Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Doctorado en Ciencias en Ciencias de la Computación

Modelo descriptivo basado en redes de similitud molecular para el análisis visual de un espacio químico-biológico de péptidos bioactivos

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de Doctor en Ciencias

Presenta:

Longendri Aguilera Mendoza

Ensenada, Baja California, México

2020

Tesis defendida por

Longendri Aguilera Mendoza

y aprobada por el siguiente Comité

Dr. Carlos Alberto Brizuela Rodríguez

Codirector de tesis

Dr. Yovani Marrero Ponce

Codirector de tesis

Dr. Israel Marck Martínez Pérez

Dr. Hugo Hidalgo Silva

Dr. Gabriel Del Rio Guerra

Dr. Israel Marck Martínez Pérez

Coordinador del Posgrado en Ciencias de la Computación

Dra. Rufina Hernández Martínez

Directora de Estudios de Posgrado

Resumen de la tesis que presenta Longendri Aguilera Mendoza como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la Computación.

Modelo descriptivo basado en redes de similitud molecular para el análisis visual de un espacio químico-biológico de péptidos bioactivos

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodríguez
Codirector de tesis

Dr. Yovani Marrero Ponce
Codirector de tesis

En la última década, el creciente interés por los péptidos bioactivos con potencial terapéutico se ha reflejado en una gran variedad de repositorios biológicos. Por tal motivo, resulta de provecho científico obtener nuevos conocimientos mediante el análisis de la información que actualmente se encuentra dispersa en fuentes heterogéneas de datos. Sin embargo, el proceso de extracción de conocimiento en bases de datos no es una tarea trivial, por lo que se convierte en la esencia de nuestro esfuerzo de investigación. Para afrontar esta problemática, desarrollamos un flujo de trabajo que emplea el aprendizaje no supervisado para obtener un modelo basado en redes de similitud de péptidos bioactivos. En la primera fase, se realiza una integración de datos basada en grafos para lograr una vista unificada de 40 bases de datos biológicas existentes. Esta colección integrada de 45120 péptidos bioactivos es una de las fuentes de datos más completas y diversas en su campo, hoy en día, con un conocimiento implícito que debe ser descubierto. Es por ello que se calculan descriptores moleculares a partir de los péptidos en estudio, aplicando distintos operadores de agregación a vectores de propiedades de aminoácidos. Luego, se selecciona un subconjunto optimizado de descriptores utilizando los conceptos de entropía de Shannon e información mutua, con el fin de retener los rasgos de alta relevancia y baja redundancia. En esta estrategia de selección de rasgos, se diseñó una función objetivo que constituye uno de nuestros principales aportes para guiar la búsqueda bajo el enfoque no supervisado. Su importancia se debe a que los descriptores optimizados definen un espacio métrico del cual se derivan las redes de similitud molecular, donde los nodos representan péptidos bioactivos, y las aristas denotan sus relaciones de distancia/similitud en el espacio métrico definido. A efectos prácticos, la generación automática de estas redes de similitud se ha implementado en una herramienta informática de análisis visual denominada “starPep toolbox”, permitiendo a los investigadores extraer información útil de la colección integrada de péptidos bioactivos. De esta manera, es posible una representación gráfica y analítica de un espacio químico-biológico ocupado por péptidos bioactivos conocidos hasta la fecha. Además, al combinar técnicas de agrupamiento y análisis de redes con la percepción visual, se aprovechan las habilidades cognitivas del investigador y el poder computacional actual para descubrir patrones ocultos. Finalmente, como caso de estudio, ilustramos la aplicabilidad de la propuesta de minería de datos no supervisada para detectar comunidades e identificar nodos centrales en redes de similitud de péptidos anticancerígenos.

Palabras clave: péptidos bioactivos, descriptores moleculares, aprendizaje no supervisado, selección de rasgos, entropía de Shannon, información mutua, redes de similitud, análisis visual, detección de comunidades, análisis de centralidad

Abstract of the thesis presented by Longendri Aguilera Mendoza as a partial requirement to obtain the Doctor of Science degree in Computer Science.

A descriptive model based on similarity networks for supporting visual analysis in a chemical-biology space of bioactive peptides

Abstract approved by:

Dr. Carlos Alberto Brizuela Rodríguez
Thesis Co-Director

Dr. Yovani Marrero Ponce
Thesis Co-Director

In the last decade, the growing interest in bioactive peptides with therapeutic potential has been reflected in a wide variety of biological repositories. Consequently, there is a scientific benefit in analyzing the currently dispersed information in heterogeneous data sources. However, the process of knowledge discovery in databases is a nontrivial task, so it becomes the essence of our research endeavor. To face these issues, we developed a workflow that uses unsupervised learning to obtain a model based on similarity networks of bioactive peptides. A graph-based data integration was performed in the first stage to achieve a unified view of 40 existing biological databases. This integrated collection of 45120 bioactive peptides is one of the most comprehensive and diverse data sources in its field nowadays, with an implicit knowledge that must be discovered. That is why molecular descriptors are calculated from the peptides understudy, applying different aggregation operators to vectors of amino acid properties. An optimized subset of descriptors is then selected using the concepts of Shannon entropy and mutual information, in order to retain descriptors having high relevance and low redundancy between them. In this feature selection strategy, an objective function was devised that constitutes one of our main contributions under the unsupervised approach. Its importance is due to the fact that the optimized descriptors define a metric space from which the molecular similarity networks are derived. In these networks, nodes represent bioactive peptides, and the edges denote their distance/similarity relationships in the defined metric space. For practical purposes, the automatic generation of these similarity networks has been implemented in a visual analytics software tool called “starPep toolbox”, enabling researchers to extract useful information from the integrated collection of bioactive peptides. In this way, a graphical and analytical representation of a chemical-biological space occupied by bioactive peptides known to date is possible. Furthermore, by combining clustering and network analysis techniques with visual perception, the researcher’s cognitive skills and current computational power are harnessed to uncover hidden patterns. Finally, as a case study, we illustrate the applicability of the unsupervised data mining proposal to detect communities and identify central nodes in similarity networks of anticancer peptides.

Keywords: bioactive peptides, molecular descriptors, unsupervised learning, feature selection, Shannon entropy, mutual information, similarity networks, visual analysis, community detection, centrality analysis

Dedicatoria

A mi esposa, mis hijos, y mi madre.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada por darme la oportunidad de ser egresado de tan prestigiosa institución.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de doctorado. No. de becario: 600824.

Al apoyo otorgado mediante el proyecto FORDECYT 296737 “Consorcio en Inteligencia Artificial” para el desarrollo de la presente tesis.

A mis asesores de tesis, Dr. Carlos Alberto Brizuela Rodríguez y Dr. Yovani Marrero Ponce, por permitirme llevar a cabo la presente investigación, aportando sus conocimientos, tiempo y dedicación. También, a los miembros del comité de tesis, Dr. Israel Marck Martínez Pérez, Dr. Hugo Hidalgo Silva, y Dr. Gabriel Del Rio Guerra, por sus valiosas opiniones que me ayudaron a reflexionar en la forma de presentar este proyecto de tesis doctoral.

A mis compañeros del posgrado por las sesiones de debate profesional y los momentos de ocio compartidos. También, a los doctores Edgar Chávez, Cesar García-Jacas, y Hugo Guillen-Ramírez, que en las postrimerías de este documento final ayudaron con sus comentarios y observaciones.

En especial, a mi esposa Mónica por hacer suyas mis alegrías y apoyarme en los momentos difíciles. Ella también ha dedicado estos cuatro años de su vida a esta tesis, por el tiempo que le he robado con las horas de desvelo invertidas en este proyecto y que no he podido dedicarle. Tampoco puedo dejar de mencionar a mi mamá, que ella más que nadie sabe lo que nos ha costado llegar a este momento, luego de tantos años de estudio y sacrificio. Mima, esta tesis también es tuya. Por último, a mis niños Malena y Mauro, que me dieron las fuerzas para seguir adelante con ímpetu, y con la ilusión de que algún día vean el granito de arena que dio su papá en los quehaceres científicos.

Tabla de contenido

Página

Resumen en español ii

Resumen en inglés iii

Dedicatoria iv

Agradecimientos v

Lista de figuras ix

Lista de tablas xi

Capítulo 1. Introducción

1.1. Antecedentes 2
1.1.1. Espacio químico-biológico 3
1.2. Planteamiento del problema 4
1.2.1. Objetivos de la investigación 6
1.2.1.1. Objetivo general 6
1.2.1.2. Objetivos especificos 7
1.3. Contribuciones 7
1.4. Organización de la tesis 8

Capítulo 2. Marco Teórico

2.1. Péptidos bioactivos 10
2.1.1. Potencialidades terapéuticas 11
2.1.2. Repositorios biológicos 12
2.2. Explorando un espacio químico-biológico 13
2.2.1. Caracterización mediante descriptores moleculares 15
2.2.2. Identificación y selección automática de descriptores 18
2.2.3. Relaciones de (di)similitud entre los péptidos 20
2.2.4. Representación visual 21
2.2.4.1. Mapas basados en coordenadas 22
2.2.4.2. Representación basada en grafos 22
2.3. Ciencia de redes: marcando una nueva era 23
2.3.1. Definición formal de una red 24
2.3.2. Modelos de redes complejas 24
2.3.2.1. Redes de mundo pequeño 26
2.3.2.2. Redes libres de escala 26
2.3.3. Trabajos relacionados 27
2.4. Conclusiones parciales 28

Capítulo 3. Metodología

3.1. Recopilación e integración de datos 31
3.1.1. Flujo de extracción, transformación y carga 32
3.2. Selección, limpieza y transformación 35
3.2.1. Espacio de descriptores 35

Tabla de contenido (continuación)

3.3. Minería de datos . . . 38
3.3.1. Selección no supervisada de rasgos . . . 38
3.3.1.1. Fase I: filtrado inicial de descriptores . . . 41
3.3.1.2. Fase II: optimización del conjunto candidato . . . 42
3.3.2. Construcción del modelo basado en redes de similitud . . . 43
3.3.2.1. Detección de comunidades . . . 45
3.3.2.2. Identificando los nodos más relevantes . . . 47
3.4. Evaluación e interpretación . . . 49
3.4.1. Evaluación de los descriptores calculados . . . 49
3.4.2. Evaluación de los descriptores seleccionados . . . 50
3.4.3. Interpretación del modelo basado en redes de similitud . . . 51
3.5. Difusión y uso . . . 52
3.5.1. Herramienta informática “starPep toolbox” . . . 53
3.5.1.1. Principales funcionalidades . . . 53
3.6. Conclusiones parciales . . . 57

Capítulo 4. Resultados y discusión
4.1. Base de datos integrada de péptidos bioactivos . . . 60
4.2. Estudio comparativo de los descriptores calculados . . . 61
4.3. Evaluando y ajustando el filtrado de descriptores . . . 65
4.4. Evaluando el subconjunto optimizado de descriptores . . . 67
4.5. Generando redes de similitud molecular . . . 69
4.6. Caso de estudio: navegando un espacio químico-biológico de péptidos anticancerígenos . . . 72
4.6.1. Exploración y detección de comunidades . . . 74
4.6.2. Identificación de compuestos biológicamente relevantes . . . 74
4.6.3. Identificación de péptidos relevantes pero no redundantes . . . 77
4.6.4. Proyección de nuevos compuestos en el modelo descriptivo . . . 79
4.7. Conclusiones parciales . . . 80

Capítulo 5. Conclusiones y trabajo futuro
5.1. Sumario . . . 82
5.2. Conclusiones generales . . . 83
5.3. Propuesta de trabajo futuro . . . 83

Literatura citada . . . 86

Anexo A. Algoritmo de filtrado inicial de descriptores . . . 105

Anexo B. Algoritmo de optimización del conjunto candidato . . . 106

Anexo C. Algoritmo para construir la red HSP . . . 107

Lista de figuras

Figura Página

1. Estructura general de los 20 aminoácidos esenciales 10
2. Formación de un enlace peptídico entre dos aminoácidos 11
3. Línea de tiempo de los repositorios biológicos 13
4. Ciclo de exploración y evaluación en el diseño *in silico* de nuevos candidatos a fármacos. 14
5. Esquema m-dimensional basado en descriptores de un EQB 15
6. Visualización de un espacio químico-biológico 22
7. Proceso de extracción de conocimiento que ha sido llevado a cabo en la presente investigación. 30
8. Modelo de red basado en un esquema en estrella. 33
9. Esquema de selección de rasgos basado en dos etapas. 39
10. Relación entre la entropía $H(\cdot)$ y la información mutua $I(\cdot,\cdot)$ 39
11. Entropía de Shannon (H) calculada para dos rasgos a partir de un esquema de discretización 40
12. Ejemplo de construcción de la red HSP 45
13. Ilustración de la densidad de red a distintos umbrales de similitud 52
14. Captura de pantalla de starPep toolbox v0.8. 54
15. Diagrama de flujo para guiar la construcción automática y el análisis visual de las redes de similitud molecular 59
16. Histograma acerca de los péptidos recopilados e integrados 60
17. Mapa de calor para visualizar el traslape entre bases de datos 61
18. Distribución de entropía de Shannon para los descriptores de iFeature versus los descriptores de starPep. 63
19. Diagramas de caja que muestran las distribuciones de la entropía de Shannon para los descriptores de iFeature versus los descriptores de starPep 64
20. Explorando el efecto de cambiar el umbral de similitud 66
21. Diagrama de caja de los valores de mérito para diferentes subconjuntos de descriptores. 69
22. Número de conjuntos de datos distintos que incluyen un descriptor seleccionado. 70
23. Densidad de la red versus umbral de similitud 71
24. Visualización de la red HSP para un conjunto amplio de péptidos bioactivos 72
25. Visualización de la red de similitud de péptidos anticancerígenos 73

Lista de figuras (continuación)

Figura Página

26. Visualización de la subred que contiene el Top 1000 y el Top 500 75
27. Un acercamiento a una comunidad detectada para resaltar el nodo más central 76
28. Visualización de la subred de péptidos anticancerígenos 78
29. Visualización del grafo de los *k* vecinos más cercanos 80
30. Ciclo de visualización, exploración y evaluación en el diseño e identificación *in silico* de nuevos candidatos a fármacos. 84

Lista de tablas

Tabla Página

1. Lista de los 20 aminoacidos esenciales 10
2. Bases de datos de péptidos bioactivos 31
3. Etiqueta de los nodos de metadatos y sus relaciones 33
4. Propiedades fisicoquímicas que fueron calculadas. 36
5. Propiedades de la cadena lateral de los aminoácidos. 37
6. Operadores de agregación. 38
7. Explorando el efecto de cambiar el método de correlación y el valor de corte de similitud 66
8. Optimización del subconjunto de rasgos 67
9. Comparación entre los valores de mérito de los subconjuntos optimizados y los descriptores sin optimizar 68
10. Redes HSP conexas generadas con el umbral de similitud $t = 0$. 70
11. Una familia de péptidos centrales en el interior de una comunidad . . 76

Capítulo 1. Introducción

La era moderna en el uso de los antibióticos comienza en el siglo XX con el descubrimiento de compuestos pioneros que sentaron las bases para futuras investigaciones en el campo de la medicina (Zaffiri *et al.*, 2012; Gaynes, 2017). No obstante, a pesar del desarrollo alcanzado a lo largo de todos estos años, aún existen agentes patógenos que han tenido el "don natural" de adaptarse y ser resistentes a los fármacos convencionales que se han creado (Huttner *et al.*, 2013; Aslam *et al.*, 2018). A tal punto, que en la actualidad continúa latente la creación de nuevos tipos de fármacos para combatir microorganismos causantes de determinadas enfermedades infecciosas, como puede ser el virus SARS-CoV-2, causante de la enfermedad COVID-19 (Yang y Wang, 2020) y la pandemia mundial que hoy vivimos.

Una alternativa posible lo constituyen los péptidos con actividad biológica capaz de ejercer un efecto terapéutico en el organismo (Lau y Dunn, 2018; Henninot *et al.*, 2018; Angell *et al.*, 2018). Estos péptidos bioactivos pueden ser de origen natural o sintético. Los de origen natural son obtenidos directamente de un organismo, mientras que los sintéticos pueden ser análogos a los naturales pero producidos químicamente con posibles modificaciones. Por ejemplo, péptidos antimicrobianos (AMPs, por sus siglas en inglés) de origen natural pueden ser aislados de diversos organismos, que van desde las bacterias hasta los mamíferos, incluidos los humanos, y modificados químicamente para combatir a varios agentes patógenos (Wang, 2017). Dichos AMPs constituyen un mecanismo natural de defensa, lo cual ha inspirado a la comunidad científica y a la industria farmacéutica en la búsqueda de nuevos tipos de fármacos basados en esta clase de compuestos (Chen y Lu, 2020).

Sin lugar a dudas, el uso clínico de nuevos péptidos como agentes terapéuticos es un campo prometedor (Lau y Dunn, 2018; Henninot *et al.*, 2018; Angell *et al.*, 2018). Además de la actividad antimicrobiana, algunos péptidos han mostrado tener un enorme potencial como productos anticancerígenos (Felício *et al.*, 2017; Zhang *et al.*, 2019; Hilchie *et al.*, 2019). Otro posible uso de los péptidos es como vehículos para transportar fármacos hacia un destino específico, ya que algunos de ellos tienen la capacidad de atravesar la membrana celular (Wang *et al.*, 2014; Ramsey y Flynn, 2015). Incluso, péptidos sintéticos se emplean con la intención de encontrar vacunas contra determinadas enfermedades infecciosas y otras degenerativas como el Alzheimer y cáncer

(Malonis *et al.*, 2019), así como para combatir al virus SARS-CoV-2 (Bhattacharya *et al.*, 2020; Kalita *et al.*, 2020).

A pesar del interés suscitado, el camino no está libre de obstáculos. Una de las principales desventajas que tienen los péptidos bioactivos es su baja estabilidad metabólica, teniendo efectos biológicos de corta duración, consecuencia de la rápida degradación por peptidasas (Lau y Dunn, 2018; Henninot *et al.*, 2018). Por otra parte, el desarrollo de un nuevo fármaco es un proceso complejo que puede durar más de una década, desde el surgimiento de la idea inicial hasta que el producto final es lanzado en el mercado, y con un costo que puede alcanzar los millones de dólares (Sertkaya *et al.*, 2016; Moore *et al.*, 2018). Para ilustrar lo complejo que puede ser el proceso de descubrimiento de un nuevo fármaco, un estudio realizado (Lau y Dunn, 2018) muestra que hasta marzo de 2017, solo 60 péptidos habían sido aprobados para su uso clínico en el mercado farmacéutico de Estados Unidos, Europa y Japón, 155 se encontraban en fases clínicas, 8 habían sido retirados, y 261 fueron descontinuados sin llegar a ser aprobados. Otros estudios más recientes (Koo y Seo, 2019; Chen y Lu, 2020; Bhopale, 2020) igualmente muestran que aún no se ha llegado a la cúspide en este campo de la medicina. Razón por la cual es necesario seguir dedicando esfuerzos para encontrar los fármacos peptídicos con mayor eficacia.

1.1. Antecedentes

El proceso de descubrimiento de fármacos transita por varias etapas: investigación, ensayos preclínicos y clínicos, así como la aprobación y seguimiento por parte de una entidad regulatoria, por ejemplo, la FDA en Estados Unidos (`https://www.fda.gov`). En cada una de estas etapas existe un riesgo de que el compuesto químico no cumpla los requisitos necesarios para seguir adelante con el proceso. En tal caso, es posible que se tenga que abandonar el candidato a fármaco en el cual se ha realizado una inversión importante. De ahí que sean necesarios los métodos de Aprendizaje Automatizado (en inglés *Machine Learning*) durante la etapa inicial, donde está enmarcado este trabajo, para contribuir a la toma de decisiones en la identificación de compuestos candidatos a fármacos peptídicos (Wu *et al.*, 2019; Basith *et al.*, 2020).

En la etapa inicial del diseño de fármacos, una valiosa fuente de información son

los repositorios biológicos con datos de péptidos bioactivos que han sido publicados en la última década (Porto *et al.*, 2017; Usmani *et al.*, 2018a). Naturalmente, los expertos del dominio necesitan extraer conocimientos ocultos en esas fuentes heterogéneas de datos para su uso posterior. Esto con el propósito de explorar y evaluar "*in silico*" un gran número de sustancias orgánicas; evitando así los altos costos experimentales ("*in vitro*" o "*in vivo*") de aquellos compuestos que, según el conocimiento extraído, no resulten de interés farmacológico. Un ejemplo sería buscar moléculas que tengan una actividad biológica deseada dentro de colecciones de compuestos existentes (Henninot *et al.*, 2018). Sin embargo, identificar un compuesto candidato a fármaco es un objetivo extremadamente complejo, que requiere por parte del investigador un profundo análisis de los resultados obtenidos por los métodos de cómputo.

En cuanto al uso del aprendizaje automatizado, se puede afirmar que los modelos predictivos de la actividad biológica que se precisa, construidos con entrenamiento supervisado, son los que mayormente han sido reportados en la literatura para identificar moléculas peptídicas sobre las cuales se deben centrar las pruebas experimentales (Maurya *et al.*, 2019; Torres y de la Fuente-Nunez, 2019; Lee *et al.*, 2019). A pesar de ello, la falta de efectividad en la búsqueda de nuevos fármacos basados en péptidos denota la complejidad de este proceso. Se conoce que son muchos los retos a enfrentar, pero los mismos fomentan la creación y el mejoramiento de métodos para apoyar la adecuada selección de compuestos candidatos con mayor efectividad, menor índice de toxicidad, y pocos o nulos efectos colaterales indeseables. Es de esperar que este proceso se acelere gracias al desarrollo y uso de los métodos computacionales, en la medida que permitan examinar un espacio químico-biológico de péptidos bioactivos (Lipinski y Hopkins, 2004).

1.1.1. Espacio químico-biológico

El hecho de que la naturaleza no cubre el amplio espectro de todos los compuestos orgánicos posibles, ha dado lugar a que se manejen términos como el de "espacio químico-biológico" (EQB). La noción del EQB, en su totalidad, sería el conjunto de todas las moléculas orgánicas posibles, es decir, tanto aquellas que existen de forma natural como las que son sintéticamente accesibles, ya sean compuestos químicos de pequeño a mediano tamaño o macromoléculas (Dobson, 2004). Sin embargo, en lugar

de considerar todo el EQB, que teóricamente es vasto, resulta factible trabajar con un subespacio químico-biológico conocido (Reymond y Awale, 2012), como por ejemplo, los péptidos bioactivos reportados hasta la fecha.

En primer lugar, para el estudio de un EQB, es útil caracterizar a cada compuesto mediante un conjunto de rasgos numéricos, llamados descriptores moleculares (Todeschini y Consonni, 2009; Jenssen, 2011). De este modo, a cada compuesto del EQB le corresponde un punto en un espacio multidimensional, que requiere una adecuada técnica de reducción de dimensionalidad para una visualización efectiva en un mapa ($2D$ o $3D$) basado en coordenadas (Opassi *et al.*, 2018). Como alternativa libre de coordenadas, se ha propuesto un esquema diferente basado en redes de similitud (Maggiora y Bajorath, 2014; Vogt *et al.*, 2016), donde los nodos de la red representan entidades moleculares, y las aristas indican relaciones de similitud entre los compuestos.

Un punto clave a favor de las redes de similitud se basa en la posible aplicación de algoritmos basados en grafos, que junto con el análisis visual (von Landesberger *et al.*, 2011) pueden dotar al experto del dominio de información útil para asistir al descubrimiento de fármacos. Por un lado, la visualización interactiva de datos (Shneiderman, 1996; Ware, 2019) facilita incorporar el factor humano con mayor protagonismo en el proceso de extracción de conocimientos (Holzinger *et al.*, 2014; Holzinger, 2016). Por otro lado, el área de ciencia de redes (Estrada, 2012; Newman, 2018) ofrece un enfoque moderno, con base en la teoría de grafos, que permite representar y analizar determinadas relaciones de similitud entre los péptidos a ser estudiados (Csermely *et al.*, 2013; Recanatini y Cabrelle, 2020).

1.2. Planteamiento del problema

En la actualidad existen microbios con una capacidad acentuada de causar enfermedad y mortalidad debido a que ofrecen altos niveles de resistencia a los antibióticos convencionales (Huttner *et al.*, 2013; Aslam *et al.*, 2018). Este mecanismo de resistencia ha llevado a la búsqueda de una nueva alternativa para combatirlos, siendo el uso de péptidos bioactivos una posible solución (Lau y Dunn, 2018; Henninot *et al.*, 2018; Angell *et al.*, 2018). Por lo que aún está latente la necesidad de descubrir nuevos fármacos basados en esta clase de compuestos, capaces de combatir determinadas

enfermedades infecciosas. En ese contexto, las técnicas asistidas por computadoras aún se continúan desarrollando, y juegan un papel crucial en la identificación *in silico* de aquellos compuestos candidatos que deben ser sintetizados y evaluados experimentalmente.

Una forma de apoyar la toma de decisiones, durante la identificación *in silico* de compuestos candidatos a fármacos, es explotando la información que actualmente se encuentra dispersa en diversos repositorios biológicos de péptidos bioactivos. En ese sentido, el enfoque tradicional que ha sido ampliamente utilizado es aplicar el aprendizaje supervisado para construir modelos capaces de predecir la actividad biológica que se desea, a partir de un conjunto de datos de entrenamiento, etiquetados para la clase positiva y negativa (Usmani *et al.*, 2018a; Wu *et al.*, 2019; Basith *et al.*, 2020). Sin embargo, el principal inconveniente que se tiene en este campo de estudio radica precisamente en la calidad del etiquetado de las instancias, lo cual repercute en la preparación de los datos para garantizar con eficacia la construcción de un modelo predictivo con mayor dominio de aplicación.

En primer lugar, son pocas las muestras negativas validadas experimentalmente (Basith *et al.*, 2020). Ante esta situación, se generan casos negativos de forma azarosa, ya sean péptidos formados aleatoriamente, o seleccionados porque tienen actividades biológicas diferentes en lugar de la función deseada (Gabere y Noble, 2017). Este enfoque ha sido aceptado en la práctica, bajo el supuesto de que es poco probable que una muestra se presente como un falso negativo, pero sin absoluta certeza de que todas las muestras consideradas son verdaderos negativos.

En segundo lugar, los repositorios biológicos de donde se recuperan los casos positivos no están exentos de incertidumbre y poca uniformidad en el etiquetado de sus datos. Estos se recopilan de artículos científicos publicados por diferentes grupos y laboratorios, que utilizan, por ejemplo, diferentes protocolos para evaluar experimentalmente a los péptidos y contabilizarlos como activos. Aunque algunos intentos han sido realizados para hacer frente a estos inconvenientes (Nagarajan *et al.*, 2019), aún no son suficientes los datos etiquetados cuyos resultados sean obtenidos y comparables siguiendo los mismos protocolos de experimentación. Además, en general, es bien conocido que un mismo péptido puede presentar más de una actividad biológica, por lo cual la naturaleza intrínseca del problema sugiere la construcción de un modelo

de clasificación multietiqueta (Gull *et al.*, 2019). Sin embargo, en la actualidad no se cuenta con un balance de clases en el etiquetado de los datos, necesario para abordar este último enfoque.

Teniendo en cuenta todo lo anterior, el área de aprendizaje no supervisado puede complementar los esfuerzos de investigación realizados en este campo. Sobre todo para inferir un conocimiento previamente desconocido a partir de la utilización de un mayor número de péptidos bioactivos existentes. Su importancia se debe a que con el aprendizaje no supervisado es posible construir un modelo que se ajuste a los datos, sin necesidad de que las instancias a *priori* tengan que estar etiquetadas. Por lo tanto, para la labor investigativa se plantea como **problema científico** a resolver: ¿Cómo obtener nuevos conocimientos a través del aprendizaje no supervisado, que puedan ser útiles para apoyar la toma de decisiones en la identificación *in silico* de compuestos candidatos a fármacos peptídicos? De este problema general se desglosan las siguientes preguntas de investigación:

- ¿Cuáles son los péptidos bioactivos existentes que forman un espacio-químico biológico conocido de interés farmacológico?
- ¿Cómo representar el espacio químico-biológico conocido de péptidos bioactivos bajo el enfoque no supervisado?
- ¿Será posible descubrir nuevos conocimientos a partir de la representación realizada del espacio químico-biológico de péptidos bioactivos?

1.2.1. Objetivos de la investigación

Para dar respuesta al problema científico y las preguntas de investigación, se trazan los siguientes objetivos.

1.2.1.1. Objetivo general

Desarrollar un modelo basado en redes de similitud molecular de péptidos bioactivos, por medio de un enfoque de aprendizaje no supervisado, con la finalidad de

descubrir patrones útiles en la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

1.2.1.2. Objetivos especificos

- Crear una nueva base de datos que sea la más completa y diversa de péptidos bioactivos conocidos hasta la fecha, a partir de la recopilación e integración de datos procedentes de varios repositorios biológicos de interés farmacológico.
- Proponer un método no supervisado para identificar un conjunto optimizado de descriptores moleculares, que caracterice de forma adecuada a los péptidos bioactivos recopilados e integrados para su posterior análisis.
- Obtener el modelo basado en redes de similitud molecular, a partir del conjunto optimizado de descriptores, para representar un espacio químico-biológico conocido de péptidos bioactivos que permita descubrir patrones mediante técnicas de agrupamiento y análisis de redes.
- Desarrollar una herramienta informática para automatizar la generación, de forma gráfica y analítica, del modelo basado en redes de similitud molecular, y asistir en la detección de patrones que sean útiles en la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

1.3. Contribuciones

Como **novedad científica**, se desarrolla un flujo de trabajo (Figura 15) empleando el aprendizaje no supervisado para obtener un modelo basado en redes de similitud molecular, que posibilita descubrir patrones ocultos mediante técnicas de agrupamiento y análisis de redes. Estos patrones a su vez son novedosos, interpretables, y potencialmente útiles en la identificación *in silico* de compuestos biológicamente relevantes en un espacio químico-biológico de péptidos bioactivos (Sección 4.6). También, se plantea una nueva función objetivo modificada para formular un problema de optimización (Ecuación 12), como método de selección de rasgos bajo el enfoque no supervisado.

Como **valor metodológico**, en el flujo de trabajo propuesto se combinan elementos matemáticos tales como: operadores de agregación (Sección 3.2.1), conceptos de entropía de Shannon e información mutua (Sección 3.3.1), y métricas de centralidad provenientes del área de ciencia de redes (Sección 3.3.2.2). Asimismo, el enfoque de aprendizaje no supervisado (Sección 3.3) se puede aplicar en el análisis de un espacio químico-biológico ocupado por otros tipos de moléculas, como son las proteínas, ADN y ARN.

Como **aporte práctico**, se obtuvo una base de datos integrada de péptidos bioactivos que es útil para el análisis de un espacio químico-biológico conocido de interés farmacológico (Sección 3.1). Además, se desarrolló la herramienta informática starPep toolbox (`http://mobiosd-hub.com/starpep`), basada en la política de software libre, para facilitar la utilización de la base de datos obtenida (Sección 3.5.1). Con esta nueva herramienta, se pretende contribuir en la toma de decisiones durante la identificación *in silico* de nuevos compuestos candidatos a fármacos peptídicos, según el flujo de trabajo propuesto en la investigación.

1.4. Organización de la tesis

Los cuatro capítulos restantes de este documento de tesis están estructurado de la siguiente manera:

En el Capítulo 2, se presenta una introducción a los péptidos bioactivos de interés en el estudio, y se abordan los conceptos relacionados con el estudio matemático de un EQB ocupado por estos compuestos orgánicos. También, se realiza una reseña bibliográfica en donde se hace un análisis crítico de la literatura que está relacionada con el tema tratado de ciencia de redes.

En el Capítulo 3, se describe como se llevó a cabo el proceso que va desde la recopilación de los datos hasta la construcción y análisis del modelo basado en redes de similitud molecular. Este capítulo le permitirá al lector comprender el estudio realizado, incluyendo como se calculan y optimizan los descriptores moleculares. Asimismo, se exhiben los algoritmos y la métrica de centralidad para generar y analizar las redes de similitud, los cuales se encuentran implementados en la herramienta informática

presentada para tal propósito.

Seguidamente, la primera parte del Capítulo 4 está dedicada a la experimentación realizada para evaluar la calidad de los descriptores calculados y optimizados. Además, se exploran y ajustan algunos parámetros del algoritmo diseñado para la selección automática de rasgos. En la segunda parte, se prueba la efectividad de un algoritmo implementado para generar redes cuando se trata de grandes conjuntos de datos. Por último, se discuten los resultados alcanzados en relación a un caso de estudio específico de minería de datos, empleando el flujo de trabajo propuesto para detectar comunidades y descubrir nodos centrales en redes de similitud de péptidos anticancerígenos.

Finalmente, el Capítulo 5 muestra las conclusiones y recomendaciones sugeridas para darle una continuidad a la presente investigación.

Capítulo 2. Marco Teórico

2.1. Péptidos bioactivos

Los péptidos son compuestos orgánicos que se forman a partir de la unión de aminoácidos (AAs), y esa unión se establece a través de un enlace covalente que recibe el nombre de enlace peptídico (Nelson y Cox, 2017). Un punto a destacar es que se ha encontrando que son 20 los tipos de AAs (Tabla 1) que comúnmente constituyen a estos compuestos orgánicos en la naturaleza (Nelson y Cox, 2017). Incluso, más notable aún es que las células de los organismos vivos pueden producir polímeros, con distintas propiedades y actividades biológicas, que estén unidos por enlaces peptídicos a partir de los mismos 20 AAs pero en varias combinaciones y secuencias diferentes (Nelson y Cox, 2017).

Tabla 1. Los nombres de los 20 aminoácidos más comunes en la naturaleza y sus respectivas codificaciones en letras.

Alanina	Ala	A	Leucina	Leu	L
Arginina	Arg	R	Lisina	Lys	K
Asparagina	Asn	N	Metionina	Met	M
Ácido Aspartico	Asp	D	Fenilalanina	Phe	F
Cisteina	Cys	C	Prolina	Pro	P
Ácido Glutámico	Glu	E	Serina	Ser	S
Glutamina	Gln	Q	Treonina	Thr	T
Glycina	Gly	G	Triptofano	Trp	W
Histidina	His	H	Tirosina	Tyr	Y
Isoleucina	Ile	I	Valina	Val	V

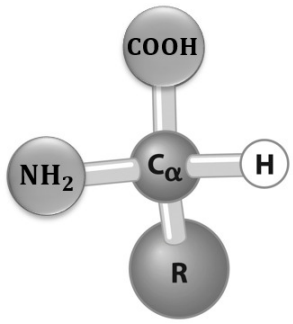


Figura 1. Estructura general de los 20 aminoácidos esenciales. El grupo R (cadena lateral) es lo que los hace diferente a cada uno de ellos.

Todos los 20 AAs esenciales encontrados en la naturaleza tienen en común una estructura general (Figura 1). En el centro de la estructura está el carbono alfa (C_α),

y este se enlaza con cuatro elementos: hidrógeno (*H*), grupo carboxilo (*COOH*), grupo amino (NH_2), y el residuo aminoacídico (*R*) que es lo que los distingue a cada uno de ellos (Nelson y Cox, 2017). Considerando a estos 20 AAs como bloques de construcción, lo que ocurre al unirse dos de ellos a través de un enlace peptídico, es que el grupo carboxilo de un primer aminoácido se une al grupo amino de un segundo aminoácido, liberándose una molécula de agua (Figura 2), y así sucesivamente se va formando la cadena peptídica.

