668, $lmax$ es el número de aminoácidos del péptido más largo, $p\left(y_{i,j}\right)$ es la probabilidad de dado un aminoácido predecir el siguiente aminoácido de la cadena y $y_{i,j}$ es un vector de unos y ceros donde se codifica con uno la posición del aminoácido a predecir.

3.3.2 Selección del algoritmo de optimización

Para elegir el algoritmo de optimización se evaluaron los algoritmos RMSProp [39] y Adam [40]. Para esta selección se realizaron un conjunto de pruebas usando como base una arquitectura de red neuronal recurrente con dos capas de celdas LSTM, cada una con 256 unidades. Según Müller et al. [19] esta configuración permite reducir la función de pérdida tanto de las secuencias de entrenamiento como de validación sin evidenciar sobre ajustes. Para estas pruebas, el número de épocas se varió entre 100 y 200, adicionalmente la tasa de aprendizaje tomó valores en el conjunto $\{0.1, 0.01, 0.001\}$. En la Tabla 3 se muestra el promedio de los valores de la función de pérdida utilizando una validación cruzada con 5-folds para cada experimento.

Los resultados de la Tabla 3 muestran que el valor de la función de pérdida para Adam no tiene mayores cambios al variar el número de épocas y la tasa de aprendizaje. Esto permite inferir que Adam, en este problema de aplicación, genera una convergencia temprana de los parámetros a mínimos locales, lo cual es una limitación del algoritmo. Lo contrario sucede con RMSProp, para el cual se observa una disminución en el valor de la función de pérdida al aumentar el número de épocas y al disminuir la tasa de aprendizaje. Con base en esto, se selecciona el algoritmo de optimización RMSProp.

3.3.3 Selección del tamaño del lote de ejecución

Brownlee [41] argumenta que lotes de ejecución pequeños suelen converger rápidamente; aunque, esto afecta la capacidad de generalización del modelo. Esta situación se puede solucionar al ir aumentando el tamaño del lote, poco a poco.
En este sentido, para determinar la mejor configuración para el tamaño del lote de ejecución se utilizó una estrategia de validación cruzada con 5-*folds* en la que se entrenaron las arquitecturas LSTM-Decoder y LSTM-LSTM tomando las mismas configuraciones que se utilizaron en la Sección 3.3. Adicionalmente, el tamaño del lote se estableció en 64, 128 y 256. En este caso, cada modelo se evaluó usando la precisión (o *accuracy*) del modelo de generación. Los resultados de esta prueba se registran en la Tabla 4.

Tabla 3.
Promedio de la función de pérdida para RMSprop y Adam, variando el número de épocas y la tasa de aprendizaje.

Épocas	Arquitectura	Tasa de Aprendizaje	RMSProp	Adam
100	LSTM-Decoder	0.1	51.47 ±4.91	1.10 ±2.41
100	LSTM-Decoder	0.01	0.92 ±0.11	0.99 ±0.02
100	LSTM-Decoder	0.001	0.94 ±0.02	0.95 ±0.19
200	LSTM-Decoder	0.1	34.26 ±7.02	0.95 ±0.11
200	LSTM-Decoder	0.01	0.53 ±0.06	0.82 ±0.08
200	LSTM-Decoder	0.001	0.74 ±0.13	0.76 ±0.02
100	LSTM-LSTM	0.1	20.54 ±3.05	1.10 ±2.33
100	LSTM-LSTM	0.01	0.92 ±0.15	0.91 ±0.24
100	LSTM-LSTM	0.001	0.92 ±0.10	0.91 ±0.19
200	LSTM-LSTM	0.1	21.26 ±4.02	0.94 ±0.22
200	LSTM-LSTM	0.01	0.54 ±0.16	0.81 ±0.16
200	LSTM-LSTM	0.001	0.64 ±0.23	0.77 ±0.18

Fuente: Los Autores

Tabla 4.
Promedio de precisión de predicción usando diferentes tamaños de lote.

Tamaño del Lote de Ejecución	Arquitectura	Precisión del Modelo
64	LSTM-Decoder	83.9 ±5.6 %
128	LSTM-Decoder	88.4 ±3.5 %
256	LSTM-Decoder	90.3 ±2.2 %
64	LSTM-LSTM	83.2 ±4.7 %
128	LSTM-LSTM	86.4 ±3.1 %
256	LSTM-LSTM	91.5 ±1.3%

Fuente: Los Autores

Los resultados obtenidos muestran que, en concordancia con Brownlee [41], a medida que aumenta el tamaño del lote, aumenta la precisión. En este sentido se escoge un tamaño de lote de 256.

