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

Maestría en Ciencias
en Ciencias de la computación

Aprendizaje de máquina para la identificación de péptidos inductores de autofagia.

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestro en Ciencias

Presenta:

Karen Guerrero Vázquez

Ensenada, Baja California, México

2019

Tesis defendida por

Karen Guerrero Vázquez

y aprobada por el siguiente Comité

Dr. Carlos Alberto Brizuela Rodriguez

Codirector de tesis

Dr. Gabriel del Rio Guerra

Codirector de tesis

Dra. Carolina Álvarez Delgado

Dr. Hugo Homero Hidalgo Silva

Dr. Aldo Moreno Ulloa

Dr. Ubaldo Ruiz López

Coordinador del Posgrado en Ciencias de la computación

Dra. Rufina Hernández Martínez

Directora de Estudios de Posgrado

Resumen de la tesis que presenta Karen Guerrero Vázquez como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la computación .

Aprendizaje de máquina para la identificación de péptidos inductores de autofagia.

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodriguez
Codirector de tesis

Dr. Gabriel del Rio Guerra
Codirector de tesis

La autofagia es un proceso celular involucrado en diversas patologías como el cáncer, enfermedades neurodegenerativas, infecciosas, autoinmunes, entre otras. Sin embargo, se conoce poco sobre los mecanismos de la autofagia. Para estudiar la autofagia con fines terapéuticos, es necesario inducir de forma específica este mecanismo sin alterar el sistema del organismo estudiado mediante moléculas pequeñas como lo son los péptidos. El uso de péptidos como inductores de autofagia permite especificidad en la ruta metabólica objetivo . Técnicas de aprendizaje de máquina han sido utilizadas para predecir secuencias de péptidos antimicrobianos (AMPs) y penetradores de células (CPPs), convirtiéndolas en una opción atractiva para la identificación de nuevos péptidos inductores de autofagia (AIPs). Sin embargo, la cantidad de AIPs conocidos no es suficiente para entrenar un modelo de clasifiación. Afortunadamente, existe una alternativa que se basa en la hipótesis de que los AMPs que son CPPs a la vez, son AIPs. Por tanto, prediciendo un péptido como AMP y posteriormente como CPP, se tendría un posible AIP. Aunque tanto la predicción de AMPs como de CPPs ha sido abordada desde hace más de una década, únicamente los modelos que abordan a los AMPs han alcanzado especificidades del 90 %, mientras que los de CPPs necesitan de una mejora mayor para obtener resultados confiables para la predicción de AIPs. En este trabajo se determinó que una deficiencia importante de los predictores de CPP de la literatura tiene que ver con los conjuntos de datos empleados, particularmente con los datos de casos negativos. Además, se encontraron un total de 175 potenciales AIPs provenientes de proteomas de mamíferos. Finalmente, a pesar de que los los experimentos *in vitro* permitieron observar diferencias significativas en algunos AMCPPs respecto controles de autofagia basal e inducida, es necesario realizar más experimentos que aporten evidencia a lo encontrado en este trabajo.

Palabras clave: Evaluación de predictores, Conjuntos de datos, Autofagia, Péptido antimicrobiano, Péptido penetrador de células

Abstract of the thesis presented by Karen Guerrero Vázquez as a partial requirement to obtain the Master of Science degree in Computer science .

Machine learning for autophagy-inductors peptides identification.

Abstract approved by:

Dr. Carlos Alberto Brizuela Rodriguez
Thesis Co-Director

Dr. Gabriel del Rio Guerra
Thesis Co-Director

Autophagy is an essential cellular process for eukaryotic life. it is involved in various pathologies such as cancer, neurodegenerative, infectious, autoimmune, chronic degenerative and inflammatory diseases. However, little is known about the autophagy mechanisms, making clear the need to study this cellular process. In order to apply autophagy for therapeutic purposes it is necessary to specifically induce this mechanism. The use of peptides to induce autophagy allows a high specificity in the target metabolic pathway with minor influence on other processes. Previously, machine learning techniques have been used to predict sequences of antimicrobial peptides (AMP) and cell penetrating (CPP), making them an attractive option for the identification of new autophagy inductor peptides (AIP). However, the number of known AIPs is insufficient to establish a learning model that identifies them. Fortunately, there is an alternative based on the hypothesis that AMPs that are CPP at the same time, are AIP. This means that predicting a peptide as AMP and subsequently as CPP, then this peptibe could be a candidate for an AIP. Even though the prediction of AMPs and CPPs have been addressed for more than a decade and models that deal with AMPs have achieved specificities of 90 %, CPP models have focused only on sensitivity. To address the related problems and to search for new AIPs, three approaches were followed during this work; a) a comparison between state-of-the-art CPP predictors; b) a classification of peptides by their activities as AMPs and CPPs in mammalian proteomes of high, medium and low longevity from an *in silico* proteolysis and a subsequent homology search with autophagy-related proteins and c) *in vitro* experiments over AMP that are also CPP (AMCPPs) to identify their activity as AIPs. This work identified as an important deficiency of the literature's CPPs predictors their datasets, particularly their negative datasets. In addition, a total of 175 potential AIPs were found in the analyzed mammals´ proteomes. Finally, although significant differences were observed in some AMCPPs regarding baseline and induced autophagy controls, more experiments are needed to support the results found here.

Keywords: Predictor-evaluator, datasets, autophagy, antimicrobial peptide, cell-penetrating peptide

Dedicatoria

A todo aquel que con un simple acto de bondad hace de este un lugar mejor, en especial a mi madre y hermanos que hicieron con amor y guía me permiten seguir luchando para aportar mi parte.

Agradecimientos

No existe forma en la que pueda agradecer suficiente a las personas que el día de hoy permitieron que estas palabras se plasmen en papel. Pero no puedo irme sin reconocer a aquellos que cargaron con papeles importantes tanto en este proyecto como en mi trayectoria.

Este trabajo no pudo haber sido sin mis directores de tesis el Dr. Carlos Alberto Brizuela Rodriguez y al Dr. Gabriel del Rio Guerra que con su apoyo y paciencia constante me permitieron mejorar la calidad de este trabajo, trabajar por primera vez en laboratorio experimental y ver más allá de mi propio proyecto. Así mismo, agradezco a CICESE por aceptarme y brindarme desde los estudios hasta los apoyos para seguir con los mismos, a UNAM por la estancia de dos meses que y a CONACYT por el apoyo económico que por un periodo de dos años permitió mantenerme para poder continuar con mis estudios.

Y hoy no estaría aquí sin mi mamá que siempre me apoyó, mi hermano que me introdujo a las ciencias y a mi hermana quien me dio soporte. A mi mentor de licenciatura el Mae. Enrrique Luna Taylor por quien conocí CICESE y finalmente agradecer a mis amigas Aly y Aurora que me hicieron reconocer una nueva faceta de lo que es amistad.

Tabla de contenido

Página

Tabla de contenido (continuación)

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Figura Página

Lista de tablas

Lista de tablas (continuación)

Lista de tablas (continuación)

Capítulo 1. Introducción

1.1. Antecedentes

La autofagia es un mecanismo de degradación de componentes celulares (organelos, proteínas agregadas; proteínas defectuosas; y patógenos intracelulares) necesario para mantener la integridad estructural y funcional de las células. Este mecanismo es escencial para mantener la homeostasis celular, en la eliminación de patógenos, así como en la regulación de la muerte celular (Jeong *et al.*, 2013; Madeo *et al.*, 2015). Así mismo, la autofagia influye en una vasta cantidad de enfermedades y procesos naturales como son el cáncer, el envejecimiento, enfermedades neurodegenerativas como Alzheimer, entre muchas otras (Nakamura y Yoshimori, 2018; Hernandez *et al.*, 2012; Cai y Yan, 2013; Levine y Kroemer, 2008; Maciel-Herrerías y Cabrera-Beníez, 2016). Sin embargo, su manipulación puede convertirse en arma de doble filo, ya que, por ejemplo, mientras que algunas veces la patogénesis de enfermedades pulmonares está relacionada con una activación excesiva de la autofagia, en otros casos está asociada con una disminución en este proceso (Maciel-Herrerías y Cabrera-Beníez, 2016).

Dado el impacto de la autofagia, su estudio resulta relevante para médicos e investigadores. Con el fin de estudiar este mecanismo se han invertido esfuerzos en el análisis de los genes involucrados (Tsukada y Ohsumi, 1993; Ahmed *et al.*, 2018), estudios *in silico* de su dinámica (HAN *et al.*, 2014) y bases de datos de proteínas y genes involucrados (Homma *et al.*, 2011).

Otro aspecto relevante es la inducción de autofagia, necesaria para el correcto estudio del fenómeno con fines terapéuticos (Boland *et al.*, 2008). Con el objetivo de encontrar inductores de autofagia, se ha abordado de forma experimental la búsqueda de péptidos inductores de autofagia (AIPs) a partir de proteínas involucradas en este proceso (Shoji-Kawata *et al.*, 2013; Garcia-Valtanen *et al.*, 2014; Dowaidar *et al.*, 2017).

Sin embargo, hasta ahora no se ha abordado la caracterización de AIPs, es decir, las propiedades que distinguen a un AIP de cualquier otro péptido. Estudios que abordan dicha caracterización incluyen, pero no se limitan a el análisis computacional. El

uso de herramientas computacionales para la predicción de actividad en péptidos ha sido abordada múltiples veces a lo largo de los años, sean de péptidos penetradores de células (CPPs) (Tang *et al.*, 2016; Manavalan *et al.*, 2018), antimicrobianos (AMPs) (Beltran *et al.*, 2017; Meher *et al.*, 2017) o anticancerígenos (Manavalan *et al.*, 2017; Wei *et al.*, 2018), por lo que resulta lógico pensar que la predicción in silico de AIPs será de utilidad para el descubrimiento de nuevos AIPs.

Desafortunadamente, la cantidad conocida de estos péptidos es pequeña (3 comerciales (Shoji-Kawata *et al.*, 2013), uno publicado (Dowaidar *et al.*, 2017) y uno identificado (Rodriguez Plaza *et al.*, 2014)) para permitir el entrenamiento de una máquina de aprendizaje como los que se realizan en, por ejemplo, el caso de los AMPs. Gracias a Rodriguez Plaza *et al.* (2014) se ha demostrado una relación entre los AMPs y los CPPs y se ha propuesto la hipótesis de que: "Péptidos que son AMPs y CPPs a la vez, son AIPs", entre otras cosas debido a que se ha encontrado que los CPPs pueden activar autofagia en las células penetradas (Dowaidar *et al.*, 2017). Identificar péptidos *in siico* que sean AMP y CPP ayudaría entonces a proponer nuevos AIPs.

Los AMPs han sido ampliamente estudiados y se tienen predictores con certezas que superan el 90 por ciento y coeficientes de correlación de Mathews (MCC) de más de 0.8 (Beltran *et al.*, 2017; Thomas *et al.*, 2010). Para los CPPs, la historia ha sido diferente; entrenamientos con conjuntos desbalanceados, poco esfuerzo para localizar los mejores descriptores para evaluar y conjuntos negativos de baja calidad son parte de las debilidades de los predictores de CPPs.

Aunque se pueden emplear, con mucha confianza los predictores de AMPs de la literatura, no se puede decir lo mismo sobre los predictores de CPPs, por tanto, seleccionar el predictor a utilizar es crucial. Para ello, surge la necesidad de comparar los predictores de CPP antes de localizar los posibles AIPs. Anteriormente se han realizado trabajos de comparación entre los predictores de CPPs disponibles en línea (Manavalan *et al.*, 2018), sin embargo, estos análisis no contemplan la calidad de los conjuntos de datos, los cuales son parte esencial del aprendizaje supervisado (Maglogiannis, 2007), paradigma en el cual caen todos los predictores de CPP hasta ahora reportados. Este trabajo de investigación busca llenar el vacío existente en la literatura sobre la dependencia de la calidad de predicción de los diferentes conjuntos de datos utilizados durante el entrenamiento.

1.2. Objetivos

En este trabajo se propone la hipótesis de que péptidos con actividades penetrantes y antimicrobianas serán inductores de autofagia. Bajo esta premisa se preparó un análisis in silico para probar dicha hipótesis y surgieron los siguientes objetivos.

1.2.1. Objetivo general

Generar encadenamiento de paquetes y algoritmos para la comparación de predictores de CPPs y para la identificación de nuevos AIPs provenientes de proteomas de mamíferos.

1.2.2. Objetivos específicos

- Caracterización de conjuntos de datos de la literatura para el entrenamiento de predictores de péptidos penetradores de células (CPPs).
- Identificación de los mejores tres modelos de aprendizaje de máquina para la predicción de CPPs.
- Identificación de candidatos para péptidos inductores de autofagia (AIPs) bajo la hipótesis de que un AIP es CPP y péptido antimicrobiano (AMP) a la vez.
- Identificación del impacto de propiedades AMP y CPP en la identificación in silico de presencia en la autofagia en proteomas de mamíferos
- Reconocer las diferencias morfológicas entre la autofagia basal y la macroautofagia inducida en un modelo experimental de células WT BY472 PKG1-GFP + VMA1-TDimer2.
- Predecir AIPs en proteomas de mamíferos.
- Generar una lista de péptidos candidatos a ser AIPs

1.2.3. Organización de la tesis

Este trabajo de tesis está organizado en seis capítulos de contenido más tres de anexos.

El Capítulo 1 le brindó los primeros conceptos al resto de este trabajo, la motivación del mismo y los retos que se enfrentaron.

En el Capítulo 2 se abordan los conceptos tanto biológicos como computacionales que serán necesarios para este trabajo. Se presenta la definición de autofagia, sus tipos, los mecanismos y cuatro proteínas esenciales para el análisis *in silico*. Así mismo, se presentan los conceptos de aprendizaje de máquina con énfasis en la clasificación y predicción junto con sus buenas y malas prácticas.

En el Capítulo 3 se presenta la metodología de las tres estrategias utilizadas para el cumplimiento de los objetivos. Estas estrategias incluyen el análisis comparativo de predictores de péptidos penetradores de células; una búsqueda de posibles péptidos inductores de autofagia (AIPs) en proteomas de mamíferos con diferentes coeficientes de longevidad; y la comparación de posibles AIPs entre especies y funciones de los péptidos y dos experimentos *in vitro* con péptidos con actividades antimicrobiana (AMPs) y penetradora de células (CPPs).

En el Capítulo 4 pueden encontrarse los resultados obtenidos al aplicar la metodología presentada en el Capítulo 3.

El Capítulo 5 expone la discusión de los resultados abordando tanto la comparación de los predictores, como la predicción *in silico* y la experimentación *in vitro*. Así mismo, durante este capítulo se busca conectar los tres caminos para que el lector logre apreciar la relación entre ellos.

El Capítulo 6 presenta de forma concreta las conclusiones a las que se llegó durante este trabajo, separadas por las estrategias utilizadas. Además, en este capítulo pueden encontrarse las propuestas de trabajo futuro.

En los anexos puede encontrarse material complementario, donde en el primer anexo observamos la cantidad de péptidos resultado de una proteólisis *in silico* y una

predicción como AMP.

En el Anexo 2 se encuentran las contabilizaciones de cada una de las imágenes tomadas por NanoImager especificando su tratamiento, cantidad de células en morfología de aro y morfología de palomita.

Por último, en el Anexo 3 se presenta el manual de la herramienta para la predicción de CPPs mediante predictores de la literatura, EvalCPPP.

Capítulo 2. Marco Teórico

2.1. Aspectos biológicos

2.1.1. Péptidos

Los péptidos son polímeros lineales constituidos por aminoácidos como lo son las proteínas, pero de longitudes más cortas (McKee *et al.*, 2003). Sin embargo, de los más de 300 aminoácidos que existen en la naturaleza, sólo 20 constituyen a los péptidos sintetizados por los organismos vivos (Murray *et al.*, 2007). Estos 20 aminoácidos son denominados aminoácidos estándar o naturales (McKee *et al.*, 2003).

Los aminoácidos estándar contienen un átomo de carbono central, el carbono α, al que están unidos un grupo amino, un grupo carboxilo, un átomo de hidrógeno y un radical R o cadena lateral que los distingue entre sí (McKee *et al.*, 2003; Peña, 1988).

Al reaccionar el grupo amino de un aminoácido con el grupo carboxilo de otro (Peña, 1988) se produce un enlace covalente conocido como enlace peptídico (Alberts *et al.*, 2015), la formación del enlace amida a partir de una amina y un ácido, implica la eliminación de agua, podemos ver la reacción como (Griffin, 1981):

$$-NH_2 + HOOC- \rightarrow -NH-CO- + H_2O$$

Los aminoácidos forman cadenas polipeptídicas que se mantienen unidas principalmente por el enlace peptídico y el resultado, es decir, el péptido (o proteína), posee propiedades que reflejan en gran medida las propiedades de sus constituyentes (Peña, 1988).

Considerando una longitud n, podríamos tener 20^n secuencias diferentes. Los péptidos se componen de hasta 50 (McKee *et al.*, 2003), o 100 (Jenssen, 2011) aminoácidos, obteniendo hasta 20^{100} posibles péptidos. Sin embargo, no todas las secuencias resultan en péptidos con actividad biológica. Empero, de todas las secuencias posibles,

al menos 650 000 y hasta dos millones, según McKee *et al.* (2003), y un billón según Huang *et al.* (2016), son producidas realmente por los seres vivos.

Las propiedades de los péptidos han evolucionado durante millones de años, definiendo características estructurales, presencia de sitios de unión, balance entre flexibilidad y rigidez en sus estructuras así como la vulnerabilidad a las reacciones de degradación (McKee *et al.*, 2003).

2.1.1.1. Péptidos penetradores de células

Los péptidos penetradores de células (CPPs) se definen como péptidos con un máximo de 30 aminoácidos, que son capaces de entrar en las células de una manera pasiva, por lo que pueden translocarse a través de las membranas (Lundberg y Lo Langel, 2003).

Los CPP son considerados herramientas potenciales para ayudar al transporte de moléculas terapéuticas que no están inherentemente biodisponibles y, además, como posibles agentes bioactivos por sí mismos (Holton *et al.*, 2013).

Dependiendo del autor y el origen del péptido, se han propuesto diferentes nombres para estos en la literatura. La secuencia de translocación de membrana (MTS) es otro nombre utilizado frecuentemente para describir la capacidad de estos péptidos para atravesar las membranas plasmáticas (Lundberg y Lo Langel, 2003).

2.1.1.2. Péptidos antimicrobianos

Los péptidos antimicrobianos (AMPs) son un grupo de péptidos con la capacidad de inhibir los efectos patógenos de los microorganismos. Son producidos por la mayoría de los organismos vivos como parte de sus estrategias de defensa. En el 2011 se habían aislado y descrito en detalle 1600 de estos péptidos (Godballe *et al.*, 2011). Actualmente, se encuentran reportados más de 40 mil AMPs en bases de datos públicas (Aguilera-Mendoza *et al.*, 2019).

En términos generales, los AMPs son péptidos de bajo peso molecular (usualmente

menores de 10kDa) codificados en el genoma, a diferencia de otros antimicrobianos de naturaleza no peptídica, como la penicilina (Montaño-Pérez, 2002).

Los AMPs fueron reportados por primera vez a mediados de la década de los ochentas a través de dos líneas de investigación: la primera dirigida a la definición de los mecanismos de la hemolinfa (sangre) de los insectos inmunizados para combatir el crecimiento microbiano; mientras que la segunda estudiaba los mecanismos de las células fagocíticas de mamíferos para matar bacterias después de haberlas ingerido (Boman, 1995).

Actualmente se han caracterizado in sílico descriptores moleculares que logran, con modelos relativamente sencillos como KNN, alcanzar una precisión de más del 97 por ciento en la detección de AMPs (Beltran *et al.*, 2017, 2018) y existen varias bases de datos especializadas en AMPS (Waghu *et al.*, 2014; Wang y Wang, 2004; Pirtskhalava *et al.*, 2016), por tanto, este trabajo no realizará un estudio sobre la predicción de AMPs.

2.1.2. Autofagia

La autofagia es un proceso celular esencial para mantener la homeostasis celular y tisular al contribuir en la generación de energía a través de eventos de degradación; en el control de calidad de las proteínas y organelos; en la eliminación de proteínas de larga vida y patógenos; así como la regulación de la muerte celular (Maciel-Herrerías y Cabrera-Beníez, 2016) la cual está altamente conservada entre organismos modelo del dominio eucariotas (Madeo *et al.*, 2015). Alteraciones en la actividad autofágica se han asociado a diversas patologías como el cáncer, enfermedades neurodegenerativas, infecciosas, autoinmunes, crónico-degenerativas e inflamatorias (Maciel-Herrerías y Cabrera-Beníez, 2016).

La autofagia se encarga de la degradación intracelular que se caracteriza por la formación de vesículas de doble membrana denominadas autofagosomas. Estas vesículas secuestran material citoplasmático y después se fusionan con el lisosoma, formando el autolisosoma, en donde ocurre la degradación del material invaginado.

Los aminoácidos y moléculas pequeñas que son generadas por autofagia, son devueltos al citoplasma para la generación de energía y para la síntesis de nuevas proteínas y biomoléculas. Uno de los primeros estímulos reconocidos de autofagia es el déficit de nutrientes, sin embargo, se ha observado que la activación de la autofagia es un mecanismo de supervivencia celular frente a diferentes condiciones de estrés, incluyendo el estrés oxidativo, la inflamación, la agregación de proteínas, etc. (Maciel-Herrerías y Cabrera-Beníez, 2016).

Existen al menos 4 tipos de autofagia reportados; 1) la macroautofagia donde existe secuestro y degradación de componentes celulares citoplasmáticos por autofagosomas de doble membrana; 2) la microautofagia que implica la captación directa de componentes citoplasmáticos mediante invaginación de la membrana limitante del lisosoma (o vacuola en caso de levadura); 3) la autofagia mediada por chaperonas que se basa en degradación lisosomal de proteínas desplegadas y; 4) la microautofagia nuclear, que es un tipo selectivo de autofagia que ocurre en las levaduras y que sirve para degradar partes del núcleo (Xie *et al.*, 2008).