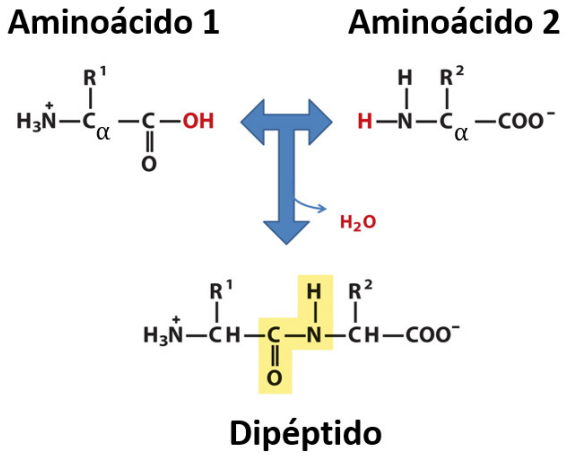


Figura 2. Formación de un enlace peptídico entre dos aminoácidos

En el péptido resultante, el aminoácido con el grupo amino libre H_3N^+ es conocido como el N-terminal (amino-terminal), y el residuo en el otro extremo con el grupo carboxilo libre COO^- es el C-terminal (carboxilo terminal). Por convención, cuando se muestra una secuencia de aminoácidos, el extremo N-terminal se coloca a la izquierda mientras que el extremo C-terminal queda ubicado a la derecha. Por lo tanto, la secuencia se lee de izquierda a derecha, comenzando con el extremo amino terminal.

2.1.1. Potencialidades terapéuticas

A partir de la creación de un péptido, la secuencia de AAs sería la estructura primaria que adopta una estructura tridimensional para una actividad biológica efectiva (Nelson y Cox, 2017). Una forma peculiar de hacerlo es la que permite la interacción del péptido con la membrana celular del patógeno invasor, con el fin de atacar y destruirlo (Madani *et al.*, 2011; Guidotti *et al.*, 2017). Por ejemplo, es una realidad objetiva que los organismos vivos están expuestos constantemente a otros microorganismos

que representan un riesgo para la vida, y se ha comprobado que vertebrados e invertebrados, como un mecanismo intrínseco de defensa, producen AMPs para enfrentar en la primera línea de combate a diversos agentes patógenos (Wang, 2017; Kumar *et al.*, 2018; Chen y Lu, 2020).

Los AMPs son considerados una alternativa viable para combatir la resistencia que ofrecen algunos microorganismos a los antibióticos convencionales (Ghosh *et al.*, 2019). Además, no solo son útiles para hacer frente a las enfermedades infecciosas causadas por bacterias, hongos, parásitos, y virus (Wang *et al.*, 2019; Lei *et al.*, 2019); sino también a aquellas que son provocadas por células tumorales (Wang *et al.*, 2017b; Conibear *et al.*, 2020). Incluso, se ha llegado a comprobar que algunos péptidos pueden tener la dualidad de ser antimicrobianos y anticancerígenos (Felício *et al.*, 2017). Si a esto se le suma el hecho de que es posible el diseño y síntesis de nuevos compuestos análogos a los que existen en la naturaleza (Fjell *et al.*, 2012b; Porto *et al.*, 2018; Gabernet *et al.*, 2019), entonces en un futuro se pudiera contar con un arsenal de fármacos basados en esta gran variedad de moléculas (Chen y Lu, 2020).

2.1.2. Repositorios biológicos

Los repositorios biológicos son creados, en su mayoría, a partir de la recopilación de datos dispersos en la literatura científica como fuente primaria (Bourne, 2005; Howe *et al.*, 2008). En particular, aquellos que almacenan péptidos bioactivos de interés farmacéutico representan una valiosa fuente de información para extraer conocimiento útil en el diseño *in silico* de fármacos (Usmani *et al.*, 2018a; Basith *et al.*, 2020). A pesar de esto, el aumento en el tamaño y variedad de estos repositorios, hace que la tarea de acceder a todos ellos sea cada vez más difícil para un ser humano. Situación que no parece mejorar con el transcurso del tiempo (Figura 3).

La Figura 3 muestra cómo ha aumentado el interés por el desarrollo de nuevos repositorios biológicos en los últimos años. Estos repositorios pueden ser catalogados como generales o específicos, en dependencia de la naturaleza de los péptidos almacenados (Porto *et al.*, 2017). Los de propósito general contienen péptidos de diversos organismos, mientras que los específicos sólo abarcan péptidos de una determinada especie de organismo o familia en particular. Por esta razón es lógico pensar que los

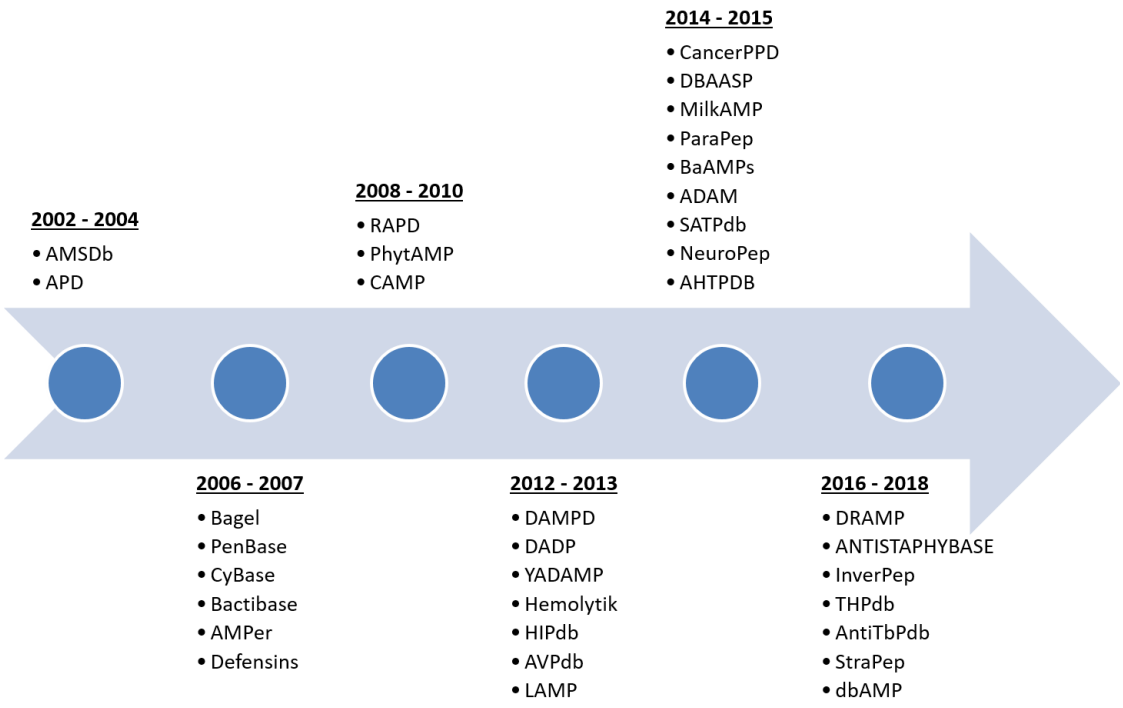


Figura 3. Línea de tiempo que muestra el incremento, en las dos últimas décadas, de los repositorios biológicos que almacenan péptidos bioactivos de interés farmacéutico..

repositorios generales comparten secuencias idénticas de péptidos con otros más especializadas. Esto se comprobó en un estudio previo (Aguilera-Mendoza *et al.*, 2015), donde se evidenció que la mayoría de los repositorios contienen datos únicos, es decir, secuencias de péptidos que no están presentes en ninguna otra fuente de datos; pero también se encontró que existen porcentajes de solapamiento entre ellas. Además, dicho estudio reveló que no existe un único repositorio que supondría una fuente universal de datos; lo cual sería especialmente útil para trazar, por ejemplo, un mapa del EQB de los péptidos bioactivos conocidos hasta la fecha (Reymond y Awale, 2012).

2.2. Explorando un espacio químico-biológico

En general, con el propósito de expandir un EQB, los 20 AAs esenciales (Tabla 1) pueden ser utilizados como bloques de construcción para generar virtualmente un gran número de compuestos peptídicos (Fjell *et al.*, 2012a; Torres *et al.*, 2019). Esta técnica, que consiste en la composición o reemplazo de los AAs, es una herramienta valiosa ya que estos bloques de construcción por si solos o en sinergia con otros re-

siduos contribuyen a la función biológica de los péptidos. No obstante, con tan solo permitir péptidos cuya longitud sea de n aminoácidos, el espacio de todas las secuencias posibles sería de 20^n. Es decir, el tamaño de un EQB puede ser tan grande, que en la práctica una búsqueda exhaustiva de los mejores candidatos a fármacos en todo el espacio teórico es inadmisible. Para ello, se debe considerar que no todo el EQB es biológicamente relevante (Dobson, 2004), y que es necesaria la exploración "inteligente" para encontrar compuestos orgánicos de gran relevancia biológica (Fjell *et al.*, 2012a; Torres *et al.*, 2019).

Como primer ejemplo de lo anterior, en (Kim *et al.*, 2013) se generaron 11 secuencias peptídicas, con actividad antibacteriana, siguiendo el patrón $L_5K_5W^n$, $n = 1 \ldots 11$, donde cada secuencia tenía 5 aminoácidos de tipo Leucina (L), otros 5 de tipo Lisina (k) y un Triptófano (W) en la posición n. Como segundo ejemplo, otros reportes (Deslouches *et al.*, 2005, 2013) han mostrado que péptidos catiónicos antimicrobianos pueden ser diseñados principalmente con Arginina y Valina, incluyendo como sustituto al Triptófano, o solamente con Arginina y Triptófano para potenciar la actividad biológica. Por otra parte, un importante campo de investigación llamado "Peptidomimetics" (Avan *et al.*, 2014; Mojsoska y Jenssen, 2015), por su nombre en inglés, promueve la utilización de péptidos conocidos para modificar su estructura química con el propósito de diseñar nuevos péptidos miméticos como agentes terapéuticos.

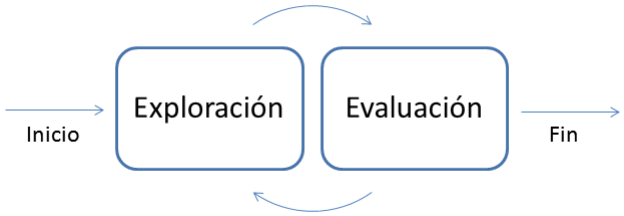


Figura 4. Ciclo de exploración y evaluación en el diseño *in silico* de nuevos candidatos a fármacos.

Precisamente, el principal reto computacional a enfrentar en el diseño e identificación *in silico* de nuevos candidatos a fármacos, es lo grande que resulta explorar el espacio combinatorio de todos los ordenamientos posibles de AAs. La gran variedad de combinaciones hace que la probabilidad de encontrar al péptido idóneo sea muy pequeña. No obstante, varios métodos computacionales han sido propuestos para lidiar con este problema (Fjell *et al.*, 2012a; Porto *et al.*, 2018; Capecchi *et al.*, 2020). La estrategia que han seguido la mayoría de ellos consiste básicamente en un ciclo

de dos etapas (Figura 4). La primera fase es la de exploración, donde se generan las secuencias de aminoácidos o se modifican péptidos conocidos, seleccionados como plantilla; mientras que en una segunda fase de evaluación una función objetivo va guiando la búsqueda hasta encontrar los compuestos líderes deseados.

Lo cierto es que con la expansión *in silico* de un EQB, es posible generar una extensa biblioteca virtual de compuestos químicos diferentes a los que existen en la naturaleza (Porto *et al.*, 2018; Capecchi *et al.*, 2020). Esto sugiere una representación análogo a un “mapa geográfico” para ilustrar la distribución de las moléculas y sus propiedades (Oprea y Gottfries, 2001; Reymond y Awale, 2012; Reymond, 2015). Un punto clave para ello es la representación mediante descriptores, que junto a una función de distancia permite transformar un conjunto de moléculas en un espacio métrico (Chávez *et al.*, 2001). A partir de dicho espacio métrico, se facilita analizar las relaciones de proximidad entre las moléculas que conforman el EQB.

2.2.1. Caracterización mediante descriptores moleculares

Para el estudio matemático de un EQB, es muy útil su representación mediante un vector de descriptores moleculares (Todeschini y Consonni, 2009), donde a cada compuesto químico i le corresponde las coordenadas $X^{(i)} = (x_{i1}, x_{i2}, \ldots, x_{im})$ en un espacio m-dimensional (Figura 5). Por lo tanto, dependiendo de cuál haya sido la representación molecular que sirvió de origen para el cálculo, los descriptores buscan codificar la información contenida en cada molécula de manera unívoca (Todeschini y Consonni, 2009; Jenssen, 2011).

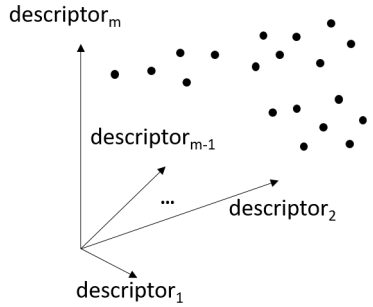


Figura 5. Esquema m-dimensional basado en descriptores de un EQB. Lo puntos representan moléculas a las cuales les corresponde una posición en el en el espacio m-dimensional.

En nuestro caso, resulta de interés considerar la secuencia de AAs como la repre-

sentación molecular que sirve de origen para el cálculo. Entonces, un enfoque simple consiste en derivar descriptores a partir de los aminoácidos constituyentes de manera individual, así como de la secuencia peptídica vista como un todo (Todeschini y Consonni, 2009; Jenssen, 2011). En ese sentido, para un péptido en particular es posible generar un vector de propiedades de AAs, y a partir de ese vector computar el valor de un descriptor mediante algún procedimiento matemático u operador de agregación. De este modo, los descriptores pueden cuantificar propiedades fisicoquímicas de la secuencia ordenada de AAs. A modo ilustrativo, a continuación se presenta el cálculo de la carga neta y la hidrofobicidad, dos propiedades importantes que caracterizan a los AMPs (Torres *et al.*, 2019).

Dada una secuencia de longitud l: $R_1, R_2, \ldots, R_l$, su carga neta puede calcularse como la suma de las cargas de los residuos aminoacídicos R_i en una determinada escala. Por ejemplo, si se supone la escala (Klein *et al.*, 1984) cuya entrada es KLEP840101 en la base de datos AAindex (Kawashima *et al.*, 1999, 2007), la carga neta es

$$net\_charge = \sum_{i=1}^{l} charge(R_i), \tag{1}$$

donde

$$charge(R_i) = \begin{cases} 1 & \text{si } R_i \in \{R, K\} \\ -1 & \text{si } R_i \in \{D, E\} \\ 0 & \text{en otro caso.} \end{cases} \tag{2}$$

Como alternativa también se podría emplear otro procedimiento matemático para calcular la carga neta a diferentes valores de pH (Piotto *et al.*, 2012). Esto se debe a que los péptidos pueden tener comportamientos diferentes a distintos valores de pH. Por lo que se pudiera cuantificar la misma propiedad fisicoquímica con varios descriptores en lugar de uno solo, sin saber a priori cuál es la codificación que mejor caracteriza a la molécula en todos los casos.

Un segundo ejemplo es el cálculo de la hidrofobicidad promedio (Bigelow, 1967; Klein *et al.*, 1984; Gasteiger *et al.*, 2005). Para ello, se puede utilizar una escala de

hidropatía (Kyte y Doolittle, 1982) en la que a cada aminoácido se le asigna un valor que refleja su hidrofobicidad e hidrofilicidad (escala KYTJ820101 en AAindex). De esta manera, valores de hidropatía muy positivos indican la presencia de AAs hidrofóbicos, mientras que valores muy negativos señalan que los AAs se consideran hidrofílicos. Así, deben existir secuencias de números positivos y negativos que representan las regiones hidrofóbicas e hidrofílicas al interior y exterior de la molécula, respectivamente (Kyte y Doolittle, 1982). Por lo que el promedio como operador de agregación de estos valores es un descriptor del péptido, y se calcula como:

$$AVG_H = \frac{1}{l}\sum_{i=1}^{l} H(R_i), \tag{3}$$

donde $H(R_i)$ es el valor en la escala de hidropatía para el residuo R_i. Opcionalmente, al igual que en el caso anterior, se pudiera generar más de un descriptor aplicando varias escalas o procedimientos matemáticos (Simm *et al.*, 2016).

En realidad, son varias las herramientas informáticas que se han publicado para calcular decenas, cientos, o miles de descriptores derivados de las secuencias de AAs. Algunas de ellas son las siguientes: **PROFEAT** (Li *et al.*, 2006; Rao *et al.*, 2011), **PseAAC** (Shen y Chou, 2008; Du *et al.*, 2012), **Propy** (Cao *et al.*, 2013), **PseAAC-General** (Du *et al.*, 2014), **protr/ProtrWeb** (Xiao *et al.*, 2015), **Rcpi** (Cao *et al.*, 2015), **ProtDCal** (Ruiz-Blanco *et al.*, 2015; Romero-Molina *et al.*, 2019), **PseKRAAC** (Zuo *et al.*, 2017), **POSSUM** (Wang *et al.*, 2017a), **modlAMP** (Müller *et al.*, 2017), **PyBioMed** (Dong *et al.*, 2018), **BioSeq-Analysis** (Liu, 2019; Liu *et al.*, 2019), **Seq2Feature** (Nikam y Gromiha, 2019), **iFeature** (Chen *et al.*, 2018), e **iLearn** (Chen *et al.*, 2020).

Lo anterior es una muestra de que, desde el año 2006, donde surge la aplicación web PROFEAT, el aumento en el poder computacional ha permitido más opciones de cálculo para describir numéricamente a un péptido. Sin embargo, el inconveniente que tiene utilizar un número grande de descriptores es que algunos de ellos pueden ser redundantes, irrelevantes o ambos, afectando la calidad en la representación y el análisis de los datos. Además, a medida que la dimensionalidad de los datos aumenta, muchos tipos de análisis se vuelven computacionalmente difíciles de tratar (“the curse of dimensionality”, en inglés) (Beyer *et al.*, 1999; Chávez *et al.*, 2001).

Una estrategia para evitar trabajar con un número grande de descriptores, es hacer una selección empírica de un conjunto de ellos. Dicha selección puede estar basada en descriptores para los cuales ya se demostró que funcionan bien en cierto tipo de problema, por ejemplo, en la clasificación de AMPs (Beltran *et al.*, 2018, 2020). Otra estrategia diferente es calcular primero un gran número de descriptores, para luego identificar y seleccionar de forma automática una combinación optimizada de los descriptores que garanticen un desempeño adecuado según cierta función objetivo. Esta segunda estrategia representa un reto computacional, ya que no es viable explorar el espacio de todos los subconjuntos posibles de descriptores, por lo que a continuación se describe una manera en que se puede llevar a cabo.

2.2.2. Identificación y selección automática de descriptores

Los métodos de selección de variables (rasgos, características) constituyen procedimientos efectivos para seleccionar los descriptores más útiles en aras de identificar y seleccionar características relevantes (Guyon y Elisseeff, 2003). Estos métodos se clasifican en tres grupos: supervisados (Kotsiantis, 2011; Tang *et al.*, 2014; Cai *et al.*, 2018), semi supervisados (Sheikhpour *et al.*, 2017; Cai *et al.*, 2018), y no supervisados (Cai *et al.*, 2018; Solorio-Fernández *et al.*, 2020). Dependiendo del método en particular, se requiere que cada instancia en el conjunto de datos tenga asignada una etiqueta que puede ser un valor entero, correspondiente a una categoría, o un valor real. Los supervisados requieren un conjunto de datos etiquetados. Los semi-supervisados solo requieren que algunas instancias estén etiquetados. Por otro lado, los no supervisados no requieren que el conjunto de datos esté etiquetado, y en ellos nos enfocamos para representar y analizar el EQB.

Al igual que sucede con los métodos supervisados y semi-supervisados, los no supervisados, a su vez, también se pueden dividir en tres tipos (Cai *et al.*, 2018; Solorio-Fernández *et al.*, 2020): filtros, envolturas (*wrapper*, en inglés), y una combinación de ambos, que serían los híbridos. A continuación se describe el propósito de cada una de estas variantes:

- Los **métodos de filtro** (Solorio-Fernández *et al.*, 2020) van a permitir seleccionar un subconjunto de características a partir de propiedades intrínsecas de los

datos, eliminando aquellos rasgos que son irrelevantes y/o redundantes, independientemente del propósito que tengan los mismos. La principal ventaja que tienen estos métodos es su velocidad, por lo que son eficaces como métodos de preprocesamiento.

- Los **métodos de envolturas** (Solorio-Fernández *et al.*, 2020) evalúan la relevancia de los subconjuntos de características a partir del resultado de un algoritmo específico de agrupamiento (*clustering*, en inglés). En consecuencia, estos métodos pueden encontrar subconjuntos de características que contribuyen a mejorar la calidad de los resultados del algoritmo de agrupamiento utilizado para tal efecto (Alelyani *et al.*, 2013). Sin embargo, la principal desventaja es que generalmente tienen un mayor costo computacional, en relación a los métodos de filtro, y están limitados a ser utilizados con un algoritmo de agrupamiento en particular.
- Los **métodos híbridos** (Solorio-Fernández *et al.*, 2020) aprovechan las cualidades de ambos enfoques, filtros y envolturas, tratando de lograr un compromiso entre la eficiencia y la calidad de las variables seleccionadas para una forma específica de agrupamiento de los datos.

Por lo tanto, los métodos de filtro son una buena opción si se quiere lograr una reducción de la dimensionalidad independiente de un algoritmo particular de agrupamiento. Adentrándonos un poco más en la clasificación de dichos métodos, un algoritmo típico de filtrado puede ser univariado o multivariado (Solorio-Fernández *et al.*, 2020). En el esquema univariado se utiliza un criterio de relevancia para evaluar la calidad de cada variable y seleccionarlas según su orden de importancia. De este modo, los filtros univariados pueden identificar y eliminar las características irrelevantes, pero no pueden eliminar variables redundantes ya que ignoran la dependencia entre ellas. En cambio, los filtros basados en el esquema multivariado sí pueden eliminar tanto las características irrelevantes como las redundantes.

A diferencia de los métodos supervisados, en los filtros no supervisados se debe considerar la importancia de una variable sin tener que recurrir a su relación con una posible etiqueta asignada (Cai *et al.*, 2018; Solorio-Fernández *et al.*, 2020). Para ello, varios criterios de evaluación han sido introducidos en la literatura científica (Solorio-Fernández *et al.*, 2020), algunos basados en la teoría de la información (Cover

y Thomas, 2012). Por ejemplo, la relevancia de una variable ha sido cuantificada en términos de la ecuación de entropía de Claude Shannon (Shannon, 1948), esto es de disímiles maneras usando la estrategia *leave-one-out* (Dash *et al.*, 1997; Varshavsky *et al.*, 2006; Rao y Sastry, 2012). En otras palabras, primero se realiza el cálculo de la entropía para el conjunto de variables originales, y luego se repite sin incluir a una variable en particular para determinar su importancia. La intuición indica que si no se distinguen cambios, es porque la variable no es importante. Otra manera de calcular la entropía basada en histogramas permite capturar la variabilidad mostrada por cada variable como criterio de relevancia (Godden *et al.*, 2000). Por otro lado, la redundancia entre pares de variables ha sido evaluada usando la información mutua (Peng *et al.*, 2005) y coeficientes de correlación (Yu y Liu, 2004).

2.2.3. Relaciones de (di)similitud entre los péptidos

Una vez que se hayan identificado y seleccionado las variables que “mejor” caracterizan un EQB, se pueden definir métricas de similitud entre los compuestos a partir de una función de distancia en el espacio de descriptores (Todeschini *et al.*, 2020). Su importancia se debe a que el principio de similitud-propiedad plantea: para compuestos similares se espera que exhiban propiedades similares, siendo la actividad biológica una de las propiedades más estudiadas (Klopmand, 1992). Este principio ha sido comprobado en la práctica (Martin *et al.*, 2002; Skolnick y Fetrow, 2000) y sustenta teóricamente un grupo de técnicas computacionales para la búsqueda por similitud, el agrupamiento y análisis de la diversidad molecular (Willett, 2014). Sin embargo, otros hallazgos sugieren que eventualmente moléculas estructuralmente similares exhiben comportamientos disimilares (Stumpfe y Bajorath, 2012; Hinz *et al.*, 2010), así como otras estructuralmente disimilares exhiben comportamientos similares (Renner y Schneider, 2006; Laskowski y Thornton, 2008). Estas incongruencias con el principio de similitud-propiedad, lejos de ser un obstáculo, conforman la lógica de base de un grupo de técnicas para el análisis y visualización de los acantilados de actividad (Stumpfe y Bajorath, 2012), cuando moléculas similares tienen diferencias en sus actividades biológicas; así como la búsqueda de patrones estructurales (“Scaffold Hopping”), cuando clases estructuralmente diferentes de compuestos activos presentan la misma actividad (Mannhold *et al.*, 2013).

Acerca del concepto de similitud, algunos autores plantean que es algo subjetivo y depende de juicios comparativos que, al igual que la belleza, están en los ojos del espectador (Maggiora y Shanmugasundaram, 2011; Maggiora *et al.*, 2013). A pesar de que no existe un criterio estandarizado de similitud para todos los casos, en la literatura científica se encuentran varios trabajos que describen cómo puede ser cuantificada la similitud molecular a través de un valor numérico, conocido como medida de similitud (Maggiora y Shanmugasundaram, 2011). Dicho valor se calcula a partir de la representación realizada de las estructuras químicas a comparar y si excede un umbral definido entonces se consideran a las moléculas similares. La asignación del valor de umbral también es algo subjetivo y no existe un corte universal para decidir a partir de que valor dos compuestos son similares (Maggiora *et al.*, 2013).

En el caso de las secuencias de aminoácidos, la codificación en caracteres ha permitido calcular medidas de similitud basadas en el alineamiento y comparación de cadenas de texto (Altschul *et al.*, 1997; Smith *et al.*, 1981; Pearson y Lipman, 1988). Estas medidas de similitud, en particular, han probado su valía y ser eficientes especialmente en la búsqueda en grandes de bases de datos. Sin embargo, no consideran propiedades fisicoquímicas y geométricas de las estructuras químicas que pudieran ser de gran utilidad en un contexto de aplicación determinado. Es por ello que otras medidas de similitud “libres de alineamiento” pueden ser aplicadas utilizando un vector numérico de rasgos (Galpert *et al.*, 2018).

2.2.4. Representación visual

La visualización de datos se puede definir como la representación gráfica de información con dos objetivos principales: análisis de datos y comunicación (Keim, 2002). En ese contexto, un EQB puede ser visualizado usando mapas basados en coordenadas, o utilizando la representación de grafos libres de coordenadas (Medina-Franco *et al.*, 2008; Osolodkin *et al.*, 2015). Para ambos tipos de representación, los puntos en el mapa o los nodos de la red, según sea el caso, pueden ser decorados con colores que grafiquen propiedades de los compuestos peptídicos. Esa nueva capa de información basada en colores es muy importante para enriquecer el análisis y la percepción visual, ya que los colores ayudan a romper con el camuflaje, y permiten distinguir visualmente a determinados elementos de su entorno, solo por su color (Ware, 2019).

(a) (b)

Figura 6. Ejemplo de visualización de un EQB: (a) representación basada en coordenadas, y (b) representación de grafos libres de coordenadas.

2.2.4.1. Mapas basados en coordenadas

En los mapas basados en coordenadas (Figure 6(a)), la posición de cada molécula está determina de manera unívoca a partir del vector de DMs que la caracteriza. Matemáticamente, lo que ocurre es que el espacio multi-dimensional de descriptores se reduce a uno de dos o tres dimensiones, aplicando alguna técnica de reducción de dimensionalidad (Medina-Franco *et al.*, 2008; Osolodkin *et al.*, 2015). Por ejemplo, el análisis de componentes principales (PCA, por sus siglas en inglés) es una técnica muy popular que busca un pequeño número de variables no correlacionadas, llamadas componentes principales, expresados como una combinación lineal de las variables originales (Abdi y Williams, 2010).

La utilidad de este procedimiento radica en la elección de un número pequeño de componentes principales, que sean suficientes para explicar la mayor parte de la varianza de los datos. De esta manera, el mapa 2D (3D) puede proyectar la información contenida en dos (o tres) de los primeros componentes obtenidos a partir del espacio de descriptores originales. Sin embargo, para representar un EQB, en los últimos años se ha desarrollado otro enfoque basado en grafos, que surge como una alternativa libre de coordenadas (Maggiora y Bajorath, 2014).

2.2.4.2. Representación basada en grafos

Los grafos son ubicuos, y han sido ampliamente utilizados en diversas ramas de la ciencia para modelar datos (Newman, 2018). En particular, en el campo de de descrubrimiento de fármacos se han empleado como una forma conveniente de describir

relaciones entre las entidades biológicas o químicas (Csermely *et al.*, 2013; Recanatini y Cabrelle, 2020). Un ejemplo lo constituyen las redes de similitud molecular (Maggiora y Bajorath, 2014; Vogt *et al.*, 2016), donde los nodos simbolizan a las entidades moleculares y las aristas indican las relaciones de similitud entre ellas (Figure 6(b)).

Por otra parte, un aspecto importante de las redes en general es que son un modelo matemático que se puede visualizar para el análisis (von Landesberger *et al.*, 2011), sin necesidad de construir un sistema de coordenadas ni ninguna otra forma de reducción de dimensionalidad. De tal manera que el análisis se fundamenta en un marco conceptual basado en la teoría de grafos y redes complejas (Estrada, 2012; Newman, 2018). Por consiguiente, especialmente prometedor es la representación de un EQB mediante redes de similitud, con el propósito de habilitar el uso de técnicas y algoritmos del área de ciencia de redes.

2.3. Ciencia de redes: marcando una nueva era

En la sociedad moderna es común encontrar que las personas conozcan, al menos de manera intuitiva, lo que son las redes: un grupo de elementos (nodos o vértices) interconectados entre sí mediante aristas; pero además de la simplicidad que suponen, el interés científico por las redes ha tenido un gran auge en las dos últimas décadas (Barabási, 2003; Newman *et al.*, 2006; Newman, 2018). A pesar de ello, el estudio de las redes no es algo nuevo. Sus orígenes se remontan a 1736, cuando el matemático Leonhard Euler resuelve el problema de los puentes de Königsberg (Biggs *et al.*, 1986).

La trascendencia del trabajo de Euler estuvo en resolver un problema del mundo real utilizando un modelo basado en la conectividad entre nodos, en lugar de aplicar complicados cálculos como suponía el problema en aquella época. Inaugurando así el campo de la teoría de grafos (Biggs *et al.*, 1986), que matemáticamente establece los fundamentos esenciales para el estudio de las redes. Por lo tanto, dada la estrecha relación que tienen los términos grafos y redes, es que se hace uso de ellos indistintamente.

2.3.1. Definición formal de una red

Una red simple G es un par $G = (V, E)$, donde V es un conjunto finito de nodos, y E es un conjunto de pares ordenados de elementos de V, denominados aristas. Cuando las aristas no tienen dirección se dice que el grafo es no dirigido, de lo contrario se dice que es dirigido. Además, las aristas del grafo pueden tener asociado un valor y en este caso se trata de un grafo ponderado $G = (V, E, w)$, donde $w : E \to \Re$.

Para el estudio de las redes, son varias las formas de representarlas matemáticamente, una de ellas es haciendo uso del álgebra de matrices. La matriz de adyacencia $A = [a_{ij}]_{n \times n}$ es una matriz cuadrada de dimensión $n \times n$, donde n es el número de vértices, y cuyos elementos a_{ij} indican:

$$a_{ij} = \begin{cases} 1 & \text{si } i \text{ está enlazado con } j \\ 0 & \text{en otro caso,} \end{cases} \tag{4}$$

en el caso de una red simple. Para las redes ponderadas, se asignan valores a los elementos a_{ij} igual a los pesos de las conexiones correspondientes:

$$a_{ij} = \begin{cases} w(i,j) & \text{si } i \text{ está enlazado con } j \\ 0 & \text{en otro caso.} \end{cases} \tag{5}$$

2.3.2. Modelos de redes complejas

Desde los tiempos de Euler, el campo de la teoría de grafos se ha seguido desarrollado (Chartrand *et al.*, 2010). Más de dos siglos después de sus inicios, es que se comienza a estudiar acerca de la formación de los grafos y cuáles son las leyes que gobiernan su estructura (Newman *et al.*, 2006). En este tipo de estudio, un hito histórico lo marcó la teoría de los grafos aleatorios (Erdős y Rényi, 1960), propuesta por el matemático Paul Erdös y su colaborador Alfréd Rényi. Dando a conocer el modelo Erdős–Rényi, al final de la década de 1950 e inicio de la década de 1960, para ser empleado en la generación de grafos aleatorios mediante un proceso estocástico. Esto con la particularidad de que un nuevo nodo tiene la misma probabilidad de conectarse

al resto de los nodos en la red. Bajo este supuesto, los autores muestran que ocurren cambios repentinos en las propiedades estudiadas de la red, y no gradualmente, cuando se añaden un número suficiente de nodos y aristas (Erdős y Rényi, 1960).

Sin embargo, a pesar de la novedad que todo hito representa, las redes aleatorias introducidas por Erdős y Rényi no constituyen un buen modelo para representar redes del mundo real. Por ejemplo, en las redes sociales (Wasserman *et al.*, 1994; Scott, 2000), donde los vértices simbolizan individuos u organizaciones de individuos, y las aristas sus relaciones sociales, es de sentido común observar que la conectividad no está basada en la pura aleatoriedad. Para ilustrar esta suposición, basta considerar a un individuo que es muy popular: en virtud de que él tendría muchas conexiones, es lógico pensar que también va a tener mayor chance de establecer lazos con nuevos amigos respecto a aquellos que tienen un número bajo de conexiones.

Con un razonamiento análogo al anterior, los individuos en busca de nuevos amigos estarían más propensos a conectarse con aquellos que comparten un mayor número de amigos en común. De tal forma, que los vínculos sociales entre las personas causan un efecto en la forma de la red, y la forma de la red causa un efecto en los nuevos lazos que se establecen entre las personas. Esto sería tan solo un ejemplo, pero son varias las redes del mundo real que no se ajustan al modelo Erdős–Rényi (Watts y Strogatz, 1998; Barabási y Albert, 1999). Sin embargo, las suposiciones de aleatoriedad en el modelo Erdős–Rényi fueron incuestionables durante más de cuarenta años.

En contraste con la teoría de grafos aleatorios, en la que todos los nodos tienen la misma probabilidad de conexión, es que comienza a resurgir un nuevo interés en la disciplina que se consagra en el área de ciencia de redes (Mark, 2002; Watts, 2004; Newman *et al.*, 2006). En esta nueva era que inicia a finales del siglo pasado, trabajos seminales (Watts y Strogatz, 1998; Barabási y Albert, 1999) propusieron nuevos modelos llamados “Redes de mundo pequeño” y “Redes libres de escala”, publicados en revistas prestigiosas como Nature y Science, respectivamente. Con la propuesta de estos dos modelos de red (Watts y Strogatz, 1998; Barabási y Albert, 1999), los autores alertaron acerca de las posibilidades de encontrar patrones en redes complejas, es decir, en redes con una topología ni tan regular ni tan aleatoria (Albert y Barabási, 2002). De hecho, las redes complejas están siendo cada vez más utilizadas para modelar y resolver problemas del mundo real en distintas ramas de la ciencia (Newman, 2018), y

el área de descubrimiento de fármacos no está exento de ello (Csermely *et al.*, 2013; Recanatini y Cabrelle, 2020).