3.3.4 Entrenamiento de la red recurrente

Una vez establecidos el tamaño del lote, la función de pérdida y el algoritmo de optimización, se procedió al entrenamiento de la red recurrente. En este caso se utilizaron entre 1 y 2 capas en la estructura *encoder* variando la magnitud del *dropout*. Estas pruebas se realizaron manteniendo constante el número de unidades LSTM en 256. Adicionalmente, con base en los experimentos de Müller et al. [19], se mantuvo fija la estructura *decoder* usando dos capas, cada una con 256 unidades. En la Tabla 5 se muestran los resultados de las pruebas con las diferentes configuraciones usando una estrategia de validación cruzada con 5-folds.

De acuerdo con la información de la Tabla 5, se puede observar que los tres mejores modelos seleccionados, con base en la precisión, son A1, A2 y C2. Entre estos, el que tuvo el mejor rendimiento es el modelo C2 que tiene una arquitectura LSTM-LSTM, con un *dropout* de magnitud 0.2 y una precisión de 93.8%. Los otros dos modelos seleccionados se caracterizan por tener una arquitectura LSTM-Decoder con una precisión de 93.6% y 90.3%.

Los resultados obtenidos, luego de entrenar y validar la arquitectura LSTM-Decoder, muestran que al utilizar una función de optimización RMSProp, con una tasa de aprendizaje de 0.01 y un tamaño de lote de ejecución de 256, logra un error de validación de 0.26±0.01 con una precisión de 93.6%. Al contrastar estos valores respecto a los reportados por Müller et al. [19], que fueron de 0.560±0.060, se observa como una correcta configuración de los parámetros de entrenamiento permite disminuir la función de pérdida de validación y por ende mejorar el aprendizaje y la capacidad de generalización de la red neuronal profunda.

Para el modelo con arquitectura LSTM-LSTM se destaca aquel que usa un *dropout* de magnitud 0.2, con error de validación de 0.51 ± 0.02 y una precisión de 93.8 ± 2.5%. Estos resultados muestran como la incorporación del contexto dentro del modelo y una buena configuración de parámetros ayuda a aumentar la precisión unos puntos, reducir la variabilidad y a su vez mejorar la capacidad de generalización respecto a la arquitectura LSTM-Decoder.

Tabla 5.
Resultado de las pruebas para las arquitecturas LSTM-Decoder (A-B) y LSTM-LSTM (C-D).

ID	Celdas LSTM	Dropout	Entrenamiento	Validación	Precisión
A1	256-256	NA	0.21±0.06	0.23±0.02	90.30% ±3.6%
A2	256-256	0.2	0.24±0.01	0.26±0.01	93.60% ±2.9%
A3	256-256	0.4	0.49±0.04	0.50±0.04	83.90% ±4.6%
B1	512-512-256-256	NA-NA	0.82±0.46	1.82±0.46	64.50% ±6.6%
B2	512-512-256-256	0.2-0.2	0.94±0.41	1.94±0.41	64.90% ±3.6%
B3	512-512-256-256	0.4-0.4	1.08±0.31	1.80±0.31	64.30% ±2.6%
C1	E(256)-D(256-256)	NA	0.50±0.05	0.55±0.09	90.20% ±1.5%
C2	E(256)-D(256-256)	E(0.2)-D(0.2-0.2)	0.52±0.09	0.51±0.02	93.80% ±2.5%
C3	E(256)-D(256-256)	E(0.4)-D(0.4-0.4)	0.64±0.04	0.64±0.05	83.90% ±5.6%
D1	E(256-256)-D(256-256)	NA	0.28±0.04	0.34±0.06	90.10% ±1.2%
D2	E(256-256)-D(256-256)	E(0.2-0.2)-D(0.2-0.2)	0.46±0.06	0.48±0.03	90.30% ±0.5%
D3	E(256-256)-D(256-256)	E(0.4-0.4)-D(0.4-0.4)	0.77±0.04	0.78±0.02	89.10% ±0.2%

Fuente: Los Autores

3.4 Generación de péptidos

El proceso de generación utiliza la red recurrente para predecir, uno a uno y de manera consecutiva, los aminoácidos que conforman el péptido. El proceso se ejecuta hasta que el péptido alcanza una longitud determinada. Así, el proceso inicia con una cadena que solo contiene la letra "B", el cual es el carácter que marca el inicio de un péptido y termina cuando se obtiene el carácter "_", el cual indica el fin del péptido. En caso de que este carácter se genere antes de alcanzar la longitud deseada para el péptido, el proceso se reinicia para comenzar de nuevo la generación de secuencias.