La macroautofagia es activada por el estrés y procede a través de cinco fases (Maciel-Herrerías y Cabrera-Beníez, 2016):

- Nucleación o formación de una estructura de doble membrana (fagóforo)
- Expansión de la membrana del fagóforo por la incorporación de la proteína LC3-II
- Maduración de esta estructura en el autofagosoma y el secuestro de material citoplasmático a degradar
- Fusión del autofagosoma con el lisosoma que resulta en la formación de los autofagolisosomas o autolisosomas
- Degradación del material biológico secuestrado por las enzimas hidrolíticas del lisosoma y reciclaje de moléculas.

Al final, el autofagosoma se fusiona con el lisosoma y el material secuestrado es degradado por las enzimas lisosomales (Maciel-Herrerías y Cabrera-Beníez, 2016).

Los diferentes pasos en la macroautofagia están mediados por más de 30 proteínas genéricamente conocidas como proteínas ATG en levadura (Cuervo *et al.*, 2005). Son

principalmente tres grupos; (1) Atg9, el complejo quinasa Atg1 (Atg1 y Atg13), Atg2 y Atg18 que permiten la formación del autofagosoma; (2) el complejo de fosfatidilinositol 3-OH quinasa (PI(3)K) que se encarga de la fusión del autofagosoma con el lisosoma para la isolación de la elongación de la membrana; (3) el sistema de proteínas de tipo ubiquitina (Ubl), que incluye dos proteínas Ubl (Atg8 y Atg12), una enzima activadora (Atg7), dos análogos de enzimas conjugadoras de ubiquitina (Atg10 y Atg3), una proteasa modificadora de Atg8 (Atg4), la proteína objetivo de la unión Atg12 (Atg5) y Atg16 (Xie *et al.*, 2008).

La microautofagia fue originalmente definida como una dinámica de la membrana lisosomal en la que los componentes del citosol son directamente secuestrados por la membrana lisosomal que se deforma y crea las invaginaciones que contiene el material del citosol. La confusión persistente en los mecanismos de los tipos de autofagia, la falta de marcadores específicos para este proceso, hace difícil evaluar los cambios referentes a la microautofagia (Cuervo *et al.*, 2005; Galluzzi *et al.*, 2017; Oku y Sakai, 2018).

Tanto la macro como la microautofagia están conservadas desde las levaduras hasta los mamíferos, mientras que las mediadas por chaperonas solo se ha descrito en mamíferos (Cuervo *et al.*, 2005).

La autofagia mediada por chaperonas (CMA) es una forma inducible de autofagia, preferentemente activada por diferentes estresantes como estrés nutricional, exposición a toxinas o estrés oxidativo. Las proteínas son dirigidas de manera selectiva a los lisosomas después de haber interactuado con una chaperona citosólica (Cuervo *et al.*, 2005). Solo las proteínas citosólicas y no los organelos pueden ser degradados por la CMA, esta es la única ruta autofágica por la cual las proteínas citosólicas pueden ser degradadas selectivamente por los lisosomas (Cuervo *et al.*, 2005).

Tsukada y Ohsumi (1993) mencionan varios motivos por los cuales estudiar los procesos de autofagia en levaduras, esos incluyen la facilidad de modificaciones genéticas o la homogeneidad de un cultivo para análisis. La autofagia es un proceso conservado en las células eucariontes, sin embargo, existen algunas diferencias entre la autofagia en células de levadura y células de mamífero.

En las células animales la degradación se lleva a cabo en el autolisosoma, la fusión

del autofagosoma con el lisosoma, degradando entre un tres y cuatro por ciento de las proteínas de la célula por hora de manera no selectiva, mientras que en las levaduras, las vacuolas son las responsables de esta actividad conteniendo la mayoría de las proteasas de la célula (Takeshige, 1992). Otro punto importante es que el proceso mediante el cual sobrepasan los niveles basales de autofagia varían, mientras que en levaduras la privación de nutrientes es el principal estímulo de autofagia, otras células eucariotas mantienen rutas más complicadas (Reggiori y Klionsky, 2013).

2.1.2.1. Planteamiento de un método para identificar inductores de macro y microautofagia

En el grupo del Dr. Del Río se ha desarrollado un grupo de cepas de la levadura *Saccharomyces cerevisiae* que facilita la identificación de inductores de autofagia. Estas cepas se construyen de la fusión de PGK1 y VMA1 a las proteínas fluorescentes verde (GFP) y roja (Tdimer2), respectivamente. Para distinguir la inducción de la inhibición del flujo de autofagia, estas cepas se combinaron con mutantes nulas de los genes ATG8 y PEP4. A continuación se hace una breve descripción de la función descrita de estos 4 genes para explicar la base del método.

ATG8. ATG8 es una de dos proteínas de tipo ubiquitina requerida para el proceso de formación de los autofagosomas durante la macroautofagia. ATG8 media el anclaje y la hemifusión de las membranas, que son producidas por la lipidación de la proteína y se modulan de manera reversible por la enzima de desconjugación ATG4 (Nakatogawa *et al.*, 2007). ATG8 es conjugada al lípido fosfatidiletanolamina (PE), ahí ATG8 es anclado a las membranas. Se ha encontrado que además, ATG8 está primordialmente localizado en la membrana más que en los autogafosomas, sugiriendo un papel importante en la formación de los mismos (Nakatogawa *et al.*, 2007).

La formación de complejos ATG12-ATG5 y ATG8-PE es esencial para la formación de los autofagosomas; una mutación en cualquiera de estas proteínas en un autofagosoma defectuoso (Geng y Klionsky, 2008).

La evidencia indica una relación cuantitativa entre la cantidad de ATG8 y el tamaño de la vesícula de la célula en cuestión. En levadura, la inducción de autofagia resulta en

la activación de ATG8 y la formación de autofagosomas más largos (Geng y Klionsky, 2008).

PEP4. PEP4 es el gen que codifica a una de las siete proteasas vacuolares de la *S. cerevisiae*. En el modelo de levadura *S. cerevisiae*, las proteasas vacuolares son conocidas por la degradación masiva de proteínas en exceso y senescentes de larga vida, para preservar la homeostasis de las proteínas (Kerstens y Van Dijck, 2018) y se ha utilizado como marcador del lumen vacuolar (Eastwood *et al.*, 2012).

PEP4 mantiene un papel en la activación y la maduración proteolíticas de la mayoría de las proteasas vacuolares. Se ha reportado que la permeabilidad de la membrana vacuolar aumenta en condiciones de estrés. Esto podría permitir que las proteasas escapen de la vacuola y se ubiquen en otros sitios de la célula. Además, PEP4 está asociado con la degradación mitocondrial en condiciones de apoptosis inducida por ácido acético. Se ha descrito que las nucleoporinas se degradan por PEP4 en la muerte celular inducida por H_2O_2, lo que sugiere una localización asociada al núcleo. Esta idea se ve reforzada por la importancia de PEP4 en la generación del complejo similar a SAGA (SLIK) por escisión del componente Spt7 del complejo SAGA (Kerstens y Van Dijck, 2018). Estas actividades pueden verse en la Figura 1.

PEP4 participa en un amplio marco de mecanismos de respuesta al estrés, independiente de los procesos autofágicos y afecta a los procesos reguladores, como la expresión de genes y la muerte celular programada.

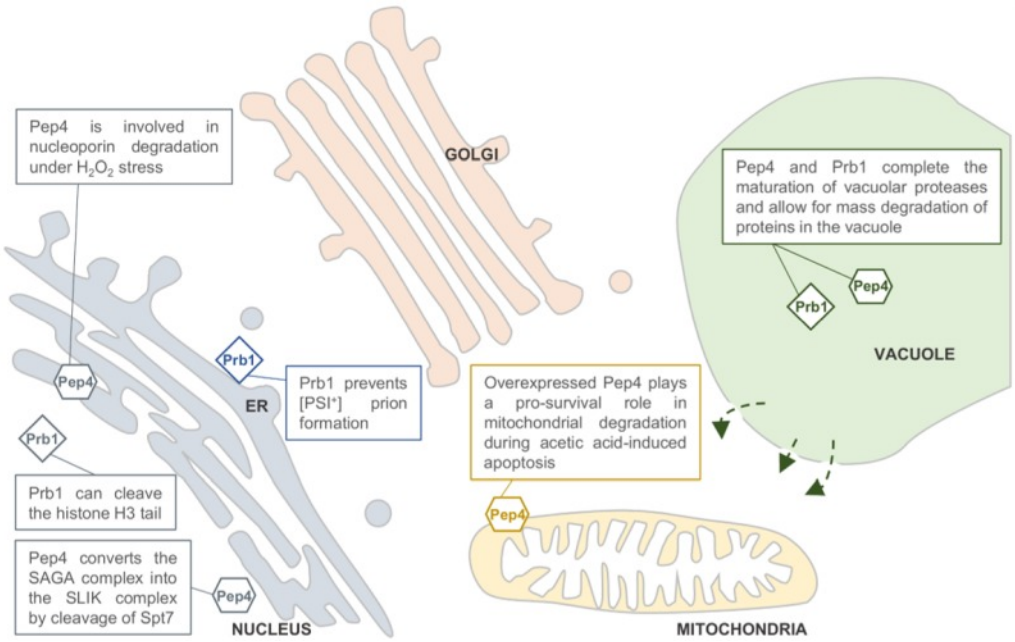

Figura 1. Diferentes mecanismos en los que participa la proteína PEP4. Imagen obtenida de Kerstens y Van Dijck (2018)

PGK1. Fosfoglicerato quinasa 1 (PGK1) es una proteína que cataliza una de las dos reacciones productoras de ATP en la ruta glicolítica (Akman *et al.*, 2011).

PGK1 está aleatoriamente distribuida en el citosol. Se expresa independientemente de la mayoría de las condiciones de cultivo y no se induce significativamente por inanición. Como la macro y micro autofagia degradan parte del material citosólico u organelos completos, también transporta PGK1 para su degradación por lo cual esta puede ser detectada mediante microscopía de fluorescencia (Welter *et al.*, 2010).

VMA1. VMA1 (ATPasas de la membrana vacuolar 1) codifica a la subunidad catalítica de la H^+ ATPasa vacuolar (V-ATPasa). Es esencial para la hidrólisis de ATP en la membrana vacuolar (Ferea y Bowman, 1996), por ello se ha empleado como marcador de la membrana vacuolar así como para detectar autofagia en las células de *S. cerevisiae* (Eastwood *et al.*, 2012).

La V-ATPasa es una enzima clave durante la homeostasis. La interacción entre la V-ATPasa y la membrana plasmática determinan la acidificación vacuolar y el pH citosólico, el cual es esencial para funciones mitocondriales y de degradación como lo es la autofagia (Aufschnaiter y Büttner, 2019). En la Figura 2 puede observarse la

composición de esta enzima.

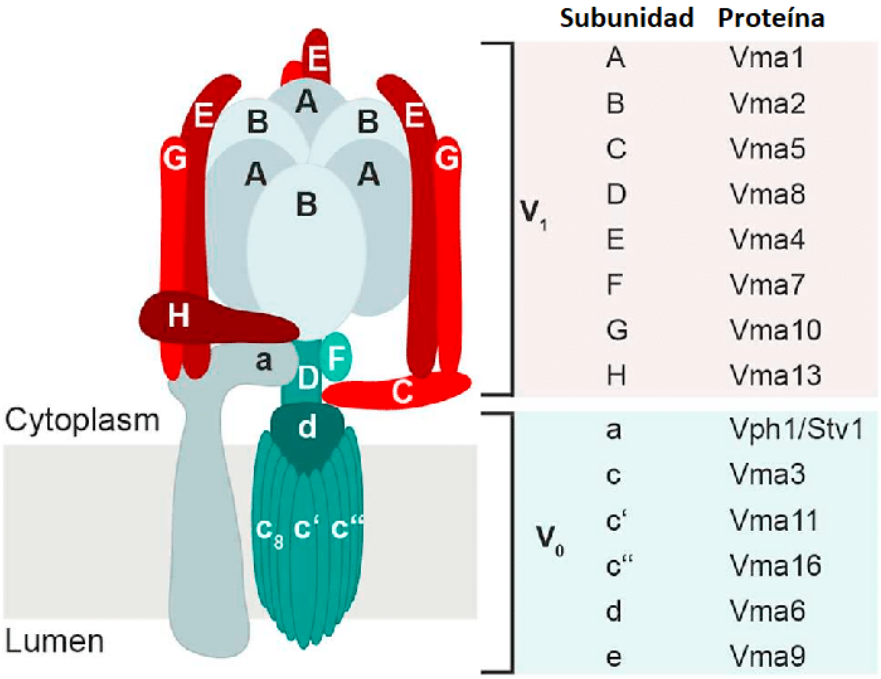

Figura 2. Proteínas que componen la enzima V-ATPasa de *S. cerevisiae*. En el dominio V_1 puede encontrarse la proteína VMA1. Imagen obtenida de Aufschnaiter y Büttner (2019).

Existen al menos 17 genes de las ATPasas de la membrana vacuolar necesarios para la formación de los complejos V-ATPasa en levadura (200, 2000). La membrana vacuolar está compuesta por una bicapa de fosfolípidos y colesterol que representa entre el 20 y 30 por ciento del volumen de la célula. Su contenido es ácido con un pH de 5.5 y contiene proteasas nucleasas, glicosidasas y fosfatasas. Las ATPasas forman parte de la membrana vacuolar de las *S. cerevisiae*. La interrupción de VMA1 genera una pobre viabilidad y lento crecimiento en un medio escaso de nutrientes (Yasuhiro *et al.*, 1989).

2.2. Aspectos computacionales

2.2.1. Aprendizaje de máquina

El campo de estudio enfocado en el desarrollo de algoritmos para transformar la información en acciones inteligentes es conocido como aprendizaje de máquina (Lantz,

2015). Se puede pensar en el aprendizaje de máquina como un conjunto de métodos y herramientas que buscan inferir patrones y extraerlos de una observación hecha en el mundo físico (Gutierrez, 2015).

El término aprendizaje es el mecanismo a través del cual se adquieren destrezas y conocimientos (Lantz, 2015). Un sistema de aprendizaje requiere reglas de comportamiento, descripciones de objetos físicos y heurísticas de resolución de problemas (Michael y Lin, 1973). El proceso de aprendizaje se compone del almacenamiento de información, la abstracción de la información, la generalización que crea el conocimiento y la evaluación (Lantz, 2015) utilizando estadísticas (Harrington, 2012). Este proceso no está completo hasta que se es capaz de usar el cocimiento abstraído para futuras acciones (Lantz, 2015). Se dice que un programa de computadora aprende de la experiencia E respecto a algún tipo de tarea T y con un desempeño medido P si su desempeño en la tarea T mejora con la experiencia E (Mitchell, 1997).

El término generalización describe el proceso de convertir el conocimiento abstraído a una forma en la que se pueda utilizar en acciones futuras en una tarea similar pero no idéntica a las vistas anteriormente (Lantz, 2015). Durante la generalización, la máquina limita los patrones que descubre a sólo aquellos con mayor relevancia. Por lo general no es factible reducir el número de patrones uno por uno, en su lugar, los algoritmos de aprendizaje de máquina emplean atajos para reducir el espacio de búsqueda, por ejemplo, mediante heurísticas (Lantz, 2015).

Aprender involucra la búsqueda de hipótesis posibles en el espacio para encontrar aquélla que mejor describa los ejemplos disponibles para entrenamiento y cumpla con las restricciones (Mitchell, 1997). Existen innumerables patrones que podrían ser identificados durante el proceso de abstracción y las miles de maneras de modelar estos patrones (Lantz, 2015), recordando que el objetivo del aprendizaje de máquina es encontrar el modelo que generalice mejor (Kelleher *et al.*, 2015).

Las entradas pueden ser cuantitativas si se trata de valores de la recta real (continuos), o cualitativas si se trata de valores finitos que pueden no tener un orden (discretos). Esto permite distinciones entre los tipos de métodos que se usarán en la predicción. Algunos métodos son definidos más naturalmente por entradas cuantitativas, otros por cualitativas y habrá métodos que se definen mejor con ambos métodos

(Friedman *et al.*, 2001). La distinción entre las salidas define el nombre convencional para la tarea de predicción; regresión si es de salidas cuantitativas y clasificación cuando son salidas categóricas. Ambas tienen en común, entre otras cosas, que se pueden ver como actividades en una tarea de aproximación a funciones (Friedman *et al.*, 2001).

Existen muchas ramas y caminos para el aprendizaje de máquina. Una forma de clasificarlo de forma general es como aprendizaje supervisado, no supervisado o el aprendizaje por refuerzo. En el aprendizaje supervisado se provee una etiqueta de categoría o costo para cada patrón en el conjunto de entrenamiento, y se busca reducir la sumatoria de los costos para ese patrón. En el aprendizaje no supervisado, también llamado conglomeración, no existe un maestro que indique etiquetas. El sistema forma conglomerados a partir de las entradas. Por último, en el aprendizaje por refuerzo, no se da señal alguna de la categoría deseada; en cambio, la única información de la enseñanza es que la categoría tentativa sea correcta o incorrecta. En clasificación, lo más común es que el refuerzo sea binario (Duda *et al.*, 1973).

Para construir modelos usados en la predicción de datos, se utiliza el aprendizaje de máquina supervisado (Kelleher *et al.*, 2015). El objetivo del aprendizaje supervisado es: para cada ejemplo, predecir los valores de las salidas según su entrada (Friedman *et al.*, 2001).

Las técnicas de aprendizaje de máquina supervisado, automáticamente aprenden de un modelo de relación entre un conjunto de características descriptivas y una característica objetivo basada en un conjunto histórico de ejemplos o casos. Automatizan este proceso buscando a través de un conjunto de modelos. El modelo seleccionado se usa para realizar predicciones de nuevos casos (Kelleher *et al.*, 2015).

Los modelos no pueden generalizar perfectamente debido al ruido, término que describe variaciones inexplicables en los datos. Los datos ruidosos son causados por problemas aleatorios como errores de medición, problemas en la calidad de los datos o fenómenos complejos que impactan en la información (Lantz, 2015), creando problemas adicionales si se sobreentrena la máquina, puesto que un modelo de predicción sólo es útil si es capaz de realizar predicciones correctas para entradas no vistas con anterioridad. Se dice que un modelo de predicciones “generaliza bien” cuando es ca-

paz de realizar predicciones correctas para entradas no vistas previamente (Kelleher *et al.*, 2015). Para generalizar se necesita adecuar suficientemente el modelo para resolver el conjunto de datos original, pero que mantenga un margen de ajuste suficiente para dar entrada a nuevos datos similares, mas no iguales.

Al proceso de adecuar un modelo a un conjunto de datos se le conoce como entrenamiento (Lantz, 2015). Cuando el modelo ha sido entrenado, la información es transformada a una forma abstracta que resume la información original (Lantz, 2015). Para esto, un algoritmo de aprendizaje utiliza dos fuentes de información para guiar su búsqueda; el conjunto de datos y el sesgo inductivo (Kelleher *et al.*, 2015).

Se dice que el algoritmo tiene sesgo si las conclusiones son sistemáticamente erróneas o incorrectas de una manera predecible (Lantz, 2015). El sesgo es la forma en la que las máquinas eligen un modelo para explicar el conjunto de datos.

En general se tienen dos tipos de errores que pueden crear un sesgo inapropiado. El sobreajuste y la sobregeneralización. La sobregeneralización ocurre cuando el modelo seleccionado es muy simplista para representar las relaciones. El sobreajuste, por el contrario, ocurre cuando el modelo es tan complejo que se vuelve muy sensible al ruido en los datos (Kelleher *et al.*, 2015). Estos errores son causados por un mal uso en los conjuntos de datos de entrenamiento. La causa puede variar desde la cantidad de elementos en cada conjunto hasta la forma en la que se distribuyen los datos (Riley, 2019).

El último paso para el proceso de aprendizaje es la evaluación o la medición del éxito a pesar de sus sesgos y de los datos de entrenamiento (Lantz, 2015). Generalmente, la evaluación ocurre después de que un modelo haya sido entrenado. Entonces, el modelo es evaluado con un conjunto nuevo para juzgar qué tan bien representó los datos (Lantz, 2015).

A grandes rasgos, si la probabilidad de obtener los datos dada alguna hipótesis nula cae por debajo de un umbral de "significancia", se rechaza la hipótesis nula a favor de la alternativa (Duda *et al.*, 1973). La estimación de la exactitud de una hipótesis es relativamente sencilla cuando los datos son abundantes. Sin embargo, cuando debemos aprender una hipótesis y estimar su exactitud futura dado sólo un conjunto limitado de datos, surgen dos dificultades claves: sesgo en la estimación y la variación

en la estimación (Mitchell, 1997). Debido a que la hipótesis aprendida deriva de los ejemplos en el entrenamiento, normalmente proporcionan una estimación optimista sesgada de la precisión de la hipótesis sobre ejemplos futuros. Esto es especialmente probable cuando la máquina considera un espacio de hipótesis muy rico, lo que le permite superar los ejemplos de entrenamiento. Para obtener una estimación imparcial de la precisión futura, típicamente probamos la hipótesis sobre un conjunto de ejemplos de prueba elegidos independientemente de los ejemplos de entrenamiento y la hipótesis (Mitchell, 1997).

2.2.2. Clasificación y predicción

Uno de los problemas clásicos que abordan los algoritmos de aprendizaje supervisado es la clasificación (Gallegos *et al.*, 2014). Se trata de un problema general que puede ser tanto de reconocimiento de patrones como de regresión ordinal (Angulo Bahón, 2001). La tarea de un clasificador es tener un sistema completo y utilizar un vector de características proveído por un extractor de características para asignar una categoría a un objeto (Duda *et al.*, 1973), los cuales son descritos por características también llamadas atributos (Kuncheva, 2004).

El objetivo general es hipotetizar la clase de estos modelos, procesar los datos detectados para eliminar el ruido, y para cualquier patrón detectado elegir el modelo que mejor corresponda, sugiriendo acciones cuando se presentan con nuevos patrones (Duda *et al.*, 1973) y así predecir a qué clase pertenece un caso dado (Harrington, 2012).

Los problemas centrales en el reconocimiento de patrones estadísticos son cuantificar y favorecer clasificadores más simples, determinar automáticamente una curva que separe los datos y lograr predecir qué tan bueno será el sistema para generalizar nuevos patrones (Duda *et al.*, 1973). El problema de clasificación se define como dado un patrón no etiquetado **x**, asignar la etiqueta de clase ω_i apropiada para **x** (Murty y Devi, 2015).