2.3.2.1. Redes de mundo pequeño

Un sorprendente experimento fue realizado en 1967, en esta ocasión por el psicólogo social Stanley Milgram (Milgram, 1967). Este experimento fue realizado para comprobar la hipótesis de que dos personas cualesquiera, a pesar de lo lejos que estuviesen físicamente, podían establecer contacto a través de una red de amigos en tan solo unos pocos pasos. El resultado fue que dos personas cualesquiera, aparentemente distantes, como promedio estaban a solo seis conexiones de amistad. Lo cual fue inspiración para que se acuñara la frase “seis grados de separación” (Watts, 2004; Guare, 1990), pero más interesante aún fue que estos resultados permitieron establecer una analogía con las redes de mundo pequeño.

En 1998, Duncan Watts y Steven Strogatz introducen el concepto de red de mundo pequeño (Watts y Strogatz, 1998) para hacer alusión a cierto tipo de redes, con una topología irregular y no tan aleatoria, que combinan un número relativamente pequeño de conexiones entre nodos con una tendencia a formar grupos (clusters). Evidenciando, por primera vez, que el fenómeno de “mundo pequeño” y la formación de grupos no era exclusivo de las redes sociales, sino que probablemente era aplicable a otras muchas redes que se pudieran encontrar en la naturaleza.

2.3.2.2. Redes libres de escala

Una limitante en la segundad mitad del siglo pasado fue la carencia de datos y poder de cómputo para analizar grandes redes (Newman *et al.*, 2006). Prácticamente, la recolección de datos se hacía mediante encuestas para el análisis de redes sociales (Newman *et al.*, 2006). Esto ha cambiado en las últimas décadas, sobre todo, con el desarrollo de los procesos de informatización y captura automática de datos. Por lo que el escenario estaba creado para que en 1999, con base en el análisis de tres redes del mundo real, los autores Albert-László Barabási y Réka Albert mostraran evidencias empíricas de que la función de distribución del grado de los vértices se ajusta a una ley de potencia (Barabási y Albert, 1999). En aquel entonces, este fue un hallazgo sin

precedentes, dado que el modelo predominante era el propuesto por Erdős y Rényi, que predecía una distribución de Poisson para el grado de los vértices (Barabási y Albert, 1999).

La relevancia del nuevo modelo propuesto es que indica que la probabilidad $P(k)$ de que un vértice elegido al azar tenga grado k decae como una ley de potencia, es decir, $P(k) \sim k^{-\gamma}$, para cierto valor constante γ (Barabási y Albert, 1999). Dicha ley de potencia es conocida por ser una distribución libre de escala (Newman, 2005). Por lo tanto, si se escala el grado de los vértice se sigue manteniendo la misma distribución. De ahí que los autores nombraran a las redes que cumplen con esta propiedad como redes libres de escala (scale-free networks en inglés). Además, a partir de la curva de distribución descrita por la ley de potencia, se interpreta que van a existir muchos nodos con un número pequeño de enlaces, que coexisten con pocos nodos altamente conectados (Barabási y Albert, 1999). Estos últimos actúan como nodos centrales, también llamados *hubs* en inglés, e inciden en la topología y el mecanismo de atracción preferencial que pueden tener los nuevos nodos en la red (Barabási y Albert, 1999). No obstante, más allá de la importancia de estos resultados para explicar las leyes que gobiernan la evolución de la redes complejas, su generalización aún es controversial (Broido y Clauset, 2019; Holme, 2019; Voitalov *et al.*, 2019).

2.3.3. Trabajos relacionados

La disciplina de ciencia de redes ha sido aplicada al descubrimiento de fármacos de disímiles maneras. Un ejemplo han sido las redes de interacción de proteínas (Meyer *et al.*, 2013; Mosca *et al.*, 2013), que se consideran macromoléculas dentro del EQB. Un segundo ejemplo de redes de interacción son las que contemplan como posibles nodos tanto a moléculas pequeñas como a proteínas, donde las moléculas pequeñas con potencial farmacológico pueden estar conectadas a múltiples dianas proteicas (Vogt y Mestres, 2010; Gergely *et al.*, 2020). Más relacionado con nuestro trabajo se encuentran las redes de similitud molecular, conocidas como *Chemical Space Networks* (CSNs), por su nombre en inglés, que fueron propuestas por Maggiora y Bajorath en 2014 para representar un espacio químico libre de coordenadas.

En los primeros reportes acerca de las CSNs, los autores dedicaron esfuerzos al di-

seño y caracterización de este tipo de redes para el estudio de moléculas pequeñas (Vogt *et al.*, 2016). Dichas redes fueron diseñadas inicialmente utilizando el coeficiente de Tanimoto para evaluar la similitud entre compuestos químicos (Zwierzyna *et al.*, 2015; Zhang *et al.*, 2015b). Luego se exploraron y evaluaron otras medidas de relaciones de similitud, que incluye, entre otras (Zhang *et al.*, 2015a), el índice de Tversky (Wu *et al.*, 2016) como medida de similitud molecular asimétrica.

Para ilustrar la utilidad de las CSNs, a continuación se mencionan algunas de sus aplicaciones recientes para descubrir conocimientos en espacios químicos ocupados por moléculas pequeñas. Por solo citar tres ejemplos: i) redes dirigidas fueron diseñadas utilizando el índice de Tversky para encontrar rutas de similitud entre los compuestos estudiados (Kunimoto *et al.*, 2017), ii) se combina la búsqueda por similitud con el análisis de redes para identificar compuestos biológicamente relevantes (Kunimoto y Bajorath, 2018), y iii) se visualizan las redes y se analiza una estructura de clusters para el diseño de semiconductores orgánicos cristalinos (Kunkel *et al.*, 2019). Sin embargo, no existen a nuestro conocimiento estudios que aborden este tema a profundidad en el análisis de un EQB de péptidos bioactivos.

2.4. Conclusiones parciales

Los repositorios biológicos de péptidos bioactivos constituyen una valiosa fuente de información para el estudio de estos compuestos orgánicos. Sin embargo, en lugar de tenerse un único modelo unificado de datos, lo que existen son varias bases de datos aisladas que han ido en aumento en los últimos años. Por lo que debe realizarse un proceso de integración de datos si se pretende llevar a cabo un estudio que involucre a todos los péptidos que se encuentran almacenados en estos repositorios. Esto adquiere mayor importancia si se considera a todos los péptidos publicados como un EQB conocido de interés farmacológico.

Por otro lado, un aspecto básico para el análisis de un EQB es la codificación numérica de las secuencias de AAs. Para lo cual, la búsqueda de nuevos DMs sigue siendo de interés científico en aras de mejorar la diversidad de estos y, complementar la información estructural no codificada por el conjunto de DMs existentes. Además, un paso fundamental es la identificación y selección automática de un conjunto reducido

de variables para definir el espacio de descriptores.

Por último, hemos enfatizado en el diseño de un espacio de descriptores, que debe ser intencionado y no accidental, para derivar redes que describan relaciones de distancia/similitud entre los péptidos. Luego, en el análisis de los datos, sugerimos el empleo de técnicas del área de ciencia de redes tanto por razones prácticas como teóricas. Bajo el supuesto de que las redes complejas, en general, cumplen con los modelos de redes de mundo pequeño y redes libres de escala, es de esperar que se encuentren estructuras similares a las de clusters y nodos centrales en las redes de similitud en estudio. Esto último tendría relevancia biológica si se utilizan las métricas y algoritmos adecuados para ello.

Capítulo 3. Metodología

El descubrimiento de conocimiento a partir de datos engloba varias fases e incorpora técnicas de Aprendizaje Automatizado en una de ellas, la de minería de datos. Aunque la minería de datos es una etapa fundamental, es tan sólo una parte, y no se debe descuidar lo realizado en todo el proceso, conocido como KDD, del inglés Knowledge Discovery from Databases (Fayyad *et al.*, 1996). En una de las definiciones más aceptada, dada por Fayyad *et al.*, se plantea que KDD es "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles, y en última instancia comprensibles a partir de los datos". Este proceso, según lo descrito en este capítulo, fue llevado a cabo para cumplir con los objetivos planteados en la presente investigación.

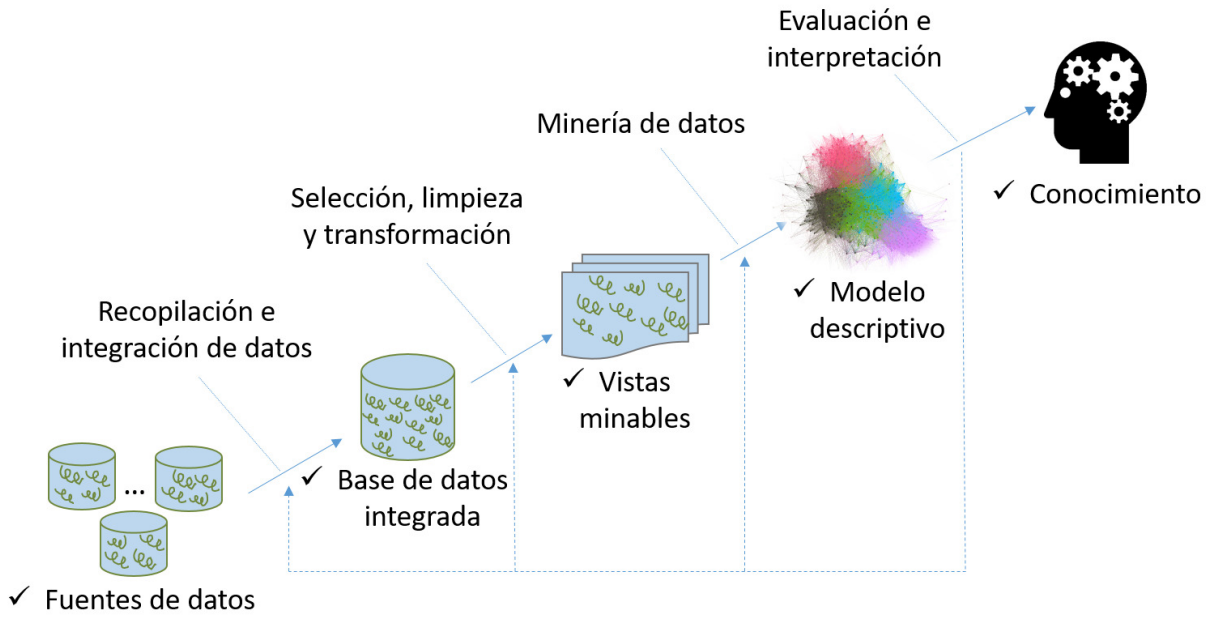


Figura 7. Proceso de extracción de conocimiento que ha sido llevado a cabo en la presente investigación.

Como se muestra en la Figura 7, una forma de apoyar la toma de decisiones sería a partir del análisis visual de un modelo descriptivo proveniente del aprendizaje no supervisado (Keim, 2002; Simoff *et al.*, 2008). Su importancia radica en facilitar el uso de la percepción visual (Shneiderman, 1996; Ware, 2019) para detectar patrones útiles, formular hipótesis y arribar a conclusiones en el dominio del problema. Bajo este enfoque, la representación visual y el análisis de un EQB (Dobson, 2004) puede jugar un papel esencial en la búsqueda de nuevos candidatos a fármacos basados en péptidos (Oprea y Gottfries, 2001; Medina-Franco *et al.*, 2008; Reymond y Awale,

2012).

3.1. Recopilación e integración de datos

El primer paso en el proceso KDD es identificar las fuentes de datos con las que se va a trabajar, para unificarlos en un nuevo repositorio integrado. Para ello, identificamos como fuente de datos a los 40 repositorios biológicos presentados en la Tabla 2. Estos repositorios fueron seleccionados ya que contienen gran variedad de péptidos bioactivos evaluados experimentalmente, cuyas actividades biológicas son de interés farmacológico. No obstante, por limitaciones de tiempo y propósito de la investigación, no todas las bases de datos existentes de péptidos bioactivos se consideraron en este estudio. A pesar de ello, la selección realizada es diversa e incluye 40 fuentes de datos luego de revisar la literatura desde el año 2002 hasta enero de 2019.

Tabla 2. Listado cronológico de las bases de datos de péptidos bioactivos utilizadas en este estudio.

			Última	Nodos extraídos por etiquetas					
Año	Base de Datos	Péptidos contenidos	actualización	*Peptide*	*Function*	*Origin*	*Target*	*Cross-Ref*	Literatura
2002	UniProtKB_KW-0929	AMPs	2019	2206	12	428	N/A	4772	Consortium (2017)
2002	AMSDb	AMPs	2004	719	12	150	97	719	Tossi y Sandri (2002)
2004	APD	AMPs	2019	2949	20	541	214	2161	Wang *et al.* (2016)
2006	Bagel_I	Bacteriocinas	2018	422	1	125	N/A	379	van Heel *et al.* (2013)
2006	Bagel_II	Bacteriocinas	2018	206	1	60	N/A	112	van Heel *et al.* (2013)
2006	PenBase	AMPs de camarones	2008	28	1	1	N/A	N/A	Gueguen *et al.* (2006)
2006	CyBase_Cyclotides	Ciclótidos	2016	629	7	20	N/A	310	Wang *et al.* (2008)
2007	Bactibase	Bacteriocinas	2017	202	3	74	219	549	Hammami *et al.* 2010
2007	AMPer	AMPs	2007	749	7	150	N/A	1213	Fjell *et al.* (2007)
2007	Defensins	Defensinas	2007	458	6	95	96	767	Seebah *et al.* (2007)
2008	RAPD	AMPs	2010	118	1	22	N/A	214	Li y Chen (2008)
2009	PhytAMP	AMPs	2012	272	6	62	N/A	379	Hammami *et al.* (2009)
2010	CAMP_Patent	AMPs patentados	2016	1917	1	N/A	N/A	N/A	Waghu *et al.* (2016)
2010	CAMP_Structure	AMPs con estructuras 3D	2016	427	6	N/A	N/A	1090	Waghu *et al.* (2016)
2010	CAMP_Validated	AMPs	2016	2581	7	482	310	2587	Waghu *et al.* (2016)
2012	DAMPD	AMPs	2011	974	8	245	N/A	1697	Seshadri Sundararajan *et al.* (2012)
2012	DADP	AMPs de anuros	2012	2211	10	160	3	2150	Novković *et al.* (2012)
2012	YADAMP	AMPs	2013	2523	4	417	190	843	Piotto *et al.* (2012)
2013	Hemolytik	Hemolíticos y no hemolíticos	2013	1726	16	N/A	14	443	Gautam *et al.* (2013)
2013	HIPdb	Anti-HIV	2013	887	1	N/A	1	109	Qureshi *et al.* (2013)
2013	AVPdb	Antivirales	2013	2074	1	N/A	63	338	Qureshi *et al.* (2014)
2013	LAMP_Experimental	AMPs	2013	3190	7	473	123	1299	Zhao *et al.* (2013)
2013	LAMP_Patent	AMPs patentados	2013	1490	3	N/A	N/A	N/A	Zhao *et al.* (2013)
2014	CancerPPD	Anticancerígenos	2015	577	10	50	23	151	Tyagi *et al.* (2015)
2014	DBAASP	AMPs	2019	9726	9	571	577	2777	Pirtskhalava *et al.* (2016)
2014	MilkAMP	AMPs de origen lácteo	2013	310	4	7	108	121	Théolier *et al.* (2014)
2014	ParaPep	Antiparasitario	2014	265	16	N/A	8	93	Mehta *et al.* (2014)
2015	BaAMPs	Antibiofilm	2017	196	1	7	41	101	Luca *et al.* (2015)
2015	ADAM	AMPs	2015	3256	7	700	175	1005	Lee *et al.* (2015)
2015	SATPdb	Terapéuticos	2015	17275	25	N/A	N/A	N/A	Singh *et al.* (2015)
2015	NeuroPep	Neuropeptidos	2015	3658	1	328	N/A	3144	Wang *et al.* 2015
2016	DRAMP_Clinical	AMPs	2019	27	3	N/A	N/A	29	Fan *et al.* (2016)
2016	DRAMP_Patent	AMPs patentados	2019	12788	N/A	127	N/A	N/A	Fan *et al.* (2016)
2016	DRAMP_General	AMPs	2019	4990	25	667	378	2118	Fan *et al.* (2016)
2016	ANTISTAPHYBASE	Anti-S. aureus	2017	582	1	99	2	169	Zouhir *et al.* (2016)
2016	InverPep	AMPs de invertebrados	2017	721	12	106	N/A	330	Gómez *et al.* (2017)
2017	THPdb	Terapéuticos	2017	34	6	N/A	N/A	N/A	Usmani *et al.* (2017)
2018	AntiTbPdb	Antituberculosos	2018	343	1	9	11	70	Usmani *et al.* (2018b)
2018	StraPep	Bioactivos	2018	1026	5	261	N/A	3230	Wang *et al.* (2018)
2018	dbAMP	AMPs	2019	3902	22	642	309	3524	Jhong *et al.* (2018)
	Overall			45120	48	3805	1311	19074	

Teniendo en cuenta el tipo de acceso que se tiene a las fuentes de datos, la tarea principal en esta fase se centra en procesar el contenido semi-estructurado procedente

de páginas Web. Para hacer frente a esta situación, se realizó el diseño e implementación de flujos ETL (Vassiliadis *et al.*, 2002), del inglés Extract, Transform and Load, para extraer, transformar y cargar el contenido de cada una de las 40 fuentes de datos de origen al nuevo destino con los datos unificados. Estos flujos ETL se implementaron en la herramienta Pentaho Data Integration, también conocida como Kettle (`http://community.pentaho.com/`).

3.1.1. Flujo de extracción, transformación y carga

En la etapa de extracción se descargaron las páginas web (archivos HTMLs) y recursos online (archivos XML, Excel, etc.) disponibles en las fuentes de datos. Luego, se transformaron los datos recopilados a un formato común, detectando las instancias de péptidos repetidas y mezclando sus anotaciones. Por último, se realizó la carga de los datos unificados hacia un repositorio destino para su posterior análisis.

Debido a que los datos provienen de diferentes fuentes, integrarlos no es una tarea sencilla (Hernández y Stolfo, 1998). Lo primero que se hizo fue elegir el modelo de datos destino. En nuestro caso, optamos por un modelo de red que a su vez está inspirado en un modelo multidimensional de los almacenes de datos, el esquema en estrella (Kimball y Ross, 2011). Este modelo es intuitivo ya que los nodos que representan a los péptidos se encuentran en el centro, teniendo como sus nodos vecinos a los metadatos (anotaciones) recopilados y unificados de las fuentes de datos de origen (Figura 8). Además, se adiciona la relación *is_a* para definir conexiones jerárquicas entre determinados tipos de nodos.

Entre las anotaciones que se recopilan y unifican como metadatos (Tabla 3), tenemos a los nombres de las fuentes de datos (Database), actividades biológicas (Function), nombres científicos de los organismos del cual se aisló el péptido (Origin), nombres científicos de los microorganismos contra los cuales se evaluó la actividad del compuesto (Target), modificaciones químicas realizadas (Nterminus, UnusualAA, Cterminus), y referencias a registros relacionados en otras bases de datos (Crossref), como pueden ser UniProtKB (`https://www.uniprot.org/uniprot/`), PDB (`https://www.rcsb.org/`), DOI (`https://www.doi.org`), y PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`).

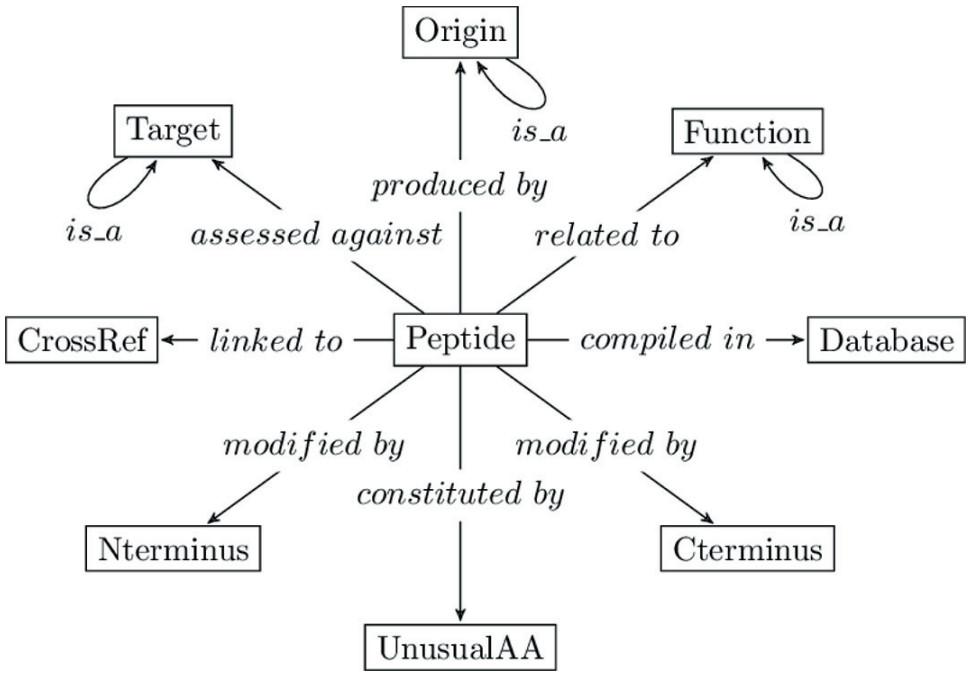


Figura 8. Modelo de red basado en un esquema en estrella.

Tabla 3. Etiqueta de los nodos de metadatos y sus relaciones con los nodos que representan péptidos.

Etiqueta	**Jerarquía (*is_a*)**	**Relación**
Origin	✓	*produced_by*
Target	✓	*assessed_against*
Function	✓	*related_to*
Database		*compiled_in*
Crossref		*linked_to*
Nterminus		*modified_by*
UnusualAA		*constituted_by*
Cterminus		*modified_by*

En cuanto al repositorio destino para dar persistencia al modelo de red diseñado, se utilizó a la base de datos Neo4j (https://neo4j.com/). Neo4j es una base de datos orientada a grafos, donde los nodos y aristas tienen etiquetas, y además se les pueden asignar propiedades que los describan (pares clave-valor). Por ejemplo, los nodos etiquetados como péptidos tienen propiedades con las claves “identificador”, “secuencia” y “longitud”; mientras que el resto de los nodos y aristas tienen propiedades cuyas claves son, respectivamente, “nombre” y “db-ref”. Esta última propiedad indica la entrada en la fuente de datos de donde se extrajo la relación entre el péptido y el metadato descrito.

Una posibilidad que ofrece el modelo de red adoptado, es la de organizar de forma

unificada los datos procedentes de diversas fuentes. En primer lugar, al fusionar todas las secuencias idénticas en un solo nodo, entonces se pueden congregar todos los metadatos reportados alrededor de él. Luego, se procede a unificar los metadatos. Por ejemplo, los nombres “E. coli” y “Escherichia coli” deben consolidarse en un solo nodo de metadatos, y este nodo debe conectarse a todos los péptidos que potencialmente pueden atacar y destruir a esta bacteria. Sin embargo, esta tarea se torna difícil debido a que las fuentes de datos no tienen una terminología estandarizada. Además de que existen errores tipográficos, que son introducidos al escribir las anotaciones de la literatura. Por lo que se llevaron a cabo las siguientes acciones para lidiar con dicho inconveniente.

- Adopción de una terminología para nombrar las actividades biológicas, así como los nombres binomiales científicos para las especies (`https://www.itis.gov/`).
- Identificación de los metadatos a partir del texto fuente utilizando expresiones regulares.
- Creación de un diccionario de sinónimos para mapear los metadatos encontrados con la terminología adoptada, según sea el caso. En caso de que el metadato no coincida exactamente con algún término del diccionario, se supone algún tipo de anomalía (errores ortográficos), y se utiliza la distancia entre palabras (Levenshtein, 1966) para recuperar los términos más cercanos, cuya distancia sea menor que tres; lo cual es seguido de una revisión manual para transformar el metadato al término correcto, si es que procede.
- Unificación de las fuentes de datos de origen con base en el modelo de red diseñado (Figure 1). En esta etapa del proceso, diferentes metadatos pueden fusionarse en uno solo utilizando el diccionario de sinónimos que asigna las variantes gramaticales y los errores ortográficos a las palabras correctas.

También se procesaron de forma automática las referencias bibliográficas incluidas en la fuente original. En primer lugar, se extrajeron los componentes de las citas (si estaban disponibles): nombres de los autores, título del artículo, nombre de la revista, número de volumen, paginación, año de publicación, DOI e identificador de PubMed (PMID). Luego, si el identificador DOI o PMID estaban disponibles, estos se consideraron para los metadatos que indican las referencias cruzadas. En caso contrario, se

usaron los componentes identificados de la cita para recuperar el PMID mediante el módulo *Bio.Entrez* implementado en Biopython (Cock *et al.*, 2009). De este modo, para muchas referencias bibliográficas sin los identificadores PMID o DOI, se lograron recuperar satisfactoriamente los PMID y agregarlos como metadato de referencia cruzada.

3.2. Selección, limpieza y transformación

Esta fase consiste en seleccionar el conjunto inicial de péptidos, eliminar secuencias redundantes, y transformar los datos al espacio de descriptores para la tarea de minería que se desea realizar. El objetivo es trabajar con una vista minable y no con todos los datos recopilados e integrados. Esto debido a que una alta redundancia en las secuencias de péptidos puede afectar la calidad de los datos minados, y a su vez la calidad del conocimiento descubierto. Por lo que se implementó el algoritmo UClust (Edgar, 2010) para detectar clústeres de secuencias que son muy parecidas, bajo cierto umbral de identidad definido por el usuario final, y así seleccionar los representantes de cada cluster como el conjunto no redundante de secuencias.

3.2.1. Espacio de descriptores

Una vez que los péptidos hayan sido seleccionados y limpiados, estos se codifican en vectores de descriptores moleculares para su uso posterior en la fase de minería de datos. Primeramente, se consideraron 29 descriptores que representan propiedades fisicoquímicas de los péptidos (Tabla 4). Según estudios realizados (Beltran *et al.*, 2018, 2020), estos descriptores mostraron buen desempeño en la evaluación de la actividad antimicrobiana de compuestos peptídicos. Por otra parte, en esta investigación se implementaron nuevos descriptores calculados a partir de vectores de propiedades de AAs, con la intención de codificar información diferente. Para ello, lo primero es precisar cómo la secuencia de AAs se representa matemáticamente mediante un vector molecular.

A partir de las propiedades de AAs consideradas en este estudio (Tabla 5), los componentes de un vector molecular no son más que valores numéricos correspondientes

Tabla 4. Propiedades fisicoquímicas que fueron calculadas.

Propiedad fisicoquímica	**Símbolos**	**Referencia**
Carga neta a distintos valores de pH (5, 7, 9)	Z(pH=5), Z(pH=7), Z(pH=9)	(Piotto *et al.*, 2012)
Carga neta a distintas escalas (KLEP840101, CHAM830107, CHAM830108)	NetCharge(KLEP840101), NetCharge(CHAM830107), NetCharge(CHAM830108)	(Klein *et al.*, 1984; Charton y Charton, 1983)
Carga promedio a distintas escalas (KLEP840101, CHAM830107, CHAM830108)	AvgNetCharge(KLEP840101), AvgNetCharge(CHAM830107), AvgNetCharge(CHAM830108)	(Klein *et al.*, 1984; Charton y Charton, 1983)
Punto Isoeléctrico	pI	(Bjellqvist *et al.*, 1994)
Masa Molecular	mw	(Gasteiger *et al.*, 2005)
Índice de Boman	boman	(Boman, 2003)
Momento Hidrofóbico	μH(angle=100, scale=KYTJ820101) μH(angle=160, scale=KYTJ820101) μH(angle=180, scale=KYTJ820101) μH(angle=100, scale=Tossi12) μH(angle=160, scale=Tossi12) μH(angle=180, scale=Tossi12) μH(angle=100, scale=EISD840101) μH(angle=160, scale=EISD840101) μH(angle=180, scale=EISD840101) maxAvgH(KYTJ820101) maxAvgH(Tossi12) maxAvgH(EISD840101)	(Kyte y Doolittle, 1982; Eisenberg, 1984; Tossi *et al.*, 2003; Gasteiger *et al.*, 2005)
Promedio de Hidrofilicidad	AvgHydrophilicity(KUHL950101) AvgGRAVY	(Kuhn *et al.*, 1995; Kyte y Doolittle, 1982; Eisenberg, 1984)
Periodicidad hidrofóbica	A(m=10)	(Klein *et al.*, 1984)
Índice Alifático	AI	(Kuhn *et al.*, 1995)
Índice de Inestabilidad	II	(Guruprasad *et al.*, 1990)

a una determinada propiedad x. Por lo tanto, un péptido de longitud n puede ser representado por un vector $v = (x_{a_1}, \ldots, x_{a_n})$, donde cada x_{a_i} sería el valor numérico de la propiedad x para el residuo a_i. De esta manera, pueden generarse diferentes descriptores teniendo como origen un vector molecular. Por ejemplo, la carga neta y la carga promedio (Tabla 4) serían propiedades fisicoquímicas del péptido que consideran, respectivamente, la suma y el promedio de un vector molecular correspondiente a la carga de los AAs en cierta escala. Desde luego, una mayor variedad de descriptores se pueden calcular si se aplican distintos operadores de agregación (Calvo *et al.*, 2012), en lugar de considerar la simple suma o el promedio de sus partes (Martínez-López *et al.*, 2019; Marrero-Ponce *et al.*, 2020).

Tabla 5. Propiedades de la cadena lateral de los aminoácidos.

Propiedad	**Símbolo**	**Referencia**
Frecuencia relativa en giros inversos	ptt	(Prabhakaran, 1990)
Parámetros de compatibilidad geométrica L_{1-9}	gcp1	(Sak *et al.*, 1999)
Parámetros de compatibilidad geométrica ξ	gcp2	(Sak *et al.*, 1999)
Calor de formación	eps	(Sak *et al.*, 1999)
Masa	scm	(Mathews *et al.*, 2000)
Volumen	scv	(Zamyatnin, 1972)
Punto isoeléctrico	pie	(Hellberg *et al.*, 1987)
Frecuencia relativa para formar hélices-α	pah	(Prabhakaran, 1990)
Frecuencia relativa para formar hojas-β	pbs	(Prabhakaran, 1990)
Área de superficie	isa	(Collantes y Dunn III, 1995)
Escala de valores z	z1, z2, z3	(Sandberg *et al.*, 1998)
Carga	NetCharge	(Klein *et al.*, 1984)
Boman	Boman	(Boman, 2003)

En cuanto a los operadores de agregación implementados (Tabla 6), estos se clasifican en cuatro grupos. En el primer grupo se encuentra la norma N1 (suma de todos los componentes), y la norma N2 (distancia euclídea al origen). En un segundo grupo se aplican operadores de tendencia central, tales como: media aritmética (AM), media cuadrática (P2), media potencial de grado 3 (P3) y media armónica (HM). En el tercer grupo se incluyen estadísticos de dispersión y forma: varianza (V), desviación estándar (SD), rango (RA), coeficiente de variación (VC), rango intercuartil (i50), asimetría (S) y curtosis (K). En el cuarto grupo se encuentran algunos operadores clásicos que tradicionalmente se emplean para obtener descriptores, como son los de autocorrelación de Broto-Moreau (AC), índices gravitacionales (GV), y sumas totales (TS). Los operadores de este último grupo, en su definición clásica, utilizan solamente la suma para adquirir el valor total del descriptor, por lo que también fueron extendidos mediante

el empleo de los otros operadores de agregación especificados (Martínez-López *et al.*, 2019).

Tabla 6. Operadores de agregación.

Operador	**Símbolo**	**Fórmula**	**Nota**
Norma 1	N1	$\sqrt[p]{\sum_{i=1}^{n} \lvert x_{a_i} \rvert^p}$	$p = 1$
Norma 2	N2		$p = 2$
Media aritmética	AM	$\left(\frac{x_{a_1}^{\beta} + x_{a_2}^{\beta} + \ldots, x_{a_n}^{\beta}}{n}\right)^{\frac{1}{\beta}}$	$\beta = 1$
Media cuadrática	P2		$\beta = 2$
Media potencial de grado 3	P3		$\beta = 3$
Media armónica	HM		$\beta = -1$
Varianza	V	$\frac{\sum_{i=1}^{n} x_{a_i} - \bar{x}}{n-1}$	$\bar{x}$: promedio
Desviación estándar	SD	$\sqrt{\frac{\sum_{i=1}^{n} (x_{a_i} - \bar{x})^2}{n-1}}$	
Rango	RA	$\text{máx}(x_{a_i}) - \text{mín}(x_{a_i})$	
Coeficiente de variación	VC	$\frac{\sigma}{\bar{x}}$	σ : desviación estándar
Rango intercuartil	i50	$Q_3 - Q_1$	Q_3 : Percentil$_{75}$ Q_1 : Percentil$_{25}$
Asimetría (Skewness)	S	$\frac{n * X_3}{(n-1)(n-2)(\sigma)^3}$	$X_j = \sum_{i=1}^{n} (x_{a_i} - \bar{x})^j$
Curtosis	K	$\frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(\sigma)^4}$	
Autocorrelación de Broto-Moreau	$AC[k]$	$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{a_i} x_{a_j} \delta(d_{ij}, k)$	d_{ij} : distancia de a_i a a_j $\delta(d_{ij}, k)$: delta de Kronecker
Índices gravitacionales	$GV[k]$	$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_{a_i} x_{a_j}}{d_{ij}^k} \delta(d_{ij}, k)$	
Sumas totales	$TS[k]$	$\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{a_i} + x_{a_j}) \delta(d_{ij}, k)$	

De lo anterior, el uso de los operadores matemáticos de la Tabla 6 posibilitan obtener una mayor variedad de descriptores para caracterizar a un péptido. Para una mayor comprensión de los mismos, sus definiciones matemáticas y aplicaciones han sido abordadas recientemente en la literatura (Martínez-López *et al.*, 2019; Marrero-Ponce *et al.*, 2020). Por último, en la notación del símbolo de los descriptores calculados se combinan los símbolos de las propiedades de AAs y operadores de agregación, de la siguiente manera: Total-[propiedad]-[operador].

3.3. Minería de datos

3.3.1. Selección no supervisada de rasgos

Un aspecto esencial es la selección (no accidental) de aquellos rasgos estructurales que sean los más adecuados para el estudio. Razón por la cual, a continuación se presenta uno de nuestros principales aportes. Sea $\mathcal{D} = [x_{ij}]_{n \times m}$ la matriz de descriptores cuyas filas y columnas representan, respectivamente, instancias de pépti-

dos y los valores numéricos asociados a las características moleculares calculadas. A partir de dicha matriz $\mathcal{D}$ se realizó un proceso de selección de variables, diseñado en dos etapas (Figura 9), con el objetivo de identificar un subconjunto optimizado $F^* = \{f_j \mid j \in I \subseteq \{1, 2 \ldots m\}\}$, tal que $f_j = \mathcal{D}^{(j)}$ es la j-ésima columna de la matriz $\mathcal{D}$. Esto fue realizado aplicando el concepto de entropía como criterio univariado de relevancia, lo cual se combina con la medida multivariada de información mutua para captar la redundancia entre variables (Figura 10).