Para el muestreo de aminoácidos se fijó el factor de temperatura en 0.05. De acuerdo con Müller et al. [19] éste influye en la diversidad del tipo de aminoácidos que se generan para construir el péptido [19]. Entre menor es el factor de temperatura, menor es la diversidad entre los aminoácidos, lo que favorece la probabilidad de que las secuencias que se generan tengan mayor actividad antimicrobiana. Por el contrario, valores mayores para el factor de temperatura producen secuencias con mayor diversidad; sin embargo, la probabilidad de que estas secuencias tengan actividad antimicrobiana se reduce. A pesar de lo anterior, la diversidad es una característica deseable puesto que no tiene sentido generar péptidos compuestos por solo uno o dos aminoácidos distintos.

4. Experimentos, resultados y discusión

En los experimentos se generaron péptidos sintéticos de longitudes cortas (de 12 a 20 aminoácidos) dado que en la práctica este tipo de péptidos son, económicamente, más viables

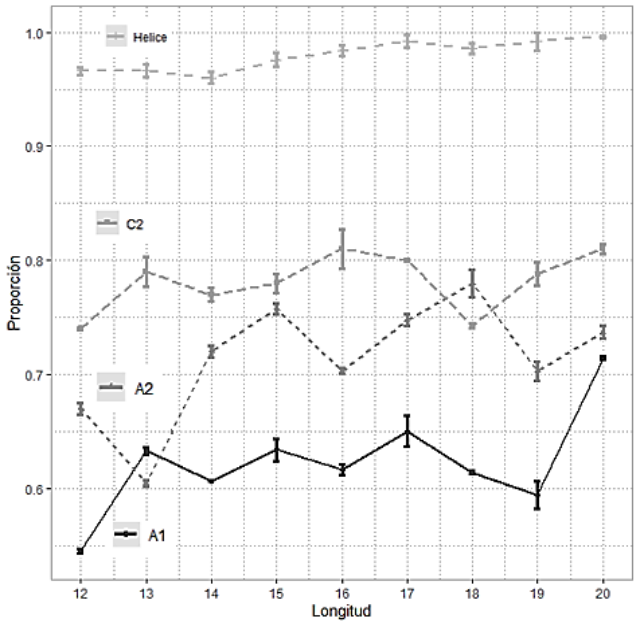

Figura 2. Comparación de la proporción de AMPs generados mediante modelos de aprendizaje profundo y el método de plantillas hélice.
Fuente: Los Autores

de sintetizar. El experimento de generación se repitió 5 veces. En cada repetición se generaron 100 péptidos de cada longitud, es decir, 500 péptidos sintéticos por cada una de las 9 longitudes posibles, para un total de 4500 péptidos sintéticos.

Las secuencias obtenidas se validaron usando el algoritmo de clasificación de actividad antimicrobiana propuesto por Veltri et al. [28] y, al igual que Müller et al. [19], en este trabajo se consideró que un péptido es antimicrobiano cuando su probabilidad de ser "activo" es mayor a 0.5 ($P(AMP) > 0.5$).

Adicionalmente, para comparar la calidad de los péptidos generados por la red recurrente, se procedió a realizar un proceso de generación usando una aproximación basada en plantillas con estructura hélice. Esto se hizo usando la clase modlamp.sequences. Helices del paquete modLamp, el cual fue desarrollado por Müller et al. en 2017 [42]. Luego, de entre todos los péptidos generados, se seleccionaron y cuantificaron aquellos clasificados como antimicrobianos, es decir, aquellos con $P(AMP) > 0.5$. La Fig. 2, muestra la proporción de AMPs generados usando plantilla con estructura hélice (modelo hélice) y los modelos con arquitecturas LSTM-Decoder (A1, A2) y LSTM-LSTM (C2).

De acuerdo con la Fig. 2, al usar una plantilla tipo hélice, aproximadamente el 98.4 ±1.3% de los péptidos generados son AMPs, mientras que los modelos de aprendizaje profundo, en promedio, el 70.8 ±4.2% de los péptidos generados son clasificados como AMPs. Es decir, la generación manual basada en la plantilla hélice es superior a la generación usando los modelos de aprendizaje profundo. Dados estos resultados, se analizaron los péptidos generados con base en sus propiedades fisicoquímicas, las cuales se resumen en la Tabla 6.

Tabla 6.
Propiedades fisicoquímicas de los péptidos sintéticos generados.