Cada objeto a ser clasificado pertenece a una clase particular w_i, las clases están

etiquetadas de ω_1 a ω_c, organizadas en un conjunto de etiquetas $\Omega = \{\omega_1, ..., \omega_c\}$ (Kuncheva, 2004). Los valores de las características de un objeto dado son arreglados en un vector n-dimensional $\mathbf{x} = [x_1, ..., x_n]^T \in \mathbb{R}^n$. El espacio real $\mathbb{R}^n$ es llamado el espacio de características (Kuncheva, 2004).

Z es el conjunto de entrenamiento y los miembros de **D** son llamados patrones etiquetados porque cada patrón tiene una etiqueta asociada. Si cada patrón $\mathbf{x}_j \in \mathbf{D}$ es n-dimensional, entonces se dice que el conjunto **D** es n-dimensional o equivalente.

Intuitivamente, una clase contiene objetos similares, mientras que objetos de clases diferentes son diferentes (Kuncheva, 2004).

La región de decisión de la clase ω_i es el conjunto de puntos para el cual la i-ésima función de discriminación tiene el valor más alto. De acuerdo con la regla de la membresía máxima, todos los puntos en una región de decisión $\mathcal{R}_i$ son asignados a la clase ω_i (Kuncheva, 2004). Si la región de decisión $\mathcal{R}_i$ contiene puntos de datos del conjunto etiquetado **Z** con las etiquetas verdaderas ω_j, $j \neq i$, las clases ω_i y ω_j están superpuestas. Si las clases en **Z** pueden ser separadas por completo por un hiperplano, se dice que dichas clases son linealmente separables (Kuncheva, 2004).

Para examinar la validez de la clasificación se utilizan métodos como Redistribución, “Hold-out”, “Repeated hold-out” (Datashuffle), validación cruzada, dejar uno fuera, entre otros (Kuncheva, 2004), sin embargo, esta separación debe evitar ser aleatoria y respetar la naturaleza de los datos (Riley, 2019).

Las pruebas deben realizarse con datos no contemplados en el entrenamiento. Todos los datos de entrenamiento deben sintonizar a todos los parámetros. Un error común en los experimentos de clasificación es seleccionar un conjunto de características usando el conjunto de datos original, y luego ejecutar experimentos para evaluar la exactitud de ese conjunto. Este problema, denominado “peeking”, está muy extendido en bioinformática y neurociencias. Al utilizar los mismos datos es probable que conduzca a una validación optimista del error. El utilizar diferentes datos de entrenamiento muestreados sin sesgo de la distribución del problema puede mejorar la calidad del predictor (Kuncheva, 2004).

A menudo, los valores de los atributos y las etiquetas de las clases no tienen un

nivel adecuado de confianza debido a fuentes de información poco fiables, a los malos dispositivos de medición, a los errores tipográficos, a la confusión del usuario y muchas otras razones. Decimos que los datos son afectados por diversos tipos de ruido. El ruido estocástico es aleatorio, por el contrario, el ruido sistemático arrastra todos los valores en la misma dirección (Kubat, 2015).

Las etiquetas de clase sufren problemas similares. Las etiquetas recomendadas por un experto pueden no haber sido registradas correctamente. Alternativamente, algunos ejemplos se encuentran en un “área gris“ entre dos clases, en cuyo caso las etiquetas marcadas no son ciertas. Ambos casos representan ruido estocástico, de los cuales el último puede afectar sólo a los ejemplos extremos (Kubat, 2015).

Muchas hipótesis son específicas del dominio o del problema, y su solución dependerá del conocimiento y de las perspectivas del diseñador. Esto es, extracción de características, características faltantes, mereología, segmentación, agrupación de pruebas, costos y riesgos (Duda *et al.*, 1973). Así mismo, estas pudieran depender entre sí.

Ya que un clasificador perfecto es a menudo imposible, una tarea general es determinar la probabilidad para cada una de las posibles categorías (Duda *et al.*, 1973).

El límite conceptual entre la extracción de características y la clasificación propiamente dicha es arbitrario. Un extractor de características ideal produciría una representación que hace trivial el trabajo del clasificador; por el contrario, un clasificador omnipotente no necesitaría la ayuda de un sofisticado extractor de características. Tanto en selección como en diseño de características se busca encontrar aquellas que sean simples de extraer, invariante a transformaciones irrelevantes, insensible al ruido y útiles para discriminar patrones en diferentes categorías.

La distinción se impone por razones prácticas, más que teóricas. El clasificador se enfrenta a problemas como el ruido, el sobreentrenamiento, la selección del modelo, mismas que el extractor busca el la mejor combinación de características para representar la información (Duda *et al.*, 1973).

Capítulo 3. Metodología

Durante este proyecto fue necesario seguir tres caminos; uno computacional que consistió en un análisis exhaustivo de los predictores de CPPs propuestos en la literatura y sus respectivos conjuntos de entrenamiento y prueba; uno experimental, que consta de la selección de un subconjunto de péptidos y prueba de inducción de autofagia *in vitro*; y uno bioinformático que busca identificar péptidos inductores de autofagia (AIP) en proteomas putativos. Este proceso puede verse simplificado en la Figura 3 y se detalla en las subsecciones siguientes.

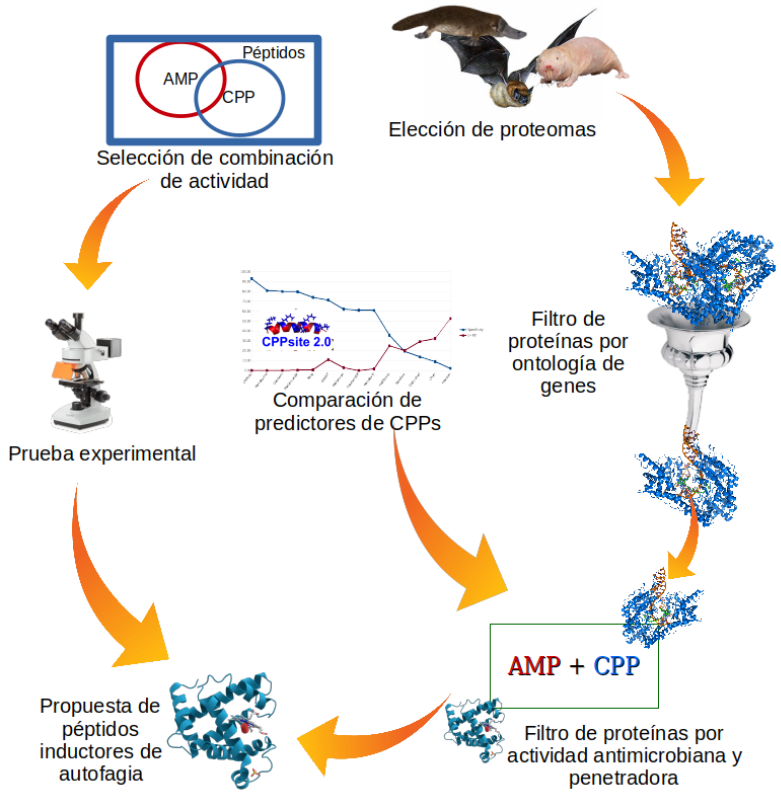

Figura 3. Metodología empleada para la propuesta de péptidos inductores de autofagia.

3.1. Evaluación de predictores de péptidos penetradores de células

Parte esencial de este trabajo es la predicción de una secuencia peptídica basada en la selección de su actividad tanto antimicrobiana como penetrante. Con la finalidad de dar lugar a una mejor evaluación de los posibles AIPs, se buscaron, probaron y

evaluaron diversos predictores disponibles en la literatura referente a predictores de CPPs. Para ello se realizó una comparación entre modelos de aprendizaje y conjuntos de datos para predicción de CPPs. Todos los predictores son importantes y de calidad digna de examinar, sin embargo, solo ocho de ellos poseen un predictor públicamente disponible y en estos se centró el análisis.

3.1.1. Comparación de predictores de CPPs

Como primer paso se realizó la evaluación entre los predictores ya existentes con la finalidad de elegir el algoritmo que presentara la mayor especificidad para discriminar al elegir péptidos como CPPs, o de forma somera, escoger el o los predictores con mejor especificidad.

La comparación se realizó tanto en los predictores, como en sus conjuntos de datos, por lo que se recopilaron de la literatura los predictores y sus conjuntos de entrenamiento y prueba según su disponibilidad. Se replicaron los trabajos de cada uno de los predictores accesibles en línea, comparando sus resultados con los obtenidos en este trabajo. Posteriormente se clasificaron todos los conjuntos de datos recopilados usando estos predictores. Se analizó el desempeño de cada uno de los algoritmos respecto a las medidas de sensibilidad (Sen), especificidad (Spc) y coeficiente de correlación de Matthews (MCC) calculada por conjunto de datos.

Se revisó el origen de los conjuntos de datos y su composición en cuanto a casos positivos, negativos y la semejanza entre los mismos. Todo esto puede verse en la Figura 4.

3.1.2. Predictores de Péptidos Penetradores de Células CPPs

Se analizaron 14 predictores de péptidos penetrantes de células en la literatura, estos son el predictor de Hällbrink (Hällbrink y Karelson, 2005), de Hansen (Hansen *et al.*, 2008b), de Dobvchef (Dobchev *et al.*, 2010), de Sanders (Sanders *et al.*, 2011),

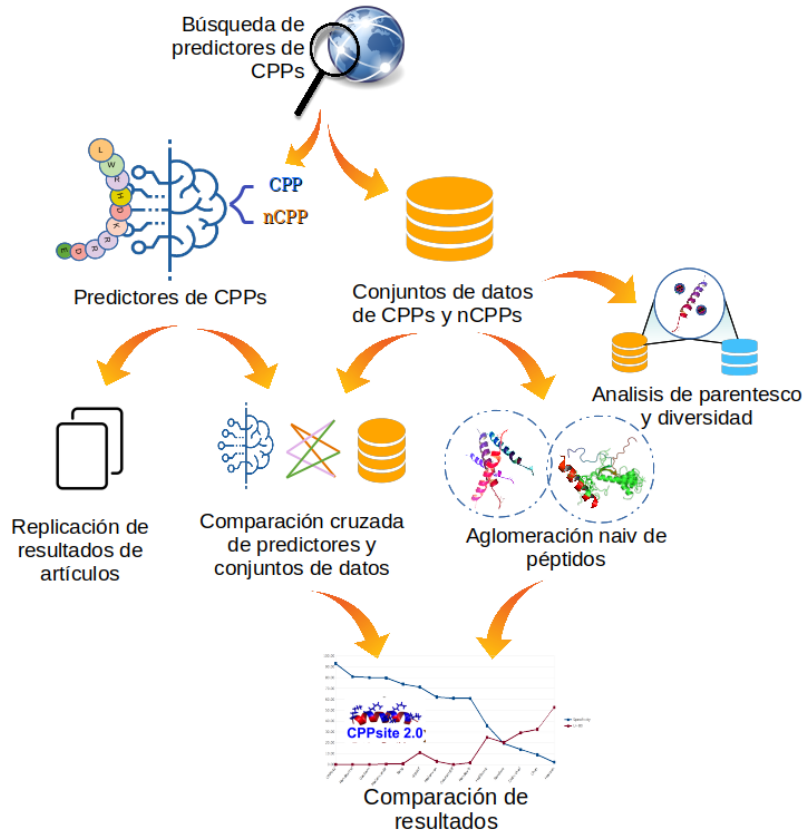

Figura 4. Diagrama de la metodología empleada para la evaluación de los CPPs

Cell-PPD (Gautam *et al.*, 2013), CPPPred (Holton *et al.*, 2013), de Chen (Chen *et al.*, 2015), C2Pred (Tang *et al.*, 2016), de Diener (Diener *et al.*, 2016), CPPPred-RF (Wei *et al.*, 2017b), Skip-CPPPred (Wei *et al.*, 2017a), KELM-CPPred (Pandey *et al.*, 2018), MLCPP (Manavalan *et al.*, 2018) y CPPred-Fl (Qiang *et al.*, 2018).

Es posible agrupar los predictores según su enfoque en métodos directos y métodos basados en aprendizaje de máquina.

Los primeros dos predictores, el de Hälbrick y el de Hansen, pertenecen al primer grupo donde se atienden diversas características fisicoquímicas (PP). Hällbrink y Karelson (2005) designan un intervalo en la escala Z, donde, si un péptido tiene un valor en dicha escala que cae dentro de este intervalo, se clasifica al péptido como CPP. Hansen *et al.* (2008b) utilizan análisis de componentes principales (PCA) con mínimos cuadrados para designar una clase a los predictores. Estos dos predictores comparten tres aspectos; a) utilizan propiedades que experimentalmente se han asociado a los CPPs (escala Z, cantidad de átomos pesados, etc), b) todos sus conjuntos de prueba y entrenamiento son péptidos validados experimentalmente, tanto CPP, como los no

CPPs (nCPPs) y c) sus conjuntos de datos son altamente desbalanceados.

En el segundo grupo, los basados en aprendizaje de máquina, todos emplean conjuntos balanceados tanto positivos como negativos con la excepción del predictor de Dobchev. Dentro de este grupo podemos volver a dividir los métodos, ahora en exploratorios y basados en composiciones.

Dentro del primer grupo están los trabajos de Dobchev *et al.* (2010), Sanders *et al.* (2011) y de Diener *et al.* (2016). Utilizan, respectivamente, redes neuronales artificiales (ANN), máquinas de soporte vectorial (SVM) y tanto SVM como bosques aleatorios (RF). Este grupo representa las etapas tempranas del aprendizaje de máquina para resolver el problema de identificación de CPPs, por lo que exploraron diferentes técnicas, conjuntos y descriptores. En general, lo que caracteriza a este grupo es que; a) usaron una amplia variedad de características fisicoquímicas de los péptidos y los primeros dos mencionados hicieron selección de las mismas, b) sus conjuntos de datos positivos son altamente redundantes y c) preservaron la relevancia de los datos validados experimentalmente.

El resto de los predictores pertenecen al segundo grupo, los basados en composiciones. En este grupo, con excepción de Chen *et al.* (2015), la validez experimental de los nCPPs no es eximio, algunos mantienen estos datos (Holton *et al.*, 2013; Pandey *et al.*, 2018) pero los mismos representan un porcentaje pequeño del conjunto negativo (menos del 1 %). No obstante, lo que caracteriza a este grupo es la aparición de CPPSite, permitiendo conjuntos positivos más diversos y extensos. A partir de esto, todos los conjuntos positivos son tratados con reducción de redundancia y para los conjuntos negativos se empiezan a utilizar péptidos de otras fuentes bajo la suposición de que la probabilidad de que un péptido aleatorio sea CPP es baja. Las propiedades fisicoquímicas pasan a segundo plano y la caracterización de los péptidos se centra en la composición de aminoácidos (ACC), la composición de dipéptidos (DAC) y en menor medida la composición de pseudoaminoácidos (Pse). Las máquinas de aprendizaje utilizadas son bosques aleatorios (RM) (en cualquiera de sus variantes), máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN). Se puede ver en la Tabla 1 los modelos de aprendizaje que cada uno de los predictores utilizó.

Tabla 1. Nombres de la herramientas, conjuntos de entrenamiento y modelos de aprendizaje de los predictores seleccionados.

Autores	Modelo de aprendizaje	Tipo de característica principal	Herramienta
Hallbrink et al 2005	Intervals	PP	-
Hansen et al 2008	PCA	PP	-
Dobvchef et al 2010	ANN	PP	-
Sanders et al 2011	SVM	PP	-
Gautam et al 2013	SVM	ALL	CPPD
Holton et al 2013	ANN	NR	CPPPred
Chen Lei et al2015	RF	PseAAC	-
Tang Hua et al 2016	SVM	DAC	C2Pred
Diener et al 2016	SVM, RF	AAC	DCF
Wei Leyi et al 2017	RF	DAC	CPPPred-RF
Wei Leyi et al 2018	RF	DAC	SkipCPP-Pred
Pandey Poonam et al 2018	Kernel extreme	AAC,DAC,PseAAC	KELM-CPPPred
Manavalan B. et al 2018	ERF, KNN,RF,SVM	DAC,PP	MLCPP
Qiang X. Et al 2018	RF	ALL	CPPred-FL

3.1.2.1. Predictores

Se seleccionaron de la literatura los predictores CPPPred, C2Pred, CPPPred-RF, KELMCPP (con sus seis modelos), SkipCPP-Pred y MLCPP, los cuales se detallaron en la sección 3.1.2. Con cada uno de ellos se evaluaron los 17 conjuntos de datos presentados en la sección 3.1.2.2 y se calcularon las pruebas diagnósticas.

Un modelo de aprendizaje puede sobre entrenarse de tal manera que sólo sea capaz de reconocer elementos similares a los que pertenecen a su conjunto de entrenamiento, entonces, si un conjunto dado contiene una cantidad considerable de péptidos que pertenecen al conjunto de entrenamiento de algún predictor, este tendría a priori una mejor evaluación. Para verificar si esto sucede o no, se evaluaron los mismos predictores únicamente con conjuntos de datos con los que el predictor no hubiera utilizado más del 10, 20, 30, 40, 50, 60, 70, 80 y 90 % de los mismos para entrenar su modelo.

3.1.2.2. Conjuntos de datos para entrenamiento y validación

Cuando se habla de predictores de CPPs, la mayoría de los autores construyen un conjunto de datos para sus respectivos modelos, ya sean para el conjunto de entrena-

miento o de prueba. En las siguientes subsecciones se explica a detalle la construcción de estos conjuntos para los distintos predictores.

Tabla 2. Elementos que contiene cada conjunto de datos separados por casos positivos (CPP) y negativos (nCPP), elementos negativos validados experimentalmente y el método usado para generar los casos negativos.

Conjuntos de datos	CPP	nCPP	nCPPs validados	Tipo de nCPP
Hällbrick	53	16	16	Validado
Hansen	66	19	19	Validado
Dobchev	77	24	24	Validado
Sanders	111	111	34	Muestreo
HoltonTr	74	100	34	Bioactivo
HoltonInd	47	47	0	Bioactivo
GautamTr	187	187	30	Reportado
GautamInd	99	99	0	Reportado
Chen	111	34	34	Validado
Tang	411	411	0	NR
Wei	649	549	0	Aleatorio
CPP924	464	462	0	Aleatorio
PandeyTr	413	413	34	Bioactivo
PandeyInd	96	96	0	Gautam
MalanvanTr	427	854	0	Reportado
MalanvanInd	311	311	0	Tang

NR:No Reportado

3.1.3. Construcción de los conjuntos de datos

Para ilustrar las características de cada uno de los 17 conjuntos se contemplarán cada uno en su parte positiva con los péptidos etiquetados como CPPs y en su parte negativa que será representativa de los péptidos etiquetados como nCPPs. Todos los conjuntos de datos pueden ser consultados en el material suplementario 1 (Datasets.zip).

3.1.3.1. Obtención de datos

En cuanto a casos positivos se refiere, la generación fue bastante uniforme en todos los conjuntos, siendo estos obtenidos a partir de la literatura o de bases de datos disponibles, de las cuales únicamente se encuentran CPPSite (Gautam *et al.*, 2012) y

CPPSite2 (Agrawal *et al.*, 2016), los dos del mismo grupo de trabajo. Sin embargo, la cardinalidad, el balanceo de clases y el tratamiento de los datos fue diferente según el autor.

Los casos negativos validados experimentalmente en ningún caso superan los 34 péptidos. Algunos autores utilizan estos péptidos para armar sus conjuntos (Hällbrink y Karelson, 2005; Hansen *et al.*, 2008b; Dobchev *et al.*, 2010; Sanders *et al.*, 2011; Chen *et al.*, 2015). Los conjuntos de Hällbrink y Karelson (2005), Hansen *et al.* (2008b) y de Dobchev *et al.* (2010) utilizan exclusivamente péptidos validados experimentalmente como nCPPs en sus casos negativos, dando como resultado conjuntos altamente desbalanceados. Por otra parte, Sanders *et al.* (2011) logran un conjunto balanceado mediante un muestreo con reemplazo, por tanto, los casos negativos de Sanders son variaciones de los 34 nCPPs originales.

La opción más popular y con la que se han elaborado una gran cantidad de péptidos catalogados como no penetrantes es mediante la construcción aleatoria de los mismos, que a su vez se pueden organizar en las siguientes categorías:

- **Random:** Generados aleatoriamente, aminoácido por aminoácido, donde uno es elegido al azar y con la misma probabilidad de entre los 20 aminoácidos proteinogénicos (Sanders *et al.*, 2011; Wei *et al.*, 2017b).
- **Bioactive:** Aleatorio de origen biológico, donde se seleccionan de manera aleatoria de bases de datos como CAMP, ADP2 o BIOPEP péptidos que tengan algún tipo de función biológica diferente a CPP (Holton *et al.*, 2013; Pandey *et al.*, 2018).
- **Reported:** Muestreo aleatorio de péptidos reportados en la literatura de bases de datos como Swiss-prot o NCBI (Gautam *et al.*, 2013; Manavalan *et al.*, 2018).

Cada una de las técnicas de generación de casos negativos tiene sus ventajas y desventajas. Las aleatorias (cualquier categoría mencionada anteriormente) aportan variedad a los conjuntos negativos. Aunque no se puede confirmar la ausencia de su actividad como penetrante, se entiende que la probabilidad de que un péptido sea CPP es baja, por tanto la probabilidad de que un péptido tomado al azar sea CPP también lo es (Sanders *et al.*, 2011). Sin embargo, no existe control sobre los datos

proporcionados y como mencionan en el trabajo de Ahmed *et al.* (2018), es crucial que los conjuntos de entrenamiento estén apegados a la realidad. Obtener péptidos validados experimentalmente como no penetradores de células es complicado y, en general, es complicado demostrar experimentalmente la ausencia de alguna actividad en un péptido (Qi *et al.*, 2006). Esto deja a este tipo de conjuntos con una cardinalidad pequeña en comparación a sus análogos positivos pero con mayor certeza de que el modelo responda positivamente en la práctica.