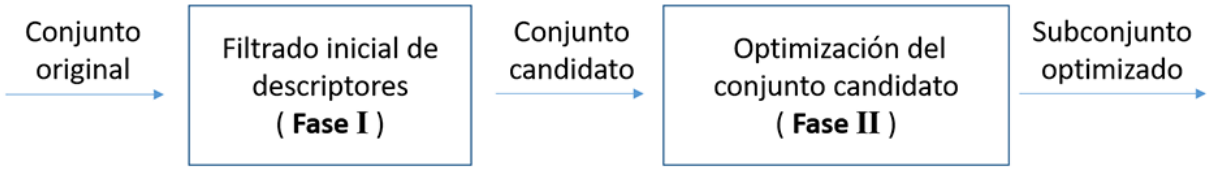


Figura 9. Esquema de selección de rasgos basado en dos etapas.

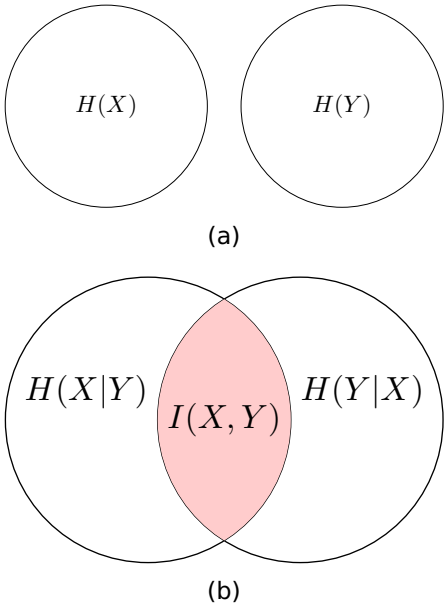
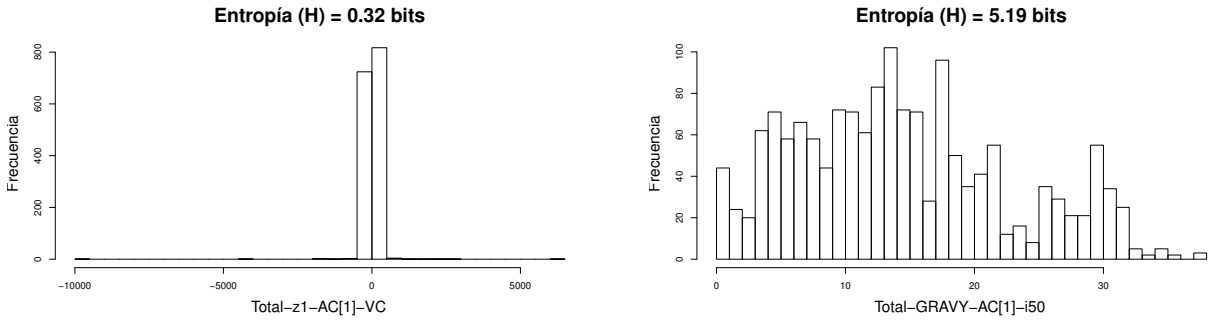


Figura 10. Esta figura muestra la relación entre la entropía $H(\cdot)$ y la información mutua $I(\cdot,\cdot)$ para dos variables X y Y (Cover y Thomas, 2012). $I(X,Y)$ mide el contenido de información compartido entre las dos variables. (a) $I(X,Y)$ es igual a cero si y solo si X y Y son estadísticamente independientes. Por otro lado, (b) $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ mide la dependencia mutua entre las dos variables, es decir, $I(X,Y)$ corresponde a la reducción en la entropía de una variable debido al conocimiento del valor de la otra. Nótese que $I(X,Y)$ puede tomar valores en el intervalo: $0 \leq I(X,Y) \leq \text{mín}\{H(X), H(Y)\}$; cuanto mayor es el valor de $I(X,Y)$, más alta es la relación entre las variables.

El cálculo de la entropía e información mutua se realizó a partir de un esquema de discretización que consiste en dividir el rango de las variables en la misma cantidad

de intervalos (*bins*), contabilizando la cantidad de valores en cada uno de ellos. Esto permite capturar el contenido de información presente en la distribución de valores del descriptor (Figura 11). De esta forma, cada péptido se ubica en un intervalo según el valor calculado por el descriptor, y los valores de entropía serán mayores en la medida en que un mayor número de intervalos sean poblados. Por el contrario, si todos los péptidos son ubicados en pocos intervalos, entonces la entropía será menor.



(a) (b)

Figura 11. Entropía de Shannon (H) calculada para dos rasgos distintos a partir de un esquema de discretización con 50 bins. La entropía máxima es de 5.64 bits.

Matemáticamente, una nueva representación de los datos es posible al dividir el rango de cada variable en el mismo número de bins n_b. Suponga que S_i^j denota el conjunto de péptidos cuyos valores se encuentra en el bin i para el rasgo j, entonces la probabilidad frecuentista $p^j(i)$ de observar un valor en el i-ésimo bin del rasgo j se calcula como:

$$p^j(i) = \frac{|S_i^j|}{\sum_{i=1}^{n} |S_i^j|} \tag{6}$$

Además, sea $S_i^j \cap S_l^k$ el conjunto intersección que contiene los péptidos cuyos valores caen en los intervalos i y l para los rasgos j y k, respectivamente, entonces la probabilidad conjunta $p^{j,k}(i, l)$ se estima como:

$$p^{j,k}(i, l) = \frac{|S_i^j \cap S_l^k|}{\sum_r \sum_s |S_r^j \cap S_s^k|} \tag{7}$$

Siendo estas probabilidades marginales $p^j(i)$ y conjuntas $p^{j,k}(i, l)$ las que se emplean

en los cálculos de la entropía $H(\cdot)$ y la información mutua $I(\cdot,\cdot)$:

$$H(f_j) = -\sum_{i=1}^{n_b} p^j(i) \log p^j(i) \tag{8}$$

$$I(f_j, f_k) = \sum_{i=1}^{n_b} \sum_{l=1}^{n_b} p^{j,k}(i,l) \log \frac{p^{j,k}(i,l)}{p^j(i) p^j(l)} \tag{9}$$

3.3.1.1. Fase I: filtrado inicial de descriptores

En la primera etapa se realiza un filtrado de los descriptores eliminando los que son irrelevantes y redundantes (ver Anexo A). Para ello, los descriptores se ordenan en orden descendente según sus valores de entropía como criterio de relevancia. De tal modo, que son eliminados descriptores irrelevantes cuyos valores de entropía están por debajo de cierto umbral. Luego, comenzando por los mejores posicionados, los descriptores son seleccionados progresivamente siempre y cuando no sean redundantes a los que ya fueron previamente seleccionados.

En esta etapa inicial, para cuantificar la redundancia entre pares de variables se utilizan coeficientes de correlación (Schober *et al.*, 2018). De esta manera, decimos que dos variables son redundantes si presentan una correlación por encima de un valor de umbral. Por ejemplo, el coeficiente de Pearson (ρ) es un buen criterio para medir la asociación lineal entre un par de rasgos f_j y f_k:

$$\rho(f_j, f_k) = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x_j})^2} \sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x_k})^2}} \tag{10}$$

donde x_{ij} y $\bar{x_j}$ indican, respectivamente, el valor i-ésimo y el promedio del descriptor f_j. Además, utilizamos el valor absoluto $|\rho(f_j, f_k)|$ ya que el signo del coeficiente solo indica que, en caso de signo positivo, ambos valores de rasgos tienden a aumentar juntos; o en caso de signo negativo, los valores de un descriptor aumentan a medida que los valores del otro disminuyen. Si dos rasgos tienen una relación de dependencia lineal exacta, $|\rho(f_j, f_k)|$ es 1. En cambio, si son totalmente independientes, $|\rho(f_j, f_k)|$ es 0. Como es de suponer, mientras mayores sean los valores en el intervalo $[0, 1]$, más

alta es la correlación lineal entre los descriptores. Sin embargo, este coeficiente de correlación no es adecuado para capturar la dependencia entre dos variables que no sea de naturaleza lineal.

Por otro lado, el coeficiente de Spearman (r_s) puede ser utilizado para medir la relación monotónica (Schober *et al.*, 2018) entre el par de rasgos f_j y f_k (ya sea lineal o no). Este coeficiente de correlación r_s se basa en la fórmula de Pearson, y utiliza el orden de los valores en lugar del valor en sí mismo:

$$r_s(f_j, f_k) = \rho(rank(f_j), rank(f_k)) \tag{11}$$

donde la función $rank(\cdot)$ denota la transformación de datos en la que los valores originales se reemplazan por su posición luego de ser ordenados. Igualmente utilizamos el valor absoluto del coeficiente de correlación $|r_s(f_j, f_k)|$ como criterio de similitud entre dos rasgos. Si ambos descriptores tienen una relación monótona perfecta, entonces $|r_s(f_j, f_k)|$ es 1; si no están correlacionados, $|r_s(f_j, f_k)|$ es 0. En otro caso, $|r_s(f_j, f_k)|$ se encuentra en el intervalo [0,1], indicando asociaciones monótonas cercanas a 0 (bajas) o próximas a 1 (altas).

3.3.1.2. Fase II: optimización del conjunto candidato

Para la segunda etapa se plantea y resuelve un problema de optimización, con el fin de encontrar un subconjunto óptimo de descriptores al maximizar una función objetivo con base en la entropía e información mutua (Cover y Thomas, 2012). El problema a resolver es el siguiente:

$$\underset{F \in \Omega}{\text{Maximize}} \quad \Phi(F) = \frac{1}{|F|} \sum_{f_j \in F} H(f_j) - \frac{1}{|F|^2} \sum_{f_j, f_k \in F} I(f_j, f_k) \tag{12}$$

donde $\Phi(F)$ es la función objetivo, y F es un subconjunto de descriptores sobre el espacio de búsqueda Ω de todos los subconjuntos posibles, a partir del conjunto candidato.

Al observar la función objetivo $\Phi(F)$, esta se define como una resta entre dos términos. El primero de ellos es el promedio de los valores de entropía, como criterio de

relevancia; mientras que el segundo término es el promedio de los valores de información mutua, como criterio de redundancia. Por lo que se puede afirmar que al resolver el problema planteado en la Ecuación 12, se pretende encontrar un conjunto solución con valor máximo de relevancia y mínimo de redundancia (Peng *et al.*, 2005).

Es importante mencionar que nos hemos basado en un estudio reportado en la literatura (Peng *et al.*, 2005) para la formulación del problema de optimización (Ecuación 12). En dicho reporte previo, consideran que se tiene una variable de respuesta (enfoque supervisado), y en la función objetivo original utilizan como primer término la información mutua entre las variable respuesta y las variables predictoras. De modo que la diferencia con nuestra propuesta radica en la forma de medir la relevancia del conjunto a evaluar. Además, hemos incorporado una etapa de filtrado que permite eliminar tempranamente a aquellos descriptores irrelevantes y redundantes. En particular, son eliminados los descriptores con valores bajos de entropía que afectan el promedio en el minuendo de $\Phi(F)$, y aquellos con alta redundancia que provocan un aumento en el valor del sustraendo.

A pesar de la eliminación temprana de las características irrelevantes y redundantes, una búsqueda exhaustiva en todos los subconjuntos posibles del conjunto candidato es impracticable. Por lo que se implementó un algoritmo heurístico que pese a no dar garantía de encontrar la solución óptima, dio buenos resultados al encontrar un óptimo local para el problema formulado. Dicha implementación es la de un algoritmo de ascenso de colina (ver Anexo B), que comienza con el conjunto completo de descriptores, y explora el espacio de solución al considerar todas las posibles eliminaciones de características (estrategia de eliminación hacia atrás), para quedarse con el subconjunto que maximice la función objetivo propuesta.

3.3.2. Construcción del modelo basado en redes de similitud

Luego de haber identificado y seleccionado al conjunto optimizado de descriptores, se procede a la construcción de una red de similitud ponderada $G = (V, E, w)$ para representar el EQB de los péptidos que sirvieron de origen para el cálculo. De tal modo que los nodos en V representan instancias de péptidos, y cada uno de ellos va a estar caracterizado por un vector de descriptores en $\mathbb{R}^m$, donde m es la cardinalidad del

vector. Además, entre dos nodos u y v, es posible establecer relaciones de similitud a partir de una función de distancia $d(u, v)$ en el espacio métrico ($\mathbb{R}^m$) definido. Para ello, se utilizó una función de transformación que permite convertir la distancia $d(u, v)$ en un valor de similitud $sim(u, v)$, según la fórmula (Todeschini *et al.*, 2020):

$$sim(u, v) = 1 - \frac{d(u, v)}{\text{máx}_{p,q \in V}\, d(p, q)} \tag{13}$$

Sobre la definición de $sim(u, v)$, se conoce de la literatura (de la Vega de León y Bajorath, 2016) que es adecuada para construir y visualizar una red de similitud a partir de un espacio de descriptores. Esto implica la generación de un enlace entre dos nodos u y v de la red, si se cumple que $sim(u, v)$ está por encima de un parámetro de umbral t. El enlace $(u, v) \in E$ se establece asignando los pesos $w : E \rightarrow [0, 1]$ para que indiquen el valor de similitud entre los nodos conectados: $w(u, v) = sim(u, v)$. Por lo tanto, a partir del cálculo de una matriz de similitud $S_M = [s_{ij}]_{n \times n}$, con los valores de similitud s_{ij} entre todos los pares de nodos, se construye la matriz de adyacencia de la red $A = [a_{ij}]_{n \times n}$, cuyos valores son:

$$a_{ij} = \begin{cases} s_{ij} & \text{if } i \neq j, s_{ij} \geq t \\ 0 & \text{en otro caso} \end{cases} \tag{14}$$

De tal manera que es posible construir y estudiar varias redes en función del parámetro t. Sin embargo, el espacio requerido en memoria principal para la matriz de similitud es $O(n^2)$, donde n es el número de nodos, lo cual puede ser muy costoso computacionalmente para grandes volúmenes de datos. Por lo que también se consideró una variante de red de similitud para estos casos: las redes HSP, del inglés Half-Space Proximal (Chavez *et al.*, 2006). Para ello, se implementó un algoritmo paralelo (ver Anexo C) que a partir del espacio métrico $\mathbb{R}^m$ extrae sólo una pequeña fracción del número máximo posible de enlaces entre nodos ($n * (n - 1)/2$), sin necesidad de calcular la matriz de similitud.

A modo de ejemplo, a continuación se describe cómo se construye la red HSP a partir de un conjunto arbitrario de puntos en el plano 2D (Figura 12(a)), suponiendo por simplicidad un espacio métrico de dos dimensiones. Como se puede apreciar en

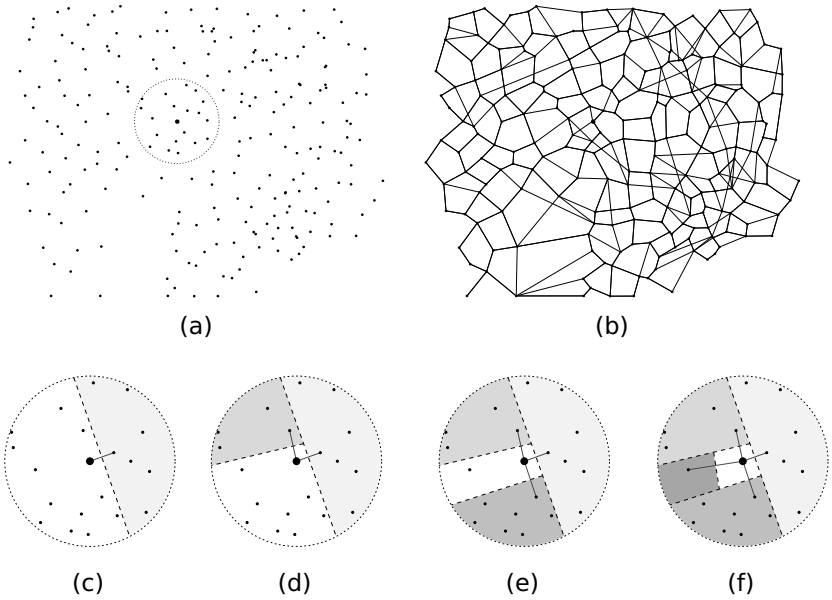


Figura 12. Construcción de la red HSP a partir de un conjunto arbitrario de puntos en el plano 2D: (a) Configuración inicial; (b) Configuración final; (c) Primer vecino; (d) Segundo vecino; (e) Tercer vecino; y, (f) Cuarto vecino.

la Figura 12(a), para un punto inicial u, se toma su vecino más cercano v y se agrega una arista de similitud entre ellos (Figura 12(c)). Suponga ahora que se divide el plano 2D en dos mitades, al trazar una línea perpendicular e imaginaria que pasa por el punto medio de la arista que conecta u y v (Figura 12(c)). Una de las mitades, la que contiene a los puntos más cercanos a v que a u, corresponde al área prohibida para u. De los puntos candidatos restantes que no pertenecen al área prohibida de u, se considera al punto más cercano como su nuevo vecino (Figura 12(d)). Una vez más, se establece el área prohibida y se selecciona entre los restantes el punto más cercano para conectarse a u (Figura 12(e)). Lo cual se repite hasta que el conjunto de puntos candidato esté vacío (Figura 12(f)). Estos pasos se realizan de manera independiente para todos los puntos del espacio (Figura 12(b)), por lo que fueron implementados para ser ejecutados en paralelo, sincronizando el acceso a la estructura de datos para representar el grafo en memoria.

3.3.2.1. Detección de comunidades

Los algoritmos de agrupamiento (en inglés, *clustering*) son una técnica fundamental en el aprendizaje no supervisado. Esta técnica consiste en separar los datos en varios grupos, de modo que los elementos en el mismo grupo son similares y los ele-

mentos en diferentes grupos son diferentes entre sí. Los grupos resultantes se denominan grupos o comunidades en el caso de las redes (Newman, 2018). En particular, las comunidades detectadas en las redes de similitud pueden representar grupos de compuestos ubicados en distintas regiones del EQB.

Un procedimiento efectivo para la detección de comunidades ha sido maximizar la métrica de modularidad de Newman (Newman, 2006). Formalmente, la modularidad (Q) para una red ponderada se define como (Newman, 2004):

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{15}$$

donde a_{ij} es el peso de la arista (en nuestro caso, el valor de similitud) entre los nodos i y j, o 0 si no existe tal conexión, $m = \frac{1}{2} \sum_{ij} a_{ij}$ es la suma total de pesos en toda la red, $k_i = \sum_j a_{ij}$ es la suma de los pesos de las aristas que inciden en el nodo i , c_i es la comunidad a la que se asigna el nodo i y $\delta(c_i, c_j)$ es 1 si ambos vértices i y j son miembros de la misma comunidad ($c_i = c_j$) o 0 de lo contrario. En otras palabras, la modularidad mide la fracción de aristas que se encuentran conectando nodos en una misma comunidad, menos el valor esperado de dicha magnitud en una red con idéntica estructura de comunidades y grados de vértices, pero donde las aristas se colocan al azar.

La modularidad Q puede ser positiva o negativa, en dependencia de si hay menos o más enlaces entre vértices de una misma comunidad que los esperados al azar, y su valor máximo es 1 (Newman, 2006). Por lo tanto, al maximizar el valor Q, se pueden detectar comunidades entre las posibles particiones de una red. El método utilizado para ello fue el conocido algoritmo de *clustering* de Louvain (Blondel *et al.*, 2008), que tiene un funcionamiento rápido y eficiente.

El algoritmo de Louvain ha logrado buenos resultados en términos de precisión y tiempo de cálculo si se compara con otros métodos disponibles en la literatura (Blondel *et al.*, 2008; Yang *et al.*, 2016). Inicialmente, asigna una comunidad diferente a cada nodo de la red. Luego, para cada nodo u, evalúa la ganancia de modularidad que tendría lugar al mover el nodo u a la comunidad de sus vecinos. El movimiento de colocar el nodo u en la comunidad del nodo v se realiza si la ganancia logra la mayor

contribución positiva a la modularidad. Si no se logra una contribución de ganancia positiva, el nodo u permanece en su comunidad original. Esta primera fase se repite para todos los nodos hasta que no sea posible una mejora adicional. Más tarde, en una segunda fase, el algoritmo construye una nueva red cuyos nodos son las comunidades encontradas en la fase anterior. Una vez más, la primera fase se vuelve a aplicar a la nueva red y se detiene cuando no hay un movimiento que pueda mejorar la modularidad.

3.3.2.2. Identificando los nodos más relevantes

El concepto de centralidad en una red ha sido ampliamente utilizado en diversas ramas de la ciencia, como por ejemplo, en el análisis de redes sociales para identificar las entidades influyentes (Landherr *et al.*, 2010; Lü *et al.*, 2016; Newman, 2018). Por supuesto, hay muchas definiciones posibles de lo que significa ser un nodo central y, en consecuencia, hay muchas formas de calcular la centralidad. Formalmente, una medida de centralidad $C(u)$ es una función que asigna un número real no negativo a cada nodo u en la red (Newman, 2018). De este modo, la tarea analítica consiste en encontrar aquellos nodos que despiertan mayor interés según los valores de $C(u)$ (Csermely *et al.*, 2013). Con este fin, se pueden utilizar métricas globales que consideren a toda la red, o locales que utilicen solo información del vecindario (Lü *et al.*, 2016).

La centralidad armónica (Boldi y Vigna, 2014) es un ejemplo de métrica global, y para un nodo i se define como:

$$C_H(i) = \sum_{j \neq i} \frac{1}{d_g(i,j)} \tag{16}$$

donde la distancia geodésica $d_g(i,j)$ es la longitud de la ruta más corta de i a j. Por convención, $d_g(i,j) = \infty$ y $1/\infty = 0$ si no existe dicha ruta cuando se trata de nodos inalcanzables en redes desconectadas. Por otro lado, en caso de existir la ruta más corta, su longitud se calcula utilizando el algoritmo de Dijkstra. Eso significa que la complejidad computacional en el peor de los casos es cuadrática en el número de nodos, lo cual puede resultar ineficiente para analizar grandes redes (Brandes y Pich, 2007).

Por otra parte, a diferencia de las globales, las métricas locales solo se basan en información alrededor de los nodos (Lü *et al.*, 2016), y aquí mostramos su utilidad al explotar la estructura de comunidades. De hecho, es razonable identificar cuáles serían los péptidos más relevantes al analizar las comunidades detectadas, ya que estos pueden desempeñar su papel dependiendo de la posición que ocupan en la comunidad a la que pertenecen (Csermely *et al.*, 2013; Lepp *et al.*, 2009). Por ejemplo, los nodos pueden actuar como centros locales (en inglés, *hubs*) si se encuentran conectando a muchos nodos internos en sus propias comunidades, mientras que los que están ubicados en la frontera de la comunidad pueden servir como puentes (en inglés, *bridges*) entre diversos grupos.

Al considerar una red con estructura de comunidades, el número de enlaces intra e inter-comunitarios pueden ser útiles para identificar aquellos nodos que actúan como centros o puentes. Primero, la fuerza de conectividad interna y externa para cada nodo i, en su comunidad c_i, se mide usando la fórmula:

- Fuerza interna: $k_i^{int} = \sum_{j \in c_i} a_{i,j}$
- Fuerza externa: $k_i^{ext} = \sum_{j \notin c_i} a_{i,j}$

donde $a_{i,j}$ contiene el peso de la arista (i, j), es decir, el valor de similitud entre los nodos i y j, o 0 si no existe dicha arista.

Cabe señalar que si los nodos i y j son adyacentes, el valor de $a_{i,j}$ sería 1 para una red no ponderada, 0 en caso contrario. Por lo tanto, en la contraparte no ponderada, la fuerza interna k_i^{int} sería análoga a la cantidad de vecinos que pertenecen a la misma comunidad c_i, mientras que la fuerza externa k_i^{ext} sería el número de vecinos que no pertenecen a la misma comunidad. Siguiendo esta analogía, la fuerza total $k_i^{int} + k_i^{ext}$ sería el grado del nodo i.

Con base en la idea anterior, las medidas k_i^{int} y k_i^{ext} se pueden combinar para calcular una métrica de centralidad propuesta recientemente, llamada por su nombre en inglés: “Community Hub-Bridge centrality” (Ghalmane *et al.*, 2019). Esta métrica se define como:

$$C_{HB}(i) = k_i^{int} * card(c_i) + k_i^{ext} * nnc(i) \tag{17}$$

donde $card(c_i)$ es el tamaño de la comunidad a la que pertenece el nodo i, y $nnc(i)$ es el número de comunidades vecinas a la que un nodo i puede alcanzar mediante sus enlaces externos. Al ordenar los vértices en orden descendente de importancia según los valores de $C_{HB}(i)$, se estarían priorizando los nodos *hubs* y *bridges*. Por un lado, los nodos *hubs* pueden representar moléculas centrales en distintas regiones del EQB. Por otro lado, los nodos *bridges* pueden simbolizar estructuras intermedias que enlazan distintos grupos o familias de péptidos en el EQB.

3.4. Evaluación e interpretación

3.4.1. Evaluación de los descriptores calculados

Un estudio comparativo se realizó para analizar la pertinencia de los descriptores que se calculan con base en los distintos operadores de agregación. La comparación fue hecha en relación con otras familias de descriptores disponibles en el paquete de software iFeature recientemente publicado (Chen *et al.*, 2018). El estudio no supervisado consistió en un Análisis de Variabilidad (AV) (Godden *et al.*, 2000; Godden y Bajorath, 2002) y un Análisis de Componentes Principales (ACP) (Hair *et al.*, 1998; Jolliffe y Cadima, 2016).

Primero, el AV nos permitió cuantificar y analizar el contenido de información presente en los conjuntos de descriptores por medio de las distribuciones de entropía de Shannon (Godden *et al.*, 2000; Urias *et al.*, 2015). De esta manera, se pueden observar y comparar las distribuciones de entropía obtenidas con los descriptores calculados y con los índices de iFeature. Además, los mejores descriptores según sus valores de entropía fueron analizados en un ACP.

El ACP es un procedimiento matemático que permite describir un conjunto de variables en términos de nuevas variables no correlacionadas, llamadas componentes principales. El primer componente identificado será aquel que explique la mayor varianza de los datos. Los subsecuentes explicaran una menor porción de la varianza en orden decreciente. Dado que los componentes extraídos tienen la propiedad de ser ortogonales entre sí, es útil observar las variables que se cargan en componentes distintos, indicando la presencia de no-colinealidad entre los descriptores analizados.

3.4.2. Evaluación de los descriptores seleccionados

En la primera etapa, los descriptores filtrados forman parte de un conjunto candidato en representación del original, y para evaluar su calidad se realiza un análisis de Procrustes (Gower, 1975). El cual nos permite medir las proximidades entre los descriptores originales y los retenidos en el conjunto candidato (King y Jackson, 1999; Ballabio *et al.*, 2014). Siendo el objetivo de esta fase lograr un conjunto reducido de descriptores que sea representativo y no diste mucho del conjunto original.

Para el análisis de Procrustes, se utilizó la implementación disponible en la biblioteca SciPy de Python (Virtanen *et al.*, 2020). Además, debido a que este análisis se debe realizar con matrices de igual dimensión, un buen criterio es utilizar los primeros componentes principales (Jolliffe y Cadima, 2016), en igual número, para ambos espacios de descriptores: el original y el reducido (King y Jackson, 1999; Ballabio *et al.*, 2014). Esto con la condición de que los componentes principales elegidos capturen, al menos, el 80 % de la variabilidad explicada en el espacio original.

Una vez seleccionados los componentes principales, siguiendo con el análisis de Procrustes, se realizan transformaciones a una de las matrices hasta lograr un ajuste óptimo entre los puntos de los dos espacios de descriptores. La transformación óptima se realiza aplicando cambios en la escala, rotación y traslación, para minimizar la suma de los errores al cuadrado (criterio de bondad de ajuste) entre los puntos definidos por las matrices a comparar. Un criterio de bondad de ajuste igual a 0 indica que ambos espacios descriptores coinciden, mientras que un valor igual a 1 expresa que los dos espacios descriptores son completamente diferentes.

Finalmente, dado que el algoritmo heurístico implementado no da garantía de encontrar una solución óptima, se realiza un análisis del impacto que tiene el subconjunto optimizado frente al conjunto candidato de descriptores. La primera parte de este análisis consiste en una comparación estadística para detectar diferencias entre los valores de la función objetivo. Para ello, se realiza una prueba de rangos de Wilcoxon para datos apareados (Hollander *et al.*, 2013), con corrección Hochberg (Hochberg, 1988), usando el paquete de software R. Además, en la segunda parte de este estudio se analiza la interpretabilidad del modelo obtenido al usar los conjuntos de descriptores optimizados, y sin optimizar.

3.4.3. Interpretación del modelo basado en redes de similitud

Teniendo en cuenta que la percepción de la similitud está en los ojos del espectador (Maggiora *et al.*, 2014), consideramos que el umbral $t \in [0, 1]$ que se utiliza para construir las redes (ver sección 3.3.2), es un parámetro que debe ser explorado de forma interactiva en el contexto donde el modelo se va a emplear. Por un lado, un valor constante predefinido para t no es útil en la práctica porque la distribución de los valores de similitud varía según el conjunto de péptidos que se estén analizando. Por otro lado, cuando se modifica el umbral t, se espera que ocurran cambios en el número de conexiones de similitud, mientras que la cantidad de nodos permanece inalterable. Por lo tanto, algunas propiedades de la red dependen del parámetro t, como por ejemplo, su densidad, que se calcula como:

$$network\_density(t) = \frac{2m(t)}{n(n-1)} \tag{18}$$

donde n y $m(t)$ representan, respectivamente, la cantidad de nodos de la red y aristas que dependen del parámetro t.

Como puede observarse en la ecuación anterior, la densidad de red se define como la relación entre el número de conexiones reales que existen, y el número total de posibles conexiones. En el caso extremo en el que el umbral $t = 0$, todos los compuestos se consideran similares y se traza la red con el mayor número de aristas, es decir, la densidad alcanza su máximo valor. Por el contrario, cuando $t = 1$, se obtiene una red mínimamente conectada, donde la densidad es nula o muy cercana a cero. En general, cuando t aumenta, la densidad de la red tiende a disminuir y viceversa (Figura 13).

Por supuesto, la tarea de comprender lo que las redes nos pueden decir depende de un valor elegido de t. Esta tarea de interpretar las redes puede volverse más complicada con valores bajos de umbral, ya que producen subgrafos densamente conectados y una imagen con demasiadas líneas que dificultan la percepción visual (Ware, 2019). Además, las grandes redes con miles de nodos y millones de relaciones pueden requerir un alto uso de memoria (Pavlopoulos *et al.*, 2017), con una complejidad espacial cuadrática.

Una forma sencilla de lograr niveles bajos de densidad sería usar un valor alto de

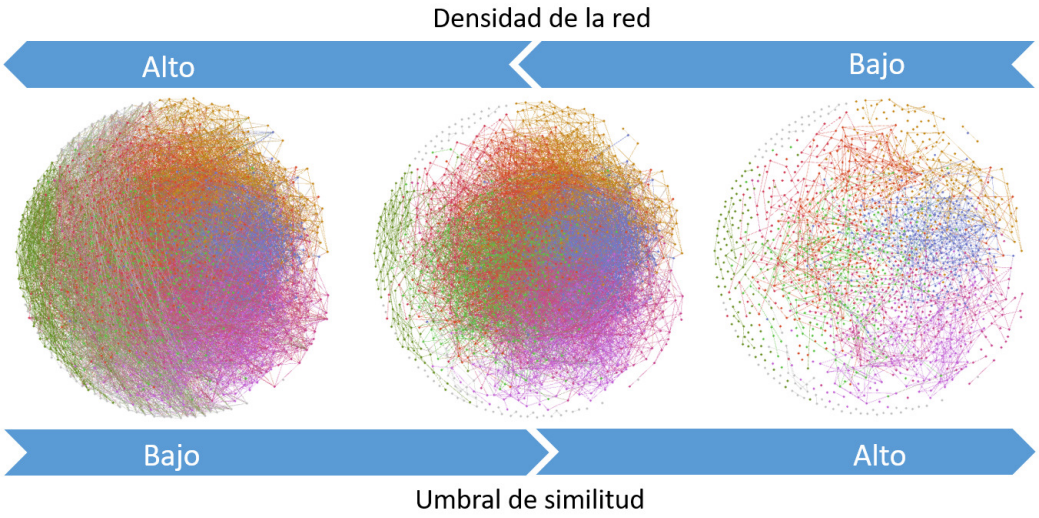


Figura 13. Ilustración de la densidad de red a distintos umbrales de similitud. Los colores representan comunidades a la que pertenecen los nodos.

umbral *t* (Figura 13). No obstante, aumentar demasiado este valor de umbral puede dar lugar a redes con pérdida de información, ya sea porque se están desconectando muchos nodos o bien porque se están aislando grupos de ellos. Por lo tanto, el parámetro *t* debe manejarse con cuidado para lograr gráficos interpretables donde la información topológica de la red no esté oculta.

Por último, una vez que se tenga un valor de umbral establecido, se procede a contrastar el conocimiento extraído de la red bajo el enfoque no supervisado. Esto comprende que los patrones descubiertos sean interpretados y, de ser necesario, se retorne a una de las fases anteriores para una nueva iteración. Además, se analizó la utilidad de los patrones detectados y su aplicabilidad para apoyar la toma de decisiones en la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

3.5. Difusión y uso

Un punto clave en la presente investigación es hacer partícipe a los investigadores de química biomédica, o potenciales usuarios, ávidos de descubrir conocimiento a partir de la colección integrada de péptidos bioactivos. Ellos no tienen que ser expertos en cada una de las fases implementadas en el proceso KDD, pero sí deben tener las habilidades cognitivas e innatas de los humanos para detectar patrones en el dominio de aplicación. Por lo tanto, se desarrolló un producto de software cuyo nombre es

“starPep toolbox”, que dota a los usuarios finales de funcionalidades necesarias para construir e inspeccionar los modelos descriptivos según sus propios intereses. Con la utilización de starPep toolbox, se pretende incorporar y facilitar el análisis visual como parte de un flujo de trabajo propuesto para asistir en la extracción de información útil.

3.5.1. Herramienta informática “starPep toolbox”

StarPep toolbox (Figura 14) ha sido diseñada para que sea fácil de usar y extensible mediante *plugins*, es decir, módulos que añaden nuevas funcionalidades a la aplicación. Su importancia se debe a que permite la generación y el análisis visual del modelo descriptivo, obtenido como resultado del flujo de trabajo implementado. La herramienta fue desarrollada en Java sobre la plataforma Netbeans (Bock, 2012), que da soporte a la arquitectura basada en *plugins*, y se encuentra disponible para su uso en `http://mobiosd-hub.com/starpep`.

En su implementación se utilizó la biblioteca proporcionada por Neo4j para interrogar la base de datos orientada a grafos (Figura 8), que se encuentra embebida dentro de la aplicación. También, se han reutilizado algunas de las funcionalidades disponibles en el proyecto de código abierto Gephi (Bastian *et al.*, 2009), que es idóneo para garantizar la visualización interactiva y el análisis de las redes (Pavlopoulos *et al.*, 2017).

3.5.1.1. Principales funcionalidades

Para ilustrar la utilidad de starPep toolbox, a continuación se mencionan algunas de sus prestaciones en posibles escenarios de uso. No es el propósito aquí enseñar cómo usar la herramienta, ya que un manual de usuario para realizar las operaciones disponibles se puede encontrar en el sitio web de la misma.

Múltiples vistas de datos. Cuando se trabaja con varias vistas de datos (a las cuales llamamos *workspaces*, en inglés), los usuarios pueden alternar entre ellas dentro de la misma ventana. Por cada vista de datos se puede tener un modelo descriptivo correspondiente a los péptidos en estudio.