Generador	Momento Hidrofóbico	Punto Isoeléctrico	Carga
A1	0.45±0.16	11.30±1.19	4.00±1.75
A2	0.37±0.15	10.80±2.02	2.00±1.51
C2	0.38±0.16	10.80±2.11	2.00±1.58
Hélice	0.89±0.10	12.50±0.45	5.00±0.86

Fuente: Los Autores

La Tabla 6 muestra que el momento hidrofóbico de los AMPs generados con el modelo A1 se encuentra cercano al valor óptimo teórico, el cual es de 0.5 [43]. Para el caso de los AMPs generados mediante una plantilla con estructura de forma hélice, se encuentra que su momento hidrofóbico tiene valores superiores, en promedio de 0.89 ±0.1. Esto indica que los AMPs generados mediante los modelos de aprendizaje profundo pueden formar una estructura de hélice−α anfipática, sin embargo, los AMPs generados mediante una plantilla pueden tener un efecto pronunciado en la actividad hemolítica, es decir que, además de atacar las bacterias, esos péptidos también pueden destruir los glóbulos rojos.

Respecto a la carga eléctrica, los AMPs generados con las redes neuronales recurrentes se encuentran entre los rangos admisibles, es decir entre +1 y +5 [46], lo que permite que los AMPs sean atraídos por la superficie aniónica de las membranas bacterianas. En contraste, los AMPs generados con la plantilla de estructura hélice tienen, en promedio, una carga mayor a +5. En cuanto al porcentaje de hidrofobicidad y el punto isoeléctrico, en su mayoría, los AMPs generados tienen los valores recomendados para estas dos propiedades.

Para estimar la capacidad de generación de AMPs, aumentando la exigencia de clasificación antimicrobiana, consideramos los péptidos con una alta probabilidad de ser antimicrobianos ($P(AMP) > 0.90\ y\ P(AMP) > 0.95$). Adicionalmente, las secuencias generadas se filtraron, con base en las propiedades fisicoquímicas de la siguiente manera.

Solo se consideraron los péptidos con un momento hidrofóbico mayor 0.4 y menor a 0.8. El momento hidrofóbico es una medida cuantitativa de la anfipaticidad de un péptido y es un indicativo de que éste puede formar una estructura de hélice anfipática. Un valor aproximado de 0.5 para esta característica se relaciona con una mayor tendencia a asumir una estructura hélice. Sin embargo, un momento hidrofóbico muy alto puede tener un efecto pronunciado en la actividad hemolítica del péptido [43].

Se descartaron los péptidos con un punto isoeléctrico menor a 9. Como se mencionó antes, un valor alto en esta característica garantiza el carácter catiónico en condiciones de pH fisiológico. Se mantuvieron los péptidos con una carga mayor o igual a +1. La carga es una de las características fisicoquímicas más importantes de los péptidos antimicrobianos, los cuales, en su mayoría, tienen valores entre +1 y +5.

El porcentaje de aminoácidos hidrofóbicos es un parámetro clave para la actividad antimicrobiana de los péptidos. El aumento de la hidrofobicidad del péptido favorece su inserción en la membrana, pero un valor muy alto de esta propiedad aumenta la citotoxicidad en células de mamíferos y disminuye su solubilidad [7]. Con base en esto, se mantuvieron los péptidos con un porcentaje de hidrofobicidad entre el 30 y el 70%.

En la Tabla 7 muestra el promedio de la proporción de péptidos

generados, antes y después de aplicar los filtros sobre las propiedades fisicoquímicas. Las proporciones se tabulan con base en diferentes valores para la probabilidad de ser AMPs. De acuerdo con estos resultados, el número de péptidos que tienen capacidad antimicrobiana disminuye considerablemente cuando se restringen los péptidos de acuerdo con sus propiedades fisicoquímicas. Esto se debe a que a pesar de que los péptidos generados son antimicrobianos, con base en el clasificador de Veltri et al. [28], estos no cumplen con las propiedades esenciales que mejoran su capacidad antimicrobiana. Adicionalmente, se puede observar que después de aplicar los filtros sobre las propiedades de los AMPs generados, el método de generación basado en una plantilla helicoidal es poco efectivo puesto que la proporción de AMPs generados con capacidad antimicrobiana es en promedio mucho más bajo que el método de generación basado en la red neuronal.

Tabla 7.
Proporción de AMPs generados usando un proceso de filtrado con base en las propiedades fisicoquímicas.