Cuando al generar los casos negativos se tiene en consideración el ambiente típico en el que la herramienta será aplicada, se esperaría que la máquina tendiera a ser de mayor utilidad que una entrenada con un conjunto que descuidó el contexto ya que se entrena con elementos similares a la realidad de la aplicación. Sin embargo, no existe forma *in silico* reportada hasta ahora de probar que los péptidos elegidos son nCPP, lo que podría resultar en máquinas capaces de distinguir sólo péptidos con características similares a los ya conocidos.

La Figura 5 presenta una caracterización de los procedimientos seguidos para obtener CPPs y nCPPs de los métodos analizados en este trabajo. Todos los CPPs son péptidos cuya actividad fue validada experimentalmente, sin embargo, la cantidad de nCPPs validados de esta manera es siempre menor a 34. En la Tabla 3 pueden verse la cantidad de péptidos utilizados, separados en CPPs y nCPPs y para el caso de los nCPPs se describe la cantidad de estos que fueron validados experimentalmente y cómo fue que se generaron.

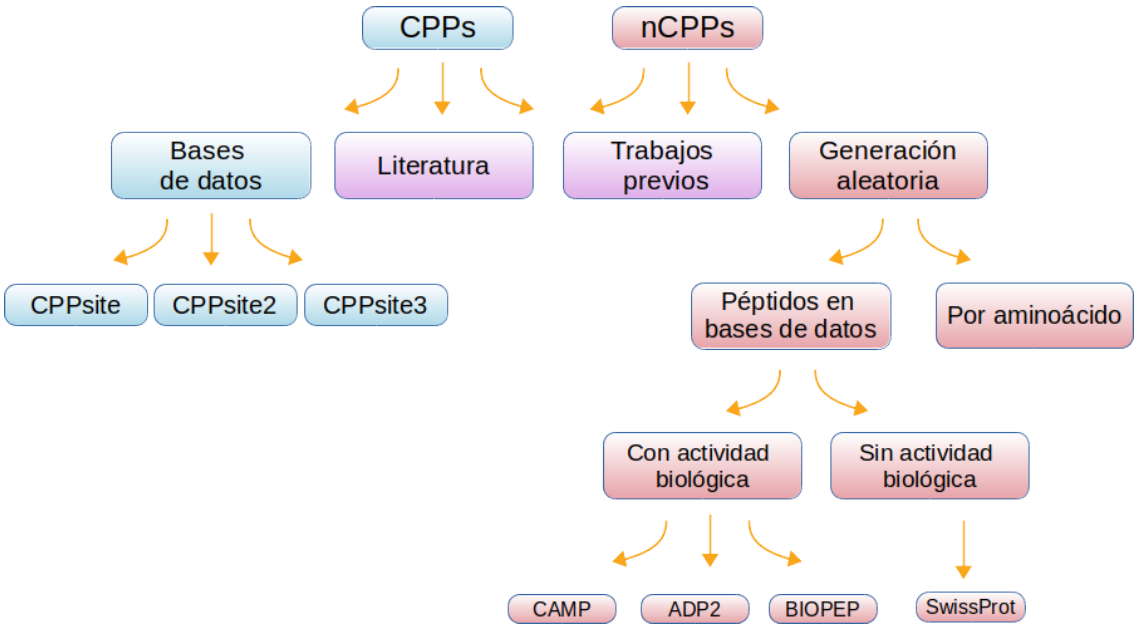

Figura 5. Esquema de obtención de los péptidos para los conjuntos de datos

3.1.3.2. Redundancia de secuencias

En el análisis computacional de secuencias biológicas se recomienda reducir la redundancia de los conjuntos de datos para disminuir espacio de almacenamiento y tiempo computacional, así como para eliminar ciertas interferencias en los métodos computacionales utilizados (Li y Godzik, 2006). En los conjuntos, la redundancia se da cuando se tienen secuencias que aportan el mismo tipo de información. Si se tiene un CPP con múltiples variaciones durante el conjunto de entrenamiento, la máquina de aprendizaje asignará un peso mayor a cualquier variación de la misma, en lugar de otras secuencias de características diferentes con menos variaciones.

Gautam *et al.* (2012) mencionan que no aplicaron reducción de redundancia debido a que el cambio de un sólo aminoácido puede eliminar la propiedad penetrante de un péptido, sin embargo, el resto de los autores que realizan predictores utilizaron reducción de redundancia.

3.1.3.3. Conjuntos de datos

De cada predictor disponible en la literatura se recopiló al menos uno de sus conjuntos de datos dependiendo de la disponibilidad de los mismos. En total se analizaron 17

conjuntos de datos positivos y 17 negativos, provenientes de los distintos predictores de la literatura.

Se renombraron los conjuntos de datos para corresponder al apellido del primer autor que propuso el predictor (con excepción del de Wei *et al.* (2017a) quien nombró a su conjunto CPP924, nombre que también se utilizó en este trabajo) y en caso de que el mismo autor hubiese tenido dos o más conjuntos, se agregó un sufijo dependiendo del uso del conjunto de datos, ya sea en entrenamiento (Tr por training) o en prueba (Ind por independent) (ver sección 3.1.2.2) para probar la calidad del predictor.

Por lo anterior los nombres que se utilizarán a lo largo de este trabajo para los conjuntos quedan como se muestra en la Tabla 3.

Tabla 3. Nombres de los conjuntos de datos analizados en este trabajo.

Referencia	Tipo	Nombre del conjunto
Hällbrick et al., 2005	Training	Hällbrick
Hansen et al., 2008	Training	Hansen
Dobchev et al., 2010	Training	Dobchev
Sanders et al., 2011	Training	Sanders
Holton et al., 2013	Training	HoltonTr
Holton et al., 2013	Independent	HoltonInd
Gautam et al., 2013	Training	GautamTr
Gautam et al., 2013	Independent	GautamInd
Chen et al., 2015	Training	Chen
Tang et al., 2016	Training	Tang
Wei et al., 2017	Training	Wei
Wei et al., 2018	Training	CPP924
Pandey et al., 2018	Training	PandeyTr
Pandey et al., 2018	Training	PandeyInd
Manavalan et al., 2018	Training	ManavalanTr
Manavalan et al., 2018	Independent	ManavalanInd

Se contemplaron los conjuntos en su parte positiva con los péptidos etiquetados como CPPs y en su parte negativa etiquetados como nCPPs. Todos los conjuntos de datos pueden ser consultados en el material suplementario 1 (Datasets.zip). Posteriormente, estos se filtraron de tal modo que cumplieran con todos los requisitos de todos los predictores, estos son: longitud entre 10 y 30 AA, únicamente aminoácidos naturales y sin caracteres especiales.

Similitud. Se definió un umbral $\alpha = 0.8$ de similitud el cual nos indica hasta qué punto dos secuencias serán consideradas diferentes. La similitud entre dos secuencias **a** y **b** se tomó como la similitud máxima alcanzada durante un alineamiento local $Al(\mathbf{a}, \mathbf{b})$ utilizando la matriz de puntuación Blosom62. La similitud de dos secuencias **a** y **b** $S(\mathbf{a}, \mathbf{b})$ se calculó como la razón de la cantidad de aminoácidos empatados al alinear las secuencias, entre la longitud del alineamiento.

$$S(\mathbf{a}, \mathbf{b}) = \frac{matches(Al(\mathbf{a}, \mathbf{b}))}{length(Al(\mathbf{a}, \mathbf{b}))}$$

Si una secuencia supera el umbral α en su medida de similitud con otra, se considera redundante.

Redundancia y diversidad. La redundancia de un conjunto se calculó como la cantidad de clusters generados al agrupar elementos con una similitud de al menos α entre ellos, dividido entre la cantidad total de elementos en el conjunto.

Se aplicó un análisis sobre la redundancia de los conjuntos en dos sentidos utilizando DoverAnalyzer v0.1.2 (Aguilera-Mendoza *et al.*, 2015). El primero se refiere a la diversidad D de conjuntos de un mismo caso, una redundancia baja resulta en una diversidad alta. Este parámetro se calcula como los r elementos resultantes de una reducción de redundancia a un umbral de similitud α entre la cardinalidad $Card(X)$ original del conjunto en cuestión X.

$$D = \frac{r(\alpha)}{Card(X)}$$

El segundo contempla la pertenencia m, un valor que representa el porcentaje de elementos k de un conjunto C_i que también pueden localizarse en un conjunto C_j. Esto puede ser calculado de la siguiente manera:

$$m_{i,j} = \frac{\sum_{k=1}^{Card(C_i)} u(i,k,j)}{Card(C_i)}$$

donde

$$u(i,k,j) = \begin{cases} 1 & si\,el\,elemento\,k\,\text{del conjunto}\,C_i\,está\,en\,C_j \\ 0 & en\,otro\,caso. \end{cases}$$

Parentesco. Finalmente, para examinar la separabilidad de los datos se calculó un valor de parentesco K entre los conjuntos de casos positivos y negativos. Para realizar esto, $K(\alpha, \mathcal{N}, \mathcal{P})$ será el total de elementos en el conjunto negativo $\mathcal{N}$ que tienen una semejanza de al menos α con al menos un elemento del conjunto positivo $\mathcal{P}$ dividido entre la cardinalidad del conjunto negativo. El caso inverso $K(\alpha, \mathcal{P}, \mathcal{N})$ no se considera dado que en algunos conjuntos de datos los casos positivos tienen cardinalidad considerablemente mayor que los negativos, mientras que los negativos siempre son menor o a lo más igual que los positivos.

$$K(\alpha, \mathcal{N}, \mathcal{P}) = \frac{\sum_{i=1}^{Card(\mathcal{N})} U(\mathcal{N}_i, \mathcal{P}, \alpha)}{Card(\mathcal{N})}$$

$U(\mathcal{N}_i, \mathcal{P}, \alpha)$ es una función que determina si existe o no un elemento positivo que sea similar en al menos α al péptido $\mathcal{N}_i$, i.e.:

$$U(\mathcal{N}_i, p, \alpha) = \begin{cases} 1 & si\,\exists\, p \in \mathcal{P} | S(\mathcal{N}_i, p) \leq \alpha \\ 0 & en\,otro\,caso \end{cases}$$

Donde $\mathcal{N}_i$ es el i-ésimo péptido negativo (nCPP) y p algún péptido del conjunto positivo. El valor de K indicará el grado de dificultad del conjunto; cuando mayor sea K, más difícil será separar casos positivos de negativos a nivel de secuencia.

Grado de separabilidad a nivel de secuencia. Adicionalmente, se llevó a cabo un SpectralClustering utilizando el paquete de SKLearn (Pedregosa *et al.*, 2011) para python, sin agregar ningún tipo de descriptor adicional a la distancia de edición entre dos secuencias calculada con pairwise2 de biopython (Cock *et al.*, 2009). Esto con el fin de observar el grado de separabilidad a nivel de secuencias, siendo este procedimiento un enfoque no supervisado.

3.1.4. Desarrollo de interfaz para la evaluación de péptidos como CPPs mediante diferentes predictores disponibles en línea

Para agilizar el proceso de evaluación *in silico*, se diseñó e implementó una aplicación para la evaluación de péptidos como CPPs mediante predictores de terceros. La aplicación fue llamada EvalCPPApp, para esto se localizaron los servidores de los predictores, el tipo de entrada que recibían y el formato de salida.

Para desarrollar esta aplicación fue necesario el empleo de jsoup ya que los formatos de salidas de todos los servidores son diferentes y se requirió un análisis particular de cada html resultante para los casos de servidores web (CPPD, CPPP, C2Pred, SKIPCPP, MLCPP, KELMCPP, y CPPFL).

En el caso de aquellos predictores que debían correrse de forma local (DCF y CPPRF) se pedía como entrada un archivo fasta que posteriormente era convertido al tipo de entrada que solicitara el predictor y una lectura posterior a la salida.

Se hicieron peticiones masivas a los servidores descubriendo que predictores como KELM no manejan un alto número de secuencias, regresando un error HTTP 504 con cantidades relativamente pequeñas (10 secuencias de longitud mayor o igual a cinco aminoácidos).

Dado que existen predictores con una cantidad limitada de datos que pueden ser accedidos sin fallas del servidor, se procedió a hacer un llamado por lotes a los servidores de forma paralela, la forma en la que esto se realizó puede verse simplificada en el pseudocódigo 3.1. Para más información sobre el funcionamiento y utilización de esta herramienta, véase el manual en el Apéndice .

Pseudocódigo 3.1. Algortimo para división de N secuencias

```
nSections=100
nThreads =10

Split N in nSections sections:
  for each section in sections:
    Split section in nThreads threads
```

```
    Send threads as multiform petition to servers
     Merge results of each thread
  Merge results of each thread section.
Ouput predictions of N
```

3.2. Identificación de péptidos inductores de autofagia de forma experimental

Debido a que este trabajo se basa en la hipótesis “Péptidos que son AMP y CPP a la vez son AIP", fue necesario hacer una selección de péptidos de la literatura que cumplieran con los requisitos AMP y CPP para posteriormente ser evaluados experimentalmente como AIP.

Antes de poder definir si los péptidos elegidos tenían actividad como inductores de autofagia, era necesario saber identificar la autofagia en la célula, para lo cual se abordaron tres técnicas diferentes. Una vez identificada la diferencia se procedió a examinar las diferencias entre los péptidos elegidos y la presencia de autofagia establecida por controles de autofagia basal (Control negativo) e inducida (Control positivo).

3.2.1. Selección de péptidos

Mediante el filtro de actividad de CPPSite (Gautam *et al.*, 2012) (Opción Major fields-Category) se seleccionaron AMPs experimentalmente validados.

Se buscaron en la literatura AMPs que no fueran CPPs (AMP+nCPPs), CPPs que no fueran AMPs (nAMPs+CPP) y péptidos sin estas actividades: todas las propiedades evaluadas experimentalmente (ver Tabla 24). En caso de no encontrarse péptidos con la actividad nAMP, estos se predijeron con el predictor CAMPred (Porto *et al.*, 2012) y para el caso de no tener AMP+nCPP la actividad penetrante se evaluó con MLCPP (Manavalan *et al.*, 2018).

Con el fin de probar si la inducción de la autofagia era función de los péptidos que fueran AMP y CPP, se mandaron a sintetizar un subconjunto de los péptidos encontrados. Estos péptidos fueron sintetizados por la empresa GenScript y por AnaSpec. A los péptidos sintetizados por la empresa GenScript se les solicitaron pruebas de solubilidad en agua, en las cuales determinó experimentalmente que los péptidos eran solubles en agua ultra pura a concentraciones menores o iguales a 5 mg/mL. Para el caso de los sintetizados por AnaSpec se calculó su solubilidad teórica a partir de sus aminoácidos.

3.2.2. Pruebas experimentales

Para realizar el análisis de los siete péptidos mostrados en la Tabla 25, se emplearon un total de cinco mutantes de la levadura *Saccharomyces cerevisiae* WT BY472.

Las mutaciones pueden verse en la Tabla 4. Todas las cepas tenían al menos una proteína fluorescente, ya sea unida a la proteína citoplasmática PGK1 o la proteína VMA1 localizada en la membrana vacuolar. Para el citoplasma se utilizó la proteína GFP y para la membrana vacuolar TDimer2.

Tabla 4. Cepas utilizadas durante la experimentación.

Nombre	**Fluoresencia**		**Auxotrofía**
	Citoplasma	**Membrana vacuolar**	
WT,PGK1,GFP	Verde	-	-URA
WT, VMA1-Tdimer2	-	Roja	-HIS
WT PGK1-GFP+VMA1-Tdimer2	Verde	Roja	-URA,-HIS
WT PGK1-GFP+VMA1-Tdimer2 ΔATG8	Verde	Roja	-URA,-HIS
WT PGK1-GFP+VMA1-Tdimer2 ΔPEP4	Verde	Roja	-URA,-HIS

3.2.2.1. Reactivación de cepas

Se sembraron por estriado cruzado las cepas de *S. cerevisiae* PGK1-GFP(WT), VMA1-Tdimer2(WT) y PGK1-GFP+VMA1-Tdimer2(WT) en cajas de Petri con medio sintético de 2.01 g de bases nitrogenadas de levadura, 0.3 g de fosfato de potasio y en correspondiente a la auxotrofía presentada previamente en la Tabla 4, 0.24 g de aminoácidos -HIS, -URA y -HIS -URA.

Tabla 5. Pesos moleculares y volumen necesario para llevar un μg de péptido a una concentración de 500μM

Nombre del péptido	Peso molecular (gr/mol)	Volumen (ml)
Bac1-15	1900	1.04
Bac15-24	1114	1.76
Bip16	538	3.59
Bradykinin	1041	1.89
CRGDK	559	3.46
Inv10	3057	0.65
MG2d	2500	0.79
SynB5	2080	0.95

Las cepas se incubaron a 30°C de 48 a 72 horas. Una vez completado el tiempo de incubación se procedió a cultivar una colonia de cada caja en medio líquido tomando en cuenta su auxotrofía. Las células se dejaron crecer entre 12 y 16 horas.

Las cepas se ajustaron a una densidad óptica (o.d.) de 600 nm a 0.12 con ayuda de un espectrofotómetro Genesys 10S UV-Vis (ThermoFisher, USA). Se evaluó la fluorescencia de las tres cepas mediante un microscopio para epifluorescencia Leica DM 1000 a 60x, con lo que se verificó que las cepas estuvieran activadas al verse fluorescencia en el citoplasma, membrana vacuolar o ambos según la cepa.

3.2.2.2. Preparación de péptidos

Para cada uno de los péptidos se pesó 1 mg y se colocaron en tubos Eppendorf de 2 ml, a cada tubo se le agregó agua des-ionizada y estéril a un volumen tal que la concentración final fuese de 500μM (estos volúmenes pueden verse en la Tabla 5).

En los casos en los que 2 ml no fue suficiente para obtener la concentración deseada, se utilizó la mitad del volumen requerido para obtener una concentración de 1000 μM.

3.2.2.3. Pruebas y controles

A lo largo de todos los experimentos realizados durante este trabajo, se contemplaron tres controles más las pruebas. Los controles se dividen en 1) control de autofagia basal, el cual se definió como la cepa en medio de cultivo rico y sin ningún péptido o inductor de autofagia, 2) control de autofagia por inducción peptídica, representado

por la cepa con tratamiento del péptido IP-1 y finalmente, 3) el control de autofagia por inducción farmacológica que fue la cepa tratada con Rapamicina.

3.2.2.4. Experimento 1: Evaluación de autofagia inducida por espectrofotometría

Se cultivaron 5 ml de medio cultivo de célula única de las cepas PGK1-GFP, VMA1-Tdimer2 y PGK1-GFP+VMA1-Tdimer2 por 12 horas.

Se ajustó el cultivo celular a 0.12 o.d. con espectrofotómetro, empezando con una relación aproximada de 1 μl de cultivo por cada 10 μl de medio.

Se analizaron los tratamientos ilustrados en la Tabla 6.

Tabla 6. Tratamientos y concentraciones utilizados para el Experimento 1.

Tratamientos	**Concentracion(μM)**
Sin tratamiento	-
Rapamicina	0.2
IP-1	10
MG2d	5
Inv10	5

Para cada tratamiento se prepararon 0.6 ml de cepa a 0.12 o.d. y se colocaron por sextuplicado en una placa de 96 pozos de fondo negro, 0.1 ml por pozo.

En el equipo Synergy Mx se monitoreó la placa bajo las siguientes condiciones:

- Versión del software: 1.01.
- Programa: Levadura fluorescencia, crecimiento 24 horas, lecturas cada 30 minutos.
- Temperatura: 30 °C
- Agitación constante.
- Lecturas de cinética: 600 nm
- Longitud de onda lectura de GFP: 488 nm

- Longitud de onda lectura de TDimer: 532 nm
- Duración: 24 horas.

3.2.2.5. Experimento 2: Evaluación de autofagia inducida por péptidos en células BY472 PKG1-GFP + VMAT1-Tdimer2 mediante microscopía de alta resolución.

La observación de autofagia se llevó a cabo mediante la reconstrucción de imágenes de súper resolución obtenidas mediante el microscopio de fluorescencia de súper resolución Nanoimager de ONI. Para ello fue necesario preparar las muestras en portaobjetos y realizar un procesamiento de los datos. Este proceso se explica a continuación.

Preparación de muestras. Se realizó el cultivo en medio líquido de la cepa BY472 PKG1-GFP + VMAT1-Tdimer2 (WT) por 12 horas. Los tratamientos (Tabla 5) se aplicaron a una concentración de 10μM (con excepción de Rapamicina cuya concentración se mantuvo a 0.2 μM) a un volumen de 1 ml con densidad óptica de 0.12 o.d.

Las células se trataron por 6 horas, después de las cuales se retiró el sobrenadante y se realizó una fijación con paraformaldehido al 4 % por una hora, finalizado el tiempo se retiró el sobrenadante una vez más.

Finalmente, se preparó un gel de agarosa al 2.5 % y se colocaron 5 μl de cepa y 5 μl de agarosa en un portaobjetos que fue cubierto con un cubreobjetos.

Lectura. Cada una de las muestras se colocó en el microscopio Nanoimager, donde se excitaron los fluoróforo GFP con una $\lambda = 473\,nm$ a 6.7 kW/cm^2 y $\lambda = 640\,nm$ a 4.7 kW/cm^2 para el fluoróforo Tdimer.

Se tomaron vídeos de 200 imágenes estroboscópicas de 3 campos por muestra enfocadas de forma manual.

3.2.2.6. Análisis de morfologías en células

Tratamiento de imágenes. Se separaron las imágenes en los dos canales previamente establecidos haciendo uso de la exportación de datos del software del microscopio, se tomaron un total de 100 imágenes por canal, sin embargo, dado que la toma fue de forma estroboscópica, uno de los láseres recibiría información del otro en cada captura, por tanto la cantidad final se duplicó y mediante un descarte intercalado se eliminaron aquellas tomas provenientes de este proceso.

Mediante el programa ImageJ (Abramoff, 2004) y el paquete de procesamiento Fiji (Schindelin *et al.*, 2012), se retiraron las imágenes provenientes del canal opuesto al correspondiente, sea para GFP o Tdimer2, quedando vídeos de 100 imágenes para cada canal.