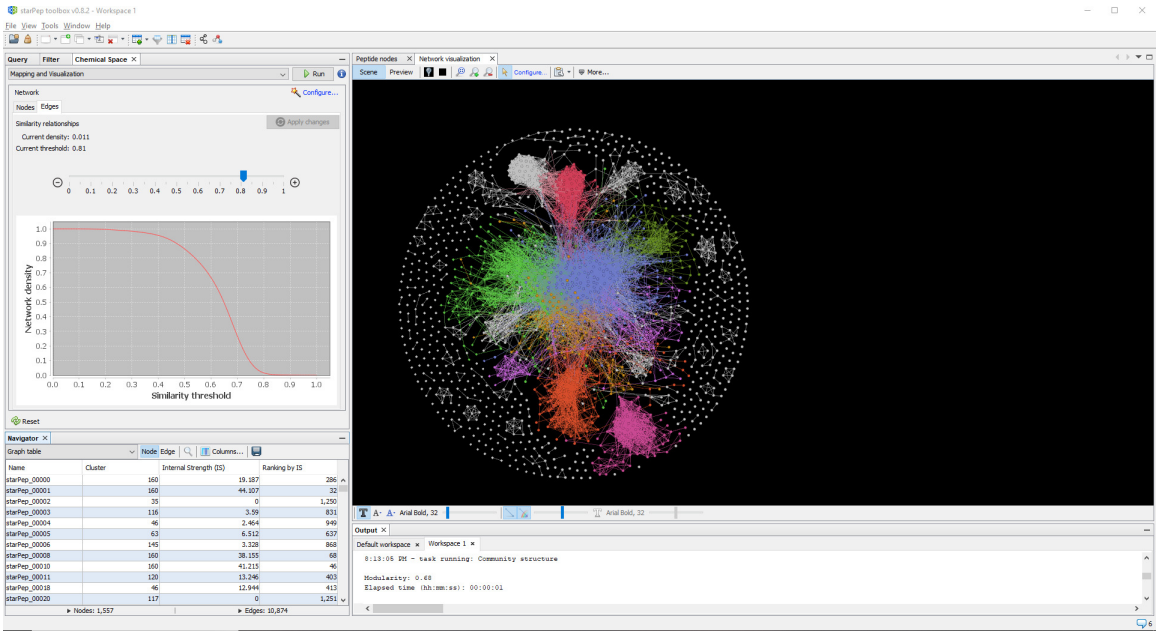


Figura 14. Captura de pantalla de starPep toolbox v0.8.

Asistente para construir el modelo descriptivo. La herramienta contiene un asistente (en inglés, *wizard*) de configuración para el cálculo del modelo descriptivo. Utilizando este *wizard* se pueden modificar parámetros con valores por omisión, según los resultados obtenidos, de los algoritmos que componen el flujo de trabajo. Por lo tanto, el cálculo se realiza de forma automática, y es posible generar una vista inicial del modelo descriptivo para los péptidos recuperados de la base de datos o importados desde un fichero externo en formato FASTA (Mount, 2004).

Proyección de nuevos compuestos peptídicos. Proyectar nuevas secuencias en un modelo descriptivo basado en redes de similitud de péptidos bioactivos conocidos, puede servir para identificar la región que ocupan los nuevos compuestos en el EQB de referencia. En primer lugar, se calculan los descriptores utilizados en la construcción de la red a cada uno de los nuevos compuestos peptídicos. Luego, se establece una conexión (relación de similitud) entre los péptidos a proyectar y sus k vecinos más cercanos en el EQB de referencia (siendo el parámetro k definido por el usuario). El cómputo de la relación distancia/similitud entre los nuevos compuestos y los k vecinos de referencia se realiza teniendo en cuenta el espacio métrico definido por los rasgos ya utilizados (Ecuación 13).

Consulta por metadatos. Los usuarios pueden recuperar de la base de datos integrada a todos los péptidos bioactivos, o en su lugar, aquellos compuestos vinculados a nodos específicos de metadatos. De forma predeterminada, todos los péptidos son recuperados a menos que el usuario proporcione nodos de metadatos para restringir el resultado de la búsqueda.

A modo de ejemplo, para recuperar a todos los péptidos producidos por bacterias, el usuario solo tiene que agregar a la consulta el metadato etiquetado como origen, cuyo nombre es "*Bacteria*". Una vez que se haya ejecutado la consulta, se recuperan todos los péptidos visitados a partir del metadato seleccionado, siguiendo las relaciones *is_a* y *produced_by*, según sea el caso (Tabla 3).

Consulta por identidad de secuencia. Igualmente es posible recuperar péptidos de la base de datos utilizando dos opciones de búsqueda por identidad de secuencia. Al realizar la consulta por una de estas opciones: i) y ii), los péptidos recuperados deben tener i) una identidad de secuencia por encima de un umbral dado respecto a una secuencia de entrada, o ii) una identidad de secuencia por debajo de un umbral dado entre pares de péptidos (conjunto no redundante).

Para llevar a cabo las comparaciones de identidad de secuencia se utilizó la biblioteca BioJava (Lafita *et al.*, 2019). En este punto, el usuario puede elegir el algoritmo de alineamiento global (Needleman y Wunsch, 1970), o local (Smith *et al.*, 1981), y la matriz de sustitución utilizada en el algoritmo. Además, dado que esta búsqueda es computacionalmente costosa, evitamos comparar una secuencia determinada con todas las demás implementando una versión Java de los algoritmos USEARCH y UCLUST (Edgar, 2010).

Filtrado de péptidos. Se pueden aplicar filtros para seleccionar los péptidos de interés luego de ejecutar una consulta por metadatos o identidad de secuencia. Algunos criterios de filtro son:

i) filtro por atributos de péptidos, tales como secuencia o longitud. Por ejemplo, la secuencia de cada péptido puede ser filtrada por un patrón dado (expresión

regular), o la longitud de cada péptido filtrado puede ser menor/mayor o igual que un valor dado. También, pueden filtrarse los péptidos según algunos atributos calculados, como son los descriptores moleculares o métricas de centralidad de las redes.

ii) por metadatos. Los péptidos filtrados deben estar vinculado a un nodo de metadatos cuyo nombre coincide o cumple ciertas reglas (contiene, termina o empieza) con una cadena de texto determinada.

Opciones avanzadas de filtrado y consulta. Estas opciones permiten combinar filtros y consultas mediante los operados lógicos Y / O / NO. Dichas combinaciones de filtros y consultas son adecuadas para muchas situaciones en las que un usuario, en su rol de analista, necesita refinar la selección de los péptidos eliminando del modelo los que resulten ser poco interesantes.

Visualizador de estructuras 3D. Las estructuras terciarias vinculadas a los péptidos pueden mostrarse dentro de la herramienta informática gracias a la reutilización del programa Jmol (Herraez, 2006). Esto sería posible en caso de que exista el vínculo entre el péptido y el metadato cuyo nombre es la referencia cruzada a la entrada en el banco de datos PDB (`https://www.rcsb.org/`). Se requiere una conexión a Internet para que el programa pueda acceder a la estructura 3D depositada en el repositorio PDB. Esta funcionalidad puede ser muy útil para adentrarse en los detalles de una molécula de interés.

Visualización interactiva de la red. Durante la visualización interactiva, la red es dibujada para que sea fácil y agradable de entender. En este proceso de dibujo, los algoritmos de diseño proporcionados por Gephi reorganizan la ubicación de los nodos de acuerdo con algunas reglas estéticas (Bastian *et al.*, 2009). También, el tamaño y color de los nodos pueden ajustarse de forma proporcional a la métrica de centralidad definida (Ecuación 17) para revelar patrones ocultos. Alternativamente, la herramienta muestra una vista de la red en forma de tabla, donde los nodos pueden ordenarse de acuerdo a la métrica de centralidad antes mencionada.

Opciones de exportación de datos. Es posible exportar la red construida en formato GraphML (Brandes *et al.*, 2001), que es un formato estándar basado en XML para representar la estructura de un grafo. Los datos tabulados de la red también se pueden exportar en formato CSV, lo que es más familiar para muchos usuarios. Además de la red, las secuencias de los péptidos y los descriptores moleculares calculados (u optimizados) pueden exportarse en formato FASTA y CSV, respectivamente.

3.6. Conclusiones parciales

En este capítulo se introdujeron las fases de un proceso de extracción de conocimientos a partir de 40 repositorios biológicos de péptidos bioactivos. En la primera etapa, se recopilan e integran las secuencias peptídicas y sus metadatos. Luego, se seleccionan las muestras de péptidos que resulten de interés para un estudio en particular. Es importante resaltar que debido a que existe cierto nivel de redundancia en las secuencias de péptidos almacenadas, se deben limpiar los datos para garantizar diversidad estructural, evitando así la sobrerepresentación de fragmentos de secuencia en el conjunto inicial que sirve de origen para el cálculo.

Vencidas las dificultades de la preparación de los datos, los péptidos se deben presentar a la fase de minería de datos, y para ello cada secuencia de AAs es codificada a través de un vector de descriptores moleculares, combinando propiedades biológicas y operadores de agregación. En esta etapa del proceso, el aumento del poder computacional ha permitido un incremento en el número de descriptores posibles a calcular. De los cuales se deben seleccionar aquellos que permitan mejorar la calidad del modelo, sin que exista un criterio a priori para saber cuáles son los mejores en todos los casos. A pesar de esto último, es de vital importancia realizar esta tarea de selección de rasgos, y para hacerlo se diseñó una estrategia no supervisada basada en dos etapas. Durante la primera etapa, se selecciona un conjunto candidato en representación del original. En la segunda etapa, se ofrece un subconjunto optimizado de rasgos, aplicando los conceptos de entropía de Shannon e información mutua en el diseño de una función objetivo modificada, para guiar la búsqueda hacia los subconjuntos de rasgos que son de alta relevancia y baja redundancia entre sí.

Los descriptores ya optimizados sirven para crear el modelo propuesto basado en

redes de similitud molecular. Dicho modelo descriptivo está sujeto a ser analizado e interpretado de manera visual por medio de la interacción hombre-computadora, utilizando la herramienta informática desarrollada con este propósito, denominada starPep toolbox. Con el análisis visual, se pretende que sea posible comprender de forma apropiada el modelo descriptivo, y que también sea viable inspeccionar los distintos conglomerados (clusters) obtenidos por el algoritmo de Louvain, e identificar los péptidos más relevantes según la métrica de centralidad definida para esto.

Capítulo 4. Resultados y discusión

El flujo de trabajo que se muestra en la Figura 15 ha sido propuesto para la exploración y el análisis visual de un EQB de péptidos bioactivos (Aguilera-Mendoza *et al.*, in-press). Todos estos pasos pueden ser realizados mediante el uso de la herramienta informática starPep toolbox (`http://mobiosd-hub.com/starpep/`), resultado de la investigación con fines prácticos. En lo que compete a este estudio, realizamos una experimentación para evaluar y ajustar la construcción del modelo basado en redes de similitud molecular. También, ilustramos la aplicabilidad del flujo propuesto para extraer información útil de un EQB de péptidos anticancerígenos, elegidos como caso de estudio a partir de la base de datos integrada de péptidos bioactivos.

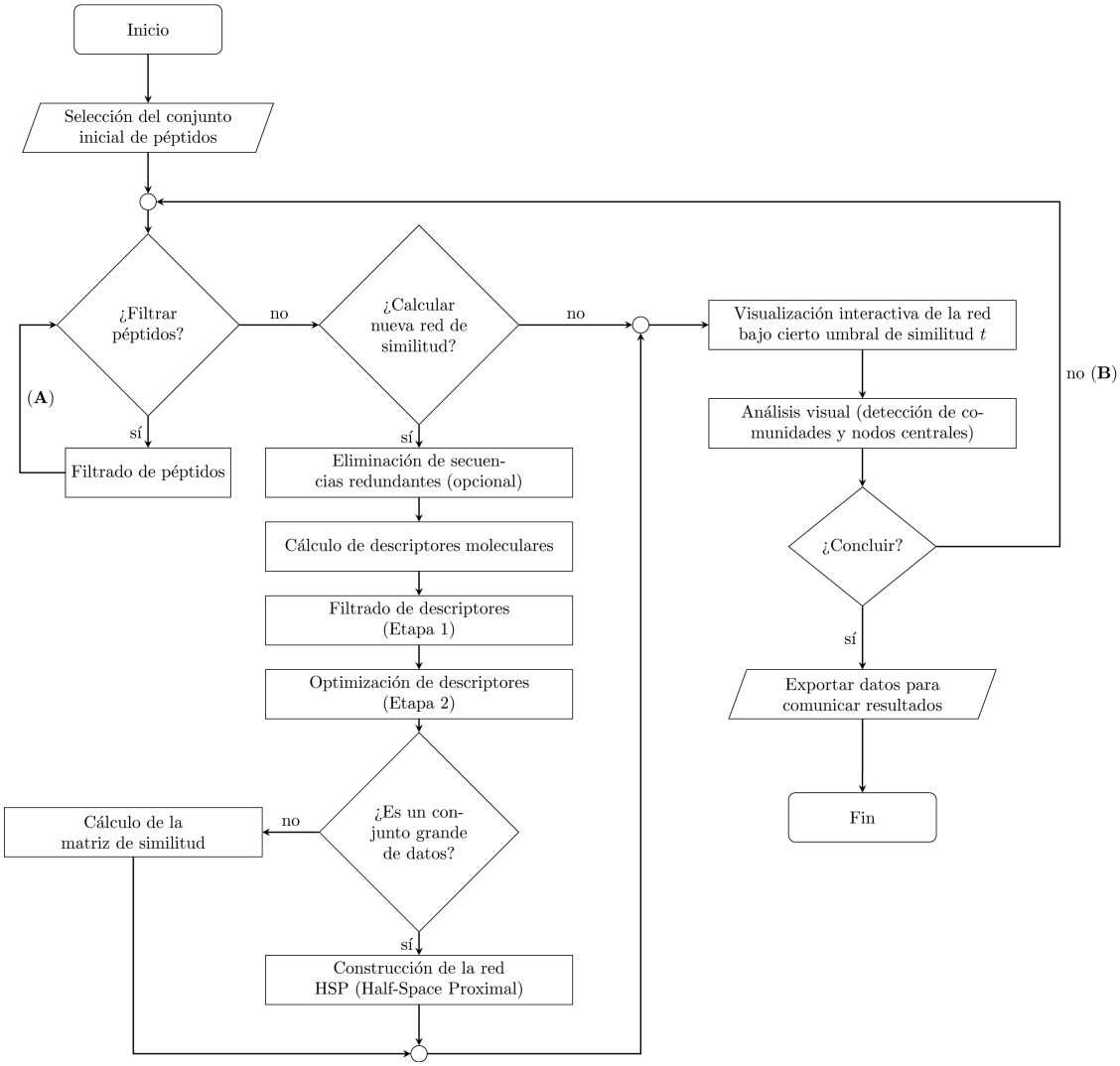


Figura 15. Un diagrama de flujo para guiar la construcción automática y el análisis visual de las redes de similitud molecular. La red de similitud puede generarse automáticamente una primera vez e incluso podría regenerarse en próximas iteraciones. El bucle más interno (A) es de manipulación de datos y el más externo (B) es de retroalimentación visual.

4.1. Base de datos integrada de péptidos bioactivos

Una nueva base de datos orientada a grafos se aporta como resultado de la investigación realizada (Aguilera-Mendoza *et al.*, 2019). Esta base de datos se denomina starPepDB, y se ha logrado luego de recopilar e integrar datos procedentes de una gran variedad de repositorios biológicos (Tabla 2). Además, se encuentra embebida en la herramienta starPep toolbox para facilitar el uso y análisis de los datos unificados. En total, starPepDB almacena 71,310 nodos y 348,505 relaciones, donde 45,120 nodos representan péptidos, y el resto de los nodos están conectados a los péptidos para describir metadatos (Figura 8). La cantidad detallada de estos tipos de nodos se puede ver en la Tabla 2.

En el llenado de starPepDB para su posterior análisis, se descartaron las secuencias peptídicas de longitud superior a 100 AAs. También, fueron rechazados los péptidos con un alto número (> 50 %) de AAs desconocidos o inusuales, que se simbolizan con una “X”. Como resultado, la Figura 16 describe los datos obtenidos en relación a estos dos criterios de longitud y composición de AAs. Por un lado, todos los 20 AAs esenciales (Tabla 1) están presentes, en general, a pesar de que no todos predominan por igual (Figura 16(a)). Por otro lado, la mayoría de los péptidos almacenados son relativamente cortos (Figura 16(b)), que es una característica intrínseca de estos compuestos bioactivos.

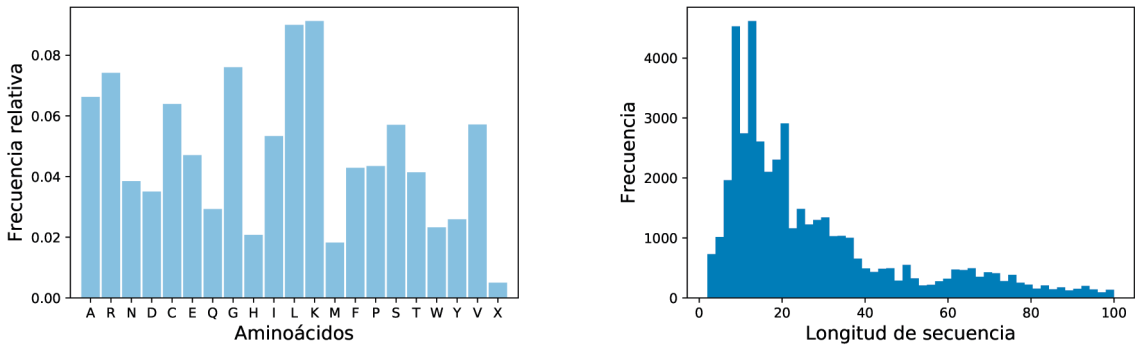


Figura 16. Histograma acerca de los péptidos recopilados e integrados: (a) muestra la frecuencia de aparición de los aminoácidos; y (b) la distribución de la longitud de secuencia.

En relación a lo que aporta starPepDB, debemos decir que es una de las bases de datos más completas y diversas en su campo, hoy en día. Como prueba de ello, la

Figura 17 presenta un mapa de calor para visualizar el traslape que existe entre las fuentes de datos utilizadas en el estudio. Se puede apreciar que existe cierto nivel de solapamiento entre las bases de datos existentes, esto si se analiza el contenido de las celdas fuera de la diagonal principal. Sin embargo, starPepDB es la única que incluye las secuencias de péptidos disponibles en todas las fuentes de datos. Por lo tanto, starPepDB constituye una valiosa fuente de información, cuyos datos pueden ser analizados mediante el flujo propuesto (Figura 15).

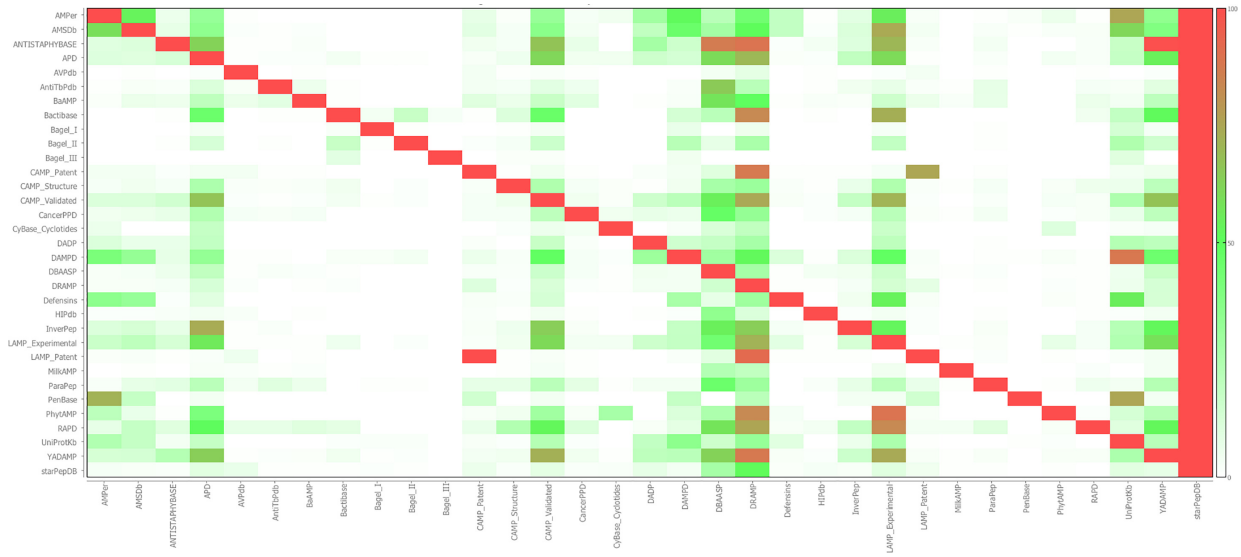


Figura 17. Mapa de calor para visualizar los porcentajes de secuencias compartidas entre bases de datos. Los valores porcentuales se sustituyen por celdas coloreadas de acuerdo con la escala de colores dada en la barra lateral derecha. Dicha escala va del blanco al rojo pasando por el verde. De tal modo, que la intensidad del color en una celda en particular indica la fracción de la base de datos de la fila que se traslapa con la base de datos de la columna. Siendo starPepDB la única que las incluye a todas (última columna a la derecha).

4.2. Estudio comparativo de los descriptores calculados

Durante la ejecución de un flujo de trabajo (Figura 15), cientos o miles de descriptores pueden ser calculados como resultado de aplicar distintos operadores de agregación a vectores de propiedades de aminoácidos (Martínez-López *et al.*, 2019; Marrero-Ponce *et al.*, 2020). Estos descriptores calculados son diferentes a los que comúnmente se emplean en la codificación de péptidos (Todeschini y Consonni, 2009; Jenssen, 2011), por lo que se procede a evaluar la pertinencia de los mismos realizando un análisis comparativo con varias familias de índices disponibles en el paquete de software iFeature (Chen *et al.*, 2018). Para dicha comparación, se seleccionaron de la

base de datos integrada a 748 secuencias de péptidos con 30 aminoácidos de longitud cada una.

Por un lado, los descriptores calculados con iFeature se componen de 3340 índices de composición de aminoácidos (AAC), 310 índices de composición de aminoácidos agrupados (GAAC), 360 índices de autocorrelación (Autocorrelation), 273 índices de composición-transición-distribución (C/T/D), 100 índices de orden de cuasi-secuencia (QSO), 85 índices de composición de pseudo-aminoácidos (PAAC), 430 índices Pse-KRAAC, 15930 índices de aminoácidos (AA), 600 índices BLOSUM62 y 150 índices de escala Z (Z-scale). Por otro lado, con starPep toolbox se calcularon los descriptores mediante la combinación de las propiedades de aminoácidos (Tabla 5), con los distintos operadores de agregación que fueron implementados.

Cabe señalar que todas las familias de descriptores de iFeature se unieron en un solo conjunto de datos para una comparación adecuada. También, se filtraron los descriptores calculados para eliminar rasgos redundantes con base en la correlación de Spearman y un umbral igual a 0,95. En consecuencia, se conservaron un total de 12018 descriptores iFeature y 8416 descriptores starPep, siendo estos dos conjuntos de descriptores los que se utilizaron en el Análisis de Variabilidad (AV).

El software IMMAN (Urias *et al.*, 2015) fue utilizado para realizar el AV de los descriptores calculados por iFeature y starPep toolbox. El objetivo del estudio es cuantificar y comparar el contenido de información codificado por los descriptores calculados. De esta manera, los descriptores relevantes pueden identificarse según el principio de que los valores de alta entropía corresponden a aquellos descriptores con una buena capacidad para discriminar entre péptidos estructuralmente diferentes, mientras que los valores de baja entropía son indicativos de lo opuesto. Aquí, el esquema de discretización adoptado para el cálculo de la entropía es igual a 748 bins, que es el número de péptidos contabilizados. Por lo tanto, la entropía máxima para cada descriptor es igual a 9,55 bits.

Como resultado del AV, la Figura 18 muestra la distribución de la entropía de Shannon correspondiente a los mejores 8000 descriptores, de acuerdo a sus valores de entropía. Como puede observarse, los descriptores de starPep presentan una mejor capacidad para discriminar entre secuencias de péptidos estructuralmente diferentes

si se comparan con los descriptores iFeature, ya que los primeros tienen una mejor distribución que los segundos. Los descriptores starPep analizados presentan un valor medio de entropía igual a 7,59 bits, mientras que los descriptores iFeature tienen un valor medio de entropía igual a 4,29 bits. En general, se puede concluir que los descriptores de starPep tienen mejor contenido de información que varias familias de índices calculadas con iFeature.

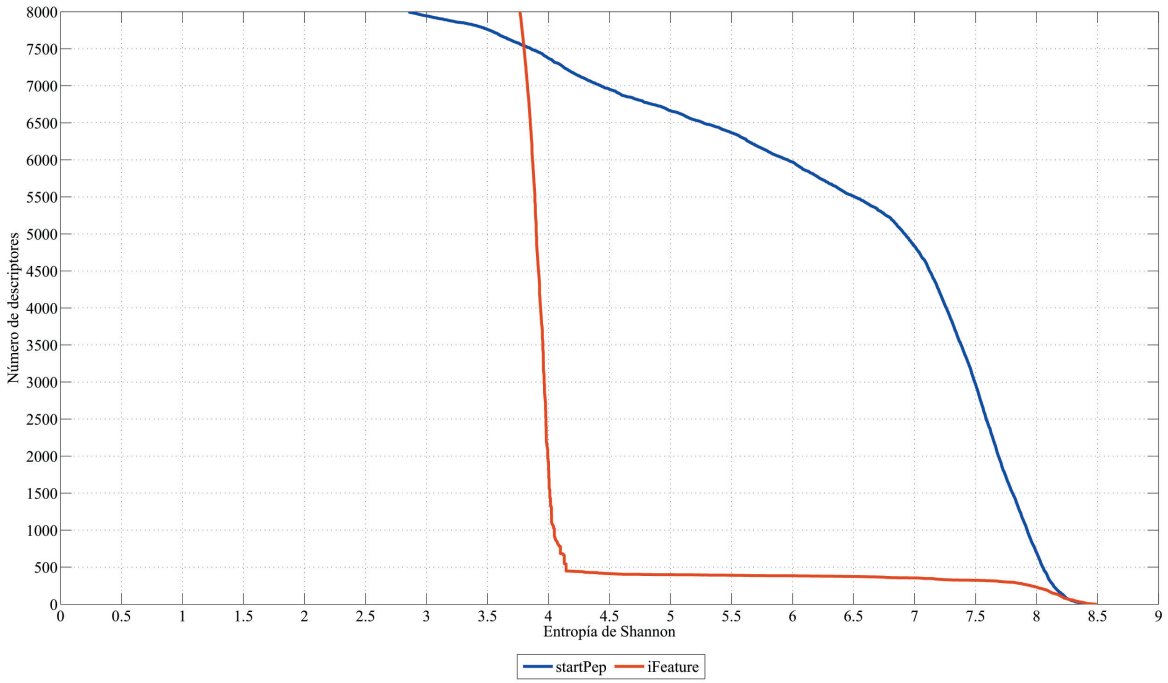


Figura 18. Distribución de entropía de Shannon para los descriptores de iFeature versus los descriptores de starPep.

Además, la Figura 19 muestra los gráficos de diagrama de caja correspondientes a cada familia de índices de iFeature (sin eliminar índices correlacionados), así como los correspondientes a los descriptores de starPep. En esa figura, se puede observar que las mejores familias de descriptores de iFeature corresponden a los índices de autocorrelación, Q-S-O y PAAC, siendo la primera la mejor de todas. Si se comparan estas familias de índices con los descriptores starPep (denotados como starPep_All), se puede observar que los 100 índices de QSO y 85 índices de PAAC presentan una distribución similar con respecto a los 8416 descriptores que fueron analizados de starPep, mientras que los 360 índices de autocorrelación de iFeature presentan una mejor distribución. Sin embargo, si se analizan los mejores 360 descriptores de starPep (denotados como starPep_360), entonces se puede observar que estos últimos presentan una mejor distribución que los 360 índices de autocorrelación de iFeature.

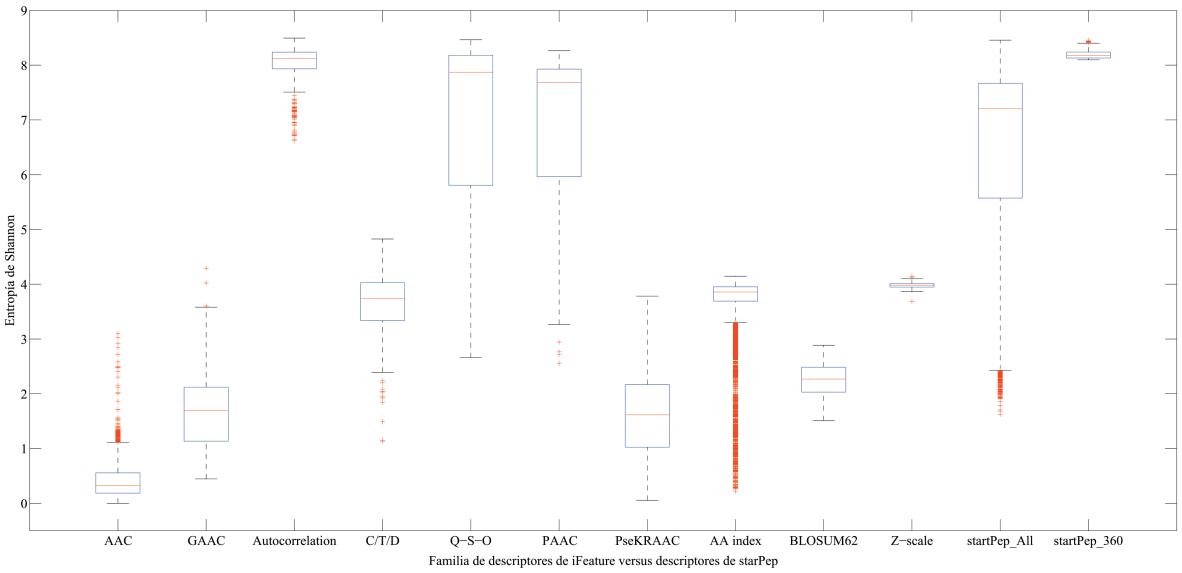


Figura 19. Diagramas de caja que muestran las distribuciones de la entropía de Shannon para las familias de descriptores iFeature, así como los descriptores starPep.

Como complemento a los resultados anteriores, se realizó un análisis de independencia lineal mediante el método de ACP (Jolliffe y Cadima, 2016). Para ello, se seleccionaron los índices de iFeature y starPep con valores de entropía superiores a 8 bits. De tal modo, que los datos de entrada para el ACP se componen de 700 descriptores starPep y 231 índices de iFeature. Al aplicar el método ACP en ese espacio de rasgos, los descriptores ortogonales están fuertemente cargados en diferentes componentes principales, mientras que los descriptores colineales se cargan en el mismo componente. Por lo tanto, a partir de los primeros 11 componentes principales obtenidos, que explican aproximadamente el 54,46 % de la varianza acumulada, observamos que algunos de los descriptores de iFeature y starPep analizados muestran colinealidad al estar cargados en los mismos componentes. Adicionalmente, los resultados obtenidos también indican que los descriptores de starPep tienen cargas exclusivas en componentes determinados, mientras que los descriptores de iFeature estudiados no presentaron cargas únicas en ninguno de los componentes obtenidos. En otras palabras, los descriptores de starPep resultaron ser útiles para la caracterización matemática de un EQB de péptidos bioactivos.

4.3. Evaluando y ajustando el filtrado de descriptores

En esta parte de nuestros experimentos, se utilizó el software starPep toolbox para calcular inicialmente 830 descriptores en diez conjuntos de datos con diferentes números de instancias (n). Para cada uno de estos conjuntos de datos, los 830 descriptores originales fueron filtrados de acuerdo a sus valores de entropía, utilizando el número de bins igual a n, y el valor de corte de entropía (θ_1) se estableció en el 10 % de la entropía máxima, es decir, $\theta_1 = 0.1 * \log n$, lo cual se hizo para eliminar características irrelevantes que estén por debajo de ese umbral θ_1 (ver Anexo A). En general, hubo como máximo 25 descriptores con un contenido de información menor que el valor de corte establecido.

Después de eliminar las características irrelevantes, se tuvieron aproximadamente 800 descriptores moleculares ordenados por sus valores de entropía, en cada uno de los conjuntos de datos de entrada. Siguiendo ese orden, se eliminan los rasgos redundantes como se describe en el algoritmo presentado en el Anexo A. Empleando ese procedimiento, para los dos métodos de correlación considerados (coeficiente de Pearson o de Spearman), se exploraron varios valores de umbral (θ_2) en el intervalo de 0 a 1, con el fin de evaluar su efecto en la eliminación de variables redundantes (Tabla 7).

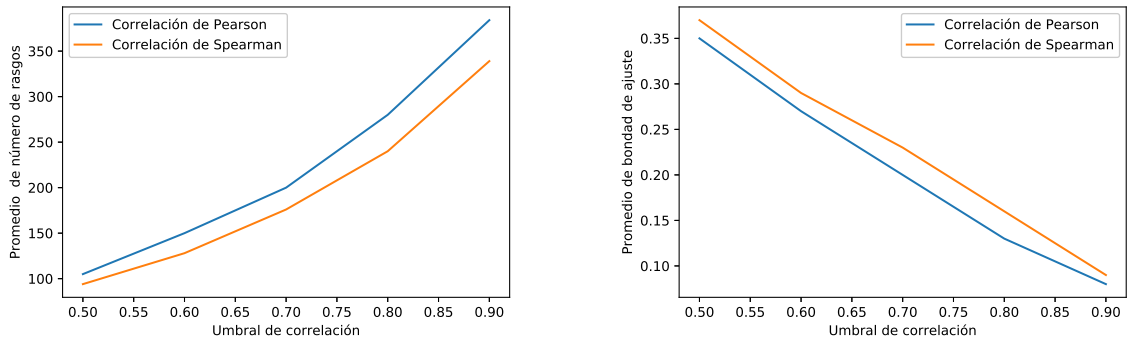
Para esta evaluación, se utilizó el análisis de Procrustes (Gower, 1975) que nos permitió comprobar cuánto se puede reducir el conjunto de descriptores mientras se conserva la estructura de los datos entre el espacio original de rasgos y el reducido. Al comparar estos dos espacios de descriptores, la bondad de ajuste de Procrustes se calcula entre los componentes principales (CPs) de los conjuntos de variables a comparar. Para ello, se utilizaron las primeros 50 CPs porque explicaban al menos el 80 % de la variación en los datos originales.

Como puede verse en la Tabla 7 y la Figura 20, la cantidad inicial de descriptores calculados se puede reducir drásticamente, eliminando cierto nivel de redundancia en el conjunto de variables resultante. Como era de esperar, una disminución en el umbral de correlación se traduce en reducir el número de rasgos filtrados, siendo el coeficiente de Spearman el que menos descriptores retuvo. Además, al observar el análisis de Procrustes, un umbral de correlación bajo afecta la bondad de ajuste entre

Tabla 7. Explorando el efecto de cambiar el método de correlación y el valor de corte para evaluar la similitud entre subconjuntos de características candidatas y las originales.