Generador	Punto de corte (P (AMP))	¿Con filtro?	Porcentaje de AMPs generados
A1	0.5	NO	61.6±4.6
A2	0.5	NO	72 ±5.3
C2	0.5	NO	78.8 ±2.6
Hélice	0.5	NO	98.4 ±1.3
A1	0.5	SI	38.2 ±11.4
A2	0.5	SI	16.7 ±5.1
C2	0.5	SI	26.2 ±6.6
Hélice	0.5	SI	6.1 ±5.6
A1	0.9	SI	22.7 ±8.9
A2	0.9	SI	12.1 ±4.5
C2	0.9	SI	18.9 ±6.6
Hélice	0.9	SI	5.8 ±5.1
A1	0.95	SI	20.4 ±9.3
A2	0.95	SI	10.7 ±4.5
C2	0.95	SI	18 ±7.1
Hélice	0.95	SI	5.7±4.1

Fuente: Los Autores

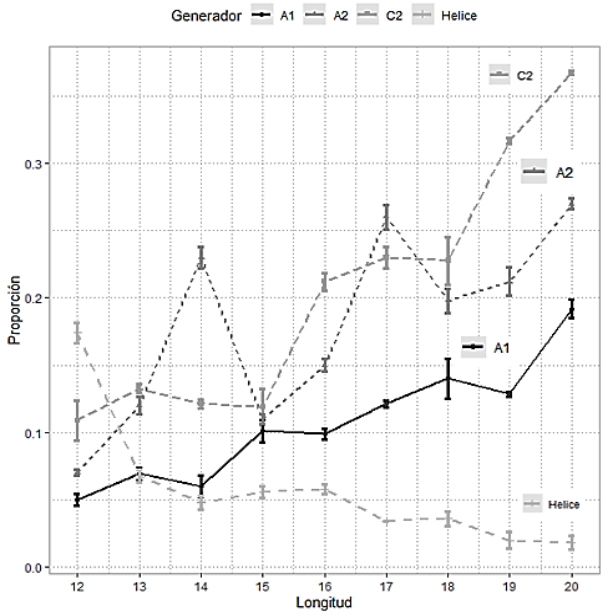

Figura 3. Proporción de AMPs generados por longitud, con P (AMP) >0.95 y con el filtro de las propiedades fisicoquímicas.
Fuente: Los Autores

Finalmente, en la Fig. 3 se muestra la proporción de péptidos generados (con $P(AMP) > 0.95$) tanto con la red recurrente como con el método de plantilla, después de aplicar los filtros, para cada posible longitud definida. De acuerdo con los resultados, el método de generación basado en plantillas tiene una proporción mayor para péptidos cortos, a medida que aumenta el tamaño del péptido su rendimiento decae. Lo contrario sucede con la generación que usa redes recurrentes, esto indica que las redes recurrentes tienen la capacidad de aprender la estructura de composición de un AMP y a partir de ella generar nuevos péptidos sintéticos con capacidad antimicrobiana.

5. Conclusión

Si bien en los últimos años, el aprendizaje profundo para la construcción de AMPs ha empezado a ser ampliamente estudiado, en el contexto de generación de secuencias cortas ha sido poco tratado, uno de los mayores aportes desarrollados constituye el entendimiento del comportamiento de generación de secuencias cortas que tienen entre 12 y 20 aminoácidos.

En este trabajo se utilizaron las redes neuronales recurrentes con arquitecturas LSTM-Decoder y LSTM-LSTM para generar péptidos sintéticos antimicrobianos. Se encontró que, si bien los modelos logran aprender la estructura de composición de los péptidos para generar nuevos AMPs, el análisis de las propiedades fisicoquímicas descarta muchos de ellos.

Esto sugiere que en trabajos futuros se deben considerar las propiedades fisicoquímicas de los péptidos en el proceso de generación como tal, a fin de mejorar la capacidad antimicrobiana de los péptidos sintéticos sin dejar de lado la exploración de nuevas arquitecturas de aprendizaje profundo.

Al comparar los péptidos generados, en términos de sus propiedades fisicoquímicas, encontramos que los péptidos que generan las redes neuronales recurrentes tienen mayor capacidad de ser antimicrobianos, comparados con los péptidos obtenidos a partir de una plantilla tipo hélice. Es por esta razón que después filtrar los péptidos generados, la proporción de AMPs aceptables que genera la red recurrente es superior a la proporción de AMPs que se generan usando la plantilla con estructura helicoidal.

Agradecimientos

Los autores quieren agradecer a la Universidad Nacional de Colombia, Sede Medellín y al Instituto Tecnológico Metropolitano de Medellín (ITM), por el financiamiento del proyectos de investigación con códigos Hermes 42990 y 47318 e ITM P19107 y y PE20202.

Referencias

[1] Fleischmann, C., Scherag, A., Adhikari, N.K., Hartog, C.S., Tsaganos, T., Schlattmann, P., Angus, D.C. and Reinhart, K., Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations, American Jorurnal of Respiratory and Criticatl Care Medicine, 193(3), pp. 259-272, 2016. DOI: 10.1164/rccm.201504-0781OC.