Se realizaron las correcciones de las imágenes generadas por la desalineación de los láseres en el microscopio por una distancia aproximada de 3.03 píxeles en el eje X y 3.27 en el eje Y. Estas correcciones se realizaron mediante el traslado de las imágenes de los canales para la lectura de GFP. Para hacer el análisis de desviación se utilizó la placa de calibración de PSFcheck (Gattass y Mazur, 2018).

Se realizó un análisis SRRF (Super-Resolution Radial Fluctuations) con el software ImageJ de cada una de las imágenes tomadas. En el caso de las imágenes provenientes de la cepa PGK1-GFP+VMA1-Tdimer2 se fusionaron los canales correspondientes a la fluorescencia de las proteínas GFP y Tdimer2.

Posteriormente, se realizó una contabilización de las diferentes morfologías de las células encontradas en cada una de las imágenes correspondientes a los tres controles: control de autofagia basal, control de autofagia por inducción peptídica y autofagia por inducción farmacéutica.

Contabilización de células. Es lógico pensar que en un cultivo no todas las células se encontrarán en la misma fase del ciclo celular ni realizando los mismos procesos, empero, al ser estimuladas con un inductor de autofagia, la probabilidad de que una célula esté realizando tal actividad, será mayor, por tanto, habrán más células que presentarán cambios en su vacuola y por tanto, en la imagen que recibimos de ella

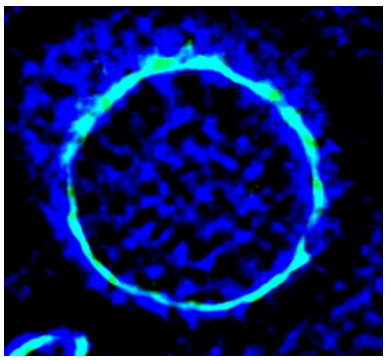

Figura 6. Fotografía de célula de Saccharomyces cerevisiae en morfología “aro".

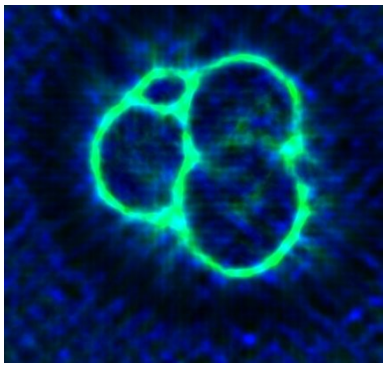

Figura 7. Fotografía de célula de Saccharomyces cerevisiae en morfología "palomita".

gracias a los marcadores en PGK1 y VMA1. Contabilizar las células nos permitirá observar la influencia de los tratamientos sobre las células.

Las células se contaron manualmente utilizando el Plugin “cell count“ de ImageJ (Abramoff, 2004). Las estructuras que se contabilizaron fueron “aro” y "palomita". La forma aro se entiende por aquellas células donde la fluorescencia de TDimer2 se encuentra como un círculo tal como se muestra en la Figura 6. Este círculo no necesariamente era perfecto, pero era importante que se encontrara solo uno en la célula a menos que esta se encontrara dividiéndose. Para la morfología de palomita, en el caso ideal se apreciaban tres o más círculos cortándose unos con otros, como lo muestra la Figura 7.

Dado que existen múltiples factores que afectan la imagen final tal como la profundidad a la que se encuentra la célula o la difracción de la luz causada por el medio, la clasificación fue subjetiva al observador, pero en caso de que la clase fuese ambigua, esta se omitía.

Una vez teniendo las células contabilizadas, se sumó la cantidad total de aros y

palomitas por cada campo y se obtuvo la relación de aros y palomitas respecto a dicho total. Las contabilizaciones completas pueden encontrarse en el Anexo .

Se agruparon las imágenes por tratamiento y se realizó un diagrama de cajas del porcentaje de aros en las células en las que se buscaron diferencias.

3.2.2.7. Prueba de toxicidad

Dado que los péptidos sintetizados provienen de predicciones in silico, se hicieron pruebas experimentales de toxicidad para evaluar el posible efecto tóxico que pudieran tener sobre la viabilidad de *S. cerevisiae*. Para ello se utilizó una placa de 96 pozos negra de fondo claro en la cual se colocaron 0.1 ml de cultivo de WT BY472 PGK1-GFP+VMA1-Tdimer a una concentración de 10 μM de cada uno de los péptidos, un control sin tratamiento y dos blancos, con cinco repeticiones por cada uno.

La placa se leyó cada media hora durante 24 horas con el equipo Synergy Mx en las mismas condiciones del Experimento 1, se tomaron los datos de la cinética y se calcularon las áreas bajo la curva para evaluar la posible toxicidad.

3.3. Búsqueda de posibles péptidos inductores de autofagia en proteomas de mamíferos

Esta sección tiene como objetivo localizar nuevos AIPs a partir de proteomas de mamíferos. Para ello se siguió la hipótesis de que “Péptidos que son AMP y CPP a la vez son AIP", a los péptidos de actividad antimicrobiana y penetradora de células encontrados en los proteomas serán desde ahora referidos como AMCPPs. Se buscaba además que los AIPs propuestos tuvieran la mayor oportunidad de ser inductores de autofagia. Se identificaron péptidos que provinieran de proteínas que tuvieran relación con la autofagia, que para sencillez de redacción, estos péptidos tendrán el subíndice r (por ejemplo, AMCPP_r). Asimismo, se evaluó el porcentaje de proteínas OP_r obtenidos de péptidos que no cumplían con ser AMCPPs, en donde se esperaría que una menor proporción de péptidos_r fuera encontrada.

3.3.1. Selección de especies

En AnAge (De Magalhães y Costa, 2009) se recopilaron los 349 registros de mamíferos que incluían tanto el peso del animal adulto (M) como la máxima longevidad reportada tanto en cautiverio como en estado salvaje (t_{max}).

De las especies recopiladas se contemplaron los datos correspondientes a su máxima esperanza de vida t_{max} y su peso promedio en gramos M. Con ellos se obtuvo un valor predicho de máxima esperanza de vida t'_{max} y se calculó el error cuadrático medio respecto a los valores reportados con la ecuación alométrica establecida por de Magalhães *et al.* (2007) en donde se buscaron diversas regresiones de la longevidad de los mamíferos y se llegó a la presentada en la siguiente ecuación:

$$t'_{max} = 4.88 * M^{0.153} \tag{1}$$

También se calculó una regresión logarítmica la cual ha sido utilizada previamente en las relaciones de longevidad y masa corporal (Temerin, 1985; K. Schmidt-Nielsen y Knut, 1986):

$$t'_{max} = -0.4742 + 2.176 * \ln M \tag{2}$$

Con ambas predicciones se obtuvo el coeficiente de longevidad ct_{max}, el cual se calcula de la siguiente manera:

$$ct_{max} = \frac{t_{max}}{t'_{max}} \tag{3}$$

A cada una de las aproximaciones se le calculó el error cuadrático medio con las 349 especies disponibles. Dando prioridad en la toma de decisiones durante el resto del procedimiento a los coeficientes de ecuación con menor error.

Se ordenaron los valores de ct_{max} de mayor a menor usando ambas ecuaciones (1 y 2) quedando las mismas especies en diferentes posiciones.

Se buscó en el NCBI (Sayers *et al.*, 2019) el proteoma de las 20 especies de mayor y de menor coeficientes de longevidad. Seleccionando solo las especies cuyo proteoma se encontrara disponible.

Con la intención de buscar un control de longevidad estándar, se seleccionaron las 40 especies más cercanas a la media de ct_{max}. De estas, se escogieron aquéllas que pertenecieran a un orden de las especies elegidas anteriormente, de estas, únicamente el orden Rodentia contó con al menos una especie con un proteoma con suficientes proteínas para realizar el análisis.

3.3.2. Filtro de proteínas por ontología

Con el fin de reducir la cantidad de proteínas a analizar, se eliminaron las proteínas cuya actividad asociada sólo ha sido reportada para el proceso de la información genética. Para esto se usó la clasificación de KEGG (Kanehisa y Goto, 2000), eliminando así todos las proteínas asociadas a la etiqueta "Genetic Information Processing". Este proceso se realizó mediante Koala y GhostKoala (Kanehisa *et al.*, 2016). Sin embargo, es importante hacer notar que esto no significa que no se puedan encontrar AIPs en estas proteínas.

3.3.3. Filtro de péptidos por actividad predicha

Las proteínas seleccionadas de cada uno de los proteomas fue sometida al software Inprot (Melendrez C., 2018) para obtener los AMPs que pudiesen existir en cada proteoma. Este software permite realizar una proteólisis *in sílico* de la cual se obtienen los péptidos de longitudes especificadas (entre 10 y 30 AA) y realiza una predicción de actividad antimicrobiana.

La realización de dicha herramienta se realizó mediante el lote de comandos mostrado en el Código 3.2, este fue ejecutado en sistema RedHat Enterprise Linux 6.7 (kernel 2.6.32-573.el6.x86_64).

Código 3.2. Bash que ejecuta a inprot

```
#SBATCH -p cicese
#SBATCH -c 24
#SBATCH -J ProtRatopin
#SBATCH -o ratopin_%j.log
#SBATCH -e ratopin_%j.err
#SBATCH --mem=120G
#SBATCH --exclusive

NAME="KeggSeqs"
NAMES=‘ls Animales/*.fasta | sed -e ’s/\.fasta$//’‘
echo "Processing at ‘date‘"
for ANIMAL in ${NAMES}; do
      BASE="${ANIMAL}"
      proteome="$PWD/${BASE}.fasta"
      mkdir "$PWD/${BASE}"
      model="$PWD/model"
      scaling="$PWD/scaling"
      num_threads=0
      ulimit -c unlimited

      ./inprot -i $proteome -m $model -s $scaling -t $num_threads -o
         ${BASE}/podado.fasta -l 10 -u 30 -v -w 1
done
```

El resultado fue obtenido del podado generado por Inprot. Los respectivos podados fueron filtrados por longitud, seleccionando solo aquellos péptidos con longitud entre 5 y 30 AA debido a las restricciones de los predictores de CPPs; CPPP, DCF y MLCPP. Para aquéllos con longitud mayor a 30 AA, se les aplicó un ventaneo de tamaño k=30. A los péptidos resultantes se les aplicó reducción de redundancia mediante CDHit (Li y Godzik, 2006) al 80 %. Finalmente, el resultado fue analizado con los predictores de CPPs previamente mencionados para obtener así los AMCPPs.

Se registró el consenso de los predictores de CPPs, se contabilizó la cantidad de

AMCPPs encontrados y esto se dividió entre la cantidad original de AMPs encontrados para esa especie, este resultado puede verse en la Tabla 7. Adicionalmente, pueden verse las proteínas del proteoma original y las proteínas *OP* de los AMCPPs. En las columnas de relación puede verse la razón de AMCPPs respecto a los AMPs y las proteínas con al menos un AMCPP respecto a las totales.

Tabla 7. Relación entre las proteínas en el proteoma, los AMPs encontrados y los AMCPPs que se derivan de esos AMPs.

Especie	Coef. longevidad	Proteínas	AMPs	AMCPPs	Proteínas *OP* de AMCPPs	Proporción	
						Péptidos	Proteínas
Myotis lucifugus	3.9	43,106	19584	583	1562	0.03	0.04
Heterocephalus glaber	2.6	41,961	19999	632	1495	0.03	0.04
Desmondus rotundus	2.5	29,845	18055	476	939	0.03	0.03
Eptesicus fuscus	1.8	49,084	55085	1428	2063	0.03	0.04
Cavia porcellus	0.6	37,360	18731	658	1492	0.04	0.04
Rattus norvegicus	0.2	66,876	59999	3724	1285	0.06	0.02
Condylura cristata	0.2	29,166	50141	1505	465	0.03	0.02

3.3.4. Búsqueda de subsecuencias de proteínas relacionadas a autofagia

Como ya se ha mencionado, los tres péptidos comerciales provienen de ATG6 (Tat-L11, Tat-L11S, Tat-D11), una proteína importante para la regulación de autofagia. Si un péptido es subsecuencia de una proteína involucrada en autofagia, esta tendría una oportunidad mayor de ser AIP que otra que no lo es ya que pequeñas subsecuencias de AA pueden jugar roles importantes en la función biológica tanto como fragmentos como en la proteína entera (Minkiewicz *et al.*, 2015). Entonces, si una secuencia de aminoácidos predicha como AMCPP fuese subcadena de una proteína relacionada a la autofagia (proteína$_r$), estaríamos añadiendo evidencia a la posible actividad como inductor de autofagia más allá de las combinaciones de actividades AMP+CPP.

Con la finalidad de identificar AIPs presentes en proteínas involucradas en autofagia, se realizó una búsqueda por homología de las proteínas *OP* con la base de datos de péptidos de autofagia Autophagy Database (Homma *et al.*, 2011). Este proceso se realizó de la siguiente manera.

A cada proteína en el proteoma original se le asignó un valor γ, el cual indica la cantidad de AMCPPs que eran subsecuencia de la proteína en cuestión. Todas las proteínas con $\gamma \geq 1$ de un proteoma conforman lo que de ahora en adelante se denominará el

conjunto *OP*.

Las *j* proteínas de *OP* se sometieron a una búsqueda por homología en la base de datos AutophagyDatabase (http://www.tanpaku.org/autophagy/index.html), la cual emplea PSI-BLAST. Una vez teniendo las proteínas que se encuentran registradas con actividad autofágica OP_r, se mapearon de vuelta a los péptidos AMCPPs, cuyo mapeo resultó en los $AMCPP_r$s. Este proceso se ilustra en la Figura 8.

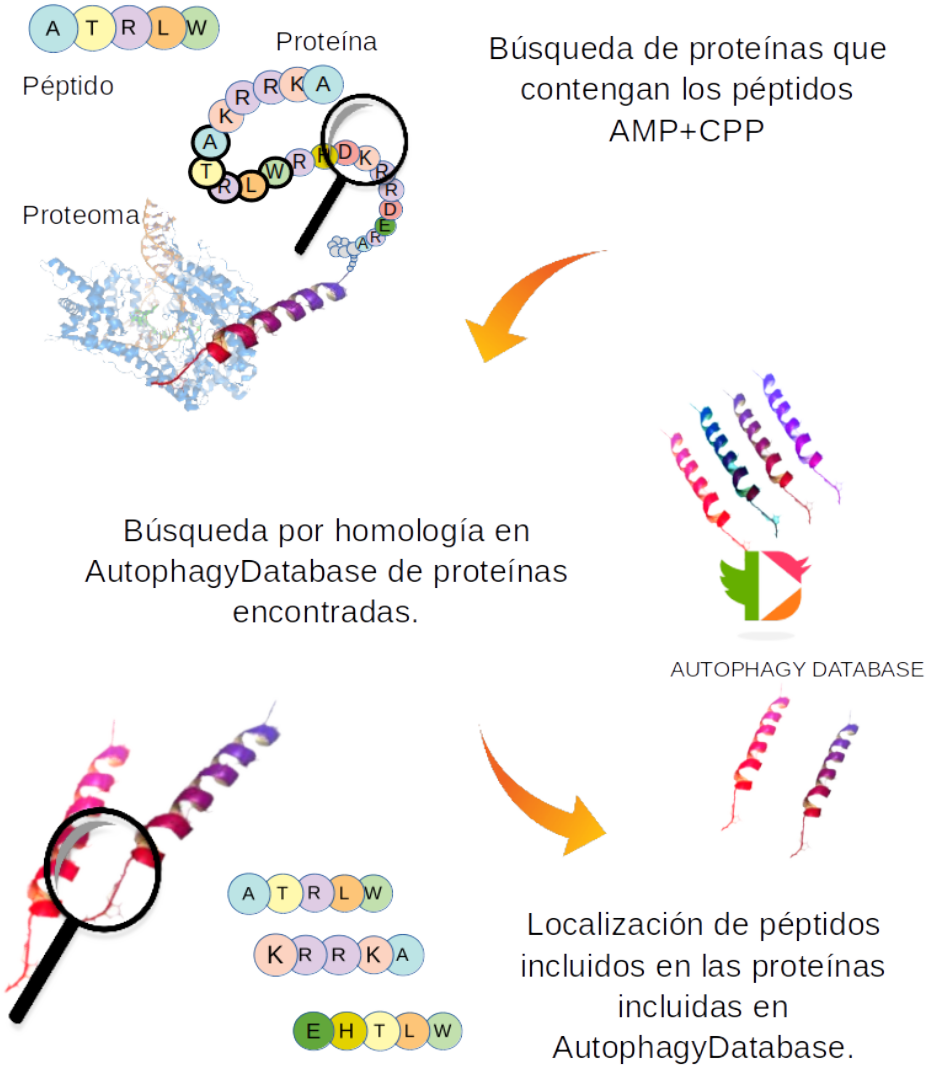

Figura 8. Proceso para la selección de péptidos AMP+CPP con evidencia de relación a la autofagia.

3.3.5. Comparación de péptidos encontrados respecto a actividades AMP y CPP

Para comparar la diferencia de la evidencia de autofagia al evaluar a los péptidos AMP y CPP respecto al resto de las combinaciones, se crearon conjuntos similares en cuanto longitudes y cardinalidades con las combinaciones AMP+nCPP (AMnCPP), nAMP+CPP (nAMCPP) y nAMP+nCPP (nAMnCPP). El procedimiento de búsqueda de evidencia se llevó a cabo de la misma manera variando únicamente la construcción de los conjuntos, las cuales se muestran a continuación.

3.3.5.1. Conjunto nAMP

Se registró la proporción de longitudes de los AMPs que salieron de inprot con la finalidad de que el conjunto nAMP se comportara de manera similar respecto a la longitud de las secuencias. Para ello se calculó la frecuencia de longitudes de 10 a 30AA, estas cantidades pueden verse en el Anexo . Respetando la cardinalidad por longitud se seleccionaron aleatoriamente las secuencias clasificadas como nAMP por inprot, quedando un archivo del mismo tamaño que el de AMPs. Posteriormente se realizó la evaluación de actividad CPP tal como se hizo para los AMPs.

3.3.5.2. Conjunto nCPP

Una vez ya obtenidos los conjuntos AMP y nAMP, para completar las combinaciones AMP+nCPP y nAMP+nCPP, era necesario de estos conjuntos sacar las intersecciones con los nCPP. Para ello, de los péptidos evaluados como AMP o nAMP dependiendo del caso, se escogieron las 500 secuencias con menor probabilidad promedio de ser CPP. Quedando al final 500 péptidos AMP+nCPP, 500 nAMP+CPP y 500 nAMP+nCPP.

Capítulo 4. Resultados

Para cumplir con los objetivos enunciados en el Capítulo 1, se abordaron tres metodologías. Cada una con su propio objetivo; 1) elegir el o los mejores predictores de CPPs que junto con 2) la prueba de concepto de la hipótesis “AMP que es CPP será AIP“ permitiría 3) la predicción de AMCPPs en proteomas de organismos longevos que tendrán una alta probabilidad de ser AIPs. Los resultados de estas partes se presentan en las secciones siguientes.

Antes de comenzar con la búsqueda de péptidos que cumplieran con las propiedades de AMP y CPP, se evaluaron los péptidos inductores de autofagia previamente reportados con diversos predictores de CPPs y CAMPred (predictor de AMPs), algunos predictores no fueron capaces de evaluar algunas secuencias dadas las restricciones impuestas por los predictores, estos casos están marcados con un guión en las tablas 9 y 10. Si todos los predictores rechazaron la secuencia, esta se omitía. No obstante, para algunos casos existe evidencia experimental de sus actividades tanto antimicrobiana como penetrante, independientemente de la predicción *in sílico* (Shoji-Kawata *et al.*, 2013).

Tabla 8. Secuencia de los péptidos conocidos AIPs.

Nombre de la secuencia	Secuencia
Tat-Beclin 1 D11 (Tat-D11)	RRRQRRKKRGYGGDHWIHFTANWV
Tat-Beclin 1 L11S (Tat-L11S)	YGRKKRRQRRRGGNWAWHDFVHIT
PepFect14 (PF14)	AGYLLGKLLOOLAAAALOOLL
Tat-Beclin 1	YGRKKRRQRRRGGTNVFNATFEIWHDGEFGT

Tabla 9. Evaluación *in silico* de los péptidos inductores de autofagia como CPPs. Se presenta la probabilidad de ser CPP. Se presentan los presentan los resultados obtenidos con los predictores evaluados en la sección 3.1.2; CPPs CPPP, C2Pred, SkipCPP, MLCPP y KELM. En el último se utilizaron los descriptores de composición de aminoácidos (ACC), composición de dipéptidos (DAC) y con pseudo aminoácidos (PSE).

Nombre de la secuencia	CPPP	C2Pred	SkipCPP	MLCPP	KELM		
					ACC	DAC	PSE
Tat-Beclin 1 L11	0.62	0.98	0.63	0.78	1	1	1
Tat-Beclin 1 L11S	0.73	0.97	0.72	0.82	1	1	1
PepFect14	0.53	-	-	-	0	1	1
Tat-Beclin 1 D11	0.49	-	0.55	0.45	1	1	1

Tabla 10. Evaluación in silico de los péptidos inductores de autofagia como AMPs. Se presenta la probabilidad de ser AMP con los modelos de máquina de soporte vectorial (SVM), bosques aleatorios (RF) y análisis discriminante (DA).

Nombre de la secuencia	SVM	RF	DA
Tat-Beclin 1 L11	0.58	0.53	0.92
Tat-Beclin 1 L11S	0.86	0.60	0.98
Tat-Beclin 1 D11	0.77	0.33	0.11

4.1. Evaluación de predictores de péptidos penetradores de células

En esta sección se presenta un estudio comparativo entre los modelos y conjuntos de datos para predicción de CPPs disponibles en la literatura.

Este estudio se compone de: i) el análisis de la diversidad de las secuencias que forman cada clase (CPP y nCPP) en cada conjunto de datos, ii) un análisis del desempeño de los 14 modelos de predicción de CPPs seleccionados (ver sección 3.1.2), iii) un análisis de la competitividad de los predictores disponibles junto con sus conjuntos de entrenamiento y prueba en cuanto parentesco entre elementos positivos y negativos y finalmente, iv) un análisis de la separabilidad de los conjuntos de datos para el entrenamiento de modelos de predicción de CPPs.