Datos[a,b]	Instancias	Umbral de correlación (θ_2)				
		Número de descriptores filtrados (bondad de ajuste)				
		0.5	0.6	0.7	0.8	0.9
				Pearson		
Overall_NR98	32300	108 (0.36)	156 (0.27)	199 (0.22)	275 (0.15)	374 (0.09)
Overall_NR90	22512	111 (0.35)	160 (0.27)	208 (0.21)	279 (0.15)	372 (0.1)
Overall_NR70	14559	108 (0.37)	150 (0.29)	207 (0.21)	276 (0.14)	382 (0.08)
Overall_NR50	9428	108 (0.37)	155 (0.28)	198 (0.21)	279 (0.14)	378 (0.09)
Overall_NR30	4735	112 (0.37)	155 (0.28)	205 (0.19)	281 (0.14)	385 (0.08)
Antibacterial_NR98	10303	94 (0.35)	147 (0.25)	191 (0.19)	282 (0.13)	379 (0.09)
Antifungal_NR98	4546	101 (0.36)	150 (0.26)	202 (0.2)	272 (0.14)	380 (0.08)
Antiviral_NR98	3849	100 (0.37)	143 (0.28)	186 (0.21)	280 (0.13)	393 (0.08)
Anticancer_NR98	1557	98 (0.35)	132 (0.28)	188 (0.19)	274 (0.12)	395 (0.07)
Antiparasitic_NR98	501	111 (0.28)	153 (0.21)	211 (0.15)	298 (0.1)	402 (0.06)
Promedio		105 (0.35)	150 (0.27)	200 (0.2)	280 (0.13)	384 (0.08)
				Spearman		
Overall_NR98	32300	94 (0.38)	126 (0.3)	177 (0.23)	235 (0.18)	338(0.09)
Overall_NR90	22512	95 (0.37)	127 (0.31)	174 (0.24)	241 (0.18)	327 (0.1)
Overall_NR70	14559	94 (0.38)	127 (0.32)	175 (0.25)	235 (0.18)	324 (0.1)
Overall_NR50	9428	95 (0.39)	132 (0.31)	174 (0.26)	241 (0.18)	327 (0.1)
Overall_NR30	4735	99 (0.37)	134 (0.29)	188 (0.22)	241 (0.18)	342 (0.1)
Antibacterial_NR98	10303	90 (0.35)	121 (0.29)	170 (0.22)	232 (0.16)	334 (0.1)
Antifungal_NR98	4546	87 (0.39)	122 (0.29)	177 (0.22)	236 (0.16)	338 (0.09)
Antiviral_NR98	3849	99 (0.35)	133 (0.3)	172 (0.22)	241 (0.16)	346 (0.09)
Anticancer_NR98	1557	92 (0.37)	128 (0.27)	172 (0.22)	245 (0.14)	352 (0.08)
Antiparasitic_NR98	501	94 (0.32)	129 (0.24)	179 (0.17)	251 (0.12)	361 (0.07)
Promedio		94 (0.37)	128 (0.29)	176 (0.23)	240 (0.16)	339 (0.09)

[a]Estos conjuntos de datos se recuperaron de la base de datos integrada. [b]NR significa “No redundante” en un porcentaje (número) de identidad de secuencia. Para realizar las comparaciones de identidad de secuencia, utilizamos un algoritmo de alineamiento local (Smith-Waterman (Smith *et al.*, 1981)) implementado en BioJava (Lafita *et al.*, 2019), con la matriz de sustitución BLOSUM62.



(a) (b)

Figura 20. Explorando el efecto de cambiar el umbral de similitud: (a) el número promedio de características retenidas se muestra como una función creciente en el umbral de similitud; y, (b) la bondad de ajuste promedio entre los 50 CPs de las variables originales y reducidas se muestra como una función decreciente en el umbral de similitud.

los datos originales y reducidos. Mientras que establecer un umbral de correlación alto da como resultado una mejor coincidencia entre la estructura de datos reducida y la original, que es lo que se desea en esta etapa temprana.

Al analizar la Tabla 7, se puede observar que el uso de Spearman con el umbral de correlación igual a 0.9 todavía produce un número alto (> 300) de descriptores. Un número menor de descriptores (< 300) se alcanza cuando el umbral de correlación es igual a 0.80, mientras que la bondad de ajuste promedio es inferior a 0.2. Este valor pequeño de bondad de ajuste expresa que el espacio reducido de rasgos representa de forma aceptable la estructura de datos del espacio original de descriptores. Por lo tanto, de ahora en adelante, el filtro basado en correlación de Spearman con umbral igual a 0.8 será el que se utilice durante la primera etapa del proceso de selección de rasgos, para calcular un conjunto candidato.

4.4. Evaluando el subconjunto optimizado de descriptores

Hasta ahora, el conjunto candidato de rasgos resultante de la fase de filtrado aún no tiene una baja dimensionalidad para la construcción de las redes de similitud. Por lo tanto, el propósito de esta segunda etapa no es la preservación de la estructura de datos del espacio original, sino la optimización de una función objetivo (Ecuación 12) para encontrar un subconjunto de rasgos adecuado para la tarea que se precisa (ver Anexo B).

Tabla 8. Optimización del subconjunto de rasgos

Datos	Instancias (n)	Número de bins[a] ($\lfloor \sqrt{n} \rfloor$)	Subconjunto ($F*$)		CPU[d] (hh:mm:ss)
			$\lvert F* \rvert$[b]	Mérito[c]	
Overall_NR98	32300	179	39	3.985	12:07:03
Overall_NR90	22512	150	40	3.87	05:17:40
Overall_NR70	14559	120	42	3.667	01:27:46
Overall_NR50	9428	97	39	3.474	00:34:04
Overall_NR30	4735	68	33	3.13	00:08:12
Antibacterial_NR98	10303	101	45	3.41	00:38:57
Antifungal_NR98	4546	67	39	3.059	00:07:17
Antiviral_NR98	3849	62	45	3.031	00:05:38
Anticancer_NR98	1557	39	36	2.588	00:00:50
Antiparasitic_NR98	501	22	40	2.11	00:00:05

[a]Número de intervalos para discretizar los descriptores, [b] cardinalidad del subconjunto optimizado, [c] valor de la función objetivo $\Phi(F)$ (Ecuación 12), [d]tiempo de uso del CPU en una *Mac Pro Server with 2 x Intel Xeon Processor 2.66 GHz 6-cores, and memory 64 GB*.

En el cómputo de dicha función objetivo (mérito), los cálculos de la entropía e información mutua se realizaron tomando la raíz cuadrada del número de instancias ($\sqrt{n}$) como regla para estimar el número de bins. Aunque no existe una respuesta "correcta" para determinar el número de bins (Maciejewski, 2011), la elección de la raíz cuadrada $\sqrt{n}$ produjo resultados aceptables en términos de la cardinalidad de los subconjuntos optimizados. La Tabla 8 muestra los resultados obtenidos, donde el número de los rasgos optimizados se encuentra entre 30 y 50, independientemente de la variación en el número de instancias de los datos de entrada.

Tabla 9. Comparación entre los valores de mérito de los subconjuntos optimizados y los descriptores más relevantes que conforman el top-*k* del conjunto candidato. La selección del top-*k* se realiza al ordenarlos en forma descendente según el criterio de entropía de Shannon.

Datos	Instancias	Mejor valor de mérito[a]	Valores de mérito de los descriptores del top-k^b				
			k				
			20	30	40	50	60
Overall_NR98	32300	3.985	3.92	3.96	3.962	3.958	3.942
Overall_NR90	22512	3.87	3.79	3.827	3.829	3.822	3.807
Overall_NR70	14559	3.667	3.577	3.608	3.616	3.607	3.589
Overall_NR50	9428	3.474	3.368	3.396	3.406	3.403	3.388
Overall_NR30	4735	3.13	3.051	3.057	3.044	3.038	3.035
Antibacterial_NR98	10303	3.41	3.336	3.361	3.364	3.356	3.334
Antifungal_NR98	4546	3.059	2.98	3.018	3.01	3.011	2.993
Antiviral_NR98	3849	3.031	2.956	2.992	3.001	3.001	2.995
Anticancer_NR98	1557	2.588	2.517	2.552	2.55	2.54	2.529
Antiparasitic_NR98	501	2.11	2.042	2.063	2.06	2.065	2.061
Promedio		3.232	3.154	3.183	3.184	3.180	3.167

[a]Valores de mérito de los subconjuntos optimizados; y [b]valores de mérito de los descriptores más relevantes que conforman el top-*k* del conjunto candidato.

Por otra parte, la Tabla 9 muestra una comparación entre los valores de mérito de los subconjuntos optimizados y los mejores rasgos del conjunto candidato sin optimizar, que son aquellos incluidos en el top-*k* de acuerdo a la entropía de Shannon en orden descendente. Para esta comparación, se utilizó una prueba de suma de rangos de Wilcoxon por pares (Hollander *et al.*, 2013) con corrección de Hochberg (Hochberg, 1988). La prueba estadística reveló que el subconjunto optimizado es significativamente mejor ($p < 0.05$) que la selección de los mejores descriptores (top-*k*) sin optimizar. Además, no se detectaron diferencias significativas entre los puntajes de mérito del top-30, top-40 y top-50.

En la Figura 21 se resume la distribución de las puntuaciones de mérito para los grupos de descriptores analizados. Se puede observar que existen puntuaciones de mérito atípicas para la mayoría de los grupos que conforman el top-*k*. La presencia de

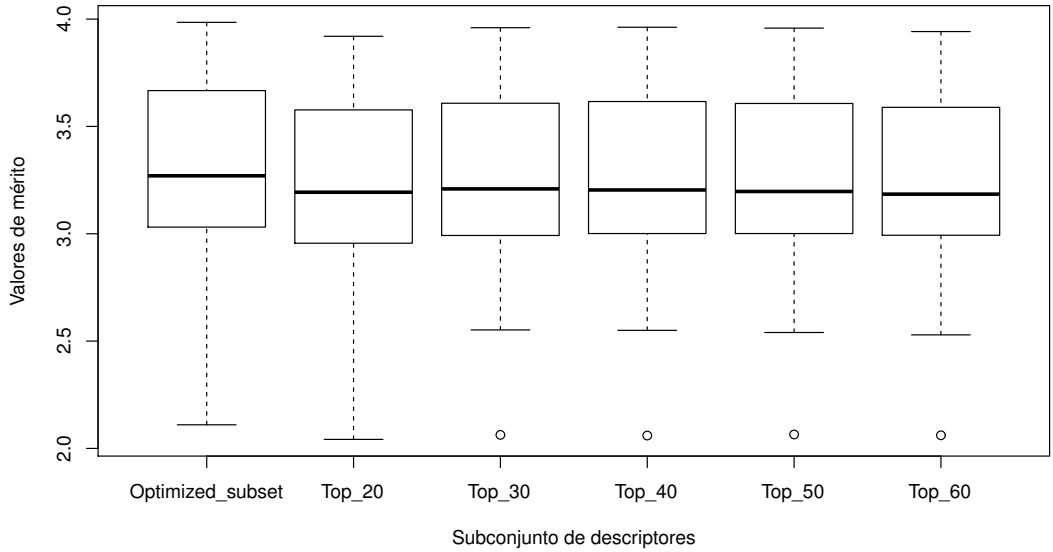


Figura 21. Diagrama de caja de los valores de mérito para diferentes subconjuntos de descriptores.

estos valores atípicos puede sugerir que la selección de los descriptores en el top-*k* puede verse afectada por una pequeña cantidad de instancias en los datos, como es el caso de Antiparasitic_NR98 (Table 9).

Para ilustrar que los descriptores optimizados caracterizan un EQB en particular, hemos contabilizado a los conjuntos de datos distintos que inducen un determinado descriptor. Los resultados se presentan en la Figura 22, donde el eje *x* contiene los descriptores comenzando por los más comunes, según la cantidad de veces que se repiten, y el eje *y* representa el número de conjuntos de datos contabilizado por descriptor seleccionado (frecuencia de aparición). Cabe señalar que solo hay tres descriptores repetidos en nueve de los diez conjuntos de datos. Además, la tendencia en el gráfico indica que la mayoría de los descriptores seleccionados reaparecen pocas veces. De hecho, muchos de ellos aparecen únicamente para caracterizar un único conjunto de datos.

4.5. Generando redes de similitud molecular

En correspondencia con los resultados anteriores, se usaron los descriptores optimizados para construir las redes de similitud molecular. Esto se realiza teniendo en

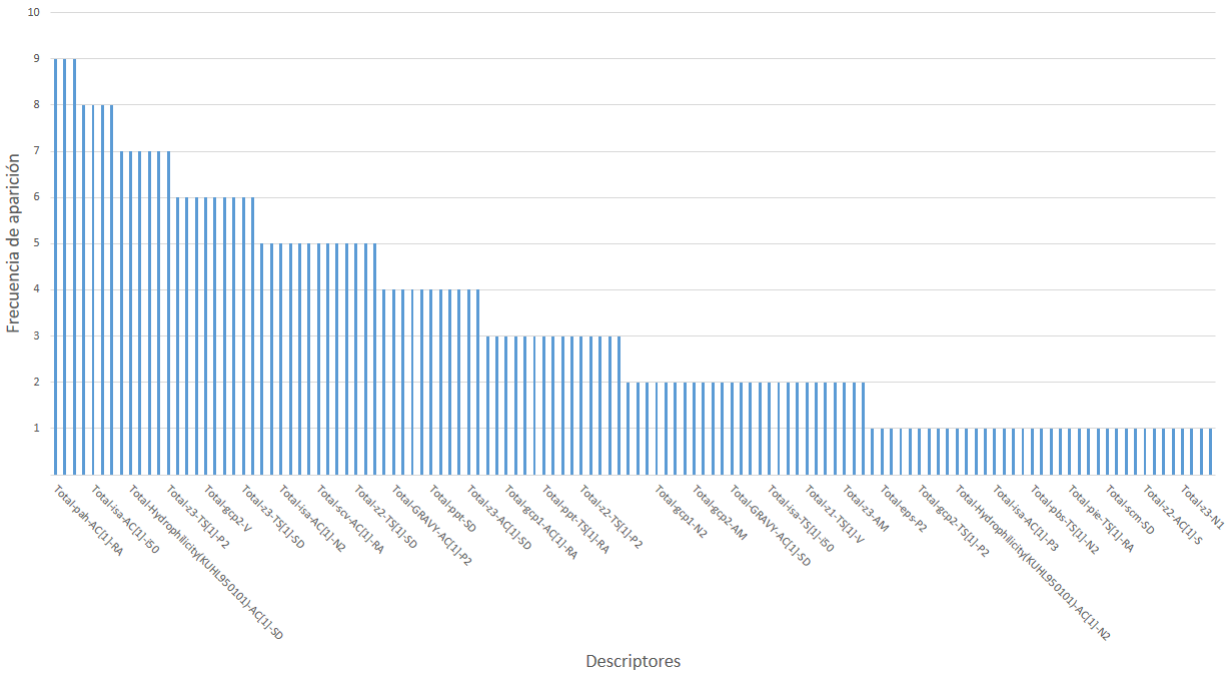


Figura 22. Número de conjuntos de datos distintos que incluyen un descriptor seleccionado.

cuenta que el cálculo de la matriz de similitud es prohibitivo para grandes conjuntos de datos. En cambio, una red alternativa basada en el algoritmo HSP (ver Anexo C) puede generarse con éxito. La viabilidad de construir las redes HSP a partir de los conjuntos de datos experimentales se ilustra en la Tabla 10. En dicha tabla se puede observar que este tipo de redes alcanza niveles de baja densidad en todos los casos, incluidos los grandes conjuntos de datos.

Tabla 10. Redes HSP conexas generadas con el umbral de similitud $t = 0$.

Datos	Nodos	Máxima distancia[a]	Relaciones de similitud	
			Aristas	Densidad[b]
Overall_NR98	32300	29.80	353 213	6.8E-4
Overall_NR90	22512	27.22	242 227	9.6E-4
Overall_NR70	14559	27.03	157 590	0.001
Overall_NR50	9428	27.01	102 226	0.002
Overall_NR30	4735	23.65	46338	0.004
Antibacterial_NR98	10303	27.95	95605	0.002
Antifungal_NR98	4546	24.00	36765	0.004
Antiviral_NR98	3849	24.12	31709	0.004
Anticancer_NR98	1557	21.56	9552	0.008
Antiparasitic_NR98	501	19.13	2560	0.02

[a]La distancia euclidiana máxima entre dos puntos en el espacio de rasgos; y [b] densidad de la red HSP conexa cuando el umbral $t = 0$.

El impacto de los descriptores optimizados también puede verse en la construcción de las redes de similitud molecular. Como se muestra en la Figura 23, al interpretar

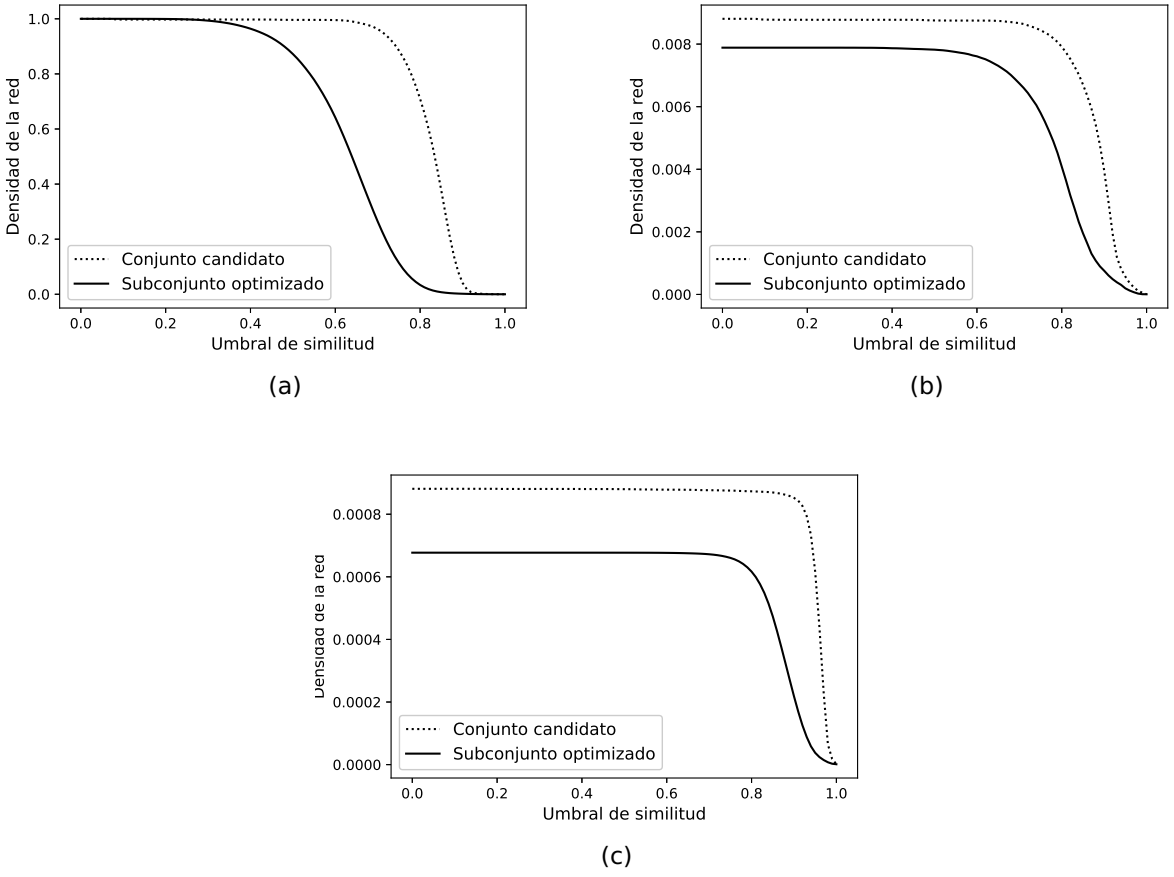


Figura 23. Ilustración de la importancia de utilizar el subconjunto de rasgos optimizados frente al conjunto de los descriptores candidatos para la construcción de redes de similitud. Las curvas de densidad se generaron variando el umbral de similitud en los siguientes casos: (a) red de similitud para Anticancer_NR98, variante de red HSP para (b) Anticancer_NR98, y (c) Overall_NR98.

el rendimiento de los descriptores en función de la densidad de la red, el subconjunto optimizado produce niveles de densidad más bajos con una apariencia más suave que las curvas de densidad del conjunto candidato. Esto muestra que los descriptores optimizados son adecuados para generar redes de similitud que puedan ser analizadas mediante inspección visual. A modo de ejemplo, en la Figura 24 se presenta la visualización de la red obtenida para el conjunto de datos con el mayor número de instancias (Overall_NR98).

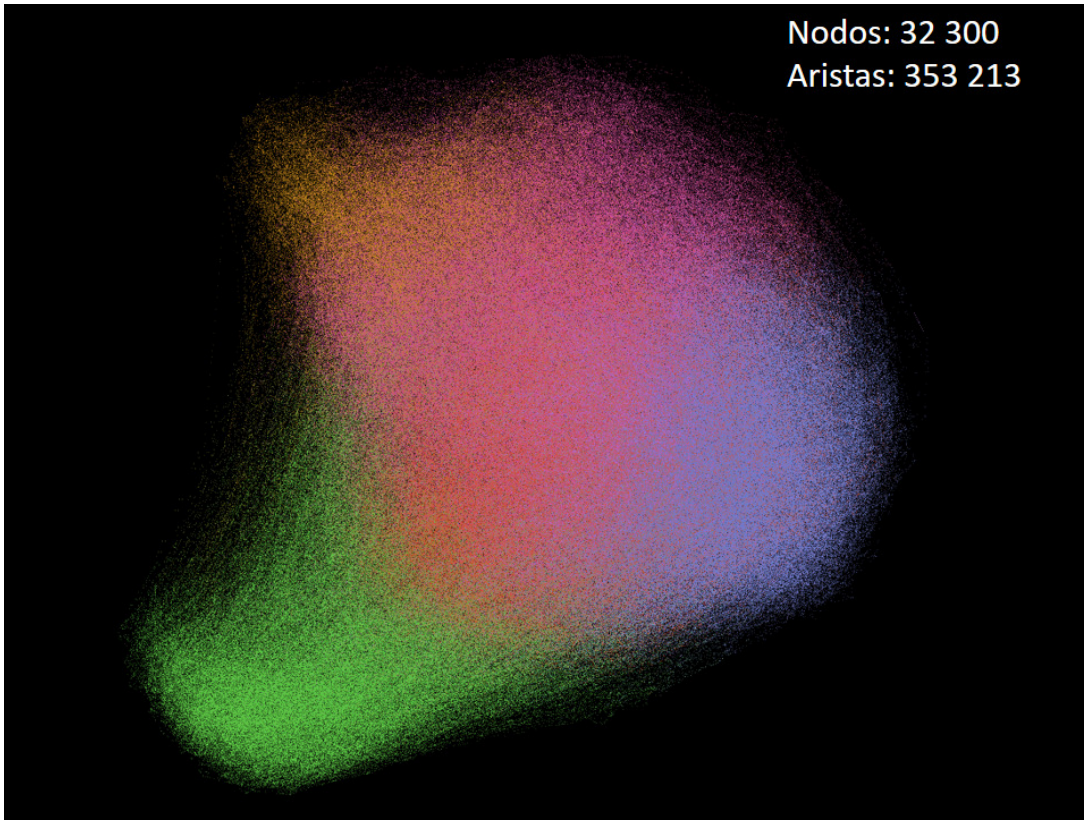


Figura 24. Visualización de la red HSP para un conjunto amplio de péptidos bioactivos (Overall_NR98) en el umbral $t = 0$, usando el algoritmo de diseño ForceAtlas (Jacomy *et al.*, 2014). Los colores representan las comunidades detectadas.

4.6. Caso de estudio: navegando un espacio químico-biológico de péptidos anticancerígenos

La visualización y la exploración interactiva de las redes de similitud generadas pueden ser útiles para facilitar un razonamiento analítico que antes no era posible. Sin embargo, comprender las redes a través de la inspección visual no es un proceso sencillo (Ware, 2019), por lo que proponemos una exploración sistemática utilizando

una combinación de técnicas de clustering y ciencia de redes (Figura 15). Durante este proceso de exploración, tres aspectos de la inspección visual son útiles para ayudar al pensamiento humano: posicionamiento, filtrado y personalización de la apariencia de los nodos (Cherven, 2013). Sobre esta base, a continuación se ilustra el uso del flujo de trabajo propuesto para el análisis visual de un EQB como caso de estudio específico.

Los 1557 péptidos anticancerígenos (Anticancer_NR98) de los datos experimentales fueron utilizados como material de estudio en esta etapa de la investigación. Dado que este conjunto de datos no es lo suficientemente grande, generamos redes de similitud tradicionales mediante el cálculo de la matriz de similitud/distancia. Para visualizar estas redes de forma útil, examinamos una familia de algoritmos de diseño que permiten espacializar la red y reorganizar los nodos (Cherven, 2013). Estos algoritmos cambian la posición de los nodos al considerar que estos se repelen entre sí, mientras que las aristas entre ellos atraen a los nodos adyacentes como resortes. En particular, el algoritmo de Fruchterman-Reingold (Fruchterman y Reingold, 1991) resultó ser el más adecuado para dibujar la red a partir del conjunto de datos en estudio (Figura 25). Con ese esquema de representación gráfica, se realiza el siguiente análisis.

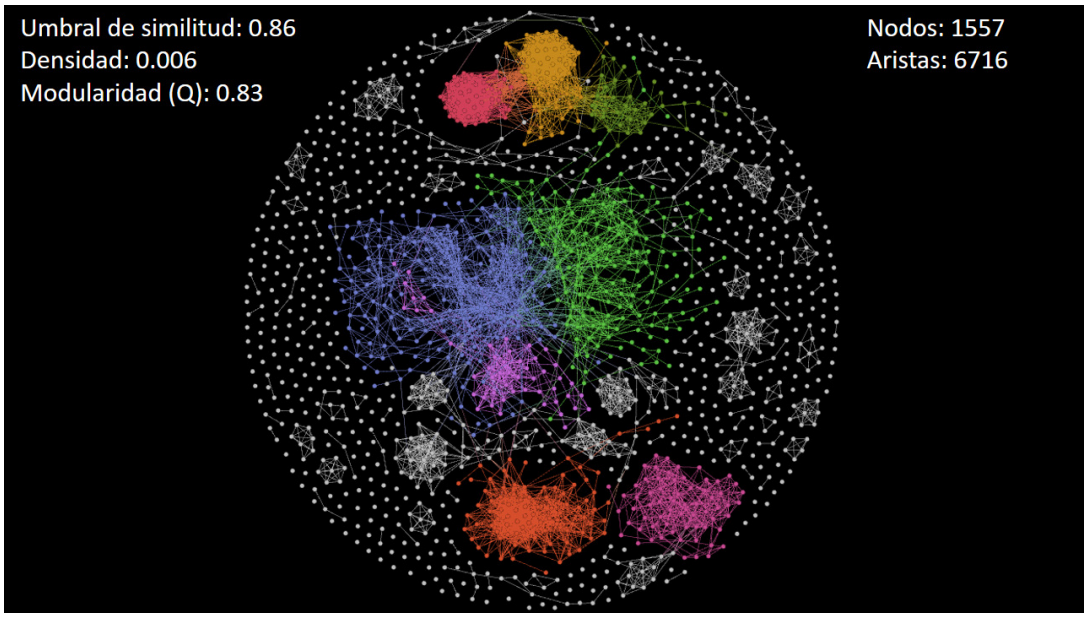


Figura 25. Visualización de la red de similitud de péptidos anticancerígenos (Anticancer_NR98) utilizando un umbral de similitud de 0,86, y el algoritmo de diseño de Fruchterman-Reingold. Los nodos están coloreados de acuerdo a las comunidades a las que pertenecen.

4.6.1. Exploración y detección de comunidades

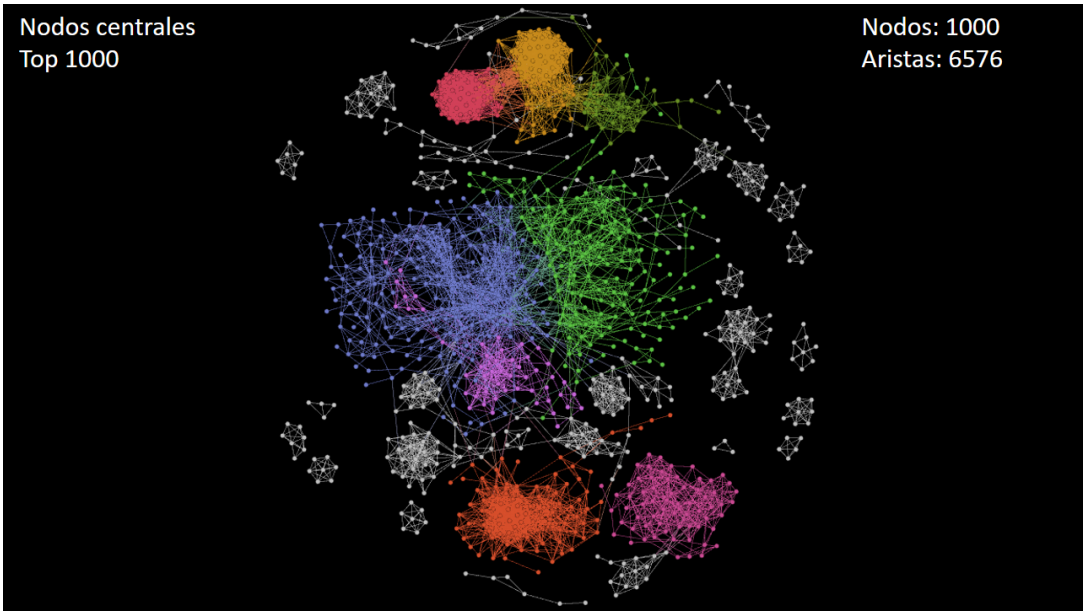
Una red con bajo nivel de densidad (< 1 %) se puede visualizar a primera vista estableciendo un valor alto de umbral (Figura 23(a)). Además, las redes generadas se tornan aún más interpretables si se detecta en ellas una estructura de comunidades (Newman, 2018). Tenga en cuenta que las comunidades en un EQB pueden representar regiones biológicamente relevantes donde residen compuestos bioactivos (Lipinski y Hopkins, 2004). Por lo tanto, se analizaron visualmente las redes variando el umbral de similitud hasta encontrar una estructura de comunidades bien definida. La red resultante fue diseñada ajustando el umbral de similitud a 0.86, lo cual produce una baja densidad de de 0.006, y una modularidad de red (Q) de 0.83 (Figura 25).

Con base en nuestra inspección visual, corroboramos lo que otros autores han hallado con estudios previos (de la Vega de León y Bajorath, 2016): se alcanza una mayor modularidad Q cuando se utiliza un alto valor de umbral de similitud. No obstante, el intervalo dentro del cual se debe explorar el valor de umbral no es fijo y depende del conjunto de datos de entrada. También, recalcamos en la importancia de que la decisión final para elegir un alto valor de umbral esté basada en la inspección visual. Esto debido a que, mediante la exploración interactiva, es posible disminuir la densidad de la red sin llegar a ocultar relaciones de similitud que visualmente ofrezcan información acerca de la topología de la red.

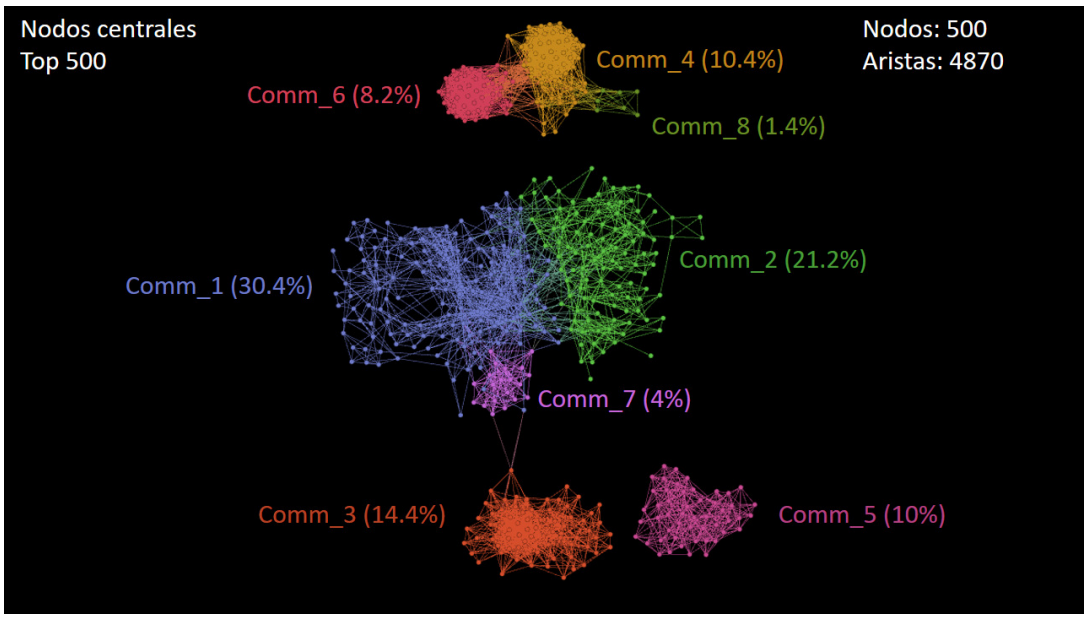
4.6.2. Identificación de compuestos biológicamente relevantes

Una vez que se ha encontrado una estructura de comunidad, se le calcula a todos los nodos la medida de centralidad de comunidad Hub-Bridge (Ecuación 17), con el propósito de retener los k mejores nodos de la red. En particular, el Top 1000 expone grupos de nodos formando cliques, que se definen como subgrafos completos (Figura 26(a)). Estas secuencias relacionadas que están densamente conectadas pueden estar formando familias en el EQB de los péptidos anticancerígenos, y su detección puede ser útil en trabajos futuros (Jachiet *et al.*, 2013; Pathmanathan *et al.*, 2017).

Por supuesto, los nodos dentro de las comunidades más grandes tienen un valor de centralidad más alto en relación a los que están dentro de grupos más pequeños.



(a)



(b)

Figura 26. Visualización de la subred que contiene el (a) Top 1000 y (b) Top 500 de los nodos más relevantes según los valores de centralidad Community Hub-Bridge.

Por ejemplo, si seleccionamos los nodos que forman parte del Top 500, estos pueden fungir como representantes de ocho comunidades líderes en el EQB de los péptidos anticancerígenos (Figura 26(b)). De todos ellos, el péptido starpep_00185, conocido como *ascaphin*-8 (Conlon *et al.*, 2004), es el nodo más central dentro de la Comunidad 1 (Figura 27), y algunos de sus péptidos vecinos son análogos con sustituciones de aminoácidos en determinadas posiciones (Tabla 11). Al buscar información adicional sobre este péptido de mayor centralidad local, notamos que su secuencia es la que ocurre naturalmente, mientras que los análogos son péptidos modificados químicamente para lograr un mayor potencial terapéutico (Michael Conlon *et al.*, 2008; Eley *et al.*, 2008).