[2] World Health Organization, Antimicrobial resistance, [online]. Feb 2018. Avialable at: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

[3] Fair, R.J. and Tor, Y., Antibiotics and bacterial resistance in the 21st Century. Perspectives in Medicinal Chemistry, 6, pp. 25-64, 2014. DOI: 10.4137/PMC.S14459

[4] Porto, W.F., Pires, A.S. and Franco, O.L., Computational tools for exploring sequence databases as a resource for antimicrobial peptides. Biotechnology Advances, 35(3), pp. 337-349, 2017. DOI: 10.1016/j.biotechadv.2017.02.001

[5] Yoshida, M., Hinkley, T., Tsuda, S., Abul-Haija, Y.M., McBurney, R.T., Kulikov, V., Mathieson, J.S., Galianes-Reyes, S., Castro, M.D. and Cronin, L., Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. Chem, 4(3), pp. 533-543, 2018. DOI: 10.1016/j.chempr.2018.01.005

[6] Brogden, K.A., Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria. Nature Reviews Microbiology, 3, pp. 238-250, 2005. DOI: 10.1038/nrmicro1098

[7] Yeaman, M.R. and Yount, N.Y. Mechanisms of antimicrobial peptide action and resistance. Pharmacological Reviews, 55(1), pp. 27-55, 2003. DOI: 10.1124/pr.55.1.2

[8] Gutierrez, P. and Orduz, S., Péptidos antimicrobianos: estructura, función y aplicaciones. Actualidades Biologicas, 25(78), pp. 5-15, 2003.

[9] Lee, E.Y., Lee, M.W., Fulan, Ferguson, A.L. and Wong, G.C., What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus, 7, pp. 1-14, 2017. DOI: 10.1098/rsfs.2016.0153

[10] Mojsoska, B. and Jenssen, H., Peptides and peptidomimetics for antimicrobial drug design. Pharmaceuticals, 8(3), pp. 366-415, 2015. DOI: 10.3390/ph8030366

[11] Zasloff, M., Antimicrobial peptides of multicellular organisms. Nature, 415(6870), pp. 389-395, 2002. DOI: 10.1038/415389a

[12] Walsh, C.T. and Wencewicz, T.A., Prospects for new antibiotics: a molecule-centered perspective. Journal of Antibiotics, 67, pp. 7-22, 2014. DOI: 10.1038/ja.2013.49

[13] Wieland, T. and Bodanszky, M., World of peptides. A Brief history of peptide chemistry. Academic Press, 1995.

[14] Téllez, G.A. and Castaño, J.C., Péptidos antimicrobianos. Infectio, 14(1), pp. 55-67, 2010. DOI: 10.1016/S0123-9392(10)70093-X

[15] Fjell, C.D., Hiss, J.A., Hancock, R.E. and Schneider, G., Designing antimicrobial peptides: Form follows function, Nature Reviews Drug Discovery. 11, pp. 37-51, 2012. DOI: 10.1038/nrd3591

[16] Loose, C., Jensen, K., Rigoutsos, I. and Stephanopoulos, G., A linguistic model for the rational design of antimicrobial peptides. Nature, 443, pp. 867-869, 2006. DOI: 10.1038/nature05233

[17] Schneider, G., Schuchhardt, J. and Wrede, P., Artificial neural networks and simulated molecular evolution are potential tools for sequence-oriented protein design. Bioinformatics, 10(6), pp. 635-645, 1994. DOI: 10.1093/bioinformatics/10.6.635

[18] Fjell, C.D., Jenssen, H., Cheung, W.A., Hancock, R.E.W. and Cherkasov, A., Optimization of antibacterial peptides by genetic algorithms and cheminformatics. Chemical Biology and Drug Design, 77(1), pp. 48-56, 2011. DOI: 10.1111/j.1747-0285.2010.01044.x

[19] Müller, A.T., Hiss, J.A. and Schneider, G., Recurrent neural network model for constructive peptide design. Journal of Chemical Information and Modeling, 58(2), pp. 472-479, 2018. DOI: 10.1021/acs.jcim.7b00414

[20] Grisoni, F., Neuhaus, C.S., Gabernet, G., Müller, A.T., Hiss, J.A. and Schneider, G., Designing anticancer peptides by constructive machine learning. ChemMedChem, 13(13), pp. 1300-1302, 2018. DOI: 10.1002/cmdc.201800204

[21] Schneider, G., Schuchhardt, J. and Wrede, P., Peptide design in machina: development of artificial mitochondrial protein precursor cleavage sites by simulated molecular evolution. Biophysical Journal, 68(2), pp. 434-447,1995. DOI: 10.1016/S0006-3495(95)80205-5