4.1.1. Filtrado de conjuntos de datos

Para que todos los predictores evaluaran los mismos péptidos y así permitir una evaluación justa, se filtraron los conjuntos de datos a utilizar a manera que todos los péptidos seleccionados cubrieran los requerimientos de todos los predictores. Las cardinalidades originales y las resultantes del filtrado pueden verse en la Tabla 11. En esta tabla se puede observar que, en promedio, los casos positivos se redujeron en un 12 % y los negativos en un 20 %.

Tabla 11. Cardinalidades para cada clase (Pos para CPPs y Neg para nCPPs) en los conjuntos empleados

Nombre del conjunto	Cardinalidad original		Cardinalidad después del filtrado	
	Pos	Neg	Pos	Neg
Hällbrick	53	16	49	14
Hansen	66	19	57	19
Dobchev	77	24	72	16
Sanders	111	111	96	95
HoltonTr	74	100	63	83
HoltonInd	47	47	38	39
GautamTr	99	99	98	50
GautamInd	187	187	156	120
Chen	111	34	94	28
Tang	411	411	396	393
Wei	649	549	540	535
CPP924	464	462	384	394
PandeyTr	96	96	95	47
PandeyInd	413	413	408	365
ManavalanTr	427	854	392	711
ManavalanInd	311	311	205	238

4.1.2. Desempeño de predictores

Siguiendo la hipótesis manejada a lo largo de este trabajo, la predicción de péptidos inductores de autofagia depende de la predicción de péptidos penetradores y péptidos antimicrobianos. Resulta necesario examinar el desempeño que los predictores disponibles logran y así hacer la elección del mejor predictor para CPPs.

En total existen siete predictores en línea (CPPP, C2Pred, DCF, CPPRF, SkipCPP, KEMLCpp y MLCPP), un diseñador de péptidos penetradores de células (CPPD) y un predictor de subcadenas penetrantes de un péptido (CPPFL).

Los resultados publicados en cuanto a las medidas de sensibilidad (Sen), especificidad (Spc) y coeficiente de correlación de Matthews (MCC), según disponibilidad del artículo donde fue publicado se muestran en la Tabla 12 bajo la etiqueta Publicado (P), y ya que muchos de los trabajos prestan especial atención a la exactitud (Acc) (Hällbrink y Karelson, 2005; Hansen *et al.*, 2008a; Gautam *et al.*, 2012; Holton *et al.*, 2013), esta fue agregada también. Sin embargo, ACC beneficia a la clase de cardinalidad mayor, entre mayor sea la diferencia, mayor será este sesgo. Dado que varios de

los conjuntos tienen desbalanceo de cargas, esta métrica favorecería a los conjuntos como Dobchev, donde clasificando a todos como CPP se alcanzan ACC>0.7. Por este motivo, esta medida no es considerada en análisis posteriores.

Podemos comparar los resultados de la tabla anterior con los obtenidos al replicar estos usando cada predictor con el conjunto de datos con el que fue entrenado, evaluación que se presenta también en la Tabla 12 bajo la etiqueta Reportado (R). En esta observamos que en general las medidas se mantienen similares en a lo publicado, siendo por lo regular mejores que las reportadas originalmente en cada trabajo.

Tabla 12. Sensibilidad (Scp), especificidad (Spc), coeficiente de correlación de Matthews (MCC) y certeza (Acc) publicado (P) u obtenidos al replicar la evaluación (R) de los diferentes predictores contemplados con sus conjuntos de entrenamiento correspondientes.

Predictor	**Sen**		**Spc**		**MCC**		**Acc**	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**
CPPD	0.92	1.00	0.94	0.95	0.86	0.96	0.93	0.98
CPPP	NR	0.81	NR	0.46	0.54	0.28	NR	0.84
C2Pred	0.97	0.98	0.77	0.96	0.75	0.94	0.92	0.97
DCF	NR	0.93	NR	0.88	NR	0.81	NR	0.91
CPPRF	0.91	1.00	0.93	0.76	0.831	0.78	0.92	0.88
KELM(AAC)	0.81	0.78	0.92	0.96	0.73	0.75	0.91	0.87
KELM(DAC)	0.81	0.99	0.89	1.00	0.71	0.99	0.91	0.87
KELM(Pse)	0.86	1.00	0.88	1.00	0.73	1.00	0.92	0.87
KELM(AACH)	0.82	0.87	0.92	1.00	0.74	0.87	0.92	0.87
KELM(DACH)	0.81	0.99	0.89	1.00	0.71	0.99	0.91	0.87
KELM(PseH)	0.86	1.00	0.88	1.00	0.73	0.99	0.92	0.87
SKIP	0.89	1.00	0.93	1.00	0.81	1.00	0.91	0.94
MLCPP	0.91	1.00	0.86	0.91	0.77	0.89	0.88	0.73
CPPFL	0.92	0.18	0.91	0.89	0.825	0.05	0.912	0.88

En contraposición podemos observar en la Tabla 13 los valores promedio de las medidas Sen, Spc y MCC, obtenidos durante este análisis considerando todos los conjuntos de datos evaluados, tanto de entrenamiento como de prueba.

Tabla 13. Sensibilidad, especificidad, coeficiente de correlación de Matthews y certeza promedio obtenidos en cada predictor de forma experimental con los 17 conjuntos de datos descritos en la sección 3.1.2.2

Predictor	Sen	Spc	MCC	Acc
CPPD	0.73	0.74	0.46	0.77
CPPP	0.75	0.51	0.26	0.67
C2Pred	0.91	0.62	0.53	0.81
DCF	0.92	0.67	0.62	0.83
CPPRF	0.97	0.49	0.49	0.79
KELM(AAC)	0.68	0.74	0.38	0.70
KELM(DAC)	0.85	0.92	0.70	0.86
KELM(Pse)	0.89	0.94	0.77	0.85
KELM(AACH)	0.80	0.95	0.70	0.79
KELM(DACH)	0.84	0.94	0.76	0.86
KELM(PseH)	0.87	0.96	0.78	0.82
SKIP	0.96	0.47	0.47	0.77
MLCPP	0.96	0.60	0.59	0.83
CPPFL	0.11	0.89	0.01	0.56

En complemento, las pruebas diagnósticas Sen, Spc y MCC por conjunto de datos pueden verse en la Tabla 14.

Tabla 14. Sensibilidad (Sen), especificidad (Spc) y coeficiente de correlación de Matthews (MCC) promedio sobre los seis predictores analizados obtenidos en cada conjunto de datos de forma experimental con los predictores seleccionados.

Conjunto de datos	Sen	Spc	MCC
Hällbrick	0.84	0.68	0.52
Hansen	0.79	*0.45*	0.22
Dobchev	0.80	0.50	0.30
Sanders	0.76	0.30	0.10
HoltonTr	0.81	0.67	0.19
HoltonInd	0.85	0.76	0.63
GautamTr	*0.83*	0.85	0.70
GautamInd	0.65	0.86	0.49
Chen	0.82	0.62	0.46
Tang	0.87	0.81	0.69
Wei	0.85	0.71	0.56
CPP924	0.85	0.86	0.71
PandeyTr	0.86	0.80	0.69
PandeyInd	0.66	0.89	0.52
ManavalanTr	0.86	0.85	0.69
ManavalanInd	0.76	0.77	0.54
Promedio	0.803	0.711	0.501

Los valores de especificidad, sensibilidad y MCC de cada predictor con cada conjunto puede ser visualizado en las figuras 9, 10 y 11. El color amarillo representa una

mejor evaluación mientras que el violeta, una peor.

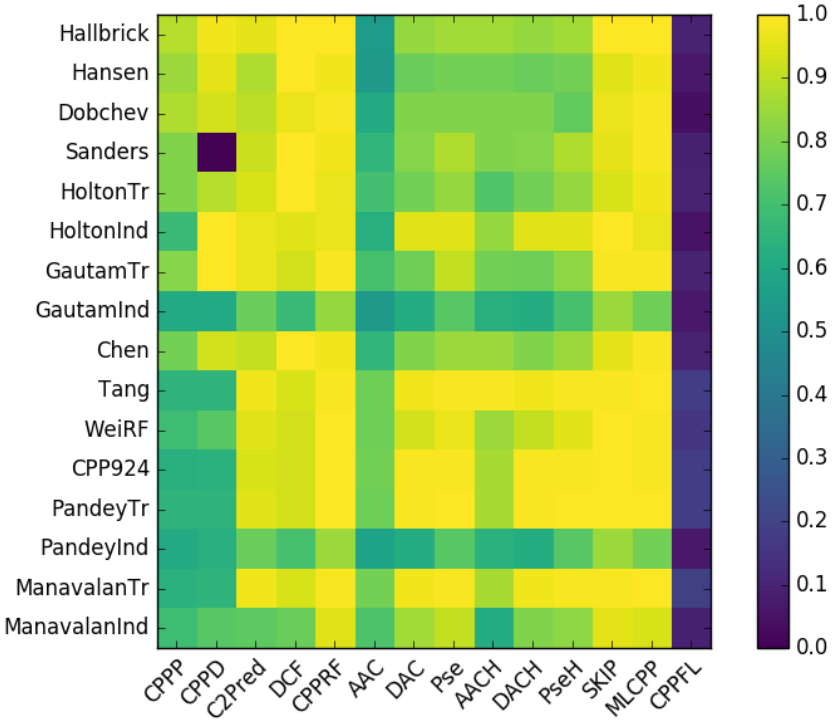

Figura 9. Mapa de calor de la sensibilidad (Sen) de los predictores con cada uno de los conjuntos de datos.

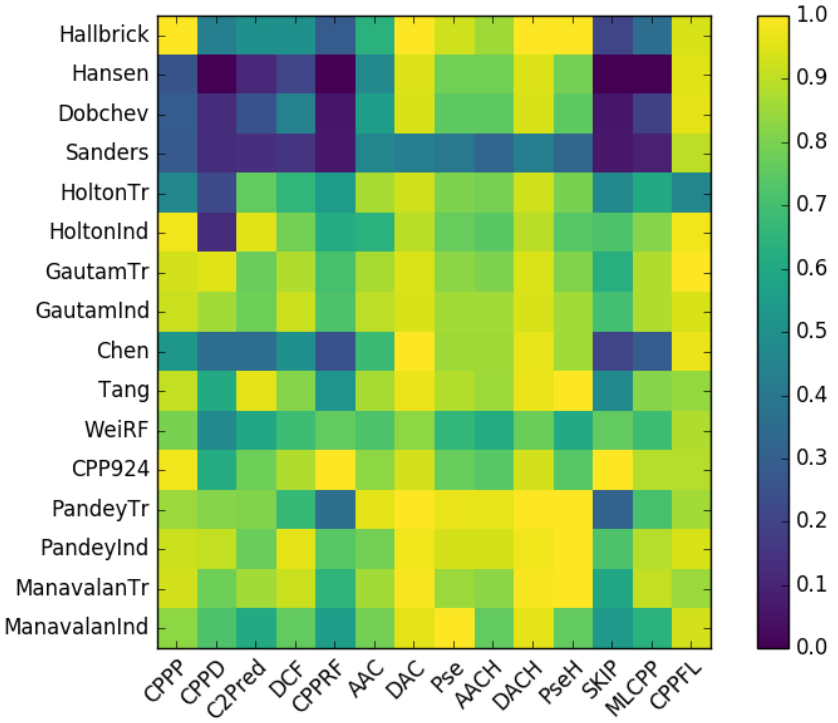

Figura 10. Mapa de calor de la especificidad (Spc) de los predictores con cada uno de los conjuntos de datos.

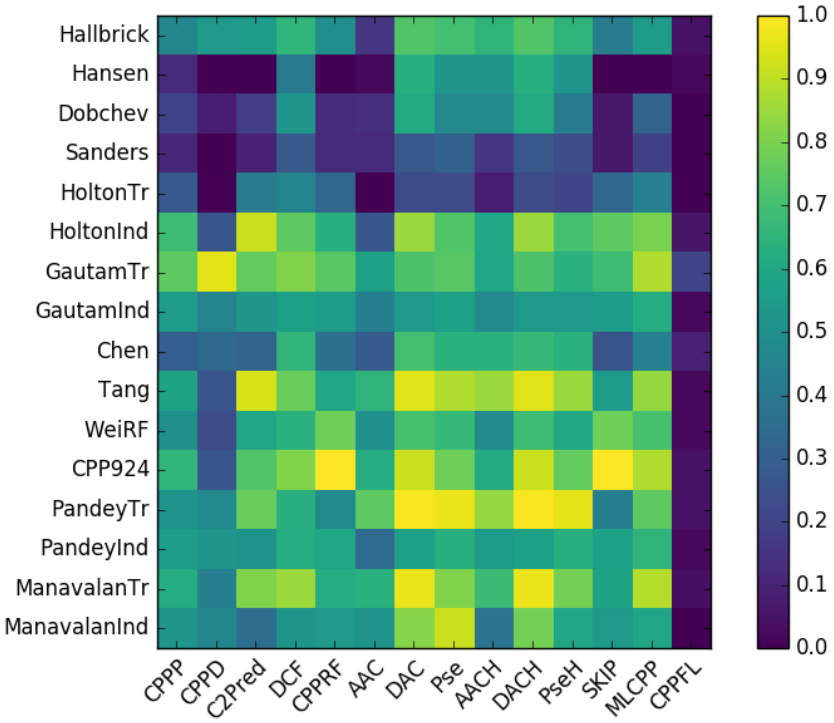

Figura 11. Mapa de calor del MCC de los predictores con cada uno de los conjuntos de datos.

4.1.3. Comparación del desempeño de los predictores de CPPs

Contemplando la medida MCC, los mejores predictores fueron KELM, DCF y MLCPP en ese orden. Podemos mencionar que los tres predictores se entrenaron utilizando datos positivos de CPPSite y de C2Pred. C2Pred diseñó sus casos negativos (nCPP) usando péptidos con actividades biológicas reportadas, con reducción de redundancia al 80 % de similitud. En consecuencia, los tres predictores mantuvieron una buena diversidad de sus conjuntos negativos.

Los predictores con mayor especificidad fueron KELM, CPPP y CPPFL. Ambos utilizaron una técnica similar a la empleada en la construcción del conjunto de C2Pred para formar sus conjuntos negativos, con la diferencia de que ambos contienen los 34 péptidos nCPP validados experimentalmente.

Con un promedio de 0.92, KELM con el modelo DAC fue el mejor predictor en especificidad, sin embargo, es importante resaltar dos cosas: KELM entrenó con la totalidad de los péptidos validados experimentalmente como negativos y tuvo un resultado de

0.43 en Sanders, conjunto que utilizó muestreo con reemplazo de los 34 negativos experimentalmente validados, aún así, este predictor obtuvo el valor más alto para dicho conjunto.

En conjuntos como Tang, Wei y CPP924 donde los predictores que los emplearon realizaron una validación cruzada de diez pliegues, el conjunto de entrenamiento es 9 veces mayor al de prueba, en la Tabla 15 pueden verse la totalidad de elementos utilizados durante el entrenamiento y prueba de los predictores. El predictor que empleo la diferencia de cardinalidad más grande fue MLCPP con un conjunto de entrenamiento 18.54 veces más grande que el de prueba.

Tabla 15. Cardinalidad de los conjuntos de entrenamiento y prueba utilizados para cada uno de los predictores. .

Predictor	**CPPP**	**CPPD**	**C2Pred**	**DCF**	**CPPRF**	**KELM**	**SKIP**	**MLCPP**	**CPPFL**
Entrenamiento	174	374	739.8	374	1078.2	826	831.6	1152.9	831.6
Prueba	94	198	82.2	94	119.8	192	92.4	62.2	92.4

4.1.4. Evaluación de los conjuntos de datos

Con el fin de establecer una relación entre los casos negativos y los positivos de los conjuntos, se buscó identificar el parentesco (K). Si los casos positivos son muy diferentes a nivel de secuencia a los negativos, entonces separarlos por similitud de secuencia o cualquier descriptor basado en el mismo, será relativamente sencillo sin realmente aportar información sobre los CPPs. Si por el contrario, los conjuntos son muy similares, el conjunto obligará al predictor a discriminar los péptidos más allá de su secuencia, a lo que llamamos aquí el nivel de separabilidad que aportan.

4.1.4.1. Diversidad y parentesco K

Al analizar los casos positivos encontramos que en promedio la diversidad es de 0.75, es decir, el 75 % de las secuencias son diferentes en al menos un 80 % entre ellas. En los casos negativos en promedio tienen una diversidad de 0.92, habiendo solo un caso, Sanders, en el que la diversidad se encuentra debajo de 0.79. Las diversidades pueden verse en la Tabla 16.

El parentesco dará una idea de la similitud de los casos positivos respecto de los negativos a nivel de secuencia. Si $K = 0$ nos indicaría que los conjuntos no tienen nada en común y en teoría serían sencillos de separar.

En la Figura 12 puede verse el parentesco (ver sección 3.1.3.3) a un umbral de $\alpha = 0.7$ y la especificidad de los conjuntos, ordenados por este último criterio en orden descendente.

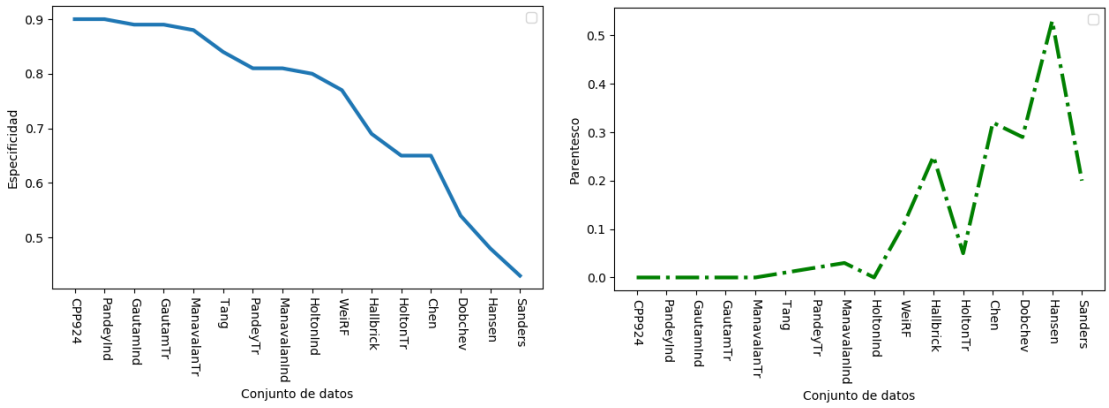

Figura 12. Comparación de especificidad (línea sólida azul) con parentesco K(α), a α =0.7 (línea discontinua verde).

Al evaluar cada par de secuencias de los 34 nCPPs validados experimentalmente se observa que el par de secuencias menos parecido mantiene una similitud del 5 %, mientras que el par más parecido llega a un 85 % de similitud. Para el conjunto Sanders estos valores son de 5 % y 95 %, respectivamente. Además, todas las secuencias en el conjunto negativo de Sanders se parecen en al menos un 80 % a al menos una secuencia de las 34 previamente mencionadas, por tanto, podemos decir que el conjunto de Sanders es similar al conjunto de los nCPPs experimentalmente validados.

Cuando se examinó el Parentesco K entre casos positivos y negativos, se buscaba la razón de la cantidad de elementos negativos representados en el conjunto positivo entre la cardinalidad del conjunto negativo a un umbral de similitud dado α. Este valor indica cierto grado de dificultad del conjunto. Entre mayor sea el parentesco, el conjunto será más difícil de separar a nivel de secuencia. En la Tabla 16 puede verse el parentesco de los conjuntos obtenidos.

Tabla 16. Tabla de las diversidades tanto de conjuntos positivos como negativos y del parentesco de los conjuntos negativos respecto a los positivos.

Conjunto de datos	Diversidad		Parentesco (%)
	Pos	Neg	
Hallbrink	0.66	0.94	25
Hansen	0.68	0.84	52.63
Dobchev	0.77	0.96	29.17
Sanders	0.68	0.32	20.18
HoltonTr	0.99	0.98	5
HoltonInd	0.72	1	0
GautamTr	0.79	1	0
GautamInd	0.52	1	0
Chen	0.68	0.91	32.35
Tang	0.97	1	0.73
Wei	0.52	0.93	10.94
CPP924	0.97	1	0
PandeyTr	0.96	0.98	1.69
PandeyInd	0.79	1	0
ManavalanTr	0.97	1	0.35
ManavalanInd	0.63	0.79	2.89

4.1.4.2. Nivel de separabilidad

El nivel de separabilidad se calculó como un apoyo al parentesco (K), donde aquellos conjuntos con un K bajo deberían tener un nivel de separabilidad alto, para ello se realizaron agrupamientos considerando únicamente la secuencia para el cálculo de distancia.

El promedio de MCC de los conjuntos evaluados con los 14 modelos diferentes fue de 0.53 y por agrupamiento de 0.33, sin embargo, hubo conjuntos en los que el MCC obtenido por aglomeración fue mayor al promedio previamente mencionado. Estos son CPP924, PandeyInd, ManavalanInd, GautamInd y ManavalanTr. El primero con un MCC de 0.57. Es importante recordar que CPP924 generó su conjunto negativo con secuencias de aminoácidos aleatorias. En la Tabla 17 se pueden apreciar los valores de Sen, Spc y MCC obtenidos en cada conjunto usando el agrupamiento a nivel de secuencia.

Tabla 17. Pruebas diagnósticas Sen, Spc y MCC de un agrupamiento espectral

Dataset	Sen	Spc	MCC
Hällbrick	0.660	0.500	0.140
Hansen	0.576	0.474	0.042
Dobchev	0.662	0.375	0.033
Sanders	0.351	0.810	0.182
HoltonTr	0.213	1.000	0.345
HoltonInd	0.333	0.739	0.079
GautamTr	0.495	0.980	0.543
GautamInd	0.406	1.000	0.505
Chen	0.667	0.324	-0.009
Tang	0.450	0.944	0.453
Wei	0.593	0.886	0.501
CPP924	0.500	0.994	0.567
PandeyTr	0.510	0.979	0.554
PandeyInd	0.538	0.852	0.411
ManavalanTr	0.473	0.964	0.534
ManavalanInd	0.511	0.977	0.552

Con estos resultados podemos concluir que conjuntos como CPP924 son fácilmente separables a nivel de secuencia.