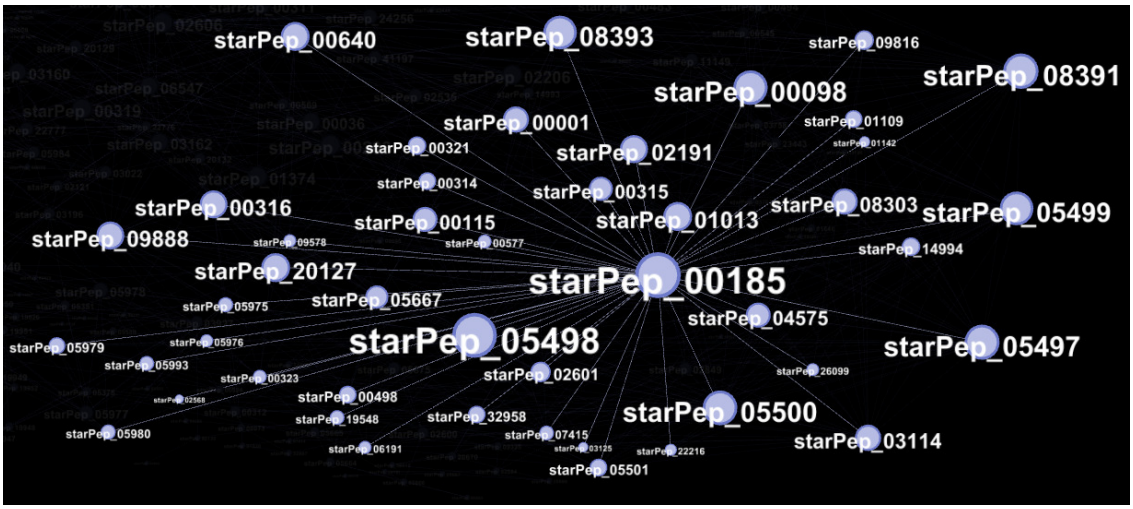


Figura 27. Un acercamiento a la Comunidad 1 para resaltar el nodo más central (starPep_00185) basado en la medida de centralidad Hub-Bridge. El tamaño de los nodos se encuentra redimensionado de acuerdo a sus valores de centralidad local.

El péptido *ascaphin*-8 tiene una longitud de 19 AAs y contiene el extremo C-terminal amidado (GFKDLLKGAAKALVKTVLF.NH2). Es uno de los ocho péptidos antimicrobianos de la familia de las *ascaphins* 1-8 (Conlon *et al.*, 2004), que se aislaron

Tabla 11. Una familia de péptidos centrales dentro de la Comunidad 1

ID[a]	Secuencia[b]	Longitud
starPep_00185	GFKDLLKGAAKALVKTVLF	19
starPep_05498	GFKDLLKGAAKALVKAVLF	19
starPep_05497	GFKDLLKGAAKALKKTVLF	19
starPep_05500	GFKDLLKGAKKALVKTVLF	19
starPep_03114	GFKKLLKGAAKALVKTVLF	19
starPep_05499	GFKDLLKGAAKALVKTVKF	19

[a]ID de los péptidos en la base de datos integrada. [b]Las sustituciones de aminoácidos en las secuencias análogas han sido señaladas con el color rojo.

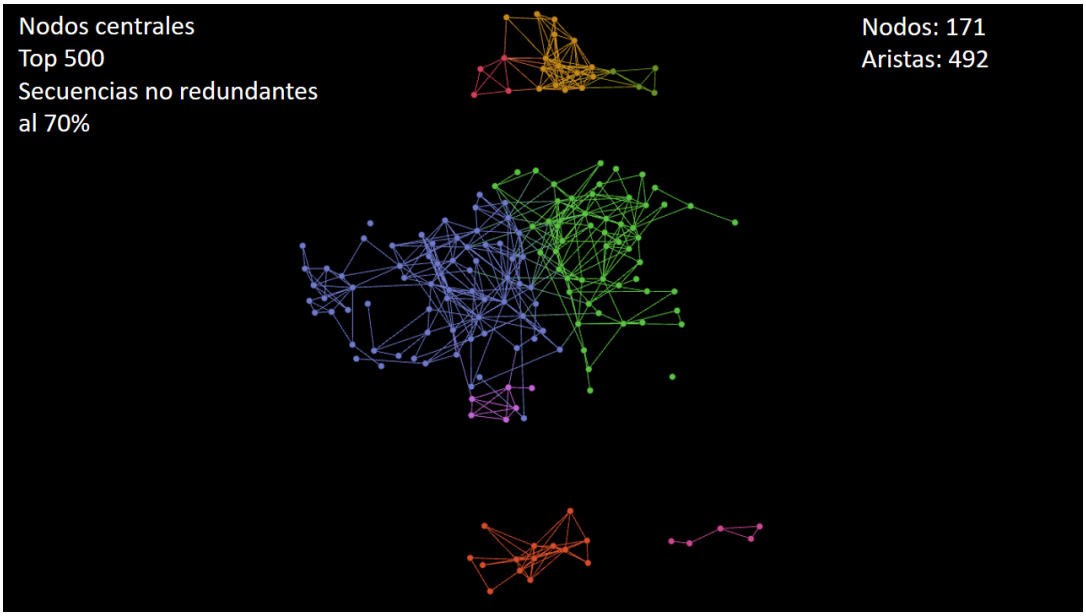
originalmente de las secreciones cutáneas de la rana con cola (Ascaphus truei). Entre los ocho péptidos purificados *ascaphins* 1-8, el *ascaphin*-8 resultó ser el compuesto más activo contra una variedad de microorganismos patógenos. A pesar de ello, presentó un potencial terapéutico limitado debido a su mayor nivel de toxicidad (actividad hemolítica) (Conlon *et al.*, 2004). Sin embargo, algunos de sus análogos han mostrado tener una actividad hemolítica baja, manteniendo una potente actividad antibacteriana contra una variedad de bacterias de amplio espectro, productoras de Betalactamasa (Eley *et al.*, 2008).

Además de la actividad antimicrobiana, la propiedad anticancerígena del *ascaphin*-8 y sus análogos ha sido evaluada en células derivadas de hepatoma humano (HepG2), donde los péptidos análogos mostraron una mayor citotoxicidad para las células HepG2 y una reducción de la toxicidad contra las células de mamíferos (Michael Conlon *et al.*, 2008). Por si fuera poco, otros estudios (Laughlin y Ahmad, 2010; Popovic *et al.*, 2012; Xu y Lai, 2015) también han revelado que el *ascaphin*-8 puede ser considerado como un punto de partida para el diseño de compuestos líderes en el desarrollo de nuevos fármacos.

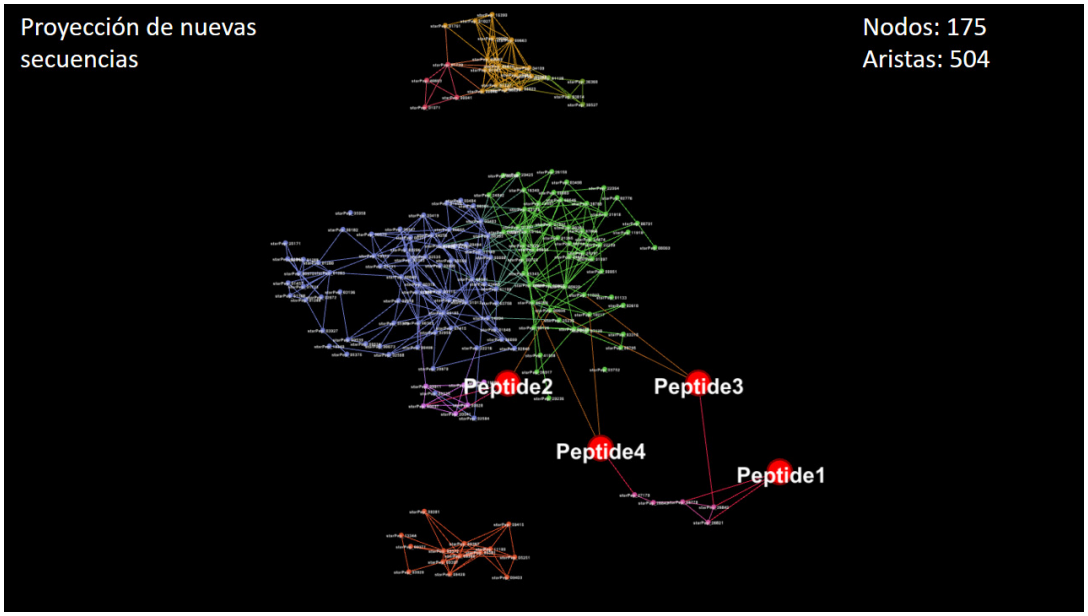
4.6.3. Identificación de péptidos relevantes pero no redundantes

Como se puede observar en la Tabla 11, algunos nodos conectados en el interior de las comunidades pueden estar representando una familia de péptidos parecidos en secuencia. Otro ejemplo de secuencias de péptidos estrechamente relacionadas se encontró en los 50 miembros de la Comunidad 5, incluidos en el Top 500 (Figura 26(b)). Todos estos 50 péptidos tienen la misma longitud de 10 AAs. También, están compuestos por solo 4 residuos diferentes (K,F,L,Q), con dos aminoácidos F y Q fijos en las posiciones 6 y 10, respectivamente. Por lo tanto, en general, es de esperar que secuencias redundantes produzcan nodos con valores similares de centralidad.

Una forma de eliminar la redundancia en las secuencias, es hacerlo a través de los algoritmos basados en alineamiento que han sido implementados en el flujo de trabajo propuesto (Figura 15). No obstante, en esa parte inicial de preparación y limpieza de datos, aún se desconoce cuál o cuáles son las secuencias más relevantes en el EQB. Por lo que se sugiere eliminar inicialmente sólo aquellas secuencias altamente



(a)



(b)

Figura 28. Visualización de la subred (a) que contiene a los péptidos anticancerígenos que son nodos centrales pero no redundantes al 70 % de identidad de secuencia, y (b) la proyección en esta red de cuatro péptidos anticancerígenos (Péptido 1-4) diseñados por métodos in silico.

redundantes, es decir, secuencias que estén por encima de un 98 % de identidad. Para luego, reducir la redundancia con la información obtenida de representar y analizar el EQB, tratando de conservar los péptidos más centrales localmente.

Por consiguiente, para extraer péptidos centrales pero no redundantes, las secuencias se eliminan siguiendo el orden decreciente de sus valores de centralidad local si cumplen un porcentaje dado de identidad, con relación a algún péptido ya seleccionado. La subred resultante se presenta en la Figura 28(a), donde hemos utilizado el 70 % de identidad de secuencia para establecer el nivel de redundancia. Esto mediante un alineamiento local (Smith *et al.*, 1981) con la matriz de sustitución BLOSUM62. Finalmente, se puede obtener la lista ordenada de los péptidos de acuerdo a sus valores decrecientes de la medida de centralidad global (Ecuación 16). En esta lista, los péptidos mejor posicionados son los que tienen rutas de similitud relativamente pequeñas hacia todos los demás nodos de la red.

4.6.4. Proyección de nuevos compuestos en el modelo descriptivo

Para ilustrar la proyección de nuevos compuestos en el modelo descriptivo obtenido (Figura 28(b)), hemos embebido a cuatro péptidos anticancerígenos que fueron diseñados por métodos *in silico* y evaluados experimentalmente en otro estudio reciente (Singh *et al.*, 2020). Estos péptidos son: Peptide1 (WLFKFLAWKKK), Peptide2 (FPK LLLKFLRLG), Peptide3 (KKFALKLFWWK) y Peptide4 (RLLRRLRIRG).

Como se puede observar en la Figura 28(b), Peptide1 y Peptide2 ocupan diferentes regiones en el EQB de referencia, pero existe una ruta de similitud que los conecta (Figura 29). Según sus creadores (Singh *et al.*, 2020), estos dos péptidos fueron los más activos y las simulaciones de dinámica molecular sugieren que tienen diferentes mecanismos de interacción con membranas de bicapa lipídica. Mientras que el Peptide1 permanece adsorbido en la superficie de la membrana, el Peptide2 tiene la capacidad de penetrar la membrana celular.

En relación al Peptide3, que también resultó ser un compuesto activo, este se encuentra en una ruta de similitud entre Peptide1 y Peptide2 (Figura 29). De acuerdo a la longitud de la ruta de similitud, Peptide3 es más similar a Peptide1 y menos similar a Peptide2, lo que es lógico ya que Peptide3 y Peptide1 tienen en común que son análogos del mismo péptido (Singh *et al.*, 2020). Por último, a pesar de que Peptide2 y Peptide4 tienen un origen común (Singh *et al.*, 2020), Peptide4 es inactivo y no está conectado a ninguno de los tres compuestos anteriores en el grafo de los k vecinos más cercanos, con $k = 3$ (Figura 29). Además, Peptide4 es el de menor

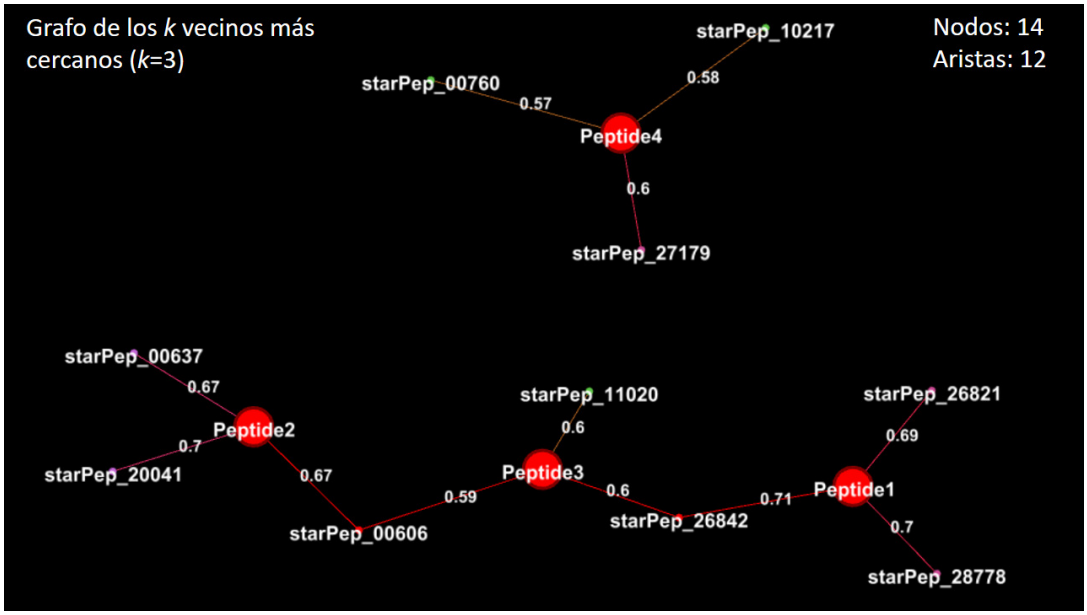


Figura 29. Visualización del grafo de los k vecinos más cercanos, con $k = 3$, para inspeccionar el EQB ocupado por los nuevos compuestos de interés (Peptide 1-4).

fuerza total (ver Sección 3.3.2.2), si se suman los pesos de las aristas incidentes a los péptidos analizados. Por lo tanto, este análisis retrospectivo señala que el enfoque presentado aquí puede ser útil para obtener información adicional acerca de nuevos compuestos peptídicos embebidos en un EQB de referencia.

4.7. Conclusiones parciales

En este capítulo, hemos diseñado e implementado un flujo de trabajo para obtener un modelo basado en redes de similitud molecular, que permite captar información visual e interpretarla a partir de datos integrados de péptidos bioactivos. También, hemos experimentado con el flujo propuesto para evaluar la construcción automática e inspección visual del modelo descriptivo, visto como una representación significativa del espacio químico-biológico de péptidos bioactivos. Los resultados experimentales mostraron la eficacia de nuestro enfoque en el uso del aprendizaje no supervisado. Esto para convertir datos existentes de péptidos bioactivos en información que no se encuentra disponible en ninguna otra fuente de datos, y que puede ser útil en la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

Como aspecto a destacar en el flujo propuesto, para cada secuencia de péptidos se realiza la codificación en descriptores aplicando distintos operadores de agregación a vectores de propiedades de aminoácidos. Luego, se selecciona un subconjunto optimizado de forma automática para conservar aquellos rasgos que son relevantes y poco redundantes entre sí, al considerar los conceptos de entropía de Shannon e información mutua, respectivamente.

En ese sentido, el método de optimización implementado permite reducir la dimensionalidad del conjunto inicial de descriptores, para aumentar la eficiencia en la definición del espacio de rasgos, y evitar así lidiar con redes de similitud molecular de alta densidad.

Por último, para ilustrar la aplicabilidad del flujo de trabajo, se generó un modelo descriptivo de forma gráfica y analítica para llevar a cabo un procedimiento sistemático basado en la detección de comunidades y el análisis de centralidad de los nodos en la red. Esta tarea de aprendizaje no supervisado, junto con el análisis visual, nos permitió navegar y minar un espacio químico biológicamente relevante de péptidos anticancerígenos, conocidos hasta la fecha, para detectar patrones que expliquen o resuman los datos. Por lo tanto, esperamos alentar a los investigadores a utilizar nuestro enfoque para convertir los datos sin procesar de las secuencias peptídicas en redes de similitud molecular, y luego en información que podría usarse en estudios futuros o llegar a convertirse en conocimiento.

Capítulo 5. Conclusiones y trabajo futuro

5.1. Sumario

A modo de sumario, se obtuvo los siguientes resultados:

1. Se desarrolló una nueva base de datos que contiene 45120 péptidos bioactivos y sus metadatos, obtenidos a partir de la recopilación e integración de datos existentes en 40 repositorios biológicos de interés farmacológico.

2. Para la caracterización numérica de los péptidos bioactivos recopilados e integrados, se propone el cálculo de descriptores moleculares mediante el uso de distintos operadores de agregación a vectores de propiedades de aminoácidos.

3. Además, un aspecto esencial es la selección automática de un subconjunto optimizado de los descriptores calculados, para lo cual se plantea una nueva función objetivo, modificada bajo el enfoque no supervisado, combinando los conceptos de entropía de Shannon e información mutua para obtener rasgos de alta relevancia y baja redundancia.

4. Para la construcción del modelo descriptivo que representa un espacio químico-biológico de péptidos bioactivos, se utilizó el algoritmo de Louvain para la detección de comunidades, y una métrica de centralidad para identificar los nodos relevantes en la red, considerando las comunidades detectadas.

5. Se desarrolló la herramienta informática starPep toolbox (`http://mobiosd-hub.com/starpep`), que permite generar el modelo basado en redes de similitud molecular de forma gráfica y analítica, facilitando al investigador la utilización de sus capacidades cognitivas para la exploración interactiva y detección de patrones, útiles en la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

5.2. Conclusiones generales

Los resultados obtenidos en la presente investigación permiten concluir lo siguiente:

1. La base de datos desarrollada es la única que ofrece una vista unificada de los datos recopilados e integrados a partir de los repositorios biológicos de interés en el estudio; y, por lo tanto, permite representar de manera más completa el espacio químico-biológico de péptidos bioactivos conocidos hasta la fecha.
2. El estudio realizado de análisis de variabilidad y componentes principales, arrojó que los descriptores calculados con base en los distintos operadores de agregación poseen mayor variabilidad y capacidad de codificar información diferente, respecto a familias de descriptores definidas en la literatura, lo cual permite una mejor caracterización del espacio químico-biológico de péptidos bioactivos.
3. Los descriptores optimizados mostraron su valía en la construcción de un modelo basado en redes de similitud molecular, logrando menor densidad en las redes si se compara con el conjunto candidato de rasgos sin optimizar, lo cual posibilita que las redes obtenidas sean analizadas e interpretadas de manera visual por medio de la interacción hombre-computadora.
4. El modelo basado en redes de similitud molecular, que se obtiene mediante una técnica de minería de datos no supervisada, constituye una alternativa que complementa a los modelos predictivos propuestos en la literatura para extraer información útil previamente desconocida, esto con el fin de apoyar la toma de decisiones durante la identificación *in silico* de compuestos candidatos a fármacos peptídicos.

5.3. Propuesta de trabajo futuro

Las técnicas computacionales empleadas en el diseño e identificación *in silico* de nuevos compuestos peptídicos, manejan estrategias que exploran el EQB de las secuencias de aminoácidos (Fjell *et al.*, 2012a; Torres *et al.*, 2019). Debido a que es prácticamente imposible explorar exhaustivamente todo ese espacio, se utilizan funciones

de evaluación para guiar la búsqueda hacia regiones factibles. Con frecuencia, dichas funciones han sido modelos predictivos donde se tiene en cuenta solo la precisión como medida de calidad, sin contemplar si los modelos obtenidos son o no de difícil comprensión por los humanos (Freitas, 2014).

Una muestra de la aplicación de modelos predictivos de tipo “caja negra” son las Redes Neuronales Artificiales (Cherkasov *et al.*, 2008; Fjell *et al.*, 2009; Torrent *et al.*, 2011) y las Máquinas de Soporte Vectorial (Lata *et al.*, 2010; Porto *et al.*, 2012; Hajisharifi *et al.*, 2014), donde los modelos adolecen de falta de interpretación en el dominio donde están siendo utilizados. Este inconveniente en algunas áreas de aplicación puede que no resulte importante, pero en la realización de pronósticos vinculados con la medicina sí lo es. Por ejemplo, comprender las explicaciones del modelo que conducen a predecir si un péptido es un buen candidato a fármaco antimicrobiano, puede ayudar a confiar en la predicción, aumentando luego las posibilidades de realizar la síntesis y experimentación *in vitro*, así como favorecer la formulación de nuevas teorías e hipótesis.

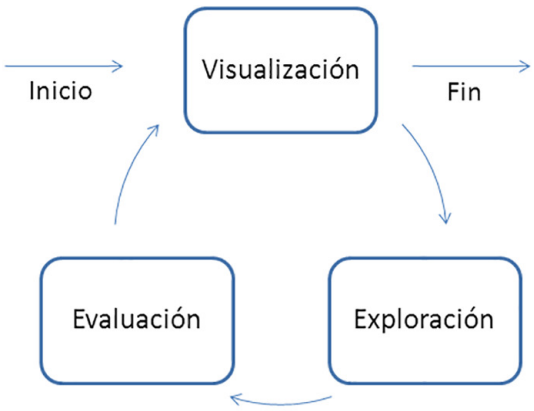


Figura 30. Ciclo de visualización, exploración y evaluación en el diseño e identificación *in silico* de nuevos candidatos a fármacos.

De lo anterior, se propone incorporar la construcción y análisis visual del modelo basado en redes de similitud molecular dentro del ciclo de diseño e identificación *in silico* de candidatos a fármacos peptídicos (ver Figura 30). Al incorporar esta nueva fase, mediante la exploración visual del EQB se combinan: i) el conocimiento y poder de análisis del investigador, con ii) la capacidad de almacenamiento y poder de cómputo de las computadoras actuales, para lograr de esta simbiosis una búsqueda eficiente (Holzinger *et al.*, 2014; Holzinger, 2016).

La idea básica del proceso ilustrado en la Figura 30 consiste en lo siguiente. Una vez que el investigador adquiere la vista de un modelo descriptivo, podrá enfocar y adentrarse en las regiones del EQB que le resulten de interés para obtener información detallada, y así filtrar los compuestos ubicados en la zona explorada. De esta manera, en la evaluación final de los compuestos explorados se tendrá en cuenta la región del EQB que estos ocupan y sus vecindades. Su importancia se debe a que el investigador puede llegar a tener más confianza en las predicciones, ya que se siente involucrado directamente en el proceso, y colabora de forma intuitiva para guiar la búsqueda de los métodos computacionales que generan los nuevos compuestos y los evalúan.

Literatura citada

Abdi, H. y Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, **2**(4): 433–459.

Aguilera-Mendoza, L., Marrero-Ponce, Y., Tellez-Ibarra, R., Llorente-Quesada, M. T., Salgado, J., Barigye, S. J., y Liu, J. (2015). Overlap and diversity in antimicrobial peptide databases: compiling a non-redundant set of sequences. *Bioinformatics*, **31**(15): 2553–2559.

Aguilera-Mendoza, L., Marrero-Ponce, Y., Beltran, J. A., Tellez Ibarra, R., Guillen-Ramirez, H. A., y Brizuela, C. A. (2019). Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. *Bioinformatics*, **35**(22): 4739–4747.

Aguilera-Mendoza, L., Marrero-Ponce, Y., García-Jacas, C. R., Chavez, E., Beltran, J. A., Guillen-Ramirez, H. A., y Brizuela, C. A. (in-press). Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Scientific Reports*.

Albert, R. y Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, **74**(1): 47.

Alelyani, S., Tang, J., y Liu, H. (2013). Feature selection for clustering: A review. En: *Data Clustering: Algorithms and Applications*.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., y Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, **25**(17): 3389–3402.

Angell, Y., Holford, M., y Moos, W. H. (2018). Building on success: A bright future for peptide therapeutics. *Protein and peptide letters*, **25**(12): 1044–1050.

Aslam, B., Wang, W., Arshad, M. I., Khurshid, M., Muzammil, S., Rasool, M. H., Nisar, M. A., Alvi, R. F., Aslam, M. A., Qamar, M. U., *et al.* (2018). Antibiotic resistance: a rundown of a global crisis. *Infection and drug resistance*, **11**: 1645.

Avan, I., Hall, C. D., y Katritzky, A. R. (2014). Peptidomimetics via modifications of amino acids and peptide bonds. *Chemical Society Reviews*, **43**(10): 3575–3594.

Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., y Todeschini, R. (2014). A novel variable reduction method adapted from space-filling designs. *Chemometrics and Intelligent Laboratory Systems*, **136**: 147–154.

Barabási, A.-L. (2003). *Linked: The new science of networks*. American Association of Physics Teachers.

Barabási, A.-L. y Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439): 509–512.

Basith, S., Manavalan, B., Hwan Shin, T., y Lee, G. (2020). Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Medicinal Research Reviews*, **n/a**(n/a).

Bastian, M., Heymann, S., Jacomy, M., *et al.* (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, **8**: 361–362.

Beltran, J. A., Aguilera-Mendoza, L., y Brizuela, C. A. (2018). Optimal selection of molecular descriptors for antimicrobial peptides classification: an evolutionary feature weighting approach. *BMC genomics*, **19**(7): 672.

Beltran, J. A., Del Rio, G., y Brizuela, C. A. (2020). An automatic representation of peptides for effective antimicrobial activity classification. *Computational and structural biotechnology journal*, **18**: 455 – 463.

Beyer, K., Goldstein, J., Ramakrishnan, R., y Shaft, U. (1999). When is "nearest neighbor" meaningful? En: *International conference on database theory*. Springer, pp. 217–235.

Bhattacharya, M., Sharma, A. R., Patra, P., Ghosh, P., Sharma, G., Patra, B. C., Lee, S.-S., y Chakraborty, C. (2020). Development of epitope-based peptide vaccine against novel coronavirus 2019 (sars-cov-2): Immunoinformatics approach. *Journal of medical virology*, **92**(6): 618–631.

Bhopale, G. M. (2020). Antimicrobial peptides: A promising avenue for human healthcare. *Current Pharmaceutical Biotechnology*, **21**(2): 90–96.

Bigelow, C. C. (1967). On the average hydrophobicity of proteins and the relation between it and protein structure. *Journal of Theoretical Biology*, **16**(2): 187–211.

Biggs, N., Lloyd, E. K., y Wilson, R. J. (1986). *Graph Theory, 1736-1936*. Oxford University Press.

Bjellqvist, B., Basse, B., Olsen, E., y Celis, J. E. (1994). Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a ph scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, **15**(1): 529–539.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., y Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008**(10): P10008.

Bock, H. (2012). *The Definitive Guide to NetBeans Platform 7*. Apress.

Boldi, P. y Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, **10**(3-4): 222–262.

Boman, H. (2003). Antibacterial peptides: basic facts and emerging concepts. *Journal of internal medicine*, **254**(3): 197–215.

Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology*, **1**(3).

Brandes, U. y Pich, C. (2007). Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, **17**(07): 2303–2318.

Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., y Marshall, M. S. (2001). Graphml progress report structural layer proposal. En: *International Symposium on Graph Drawing*. Springer, pp. 501–512.

Broido, A. D. y Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, **10**(1): 1–10.

Cai, J., Luo, J., Wang, S., y Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, **300**: 70–79.

Calvo, T., Mayor, G., y Mesiar, R. (2012). *Aggregation operators: new trends and applications*, Vol. 97. Physica.

Cao, D.-S., Xu, Q.-S., y Liang, Y.-Z. (2013). propy: a tool to generate various modes of chou's pseaac. *Bioinformatics*, **29**(7): 960–962.

Cao, D.-S., Xiao, N., Xu, Q.-S., y Chen, A. F. (2015). Rcpi: R/bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, **31**(2): 279–281.

Capecchi, A., Zhang, A., y Reymond, J.-L. (2020). Populating chemical space with peptides using a genetic algorithm. *Journal of Chemical Information and Modeling*, **60**(1): 121–132.

Charton, M. y Charton, B. I. (1983). The dependence of the chou-fasman parameters on amino acid side chain structure. *Journal of theoretical biology*, **102**(1): 121–134.

Chartrand, G., Lesniak, L., y Zhang, P. (2010). *Graphs & digraphs*, Vol. 39. CRC press.

Chávez, E., Navarro, G., Baeza-Yates, R., y Marroquín, J. L. (2001). Searching in metric spaces. *ACM computing surveys (CSUR)*, **33**(3): 273–321.

Chavez, E., Dobrev, S., Kranakis, E., Opatrny, J., Stacho, L., Tejeda, H., y Urrutia, J. (2006). Half-space proximal: A new local test for extracting a bounded dilation spanner of a unit disk graph. En: *Proceedings of the 9th International Conference on Principles of Distributed Systems*, Berlin, Heidelberg. Springer-Verlag, OPODIS'05, pp. 235–245.

Chen, C. H. y Lu, T. K. (2020). Development and challenges of antimicrobial peptides for therapeutic applications. *Antibiotics*, **9**(1): 24.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K.-C., *et al.* (2018). ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**(14): 2499–2502.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., *et al.* (2020). ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data. *Briefings in bioinformatics*, **21**(3): 1047–1057.

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

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., *et al.* (2014). Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, **57**(12): 4977–5010.

Cherven, K. (2013). *Network graph analysis and visualization with Gephi*. Packt Publishing Ltd.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., y de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11): 1422–1423.

Collantes, E. R. y Dunn III, W. J. (1995). Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogs. *Journal of medicinal chemistry*, **38**(14): 2705–2713.

Conibear, A. C., Schmid, A., Kamalov, M., Becker, C. F., y Bello, C. (2020). Recent advances in peptide-based approaches for cancer treatment. *Current Medicinal Chemistry*, **27**(8): 1174–1205.

Conlon, J., Sonnevend, A., Davidson, C., Smith, D. D., y Nielsen, P. F. (2004). The ascaphins: a family of antimicrobial peptides from the skin secretions of the most primitive extant frog, ascaphus truei. *Biochemical and Biophysical Research Communications*, **320**(1): 170 – 175.

Consortium, U. (2017). Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1): D158–D169.

Cover, T. M. y Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., y Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, **138**(3): 333–408.

Dash, M., Liu, H., y Yao, J. (1997). Dimensionality reduction of unsupervised data. En: *Proceedings ninth ieee international conference on tools with artificial intelligence*. IEEE, pp. 532–539.

Dash, M., Choi, K., Scheuermann, P., y Huan Liu (2002). Feature selection for clustering - a filter solution. En: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*. pp. 115–122.

de la Vega de León, A. y Bajorath, J. (2016). Chemical space visualization: transforming multidimensional chemical spaces into similarity-based molecular networks. *Future medicinal chemistry*, **8**(14): 1769–1778.

Deslouches, B., Phadke, S. M., Lazarevic, V., Cascio, M., Islam, K., Montelaro, R. C., y Mietzner, T. A. (2005). De novo generation of cationic antimicrobial peptides: influence of length and tryptophan substitution on antimicrobial activity. *Antimicrobial agents and chemotherapy*, **49**(1): 316–322.

Deslouches, B., Steckbeck, J. D., Craigo, J. K., Doi, Y., Mietzner, T. A., y Montelaro, R. C. (2013). Rational design of engineered cationic antimicrobial peptides consisting exclusively of arginine and tryptophan, and their activity against multidrug-resistant pathogens. *Antimicrobial agents and chemotherapy*, **57**(6): 2511–2521.

Dobson, C. M. (2004). Chemical space and biology. *Nature*, **432**(7019): 824–828.

Dong, J., Yao, Z.-J., Zhang, L., Luo, F., Lin, Q., Lu, A.-P., Chen, A. F., y Cao, D.-S. (2018). Pybiomed: a python library for various molecular representations of chemicals, proteins and dnas and their interactions. *Journal of cheminformatics*, **10**(1): 16.

Du, P., Wang, X., Xu, C., y Gao, Y. (2012). Pseaac-builder: a cross-platform standalone program for generating various special chou's pseudo-amino acid compositions. *Analytical biochemistry*, **425**(2): 117–119.

Du, P., Gu, S., y Jiao, Y. (2014). Pseaac-general: fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. *International journal of molecular sciences*, **15**(3): 3495–3506.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, **26**(19): 2460–2461.

Eisenberg, D. (1984). Three-dimensional structure of membrane and surface proteins. *Annual review of biochemistry*, **53**(1): 595–623.

Eley, A., Ibrahim, M., Kurdi, S. E., y Conlon, J. M. (2008). Activities of the frog skin peptide, ascaphin-8 and its lysine-substituted analogs against clinical isolates of extended-spectrum β-lactamase (esbl) producing bacteria. *Peptides*, **29**(1): 25 – 30.

Erdős, P. y Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, **5**(1): 17–60.

Estrada, E. (2012). *The structure of complex networks: theory and applications*. Oxford University Press.

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.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11): 27–34.

Felício, M. R., Silva, O. N., Gonçalves, S., Santos, N. C., y Franco, O. L. (2017). Peptides with dual antimicrobial and anticancer activities. *Frontiers in chemistry*, **5**: 5.

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

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.

Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A., Pante, N., Hancock, R. E., y Cherkasov, A. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine learning†. *Journal of medicinal chemistry*, **52**(7): 2006–2015.

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

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, **15**(1): 1–10.

Fruchterman, T. M. y Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, **21**(11): 1129–1164.

Gabere, M. N. y Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics*, **33**(13): 1921–1929.

Gabernet, G., Gautschi, D., Müller, A. T., Neuhaus, C. S., Armbrecht, L., Dittrich, P. S., Hiss, J. A., y Schneider, G. (2019). In silico design and optimization of selective membranolytic anticancer peptides. *Scientific reports*, **9**(1): 1–11.

Galpert, D., Fernández, A., Herrera, F., Antunes, A., Molina-Ruiz, R., y Agüero-Chapin, G. (2018). Surveying alignment-free features for ortholog detection in related yeast proteomes by using supervised big data classifiers. *BMC bioinformatics*, **19**(1): 166.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., Bairoch, A., *et al.* (2005). Protein identification and analysis tools on the expasy server. En: *The proteomics protocols handbook*. Springer, pp. 571–607.

Gautam, A., Chaudhary, K., Singh, S., Joshi, A., Anand, P., Tuknait, A., Mathur, D., Varshney, G. C., y Raghava, G. P. (2013). Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic acids research*, **42**(D1): D444–D449.

Gaynes, R. (2017). The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerging infectious diseases*, **23**(5): 849.

Gergely, Z.-K., Sheils, T., y Oprea, T. I. (2020). Smartgraph: a network pharmacology investigation platform. *Journal of Cheminformatics*, **12**(1).

Ghalmane, Z., El Hassouni, M., y Cherifi, H. (2019). Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining*, **9**(1): 45.

Ghosh, C., Sarkar, P., Issa, R., y Haldar, J. (2019). Alternatives to conventional antibiotics in the era of antimicrobial resistance. *Trends in microbiology*.

Godden, J. W. y Bajorath, J. (2002). Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by se-dse analysis. *Journal of chemical information and computer sciences*, **42**(1): 87–93.