[22] Fjell, C.D., Jenssen, H., Hilpert, K., Cheung, W.A., Pante, N., Hancock, R.E.W. and Cherkasov, A., Identification of novel antibacterial peptides by chemoinformatics and machine learning. Journal of Medicinal Chemistry, 52(7), pp. 2006-2015, 2009. DOI: 10.1021/jm8015365

[23] Lata, S., Mishra, N.K. and Raghava, G.P., AntiBP2: improved version of antibacterial peptide prediction. BMC Bioinformatics, 11(S19), pp. 1471-2105, 2010. DOI: 10.1186/1471-2105-11-S1-S19

[24] Xiao, X., Wang, P., Lin, W.Z., Jia, J.H. and Chou, K.C., IAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical Biochemistry, 436(2), pp. 168-177, 2013. DOI: 10.1016/j.ab.2013.01.019

[25] Rondon-Villarreal, P., Sierra, D.A. and Torres, R., Classification of antimicrobial peptides by using the p-spectrum kernel and support vector machines. Advances in Intelligent Systems and Computing, 232, pp. 155-160, 2014. DOI: 10.1007/978-3-319-01568-2_23

[26] Waghu, F.H., Gopi, L., Barai, R.S., Ramteke, P., Nizami, B. and Idicula-Thomas, S., CAMP: collection of sequences and structures of antimicrobial peptides. Nucleic Acids Research, 42(D1), pp. 1154-1158, 2014. DOI: 10.1093/nar/gkt1157

[27] Bhadra, P., Yan, J., Li, J., Fong, S. and Siu, S.W., AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Scientific Reports, 8, pp. 1-10, 2018. DOI: 10.1038/s41598-018-19752-w

[28] Veltri, D., Kamath, U. and Shehu, A., Deep learning improves antimicrobial peptide recognition. Bioinformatics, 34(16), pp. 2740-2747, 2018. DOI: 10.1093/bioinformatics/bty179

[29] Das, P., Wadhawan, K., Chang, O., Sercu, T., Santos, C.D., Riemer, M., Chenthamarakshan, V., Padhi, I. and Mojsilovic, A., PepCVAE: Semisupervised targeted design of antimicrobial peptide sequences. [online]. 2018. Available at: http://arxiv.org/abs/1810.07743

[30] Chung, C.-R., Kuo, T.-R., Wu, L.-C., Lee, T.-Y. and Horng, J.-T. Characterization and identification of antimicrobial peptides with different functional activities. Briefings in Bioinformatics, 21(3), pp. 1098-1114, 2020. DOI: 10.1093/bib/bbz043

[31] Lee, H.T., Lee, C.C., Yang, J.R., Lai, J.Z. and Chang, K.Y., A large-scale structural classification of Antimicrobial peptides. BioMed Research International, 2015, pp. 1-6, 2015. DOI: 10.1155/2015/475062

[32] Wang, G., Li, X. and Wang, Z., APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Research, 44(D1), pp. D1087-D1093, 2016. DOI: 10.1093/nar/gkv1278

[33] Theolier, J., Fliss, I., Jean, J. and Hammami, R., MilkAMP: a comprehensive database of antimicrobial peptides of dairy origin. Dairy Science and Technology, 94(2), pp. 181-193, 2014. DOI: 10.1007/s13594-013-0153-2

[34] Zhao, X., Wu, H., Lu, H., Li, G. and Huang, Q., Lamp: a database linking antimicrobial peptides. PLoS ONE, 8(6), pp. 1-6, 06 2013. DOI: 10.1371/journal.pone.0066557

[35] Yen, T.Y., Joshi, R.K., Yan, H., Seto, N.O., Palcic, M.M. and Macher, B.A., Characterization of cysteine residues and disulfide bonds in proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. Journal of Mass Spectrometry, 35(8), pp. 990-1002, 2000. DOI: 10.1002/1096-9888(200008)35:8<990::AID-JMS27>3.0.CO;2-K

[36] Sutskever, I., Vinyals, O. and Le, Q.V., Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, pp. 3104-3112, 2014. DOI: 10.5555/2969033.2969173

[37] Dugar, P., Attention seq2seq models. Towards Data Science, [online]. 2019. Disponible en: https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263.

[38] Gete, H., Neural natural language generation with unstructured contextual information. MSc. Thesis, Universidad del País Vasco, España, 2018.

[39] Tieleman, T. and Hinton, G., Neural Networks for Machine Learning. COURSERA, 2012.