En contraposición el peor resultado en cuanto a MCC por aglomeración fue sobre el conjunto de datos Chen con -0.01 que así como los siguientes cuatro peores resultados por aglomeración fueron los conjuntos: Hällbrick, Sanders, Hansen y Dobchev; todos ellos utilizan péptidos validados experimentalmente como nCPP. De estos, Hällbrink y Karelson (2005) obtuvieron seis secuencias provenientes de cinco proteínas del ser humano y dos aleatorias evaluadas todas como CPP y corroboradas experimentalmente. Adicionalmente, Sanders *et al.* (2011) sintetizaron cuatro péptidos predichos como CPP y dos predichos como nCPP los cuales fueron obtenidos a partir del proteoma del pollo, de los cuales tres (de cuatro) predichos como CPPs y uno de los dos predichos como nCPP fueron evaluados correctamente. Aún siendo las técnicas de aprendizaje de estos predictores menos sofisticadas que las de predictores como los de Wei *et al.* (2017b) o Wei *et al.* (2017a), mostraron ser más competitivas en general según nuestros análisis.

Vemos en general que el agrupamiento por similitud da una cota inferior para el desempeño de los predictores. Tomando esto como referencia, C2Pred y MLCPP mantienen los valores más altos mientas que Skip el más bajo. Si tomamos los valores de

MCC de cada predictor en cada conjunto y lo restamos al valor obtenido al aglomerar dicho conjunto vemos que CPPD es el que tiene una diferencia menor mientras que MLCPP realiza un trabajo más significativo que simplemente aglomerar. Con esto vemos de nuevo la presencia de MLCPP con un desempeño notable.

CPP924 es el conjunto más fácilmente separable, de parentesco 0 entre positivos y negativos y el mejor evaluado por todos los predictores, sin embargo, SkipCPP, siendo el predictor que dio origen a este conjunto tuvo el desempeño en especificidad más bajo para el resto de los conjuntos de prueba y unos de los peores MCC.

4.2. Búsqueda de posibles péptidos inductores de autofagia en proteomas de mamíferos

Las especies de mamíferos encontradas en AnAge están distribuidas en un total de 11 órdenes y 224 géneros. El orden con más representantes es Rodentia con 115 especies, mientras que Hyracoidea, Cetacea, Tubulidentata, Perissodactyla y Proboscidea cuentan con solo un representante. Los órdenes más representados son Tubulidentata, Monotremata y Sirenia con el 100 %, el 60 % y el 50 % de todas las especies vivas conocidas; el menos representado, Chiroptera, cuenta con solo una de las 32 especies conocidas respectivamente. El error cuadrático medio de la ecuación 1 fue de 84.503 y para la ecuación 2 de 100.469, por tanto, durante el resto del trabajo se utilizó la primera, cuya regresión resultante se presenta en la Figura 13. En el eje horizontal vemos la masa del animal en gramos y en el vertical la edad máxima registrada en años.

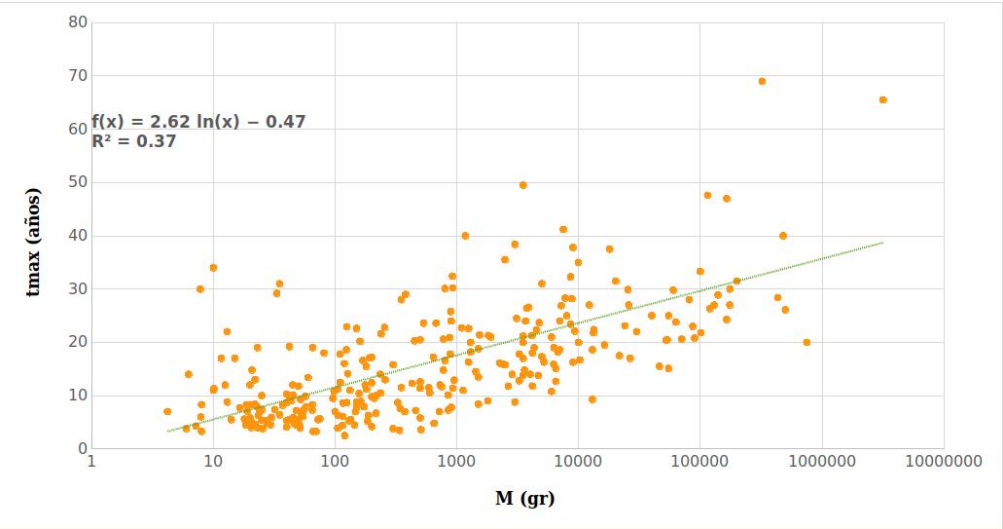

Figura 13. Comparación de $t'_{max} = -0.4742 + 2.176 * \ln M$ contra t_{max} reportado de las especies de mamíferos en AnAge (De Magalhães y Costa, 2009) en años respecto a M en gramos.

Durante este proceso se seleccionaron especies de longevidad alta y baja. Los 20 mamíferos con mayor coeficiente de longevidad pueden verse en la Tabla 18. En contraparte las 20 especies con los coeficientes más bajos pueden verse en la Tabla 19. Para ambas tablas puede verse la cantidad de proteínas totales en su proteoma en caso de que estuvieran disponible.

Tabla 18. Mamíferos de AnAge ordenados por ct_{max} descendente con la cantidad de proteínas disponibles.

Orden	Familia	Nombre científico	$c_0 t_{max}$	$c_1 t_{max}$	#Proteínas
Chiroptera	Vespertilionidae	*Myotis lucifugus*	4.90	3.90	43106
Primates	Hominidae	*Homo sapiens*	4.64	3.96	113620
Chiroptera	Vespertilionidae	*Plecotus auritus*	4.49	3.70	-
Rodentia	Bathyergidae	*Heterocephalus glaber*	3.69	2.60	41961
Chiroptera	Phyllostomidae	*Desmodus rotundus*	3.50	2.48	29845
Chiroptera	Miniopteridae	*Miniopterus schreibersii*	3.04	2.34	-
Monotremata	Tachyglossidae	*Tachyglossus aculeatus*	2.91	2.10	-
Chiroptera	Pteropodidae	*Pteropus giganteus*	2.78	1.92	-
Cingulata	Dasypodidae	*Tolypeutes matacus*	2.46	1.71	-
Chiroptera	Vespertilionidae	*Eptesicus fuscus*	2.41	1.75	49822
Primates	Cheirogaleidae	*Cheirogaleus medius*	2.39	1.61	-
Chiroptera	Phyllostomidae	*Macrotus californicus*	2.39	1.87	-
Chiroptera	Pteropodidae	*Pteropus rodricensis*	2.34	1.58	-
Primates	Lorisidae	*Perodicticus potto*	2.34	1.60	-
Carnivora	Procyonidae	*Potos flavus*	2.31	1.65	-
Chiroptera	Phyllostomidae	*Carollia perspicillata*	2.30	1.74	-
Chiroptera	Pteropodidae	*Rousettus aegyptiacus*	2.24	1.51	48803
Chiroptera	Phyllostomidae	*Artibeus jamaicensis*	2.22	1.55	-
Primates	Aotidae	*Aotus trivirgatus*	2.22	1.52	-
Pilosa	Megalonychidae	*Choloepus hoffmanni*	2.21	1.63	-

Tabla 19. Mamíferos de AnAge ordenados por ct_{max} ascendente con la cantidad de proteínas disponibles.

Orden	Familia	Nombre científico	ct_{max0}	ct_{max1}	#Proteínas
Peramelemorphia	Peramelidae	*Echymipera rufescens*	0.21	0.15	-
Didelphimorphia	Didelphidae	*Chironectes minimus*	0.23	0.16	-
Didelphimorphia	Didelphidae	*Lutreolina crassicaudata*	0.24	0.16	-
Didelphimorphia	Didelphidae	*Metachirus nudicaudatus*	0.24	0.16	-
Rodentia	Cricetidae	*Arvicola amphibius*	0.25	0.17	-
Afrosoricida	Tenrecidae	*Hemicentetes semispinosus*	0.25	0.17	-
Soricomorpha	Soricidae	*Crocidura flavescens*	0.25	0.18	-
Soricomorpha	Talpidae	*Condylura cristata*	0.28	0.19	29166
Didelphimorphia	Didelphidae	*Didelphis marsupialis*	0.28	0.20	-
Soricomorpha	Soricidae	*Blarina brevicauda*	0.28	0.21	-
Rodentia	Cricetidae	*Cricetus cricetus*	0.28	0.19	-
Rodentia	Caviidae	*Galea musteloides*	0.29	0.20	-
Rodentia	Muridae	*Rattus norvegicus*	0.33	0.22	66876
Rodentia	Cricetidae	*Lemmus lemmus*	0.35	0.24	-
Didelphimorphia	Didelphidae	*Philander opossum*	0.35	0.24	-
Rodentia	Cricetidae	*Dicrostonyx groenlandicus*	0.36	0.24	-
Rodentia	Spalacidae	*Cannomys badius*	0.37	0.25	-
Soricomorpha	Soricidae	*Suncus murinus*	0.37	0.25	-
Dasyuromorphia	Dasyuridae	*Dasyurus maculatus*	0.37	0.27	-
Rodentia	Muridae	*Rattus rattus*	0.38	0.26	-

En total, seis especies de las identificadas anteriormente tuvieron proteoma disponible en NCBI, de las cuales *Myotis lucifugus*, *Heterocephalus glaber*, *Desmodus rotundus* y *Eptesicus fuscus* son de longevidad alta y *Rattus norvegicus* y *Condylura cristata* de longevidad baja, correspondientes a 3 órdenes diferentes.

Finalmente, al agregar la especie de longevidad media, Rodentia fue el único orden que contó con al menos una especie con un proteoma con suficientes proteínas para realizar el análisis, agregando a *Cavia porcellus* al análisis. Quedando entonces siete especies seleccionadas con alrededor de 42,591 proteínas en promedio y con 12,028 de desviación estándar las cuales pueden ser apreciadas en la Tabla 20.

Tabla 20. Tipo de longevidad, posición relativa a las 349 especies contempladas por ct_{max}, orden, familia, especie y cantidad de proteínas disponibles de los animales seleccionados.

Logevidad	Posición	Orden	Especie	#Proteinas
Alta	1	Chiroptera	Myotis lucifugus	43,106
Alta	4	Rodentia	Heterocephalus glaber	41,961
Alta	5	Chiroptera	Desmodus rotundus	29,845
Alta	10	Chiroptera	Eptesicus fuscus	49,822
Media	181	Rodentia	Cavia porcellus	37,360
Baja	337	Rodentia	Rattus norvegicus	66,876
Baja	342	Soricomorpha	Condylura cristata	29,166

4.2.1. AMPs, CPP y su posible relación con autofagia

Las 287 especies de mamíferos con longevidad y peso disponibles en AnAge (De Magalhães y Costa, 2009) tenían coeficientes de longevidad (véase sección 3.3.1) distribuidos entre 0.25 y 7.51 con una media de 1.26 y desviación estándar de 0.86. Las especies seleccionadas se mostraron en la Tabla 20 junto con las proteínas disponibles en cada proteoma.

Al realizar la proteólisis *in silico* y evaluar los péptidos de longitud 10 a 30 derivados de todas las proteínas de estos organismos y predecir su actividad como AMPS, encontramos al rededor de 20 mil posibles péptidos como antimicrobianos. En el Apéndice pueden verse detalles de la distribución de las longitudes de los péptidos en cada una de las especies y de la proporción que implican en la totalidad de los péptidos de longitud entre 10 y 30. Sin embargo, se puede mencionar que a pesar de que *Rattus norvegicus* fue la especie del proteoma más amplio (66,876 proteínas contra 49,822 del segundo más extenso de *Eptesicus fuscus*) fue quien obtuvo menor cantidad de péptidos por una diferencia de casi 6 mil péptidos contra el segundo lugar *Condylura cristata*. En cuanto a la distribución de las longitudes de los péptidos, se observa que las proporciones se mantienen de forma muy similar entre especies, siendo la longitud de 11 AA la más frecuente y la de 30 la menos frecuente (Anexo).

Al evaluar a los péptidos predichos como AMPs, por actividad penetrante se obtienen entre 200 y 600 AMCPPs por especie, como puede verse en la Tabla 7. Es importante mencionar que las proteínas en los respectivos *OP*, es decir, las proteínas del proteoma original con $\gamma \geq 1$, donde γ de una proteína es la cantidad de péptidos que son subcadenas de ella, en su mayoría poseen un $\gamma > 1$.

Con el fin de realizar análisis futuros, se tomaron las proteínas de *OP*, cuya cantidad puede verse en la tercera columna de la Tabla 21, mientras que en la cuarta puede encontrarse la cantidad de estas proteínas con $\gamma > 1$. Cabe señalar que la relación péptido-proteína no es única, una secuencia de péptido puede ser subsecuencia de más de una proteína y una proteína podría contener a más de una secuencia de péptido.

Tabla 21. Cantidad de proteínas en *OP* y proteínas en *OP* con $\gamma > 1$

Especie	AMCPPs	Proteínas en OP	Proteínas en OP con $\gamma > 1$
Myotis lucifugus	583	1562	171
Heterocephalus glaber	632	1495	190
Desmodus rotundus	476	939	101
Eptesicus fuscus	662	2063	54
Cavia porcellus	658	1492	181
Rattus norvegicus	404	1285	25
Condylura cristata	232	465	8

Como una propuesta para validar que los péptidos pudieran tener actividad inductora de autofagia, se cribaron las proteínas de la base de datos AutophagyDatabase (Homma *et al.*, 2011), cuando se evaluaban únicamente las proteínas de *OP* y cuando se tomaban las proteínas de *OP* con $\gamma > 1$. Cuando se examinaron los péptidos AMCPP$_r$s, se encontró una diferencia de hasta el 49 % (*Heterocephalus glaber*) en favor de los AMCPP de *OP* con $\gamma > 1$. Esto puede verse en la Tabla 22, donde se separan los resultados cuando se toman los dos conjuntos de proteínas mencionados.

Tabla 22. Cantidad de AMCPPs (AMCPPs), proteínas en OP (OP), proteínas OP con presencia en AutohagyDatabase (*OP*$_r$), los AMCPPs que pertenecen a esas proteínas (AMCPP$_r$s), el porcentaje correspondiente de proteínas *OP*$_r$ respecto *OP* (% *OP*$_r$) y de AMCPP$_r$s respecto AMCPPs (% AMCPP$_r$s)

Proteías de *OP*

Especie	AMCPPs	*OP*	*OP*$_r$	AMCPP$_r$	% *OP*$_r$	% AMCPP$_r$s
Myotis lucifugus	583	1562	317	98	20.29	16.81
Heterocephalus glaber	632	1495	315	136	21.07	21.52
Desmodus rotundus	476	939	189	110	20.13	23.11
Eptesicus fuscus	662	2063	427	141	20.70	21.30
Cavia porcellus	658	1492	332	54	22.25	8.21
Rattus norvegicus	404	1285	286	89	22.26	22.03
Condylura cristata	232	465	107	47	23.01	20.26

Proteínas de *OP* con $\gamma > 1$

Especie	AMCPPs	*OP*	*OP*$_r$	AMCPP$_r$	% *OP*$_r$	% AMCPP$_r$s
Myotis lucifugus	80	171	48	16	28.07	20.00
Heterocephalus glaber	55	190	43	39	22.63	70.91
Desmodus rotundus	42	101	20	18	19.80	42.86
Eptesicus fuscus	114	254	14	30	5.51	26.32
Cavia porcellus	98	181	89	54	49.17	55.10
Rattus norvegicus	25	148	7	14	4.73	56.00
Condylura cristata	17	24	2	4	8.33	23.53

4.2.2. Comparación por actividades

Al realizar la búsqueda de evidencia de los péptidos en las diferentes combinaciones de actividades, se aprecia una caída en la proporción de los péptidos seleccionados

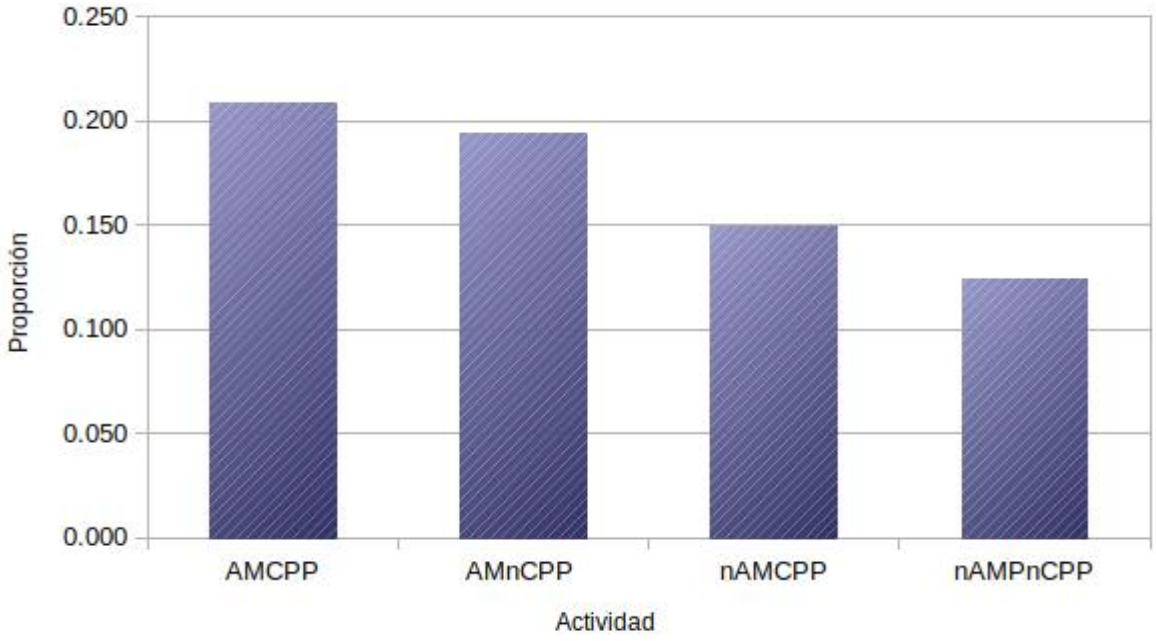

Figura 14. Proporción de péptidos$_r$ de proteínas *OP* con $\gamma > 0$ para las diferentes combinaciones de actividades AMP y CPP.

con dicha evidencia (AMCPP$_r$s entre la cantidad de AMCPPs o péptidos$_r$ dividido entre la cantidad de péptidos con las propiedades bajo estudio). Mientras que para AMP+CPP se observa que un promedio de 20 % de los péptidos elegidos son AMCPP$_r$, 19 % de los nAMCPP son nAMCPP$_r$ y este número continúa bajando como se ve en la Figura 14.

Adicionalmente, al utilizar sólo péptidos *OP* con $\gamma > 1$, se observa que mientras que las proporciones de péptidos AMCPP$_r$ se duplica, a medida que las actividades AMP y CPP desaparecen, el aumento de la relación disminuye como puede verse en la Figura 15. Es decir, escoger sólo péptidos de proteínas *OP* con $\gamma > 1$ solo mejora el porcentaje de proteínas$_r$ cuando estas son AMCPPs.

Finalmente, puede verse en la Tabla 23 cómo los AMCPPs presentan un mayor porcentaje de AMCPP$_r$s, en especial si se seleccionan péptidos de *OP* con $\gamma > 1$.

Tabla 23. Proporciones máximas, mínimas y promedio que presentaron todas las especies al tomar las diferentes combinaciones de actividades

	Todos			**Más de un péptido**		
	Min	**Max**	**Promedio**	**Min**	**Max**	**Promedio**
AMCPP$_r$	0.168	0.231	0.208	0.200	**0.709**	0.421
AMCPP$_r$	0.154	**0.238**	0.194	0.061	0.398	0.243
nAMCPP$_r$	0.114	0.198	0.149	0.056	0.229	0.132
nAMnCPP$_r$	0.108	0.142	0.124	0.067	0.190	0.129

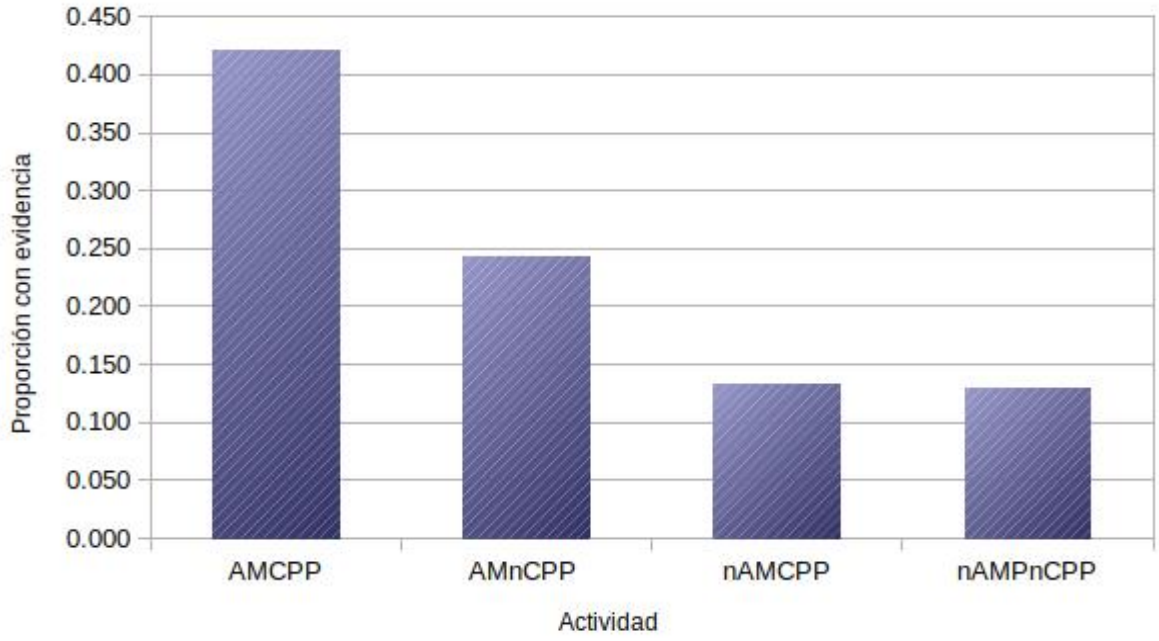

Figura 15. Proporción de péptidos$_r$ de proteínas *OP* con $\gamma > 1$ para las diferentes combinaciones de actividades AMP y CPP.