Godden, J. W., Stahura, F. L., y Bajorath, J. (2000). Variability of molecular descriptors in compound databases revealed by shannon entropy calculations. *Journal of chemical information and computer sciences*, **40**(3): 796–800.

Gómez, E. A., Giraldo, P., y Orduz, S. (2017). Inverpep: A database of invertebrate antimicrobial peptides. *Journal of Global Antimicrobial Resistance*, **8**: 13–17.

Gonzalez, M. P., Teran, C., Saiz-Urra, L., y Teijeira, M. (2008). Variable selection methods in qsar: an overview. *Current topics in medicinal chemistry*, **8**(18): 1606–1627.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, **40**(1): 33–51.

Guare, J. (1990). *Six degrees of separation: A play*. Vintage.

Gueguen, Y., Garnier, J., Robert, L., Lefranc, M.-P., Mougenot, I., De Lorgeril, J., Janech, M., Gross, P. S., Warr, G. W., Cuthbertson, B., *et al.* (2006). Penbase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Developmental & Comparative Immunology*, **30**(3): 283–288.

Guidotti, G., Brambilla, L., y Rossi, D. (2017). Cell-penetrating peptides: from basic research to clinics. *Trends in pharmacological sciences*, **38**(4): 406–424.

Gull, S., Shamim, N., y Minhas, F. (2019). Amap: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Computers in biology and medicine*, **107**: 172–181.

Guruprasad, K., Reddy, B. B., y Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, **4**(2): 155–161.

Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**(Mar): 1157–1182.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., *et al.* (1998). *Multivariate data analysis*. Prentice hall Upper Saddle River, NJ.

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., y Mohabatkar, H. (2014). Predicting anticancer peptides with chou's pseudo amino acid composition and investigating their mutagenicity via ames test. *Journal of Theoretical Biology*, **341**: 34–40.

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., Le Lay, C., Ben Hamida, J., y Fliss, I. (2010). Bactibase second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, **10**(1): 22.

He, P.-a., Wei, J., Yao, Y., y Tie, Z. (2012). A novel graphical representation of proteins and its application. *Physica A: Statistical Mechanics and its Applications*, **391**(1): 93–99.

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

Henninot, A., Collins, J. C., y Nuss, J. M. (2018). The current state of peptide drug discovery: back to the future? *Journal of medicinal chemistry*, **61**(4): 1382–1414.

Hernández, M. A. y Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, **2**(1): 9–37.

Herraez, A. (2006). Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, **34**(4): 255–261.

Hilchie, A., Hoskin, D., y Coombs, M. P. (2019). Anticancer activities of natural and synthetic peptides. En: *Antimicrobial Peptides*. Springer, pp. 131–147.

Hinz, U., Consortium, U., *et al.* (2010). From protein sequences to 3d-structures and beyond: the example of the uniprot knowledgebase. *Cellular and molecular life sciences*, **67**(7): 1049–1064.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, **75**(4): 800–802.

Hollander, M., Wolfe, D. A., y Chicken, E. (2013). *Nonparametric statistical methods*, Vol. 751. John Wiley & Sons.

Holme, P. (2019). Rare and everywhere: Perspectives on scale-free networks. *Nature communications*, **10**(1): 1–3.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, **3**(2): 119–131.

Holzinger, A., Dehmer, M., y Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, **15**(6): I1.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., *et al.* (2008). The future of biocuration. *Nature*, **455**(7209): 47–50.

Huttner, A., Harbarth, S., Carlet, J., Cosgrove, S., Goossens, H., Holmes, A., Jarlier, V., Voss, A., y Pittet, D. (2013). Antimicrobial resistance: a global view from the 2013 world healthcare-associated infections forum. *Antimicrobial resistance and infection control*, **2**(1): 1.

Jachiet, P.-A., Pogorelcnik, R., Berry, A., Lopez, P., y Bapteste, E. (2013). Mosaicfinder: identification of fused gene families in sequence similarity networks. *Bioinformatics*, **29**(7): 837–844.

Jacomy, M., Venturini, T., Heymann, S., y Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, **9**(6): 1–12.

Jenssen, H. (2011). Descriptors for antimicrobial peptides. *Expert opinion on drug discovery*, **6**(2): 171–184.

Jhong, J.-H., Chi, Y.-H., Li, W.-C., Lin, T.-H., Huang, K.-Y., y Lee, T.-Y. (2018). dbamp: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic acids research*, **47**(D1): D285–D297.

Jolliffe, I. T. y Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**(2065): 20150202.

Kalita, P., Padhi, A. K., Zhang, K. Y., y Tripathi, T. (2020). Design of a peptide-based subunit vaccine against novel coronavirus sars-cov-2. *Microbial Pathogenesis*, **145**: 104236.

Kawashima, S., Ogata, H., y Kanehisa, M. (1999). Aaindex: amino acid index database. *Nucleic acids research*, **27**(1): 368–369.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., y Kanehisa, M. (2007). Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, **36**(suppl_1): D202–D205.

Keim, D. A. (2002). Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, **8**(1): 1–8.

Kim, S.-J., Kim, J.-S., Lee, Y.-S., Sim, D.-W., Lee, S.-H., Bahk, Y.-Y., Lee, K.-H., Kim, E.-H., Park, S.-J., Lee, B.-J., *et al.* (2013). Structural characterization of de novo designed l5k5w model peptide isomers with potent antimicrobial and varied hemolytic activities. *Molecules*, **18**(1): 859–876.

Kimball, R. y Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

King, J. R. y Jackson, D. A. (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics: The official journal of the International Environmetrics Society*, **10**(1): 67–77.

Klein, P., Kanehisa, M., y DeLisi, C. (1984). Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, **787**(3): 221–226.

Klopmand, G. (1992). Concepts and applications of molecular similarity. *Journal of Computational Chemistry*, **13**(4): 539–540.

Koo, H. B. y Seo, J. (2019). Antimicrobial peptides under clinical investigation. *Peptide Science*, **111**(5): e24122.

Kotsiantis, S. (2011). Feature selection for machine learning classification problems: a recent overview. *Artificial Intelligence Review*, **42**(1): 157–176.

Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A., y Getzoff, E. D. (1995). Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Structure, Function, and Bioinformatics*, **23**(4): 536–547.

Kumar, P., Kizhakkedathu, J. N., y Straus, S. K. (2018). Antimicrobial peptides: Diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. *Biomolecules*, **8**(1): 4.

Kunimoto, R. y Bajorath, J. (2018). Combining similarity searching and network analysis for the identification of active compounds. *ACS omega*, **3**(4): 3768–3777.

Kunimoto, R., Vogt, M., y Bajorath, J. (2017). Tracing compound pathways using chemical space networks. *MedChemComm*, **8**(2): 376–384.

Kunkel, C., Schober, C., Oberhofer, H., y Reuter, K. (2019). Knowledge discovery through chemical space networks: the case of organic electronics. *Journal of molecular modeling*, **25**(4): 87.

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

Lafita, A., Bliven, S., Prlić, A., Guzenko, D., Rose, P. W., Bradley, A., Pavan, P., Myers-Turnbull, D., Valasatava, Y., Heuer, M., *et al.* (2019). Biojava 5: A community driven open-source bioinformatics library. *PLoS computational biology*, **15**(2): e1006791.

Landherr, A., Friedl, B., y Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, **2**(6): 371–385.

Laskowski, R. A. y Thornton, J. M. (2008). Understanding the molecular machinery of genetics through 3d structures. *Nature Reviews Genetics*, **9**(2): 141–151.

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

Lau, J. L. y Dunn, M. K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry*, **26**(10): 2700–2707.

Laughlin, T. F. y Ahmad, Z. (2010). Inhibition of escherichia coli atp synthase by amphibian antimicrobial peptides. *International journal of biological macromolecules*, **46**(3): 367–374.

Lee, A. C.-L., Harris, J. L., Khanna, K. K., y Hong, J.-H. (2019). A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences*, **20**(10): 2383.

Lee, H.-T., Lee, C.-C., Yang, J.-R., Lai, J. Z., y Chang, K. Y. (2015). A large-scale structural classification of antimicrobial peptides. *BioMed research international*, **2015**.

Lei, J., Sun, L., Huang, S., Zhu, C., Li, P., He, J., Mackey, V., Coy, D. H., y He, Q. (2019). The antimicrobial peptides and their potential clinical applications. *American journal of translational research*, **11**(7): 3919.

Lepp, Z., Huang, C., y Okada, T. (2009). Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. *Journal of chemical information and modeling*, **49**(11): 2429–2443.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. En: *Soviet physics doklady*. Vol. 10, pp. 707–710.

Li, Y. y Chen, Z. (2008). Rapd: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiology Letters*, **289**(2): 126–129.

Li, Z.-R., Lin, H. H., Han, L., Jiang, L., Chen, X., y Chen, Y. Z. (2006). Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, **34**(suppl_2): W32–W37.

Lipinski, C. y Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature*, **432**(7019): 855–861.

Liu, B. (2019). Bioseq-analysis: a platform for dna, rna and protein sequence analysis based on machine learning approaches. *Briefings in bioinformatics*, **20**(4): 1280–1294.

Liu, B., Gao, X., y Zhang, H. (2019). Bioseq-analysis2.0: an updated platform for analyzing dna, rna and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic acids research*, **47**(20): e127–e127.

Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., y Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, **650**: 1–63.

Luca, M. D., Maccari, G., Maisetta, G., y Batoni, G. (2015). Baamps: the database of biofilm-active antimicrobial peptides. *Biofouling*, **31**(2): 193–199.

Maciejewski, R. (2011). Data representations, transformations, and statistics for visual reasoning. *Synthesis Lectures on Visualization*, **2**(1): 1–85.

Madani, F., Lindberg, S., Langel, Ü., Futaki, S., y Gräslund, A. (2011). Mechanisms of cellular uptake of cell-penetrating peptides. *Journal of biophysics*, **2011**.

Maggiora, G., Vogt, M., Stumpfe, D., y Bajorath, J. (2013). Molecular similarity in medicinal chemistry: Miniperspective. *Journal of medicinal chemistry*, **57**(8): 3186–3204.

Maggiora, G., Vogt, M., Stumpfe, D., y Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, **57**(8): 3186–3204.

Maggiora, G. M. y Bajorath, J. (2014). Chemical space networks: a powerful new paradigm for the description of chemical space. *Journal of Computer-Aided Molecular Design*, **28**(8): 795–802.

Maggiora, G. M. y Shanmugasundaram, V. (2011). Molecular similarity measures. *Chemoinformatics and computational chemical biology*, pp. 39–100.

Malonis, R. J., Lai, J. R., y Vergnolle, O. (2019). Peptide-based vaccines: Current progress and future challenges. *Chemical reviews*, **120**(6): 3210–3229.

Mannhold, R., Kubinyi, H., Folkers, G., y Brown, N. (2013). *Scaffold Hopping in Medicinal Chemistry*, Vol. 58. John Wiley & Sons.

Mark, B. (2002). *Nexus: Small Worlds and the Groundbreaking Science of Networks*. WW Norton, New York.

Marrero-Ponce, Y., Teran, J. E., Contreras-Torres, E., García-Jacas, C. R., Perez-Castillo, Y., Cubillan, N., Peréz-Giménez, F., y Valdés-Martini, J. R. (2020). Lego-based generalized set of two linear algebraic 3d bio-macro-molecular descriptors: Theory and validation by qsars. *Journal of theoretical biology*, **485**: 110039.

Martin, Y. C., Kofron, J. L., y Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, **45**(19): 4350–4358.

Martínez-López, Y., Marrero-Ponce, Y., Barigye, S. J., Teran, E., Martínez-Santiago, O., Zambrano, C. H., y Torres, F. J. (2019). When global and local molecular descriptors are more than the sum of its parts: Simple, but not simpler? *Molecular Diversity*, pp. 1–20.

Mathews, C. K., Van Holde, K., y Ahern, K. G. (2000). Biochemistry, ed. *San Francisco: BenjaminlCummings*.

Maurya, N. S., Kushwaha, S., y Mani, A. (2019). Recent advances and computational approaches in peptide drug discovery. *Current pharmaceutical design*, **25**(31): 3358–3366.

Medina-Franco, J. L., Martínez-Mayorga, K., Giulianotti, M. A., Houghten, R. A., y Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *Current Computer-Aided Drug Design*, **4**(4): 322–333.

Mehta, D., Anand, P., Kumar, V., Joshi, A., Mathur, D., Singh, S., Tuknait, A., Chaudhary, K., Gautam, S. K., Gautam, A., Varshney, G. C., y Raghava, G. P. (2014). Parapep: a web resource for experimentally validated antiparasitic peptide sequences and their structures. *Database*, **2014**: bau051.

Meyer, M. J., Das, J., Wang, X., y Yu, H. (2013). Instruct: a database of high-quality 3d structurally resolved protein interactome networks. *Bioinformatics*, **29**(12): 1577–1579.

Michael Conlon, J., Galadari, S., Raza, H., y Condamine, E. (2008). Design of potent, non-toxic antimicrobial agents based upon the naturally occurring frog skin peptides, ascaphin-8 and peptide xt-7. *Chemical biology & drug design*, **72**(1): 58–64.

Milgram, S. (1967). The small world problem. *Psychology today*, **2**(1): 60–67.

Mojsoska, B. y Jenssen, H. (2015). Peptides and peptidomimetics for antimicrobial drug design. *Pharmaceuticals*, **8**(3): 366–415.

Moore, T. J., Zhang, H., Anderson, G., y Alexander, G. C. (2018). Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. *JAMA Internal Medicine*, **178**(11): 1451–1457.

Mosca, R., Pons, T., Céol, A., Valencia, A., y Aloy, P. (2013). Towards a detailed atlas of protein–protein interactions. *Current opinion in structural biology*, **23**(6): 929–940.

Mount, D. W. (2004). *Bioinformatics: sequence and genome analysis (2. ed.)*. Cold Spring Harbor Laboratory Press.

Müller, A. T., Gabernet, G., Hiss, J. A., y Schneider, G. (2017). modlamp: Python for antimicrobial peptides. *Bioinformatics*, **33**(17): 2753–2755.

Nagarajan, D., Nagarajan, T., Nanajkar, N., y Chandra, N. (2019). A uniform in vitro efficacy dataset to guide antimicrobial peptide design. *Data*, **4**(1): 27.

Needleman, S. B. y Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3): 443–453.

Nelson, D. L. y Cox, M. M. (2017). *Lehninger principles of biochemistry*. W. H. Freeman; Edición: Seventh.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2004). Analysis of weighted networks. *Physical review E*, **70**(5): 056131.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, **46**(5): 323–351.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, **103**(23): 8577–8582.

Newman, M. E., Barabási, A.-L. E., y Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton university press.

Nikam, R. y Gromiha, M. M. (2019). Seq2feature: a comprehensive web-based feature extraction tool. *Bioinformatics*, **35**(22): 4797–4799.

Novković, M., Simunić, J., Bojović, V., Tossi, A., y Juretić, D. (2012). Dadp: the database of anuran defense peptides. *Bioinformatics*, **28**(10): 1406–1407.

Opassi, G., Gesù, A., y Massarotti, A. (2018). The hitchhiker's guide to the chemical-biological galaxy. *Drug discovery today*, **23**(3): 565–574.

Oprea, T. I. y Gottfries, J. (2001). Chemography: the art of navigating in chemical space. *Journal of combinatorial chemistry*, **3**(2): 157–166.

Osolodkin, D. I., Radchenko, E. V., Orlov, A. A., Voronkov, A. E., Palyulin, V. A., y Zefirov, N. S. (2015). Progress in visual representations of chemical space. *Expert opinion on drug discovery*, **10**(9): 959–973.

Pande, A., Patiyal, S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S., *et al.* (2019). Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv*, p. 599126.

Pathmanathan, J. S., Lopez, P., Lapointe, F.-J., y Bapteste, E. (2017). Compositesearch: a generalized network approach for composite gene families detection. *Molecular biology and evolution*, **35**(1): 252–255.

Pavlopoulos, G. A., Paez-Espino, D., Kyrpides, N. C., y Iliopoulos, I. (2017). Empirical comparison of visualization tools for larger-scale network analysis. *Advances in bioinformatics*, **2017**.

Pearson, W. R. y Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, **85**(8): 2444–2448.

Peng, H., Long, F., y Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, **27**(8): 1226–1238.

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

Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H. L., Squires, R. B., Hurt, D. E., Grigolava, M., Chubinidze, M., Gogoladze, G., Vishnepolsky, B., Alekseev, V., Rosenthal, A., y Tartakovsky, M. (2016). Dbaasp v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Research*, **44**(D1): D1104–D1112.

Popovic, S., Urbán, E., Lukic, M., y Conlon, J. M. (2012). Peptides with antimicrobial and anti-inflammatory activities that have therapeutic potential for treatment of acne vulgaris. *Peptides*, **34**(2): 275–282.

Porto, W., Pires, A., y Franco, O. (2017). Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnology advances*, **35**(3): 337–349.

Porto, W. F., Pires, Á. S., y Franco, O. L. (2012). Cs-amppred: an updated svm model for antimicrobial activity prediction in cysteine-stabilized peptides. *PloS one*, **7**(12): e51444.

Porto, W. F., Irazazabal, L., Alves, E. S., Ribeiro, S. M., Matos, C. O., Pires, Á. S., Fensterseifer, I. C., Miranda, V. J., Haney, E. F., Humblot, V., *et al.* (2018). In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nature communications*, **9**(1): 1–12.

Prabhakaran, M. (1990). The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochemical Journal*, **269**(3): 691–696.

Qureshi, A., Thakur, N., y Kumar, M. (2013). Hipdb: A database of experimentally validated hiv inhibiting peptides. *PLOS ONE*, **8**(1): 1–5.

Qureshi, A., Thakur, N., Tandon, H., y Kumar, M. (2014). Avpdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Research*, **42**(D1): D1147–D1153.

Ramsey, J. D. y Flynn, N. H. (2015). Cell-penetrating peptides transport therapeutics into cells. *Pharmacology & therapeutics*, **154**: 78–86.

Randic, M., Zupan, J., Balaban, A. T., Vikic-Topic, D., y Plavsic, D. (2010). Graphical representation of proteins†. *Chemical reviews*, **111**(2): 790–862.

Rao, H., Zhu, F., Yang, G., Li, Z., y Chen, Y. (2011). Update of profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic acids research*, **39**(suppl_2): W385–W390.

Rao, V. M. y Sastry, V. (2012). Unsupervised feature ranking based on representation entropy. En: *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, pp. 421–425.

Recanatini, M. y Cabrelle, C. (2020). Drug research meets network science: Where are we? *Journal of Medicinal Chemistry*, **0**(0): null.

Renner, S. y Schneider, G. (2006). Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem*, **1**(2): 181–185.

Reymond, J.-L. (2015). The chemical space project. *Accounts of Chemical Research*, **48**(3): 722–730.

Reymond, J.-L. y Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*, **3**(9): 649–657.

Romero-Molina, S., Ruiz-Blanco, Y. B., Green, J. R., y Sanchez-Garcia, E. (2019). Protdcal-suite: A web server for the numerical codification and functional analysis of proteins. *Protein Science*, **28**(9): 1734–1743.

Ruiz-Blanco, Y. B., Paz, W., Green, J., y Marrero-Ponce, Y. (2015). Protdcal: A program to compute general-purpose-numerical descriptors for sequences and 3d-structures of proteins. *BMC bioinformatics*, **16**(1): 162.

Sak, K., Karelson, M., y Järv, J. (1999). Modeling of the amino acid side chain effects on peptide conformation. *Bioorganic Chemistry*, **27**(6): 434–442.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., y Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *Journal of medicinal chemistry*, **41**(14): 2481–2491.

Schober, P., Boer, C., y Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, **126**(5): 1763–1768.

Scott, J. (2000). *Social Network Analysis: A Handbook*. Sage Publications, segunda edición.

Seebah, S., Suresh, A., Zhuo, S., Choong, Y. H., Chua, H., Chuon, D., Beuerman, R., y Verma, C. (2007). Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Research*, **35**(suppl_1): D265–D268.

Sertkaya, A., Wong, H.-H., Jessup, A., y Beleche, T. (2016). Key cost drivers of pharmaceutical clinical trials in the united states. *Clinical Trials*, **13**(2): 117–126.

Seshadri Sundararajan, V., Gabere, M. N., Pretorius, A., Adam, S., Christoffels, A., Lehväslaiho, M., Archer, J. A. C., y Bajic, V. B. (2012). Dampd: a manually curated antimicrobial peptide database. *Nucleic Acids Research*, **40**(D1): D1108–D1112.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**(3): 379–423.

Sheikhpour, R., Sarram, M. A., Gharaghani, S., y Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern Recognition*, **64**: 141–158.

Shen, H.-B. y Chou, K.-C. (2008). Pseaac: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry*, **373**(2): 386–388.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. En: *Proceedings 1996 IEEE Symposium on Visual Languages*, Sep. pp. 336–343.

Simm, S., Einloft, J., Mirus, O., y Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological research*, **49**(1): 31.

Simoff, S., Böhlen, M. H., y Mazeika, A. (2008). *Visual data mining: theory, techniques and tools for visual analytics*, Vol. 4404. Springer Science & Business Media.

Singh, M., Kumar, V., Sikka, K., Thakur, R., Harioudh, M. K., Mishra, D. P., Ghosh, J. K., y Siddiqi, M. I. (2020). Computational design of biologically active anticancer peptides and their interactions with heterogeneous popc/pops lipid membranes. *Journal of Chemical Information and Modeling*, **60**(1): 332–341. PMID: 31880450.

Singh, S., Chaudhary, K., Dhanda, S. K., Bhalla, S., Usmani, S. S., Gautam, A., Tuknait, A., Agrawal, P., Mathur, D., y Raghava, G. P. (2015). Satpdb: a database of structurally annotated therapeutic peptides. *Nucleic acids research*, **44**(D1): D1119–D1126.

Skolnick, J. y Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in biotechnology*, **18**(1): 34–39.

Smith, T. F., Waterman, M. S., *et al.* (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1): 195–197.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., y Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, **53**(2): 907–948.

Stumpfe, D. y Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, **55**(7): 2932–2942.

Tang, J., Alelyani, S., y Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, p. 37.

Théolier, J., Fliss, I., Jean, J., y Hammami, R. (2014). Milkamp: a comprehensive database of antimicrobial peptides of dairy origin. *Dairy Science & Technology*, **94**(2): 181–193.

Todeschini, R. y Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*, Vol. 41. John Wiley & Sons.

Todeschini, R., Ballabio, D., y Consonni, V. (2020). *Distances and Similarity Measures in Chemometrics and Chemoinformatics*, pp. 1–40. American Cancer Society.

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

Torres, M. D., Sothiselvam, S., Lu, T. K., y de la Fuente-Nunez, C. (2019). Peptide design principles for antimicrobial applications. *Journal of molecular biology*, **431**(18): 3547–3567.

Torres, M. D. T. y de la Fuente-Nunez, C. (2019). Toward computer-made artificial antibiotics. *Current opinion in microbiology*, **51**: 30–38.

Tossi, A. y Sandri, L. (2002). Molecular diversity in gene-encoded, cationic antimicrobial polypeptides. *Current pharmaceutical design*, **8**(9): 743–761.

Tossi, A., Sandri, L., y Giangaspero, A. (2003). New consensus hydrophobicity scale extended to non-proteinogenic amino acids. En: *27th European Peptide Symposium*. Edizioni Ziino, pp. 416–417.

Tropsha, A. (2010). Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, **29**(6-7): 476–488.

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., y Raghava, G. P. (2015). Cancerppd: a database of anticancer peptides and proteins. *Nucleic Acids Research*, **43**(D1): D837–D843.

Urias, R. W. P., Barigye, S. J., Marrero-Ponce, Y., García-Jacas, C. R., Valdes-Martiní, J. R., y Perez-Gimenez, F. (2015). Imman: free software for information theory-based chemometric analysis. *Molecular Diversity*, **19**(2): 305–319.

Usmani, S. S., Bedi, G., Samuel, J. S., Singh, S., Kalra, S., Kumar, P., Ahuja, A. A., Sharma, M., Gautam, A., y Raghava, G. P. (2017). Thpdb: Database of fda-approved peptide and protein therapeutics. *PLoS One*, **12**(7): e0181748.

Usmani, S. S., Kumar, R., Bhalla, S., Kumar, V., y Raghava, G. P. (2018a). In silico tools and databases for designing peptide-based vaccine and drugs. En: *Advances in protein chemistry and structural biology*, Vol. 112. Elsevier, pp. 221–263.

Usmani, S. S., Kumar, R., Kumar, V., Singh, S., y Raghava, G. P. (2018b). Antitbpdb: a knowledgebase of anti-tubercular peptides. *Database*, **2018**.

van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J., y Kuipers, O. P. (2013). Bagel3: automated identification of genes encoding bacteriocins and (non-)bactericidal post-translationally modified peptides. *Nucleic Acids Research*, **41**(W1): W448–W453.

Varshavsky, R., Gottlieb, A., Linial, M., y Horn, D. (2006). Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**(14): e507–e513.

Vassiliadis, P., Simitsis, A., y Skiadopoulos, S. (2002). Conceptual modeling for etl processes. En: *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*, New York, NY, USA. ACM, DOLAP '02, pp. 14–21.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, **17**(3): 261–272.

Vogt, I. y Mestres, J. (2010). Drug-target networks. *Molecular Informatics*, **29**(1-2): 10–14.

Vogt, M., Stumpfe, D., Maggiora, G. M., y Bajorath, J. (2016). Lessons learned from the design of chemical space networks and opportunities for new applications. *Journal of Computer-Aided Molecular Design*, **30**(3): 191–208.

Voitalov, I., van der Hoorn, P., van der Hofstad, R., y Krioukov, D. (2019). Scale-free networks well done. *Physical Review Research*, **1**(3): 033034.

von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D., y Fellner, D. (2011). Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, **30**(6): 1719–1749.

Waghu, F. H., Barai, R. S., Gurung, P., y Idicula-Thomas, S. (2016). Campr3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, **44**(D1): D1094–D1097.

Wang, C. K. L., Kaas, Q., Chiche, L., y Craik, D. J. (2008). Cybase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, **36**(suppl_1): D206–D210.

Wang, F., Wang, Y., Zhang, X., Zhang, W., Guo, S., y Jin, F. (2014). Recent progress of cell-penetrating peptides as new carriers for intracellular cargo delivery. *Journal of Controlled Release*, **174**: 126–136.

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

Wang, G., Li, X., y Wang, Z. (2016). Apd3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, **44**(D1): D1087–D1093.

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., Song, J., Chou, K.-C., y Lithgow, T. (2017a). Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles. *Bioinformatics*, **33**(17): 2756–2758.

Wang, J., Yin, T., Xiao, X., He, D., Xue, Z., Jiang, X., y Wang, Y. (2018). Strapep: a structure database of bioactive peptides. *Database*, **2018**: bay038.

Wang, J., Dou, X., Song, J., Lyu, Y., Zhu, X., Xu, L., Li, W., y Shan, A. (2019). Antimicrobial peptides: Promising alternatives in the post feeding antibiotic era. *Medicinal research reviews*, **39**(3): 831–859.

Wang, L., Dong, C., Li, X., Han, W., y Su, X. (2017b). Anticancer potential of bioactive peptides from animal sources. *Oncology reports*, **38**(2): 637–651.

Wang, Y., Wang, M., Yin, S., Jang, R., Wang, J., Xue, Z., y Xu, T. (2015). Neuropep: a comprehensive resource of neuropeptides. *Database*, **2015**.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Wasserman, S., Faust, K., *et al.* (1994). *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press.

Watts, D. J. (2004). *Six degrees: The science of a connected age*. WW Norton & Company.

Watts, D. J. y Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, **393**(6684): 440.

Willett, P. (2014). The calculation of molecular structural similarity: principles and practice. *Molecular Informatics*, **33**(6-7): 403–413.

Wu, M., Vogt, M., Maggiora, G. M., y Bajorath, J. (2016). Design of chemical space networks on the basis of tversky similarity. *Journal of computer-aided molecular design*, **30**(1): 1–12.

Wu, Q., Ke, H., Li, D., Wang, Q., Fang, J., y Zhou, J. (2019). Recent progress in machine learning-based prediction of peptide activity for drug discovery. *Current topics in medicinal chemistry*, **19**(1): 4–16.

Xiao, N., Cao, D.-S., Zhu, M.-F., y Xu, Q.-S. (2015). protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**(11): 1857–1859.

Xu, X. y Lai, R. (2015). The chemistry and biological activities of peptides from amphibian skin secretions. *Chemical reviews*, **115**(4): 1760–1846.

Yang, P. y Wang, X. (2020). Covid-19: a new challenge for human beings. *Cellular & Molecular Immunology*, **17**: 555–557.

Yang, Z., Algesheimer, R., y Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, **6**: 30750.

Yu, L. y Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, **5**(Oct): 1205–1224.

Zaffiri, L., Gardner, J., y Toledo-Pereyra, L. H. (2012). History of antibiotics. from salvarsan to cephalosporins. *Journal of Investigative Surgery*, **25**(2): 67–77.

Zamyatnin, A. (1972). Protein volume in solution. *Progress in biophysics and molecular biology*, **24**: 107–123.

Zhang, B., Vogt, M., Maggiora, G. M., y Bajorath, J. (2015a). Comparison of bioactive chemical space networks generated using substructure-and fingerprint-based measures of molecular similarity. *Journal of computer-aided molecular design*, **29**(7): 595–608.

Zhang, B., Vogt, M., Maggiora, G. M., y Bajorath, J. (2015b). Design of chemical space networks using a tanimoto similarity variant based upon maximum common substructures. *Journal of computer-aided molecular design*, **29**(10): 937–950.

Zhang, C., Yang, M., y Ericsson, A. C. (2019). Antimicrobial peptides: potential application in liver cancer. *Frontiers in microbiology*, **10**.

Zhao, X., Wu, H., Lu, H., Li, G., y Huang, Q. (2013). Lamp: A database linking antimicrobial peptides. *PLOS ONE*, **8**(6): 1–6.

Zouhir, A., Taieb, M., Lamine, M. A., Cherif, A., Jridi, T., Mahjoubi, B., Mbarek, S., Fliss, I., Nefzi, A., Sebei, K., y Hamida, J. B. (2016). Antistaphybase: database of antimicrobial peptides (amps) and essential oils (eos) against methicillin-resistant staphylococcus aureus (mrsa) and staphylococcus aureus. *Archives of Microbiology*, **199**: 215–222.

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., y Yang, L. (2017). Psekraac: a flexible web server for generating pseudo k-tuple reduced amino acids composition. *Bioinformatics*, **33**(1): 122–124.

Zwierzyna, M., Vogt, M., Maggiora, G. M., y Bajorath, J. (2015). Design and characterization of chemical space networks for different compound data sets. *Journal of computer-aided molecular design*, **29**(2): 113–125.

Anexo A. Algoritmo de filtrado inicial de descriptores

Algoritmo 1: Feature ranking and filtering

input : A descriptor matrix $\mathcal{D} = [x_{ij}]_{n \times m}$, an entropy threshold θ_1, a correlation method *corr_method*, and correlation-based similarity threshold θ_2

output: A subset F of candidate features

```
   /* Entropy-based filtering                                          */
 1 F ⟵ {j | j = 1...m} ;                  /* Initialize the candidate set */
 2 for j = 1 to m do
 3 |  f_j.entropy ⟵ H(f_j) ;                      /* It is defined in Eq. 3 */
 4 |  if f_j.entropy < θ_1 then
 5 |  |  F ⟵ F \ {f_j} ;                  /* Removing irrelevant features  */
 6 |  end
 7 end
 8 rankedFeatures ⟵ Sort features in F by descending order of their entropy
     values;
   /* Correlation-based filtering                                      */
 9 for j = 1 to sizeOf(rankedFeatures)−1 do
10 |  for k = j + 1 to sizeOf(rankedFeatures) do
11 |  |  if corr_method = "pearson" then
12 |  |  |  sim(f_j, f_k) ⟵ |ρ(f_j, f_k)| ;          /* It is defined in Eq. 5  */
13 |  |  else
14 |  |  |  if corr_method = "spearman" then
15 |  |  |  |  sim(f_j, f_k) ⟵ |r_s(f_j, f_k)| ;      /* It is defined in Eq. 6 */
16 |  |  |  end
17 |  |  end
18 |  |  if sim(f_j, f_k) ≥ θ_2 then
19 |  |  |  F ⟵ F \ {f_k} ;                  /* Removing redundant features */
20 |  |  end
21 |  end
22 end
```

Anexo B. Algoritmo de optimización del conjunto candidato

Algoritmo 2: Feature subset optimization

```
   input  : A candidate feature set F
   output: A subset F* of optimized features
   /* Second stage: subset optimization                          */
 1 F* ⟵ F ;                                  /* Candidate feature set */
 2 best_merit ⟵ Φ(F*) ;                     /* It is defined in Eq. 7 */
 3 success ⟵ true;
 4 while success do
 5 |   success ⟵ false;
 6 |   foreach f_j ∈ F* do
 7 |   |   F' ⟵ F* \ {f_j};
 8 |   |   merit ⟵ Φ(F');
 9 |   |   if merit > best_merit then
10 |   |   |   best_merit ⟵ merit ;
11 |   |   |   best_subset ⟵ F' ;
12 |   |   |   success ⟵ true;
13 |   |   end
14 |   end
15 |   if success then
16 |   |   F* ⟵ best_subset ;
17 |   end
18 end
```

Anexo C. Algoritmo para construir la red HSP

Algoritmo 3: Parallel construction of the HSP network

input : A descriptor matrix $\mathcal{D} = [x_{ij}]_{n \times m}$, and distance function d
output: A weighted graph $G' = (V, E', w)$ with a weight $w : E' \rightarrow [0, 1]$

```
1  V ⟵ {i | i = 1...n};
2  E' ⟵ ∅;
3  maxDist ⟵ 0;
4  for u = 1 to n do in parallel
5      candidates ⟵ Array[0...n−1] ; /* Candidate neighbors to node u */
6      cursor ⟵ 0;
7      foreach v ∈ V do
8          if v ≠ u then
9              dist ⟵ d(u, v) ;  /* Calculating distance d in the space D */
10             candidates[cursor] ⟵ Candidate(v, dist) ;
11             cursor ⟵ cursor + 1 ;
12         end
13     end
14     Sort candidates by ascending order of their distances to node u;
15     largeDist ⟵ candidates[n − 1].distance;
16     cursor ⟵ 0;
17     while cursor < n − 1 do
18         if candidates[cursor] ≠ null then
19             v ⟵ candidates[cursor].node ;  /* Nearest candidate node to u
                */
20             dist ⟵ candidates[cursor].distance;
21             writeLock() ;  /* Synchronizing access to a shared resource
                E' */
22             E' ⟵ E' ∪ {(u, v)} ;
23             writeUnlock();
               /* Ignoring candidates in the forbidden area for node u */
24             for k = cursor + 1 to n − 1 do
25                 if candidates[k] ≠ null then
26                     dist ⟵ d(v, candidates[k].node ) ;
27                     if dist < candidates[k].distance then
28                         candidates[k] ⟵ null ;      /* Ignore candidate nodes
                            closer to v than to u */
29                     end
30                 end
31             end
32         end
33         cursor ⟵ cursor + 1 ;
34     end
35     writeLock() ;       /* Synchronizing access to a shared resource
        maxDist */
36     if largeDist > maxDist then
37         maxDist ⟵ largeDist;
38     end
39     writeUnlock();
40 end
41 foreach (u, v) ∈ E' do
42     w(u, v) ⟵ sim(u, v) ⟵ 1 − d(u,v)/maxDist ; /* The similarity defined in Eq.
        8 */
43 end
```