[40] Kingma, D.P. and Ba, J.A., A method for stochastic optimization. International Conference on Learning Representations, [online]. 2014, pp. 1-15. Available at: https://arxiv.org/pdf/1412.6980.pdf

[41] Brownlee, J., Difference between a batch and an epoch in a neural network. machine learning mastery, [online]. 2018. Available at: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

[42] Müller, A.T., Gabernet, G., Hiss, J.A. and Schneider, G., modlAMP: Python for antimicrobial peptides. Bioinformatics, 33(17), pp. 2753-2755, 2017. DOI: 10.1093/bioinformatics/btx285

[43] Dathe, M. and Wieprecht, T., Structural features of helical antimicrobial peptides: their potential to modulate activity on model membranes and biological cells. Biochimica et Biophysica Acta - Biomembranes, 1462(1-2), pp. 71-87, 1999. DOI: 10.1016/S0005-2736(99)00201-1
[44] World Health Organization. The Evolving threat of antimicrobial resistance: options for action. [online]. 2012. Avialable at: https://apps.who.int/iris/handle/10665/44812
[45] Osorio, D., Rondón-Villarreal, P. and Torres, R., Peptides: a package for data mining of antimicrobial peptides. The R Journal, 7(1), pp. 4-14, 2015. DOI: 10.32614/rj-2015-001
[46] Dathe, M., Nikolenko, H., Meyer, J., Beyermann, M., Bienert, M., Optimization of the antimicrobial activity of magainin peptides by modification of charge. FEBS Lett. 501(2), pp. 146-50, 2001.
[47] Su, X., Xu, J., Yin, Y., Quan, X. and Zhang, H., Antimicrobial peptide identification using multi-scale convolutional network. BMC Bioinformatics, 20(730), pp. 1-10, 2019. DOI: 10.1186/s12859-019-3327-y
[48] Yan, J., Bahadra, P., Li, A., Sethiya, P., Quin, L., Tai, H.K., Wong, K.H. and Siu, S.W.I., Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. Molecular Therapy-Nucleic Acids, 20(5), pp. 882-894, 2020. DOI: 10.1016/j.omtn.2020.05.006
[49] Burdukiewicz, M., Sidorczuk, K., Rafacz, D., Pietluch, F., Chilimoniuk, J., Rödiger, S. and Gagat, P., Proteomic Screening for prediction and design of antimicrobial peptides with AmpGram. International Journal of Molecular Sciences, 21(12), 4310, pp. 1-13, 2020. DOI: 10.3390/ijms21124310

A. Vélez, obtuvo su título de pregrado en Estadística de la Universidad Nacional de Colombia, sede Medellín, en 2016 y su título de MSc. en Ingeniería de Sistemas e Informática de la misma Universidad en 2020. Sus intereses de investigación se centran en el uso de Deep learning en campos como el procesamiento, generación de lenguaje natural y bioinformática.
ORCID: 0000-0002-0180-2246

C. Mera, recibió su título de Ing. de Sistemas en 2004 y su título de MSc. en Ingeniería de Sistemas y Computación en 2017, ambos de la Universidad del Valle, Cali, Colombia. También obtuvo el título de MSc. en Sistemas Inteligentes de la Universidad de Salamanca, España, en 2009 y en 2017 recibió el título de Dr. en Ing. de Sistemas e Informática de la Universidad Nacional de Colombia, Sede Medellín. Actualmente es profesor asistente del Instituto Tecnológico Metropolitano (ITM) de Medellín. Sus áreas de interés incluyen la visión artificial y el aprendizaje de máquina.
ORCID: 0000-0002-6513-3053

S. Orduz, obtuvo su título de pregrado en Biología de la Pontificia Universidad Javeriana, Bogotá, Colombia, en 1977. Recibió el título de MSc. en Entomología de North Carolina State University, North Carolina, USA, en 1987 y su grado de PhD. en Ciencias Básicas Biomédicas de la Universidad de Antioquia, Medellín, Colombia, en 1997. Actualmente es profesor titular de la Universidad Nacional de Colombia, sede Medellín, donde trabaja en el desarrollo de nuevos péptidos bioactivos usando estrategias bioinformáticas y computacionales.
ORCID: 0000-0001-7587-3816

J.W. Branch, recibió sus títulos como Ing. de Minas, MSc. en Ingeniería de Sistemas y Dr. en Ingeniería de la Universidad Nacional de Colombia, sede Medellín en 1995, 1997 y 2007, respectivamente. En la actualidad, es profesor titular en el Departamento de Ciencias de la Computación y de la Decisión en la Universidad Nacional de Colombia, Sede Medellín. Sus principales intereses de investigación incluyen visión por computador, reconocimiento de patrones, procesamiento de imágenes y sus aplicaciones en el campo industrial.
ORCID: 0000-0002-0378-028X