4.3. Identificación de inducción de autofagia de forma experimental

En esta sección se presentan los resultados de los análisis experimentales. Se hizo uso de un espectrofotómetro y un citómetro de flujo para la búsqueda de diferencias entre células en autofagia basal y células en autofagia inducida. La tercera parte corresponde a la microscopía de alta resolución. En los primeros dos casos no se lograron obtener diferencias entre los controles definidos en la sección 3.2.2.3, sin embargo, en el tercer análisis se encontró una diferencia en la morfología de las células en autofagia basal respecto a las tratadas con Rapamicina.

4.3.1. Selección de péptidos

Los 21 péptidos reportados en la literatura con actividad tanto antimicrobiana como penetradora de células experimentalmente validada pueden verse en la Tabla 24. De estos, con excepción de D-SynB1 y D-SynB3 todos son de quiralidad L (dado que las enzimas de eucariotes solo reconocen conformaciones L, la mayoría de los péptidos naturales son de esta quiralidad), 10 de los péptidos provienen de Bactenecin 7, 5 de

Protegrin y el resto de origen variado.

Tabla 24. Péptidos obtenidos por la base de datos CPPSite como AMP y CPPs.

ID	Secuencia	Quiralidad	Nombre	Origen
1319	RRIRPRP	L	Bac1-7	Bactenecin 7
1320	RRIRPRPPRLPRPRP	L	Bac-1-15	Bactenecin 7
1321	RRIRPRPPRLPRPRPRPLPFPRPG	L	Bac1-24	Bactenecin 7
1322	RRIRPRPPRLPRPRPRP	L	Bac1-17	Bactenecin 7
1323	PRPPRLPRPRPRPLPFPRPG	L	Bac5-24	Bactenecin 7
1324	PPRLPRPRPRPLPFPRPG	L	Bac7-24	Bactenecin 7
1325	RLPRPRPRPLPFPRPG	L	Bac9-24	Bactenecin 7
1326	PRPRPRPLPFPRPG	L	Bac11-24	Bactenecin 7
1327	PRPRPLPFPRPG	L	Bac13-24	Bactenecin 7
1328	PRPLPFPRPG	L	Bac15-24	Bactenecin 7
1380	RAGLQFPVGRVHRLLRK	L	Buforin-II	de Bufo bufo
1431	ALWMTLLKKVLKAAAKAALNAVLVGANA	L	Dermaseptin S4	Dermaseptinas
1432	ALWKTLLKKVLKA	L	S4(13)	Dermaseptina S4
1480	RGGRLSYSRRRFSTSTGR	L	SynB1	Protegrin
1481	rggrlsysrrrfststgr	D	D-SynB1	Protegrin
1482	RRLSYSRRRF	L	SynB3	Protegrin
1483	rrlsysrrrf	D	D-SynB3	Protegrin
1484	RGGRLAYLRRRWAVLGR	L	SynB5	Protegrin
1489	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTESC	L	LL-37	de catelina humana
1492	GIGKFLHSAKKWGKAFVGQIMNC	L	MG2d	Análogo de Magainin 2
1493	TRSSRAGLQWPVGRVHRLLRKGGC	L	BF2d	Análogo a Buforin 2

En la literatura se encontró el péptido Bradykinn con actividad antimicrobiana (Kowalska *et al.*, 2002) y sin actividad penetrante de células (Hällbrink y Karelson, 2005), ambas propiedades validadas experimentalmente. Del resto de las categorías no se encontraron péptidos validados experimentalmente. A aquellos péptidos que les faltaba solo una actividad por reportar se les aplicó el predictor CAMPred (Porto *et al.*, 2012) para AMP y MLCPP (Manavalan *et al.*, 2018) para CPP. Los péptidos seleccionados se muestran en la Tabla 25.

Tabla 25. Péptidos con las diferentes combinaciones de actividades AMP y CPP.

Tipo	Nombre	Actividad AMP	Referencia AMP	Referencia CPP
AMCPP	Bradykinn	AntiGram+ AntiGram- AntiFungal	(Kowalska *et al.*, 2002)	(Hällbrink y Karelson, 2005)
NAMCPP	-	-	-	-
NAMCPP	Inv10	-	Predicho	(Lu *et al.*, 2006)
	Bip16	-	Predicho	(Gomez *et al.*, 2010)
	CRGDK	-	Predicho	(Wei *et al.*, 2013)

Los péptidos seleccionados y mandados a sintetizar por ser AMCPP fueron Bac1-15, Bac15-24, MG2d y SynB. Por su actividad AMnCPP se sintetizó el péptido Bradikinn y como nAMCPP los péptidos Bip16, CRGDK e Inv10. Los primeros cinco fueron sintetizado por la empresa GenScript y los tres restantes por AnaSpec. Todos los péptidos, fuese *in silico* o *in vitro*, mostraron solubilidad en agua por lo que los péptidos sintetizados

se disolvieron en agua desionizada.

4.3.1.1. Evaluación de autofagia inducida por péptidos en células VMA1-TDIMER2 según morfología.

De cada uno de los experimentos se obtuvo un número variable de campos por muestra dada la experimentación, en total se lograron capturar un total de 192 imágenes.

Con el objetivo de determinar la tasa de cambio de fluorescencia en células de diferente tratamiento en respuesta a la autofagia asociada a la proteína VMA1, se realizó una evaluación de manera manual seleccionando las células con base en su morfología.

Contabilización. En las imágenes resultantes de células en autofagia basal se encuentra una proporción de dos células en morfología palomita (ver sección 3.2.2.6) por cada una en morfología de aro, mientras que en aquellas tratadas con Rapamicina esto se invertía, por cada tres células, dos eran de morfología de aro y una de palomita.

Con esta información se tomaron las proporciones de aros de cada una de las imágenes y estas se agruparon por tratamiento, generando el diagrama de cajas de la Figura 16 donde se presentan con * y (*) si presentan diferencias significativas respecto el control negativo y el control positivo pepídico IP-1 respectivamente.

Al aplicarse el análisis Mann-Whitney sobre los porcentajes de aros en cada uno de los 192 campos tomados se encontró una diferencia significativa de los tratamientos Rapamicina, IP-1, Bac1-15, Bradikin, Inv10 y MG2d respecto al control negativo. Rapamicina e IP-1 corresponden a los controles positivos. Los valores de U para cada uno de los tratamientos respecto al control negativo puede encontrarse en la Tabla 26, Rapamicina fue significativamente diferente al resto de controles y tratamientos. Ip-1 además de su diferencia con los controles, únicamente presentó diferencias con SynB y Bip16.

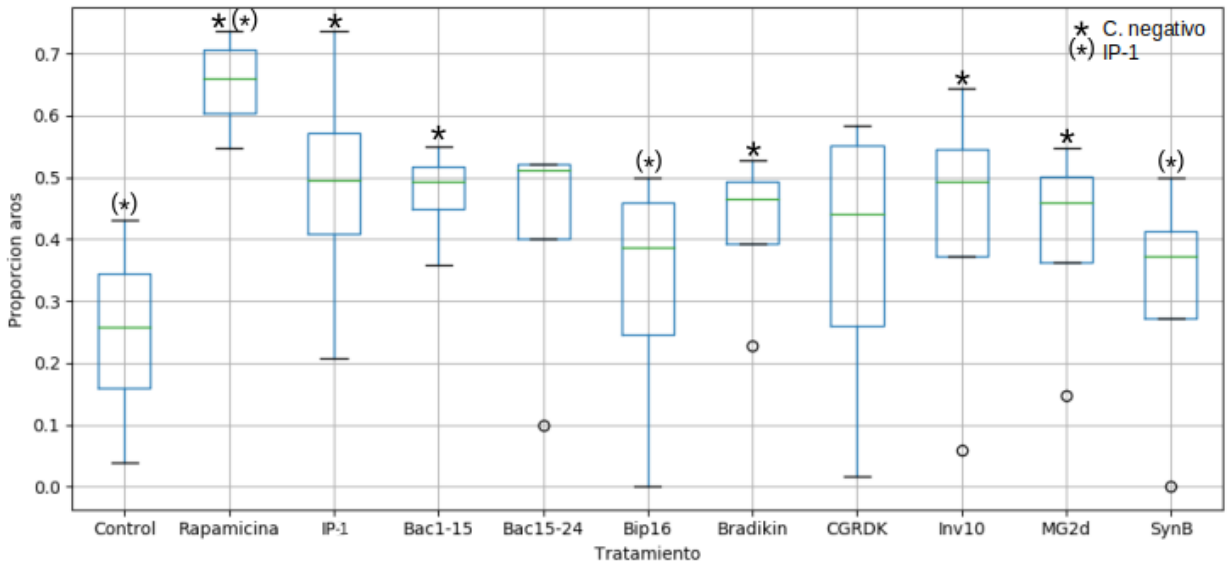

Figura 16. Diagrama de cajas del porcentaje de aros encontrados en cada tratamiento. Los tratamientos que presentan diferencias significativas con el análisis Mann-Whitney respecto al control negativo están señalados con un * y los que presentan diferencias respecto IP-1 con un (*) a una $P<0.05$.

Tabla 26. Valores de U y valores críticos para cada uno de los tratamientos respecto el control negativo. El valor de n corresponde a la cantidad de campos del tratamiento especificado. Se tomaron 9 campos en total para el control negativo.

Tratamiento	n	U	Valor Crítico	Resultado
Rapamicina	6	0	14	Diferente
IP-1	12	18	37	Diferente
Bac1-15	4	2	7	Diferente
Bac15-24	6	20	14	
Bip16	10	35	29	
Bradikin	8	12	22	Diferente
CGRDK	10	33	29	
Inv10	11	32	33	Diferente
MG2d	10	20	29	Diferente
SynB	9	35	26	

Con esto, al separar a los péptidos por actividad se observa que:

AMCPPs. Bac1-15, Bacc15-24 y MG2d fueron tratamientos que no presentaron diferencias respecto IP-1 (que fue significativamente diferente al resto de los controles), siendo tres de los cuatro AMCPPs elegidos. Sin embargo, de ellos, Bac15-24 junto con SynB no presentaron diferencias respecto el control negativo. Por tanto, es posible que Bac1-15 y MG2d tengan efectos similares a IP-1 y Bac15-24 efectos atenuados mientras que SynB no tenga efecto alguno.

AMnCPP. Solo Bradikinn pertenece a esta clase, teniendo diferencias respecto el control más no con IP-1, por tanto, se espera de este péptido lo mismo que de Bac1-15 y MG2d.

nAMCPP Los tres péptidos sintetizados tenían su actividad AMP evaluada *in silico* y de ellos se encontró que Inv10 presentaba toxicidad y se comportó como los AMCPPS Bac1-15 y MG2d, mientras que Bip16 y CGRDK no presentaron diferencias con el control negativo. Bip16 incluso queda como un no inductor al tener diferencias con IP-1 mientras que CGRDK podría tener efectos similares a Bac15-24.

4.3.2. Prueba de toxicidad

Una vez terminadas las 24 horas en el espectrofotómetro, se analizó la absorbancia 600 nm, esto para descartar que los efectos visualizados por los tratamientos no fueran debidos a toxicidad. Al graficar el área bajo la curva de las lecturas de los pozos durante las 48 muestras (dos por hora) se obtiene el diagrama de cajas de la Figura 17. Si uno de los tratamientos es tóxico, esto se verá reflejado en la vialidad de las células, teniendo un menor crecimiento que las células con tratamientos no tóxicos.

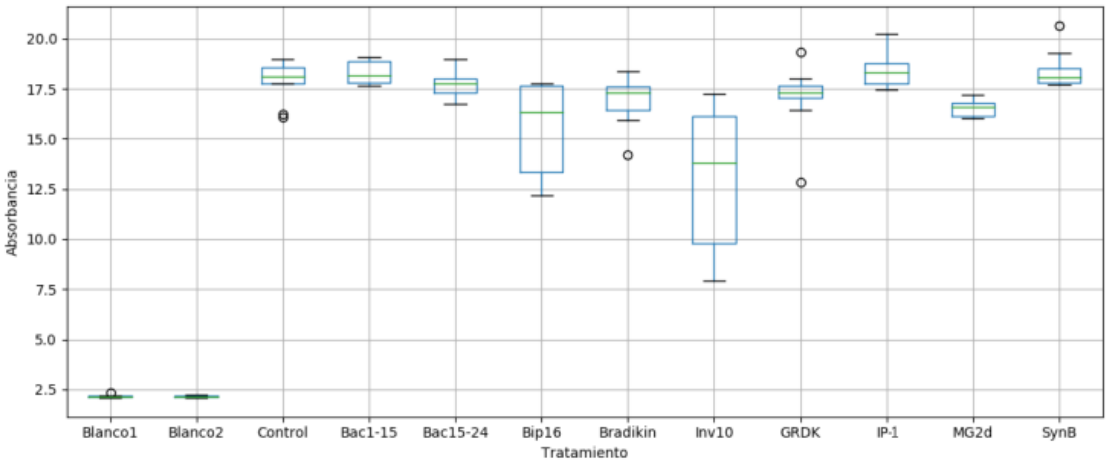

Figura 17. Diagrama de cajas del área bajo la curva de las lecturas a 600 nm tomadas con el espectrofotómetro Sinergy Mx durante 24 horas con tomas cada 30 minutos a $P<0.05$.

Capítulo 5. Discusión

En esta tesis se buscaron nuevos péptidos inductores de autofagia (AIPs). Para concretar la búsqueda, fue necesario evaluar la hipótesis en la cual se mantiene que un péptido antimicrobiano y penetrador de células (AMCPP) es un AIP.

El evaluar la hipótesis en este trabajo mediante enfoques bioinformáticos implica primero examinar las herramientas con las que se predicen las actividades individuales de péptido antimicrobiano (AMP) y péptido penetrador de células (CPP). La predicción de AMPs se ha trabajado con atención a lo largo de los años, sin embargo la predicción de CPPs necesitó de su propio análisis, con el cual se seleccionaron los predictores que se utilizaron para evaluar los péptidos de proteomas de mamíferos como AMCPPs.

A partir de la evaluación de péptidos en proteomas de mamíferos se pudo no solo hacer la selección de AIPs, sino también examinar el comportamiento de las diferentes combinaciones de actividades AMP y CPP respecto su presencia en proteínas cuya relación con la autofagia han sido identificada.

Finalmente, la prueba de concepto permite observar si existen diferencias entre la morfología de los péptidos seleccionados como AMCPPs, AMnCPP y nAMPCPP.

Cada uno de estos enfoques pueden ser analizados y discutidos independientes entre sí, discusiones que se mostrarán en las secciones por venir.

5.1. Evaluación de predictores de CPPs

En las tablas 13 y 14 se observa que SkipCPP es el predictor que obtiene la especifididad más baja a pesar de tener una sensibilidad alta en promedio, quedando en segundo lugar después de CPPRF y terminando con un MCC menor a 0.5. Asimismo, podemos ver que el conjunto con el que fue entrenado, CPP924, es el que obtiene los valores más altos tanto en especificidad como MCC por todos los predictores. Una posible causa de esto puede ser la baja dificultad que presenta este conjunto para discriminar entre casos positivos y negativos.

En contraparte, vemos que el mejor desempeño de los predictores tanto en sensibilidad como especificidad lo genera KELM en sus diferentes modelos, asimismo, su conjunto de entrenamiento PandeyTr fue elaborado con cuidado contemplando un origen de datos variado ($D = 0.96$ a umbral de 0.8) con reducción de redundancia para los casos positivos mientras que sus casos negativos incluyen nCPPs validados experimentalmente adicionados con péptidos aleatorios con alguna actividad biológica, muestreados de forma aleatoria, mientras que los casos positivos mantienen una diversidad alta. Cabe destacar que si bien PandeyTr no fue el conjunto en el que los predictores en general obtuvieron peor resultado, sí fue el de MCC más bajo que contenía algún tipo de selección aleatoria en su conjunto de entrenamiento y que ningún modelo entrenado solo con nCPPs validados experimentalmente fue contemplado en este experimento.

Esto abre la interrogante de si los modelos entrenados con conjuntos “sencillos“ sean capaces de distinguir péptidos en un entorno real. Para determinar un conjunto “sencillo“ usamos las medidas de diversidad D y parentesco K. En este trabajo mostramos que existe una diferencia en las diversidades de los casos negativos respecto los positivos en la mayoría de los conjuntos de datos evaluados. Esta diferencia se puede explicar por la generación de los casos negativos; mientras que los casos positivos se generaron a partir de la recompilación de péptidos en la literatura, como vimos en la Tabla 2, los negativos varían en origen y al no estar limitados por un conjunto definido de péptidos (como de una base de datos), su diversidad aumenta. Esto no necesariamente es bueno, debido a que los casos negativos son tan diferentes entre sí que encontrar el patrón que los relacione como nCPP se vuelve más complicado, y si el parentesco de los conjuntos de los conjuntos negativos respecto a los positivos es baja, la separabilidad aumenta. Como pudo observarse en la Tabla 16, tenemos casos de $K > 0.05$.

Si $K = 0$, entonces, a nivel de secuencia, cualquier par de elementos de casos opuestos difieren entre sí en más de un 20 % (en casos como CPP924, GautamInd y PandeyInd cualquier par de secuencias difiere en al menos 50 %). Si el valor es mayor a 0, entonces existe ese porcentaje de secuencias en el conjunto negativo tal que existe al menos una secuencia en el conjunto positivo al que se parece en al menos un 80 %.

Cuando comparamos el valor de K de los casos positivos y negativos respecto a

los resultados, con algunas excepciones, el desempeño en la especificidad tiene un comportamiento inverso al parentesco a diferentes umbrales de similitud α. Donde, claramente se ve que las especificidades mayores al promedio (0.73) pertenecen a conjuntos de datos con $K(\alpha) < 0.06$ y aquellas de especificidad menor al promedio poseen un $K(\alpha) > 0.2$ para $\alpha = 0.6, 0.7, 0.8, 0.9$, con las únicas excepciones de HoltonInd y Wei. En la Figura 12 puede verse el parentesco a $\alpha = 0.7$ y la especificidad de los conjuntos ordenados por este último campo en orden descendente.

Se esperaría que al agrupar a nivel de secuencia los conjuntos no se lograra separar de forma significativa los casos positivos de los negativos, sin embargo, al agrupar los conjuntos se puede observar que algunos de ellos son hasta cierto punto separables (MCC>0.5). Si la similitud entre datos positivos y negativos es despreciable, debería ser sencillo separarlos y si un método naíf como una aglomeración únicamente por similitud de secuencias puede separar casos positivos de negativos en un conjunto, un método que considere más factores debería tener un resultado mejor y eso es lo que vimos reflejado en la Tabla 13 .

Todos los predictores utilizan en su mayoría atributos generados a partir de la estructura de la secuencia y un conjunto fácilmente separable daría resultados buenos independientemente del predictor.

Por la naturaleza de sus datos y los resultados anteriores, podemos decir que el conjunto de entrenamiento empleado para SkipCPP fue sencillo y no le permitió generalizar a su máquina. Su alta sensibilidad puede ser debido a que prefiere clasificar a todos los péptidos de entrada como CPP. Como Wei *et al.* (2017a) generan sus secuencias negativas como secuencias de aminoácidos aleatorias podría no estar buscando o identificando CPPs, sino simplemente péptidos de origen biológico de péptidos aleatorios, de ahí el efecto en la sensibilidad.

CPP924 es el conjunto más fácilmente separable, de parentesco 0 entre positivos y negativos y el mejor evaluado por todos los predictores, además, SkipCPP, siendo el predictor que dio origen a este conjunto tuvo el desempeño en especificidad más bajo para el resto de los conjuntos de prueba y uno de los peores MCC. Esto es importante porque Wei *et al.* (2017b) consideran a CPP924 como "high-quality benchmark dataset", Qiang *et al.* (2018) utilizan este conjunto porque dicen que con este con-

junto evitan sobreestimación de sus predictores gracias a su nivel de identidad entre secuencias (que en efecto es menor al 80 % para casos positivos y negativos pero para este último era menor al 30 %) y por último, Wei *et al.* (2017a) afirman que su conjunto es altamente representativo de los CPPs de la realidad y aunque esto no se discute para los positivos, el conjunto negativo carece de los mismos adjetivos.

El utilizar péptidos con actividad biológica combinado con péptidos validados experimentalmente parece ser una técnica efectiva para la generación de casos negativos en el conjunto de entrenamiento. Estas características permiten un conjunto diverso que no sacrifica por completo la similitud que existe tanto en los casos positivos y los negativos. Sin lugar a dudas, es necesario abordar sobre el tema de los casos negativos, al momento, cualquier técnica computacional que se utilice no dará la suficiente seguridad hasta que se realice una prueba experimental con más nCPPs.

5.2. Identificación de inducción de autofagia de forma experimental

El objetivo de este apartado fue comprobar experimentalmente si los péptidos predichos *in silico* como AIP inducían autofagia en células de *S. cerevisiae in vitro*.

Pudo verse en la sección 4.3 cómo la diferencia en las morfologías de las células tratadas con rapamicina respecto a las no tratadas era marcada, al punto que en ninguno de los casos estos tratamientos tenían intersección en sus valores.

La rapamicina es un fármaco utilizado para prevenir el rechazo de órganos durante los transplantes de órganos e inhibe a mTOR (objetivo en mamífero de rapamicina) (mammalian target of rapamycin), lo cual lo convierte en un claro inductor de macroautofagia (Cai y Yan, 2013; Hernandez *et al.*, 2012).

Se espera que los tratamientos que provocan una morfología semejante a la que induce la rapamicina, específicamente, la aparición de vacuolas, estén provocando macroautofagia. Sin embargo, las vacuolas también se pueden observar en otros procesos celulares

Se esperaría que si un tratamiento provoca una morfología vacuolar semejante a

la que genera rapamicina, esta sea inductor de macroautofagia. Sin emabargo, las diferencias en la morfología vacuolar no es exclusiva de mecanismos de autofagia; Una disrupción de la vacuola puede deberse a esporulación de la célula o a un choque hiperosmótico (Eastwood *et al.*, 2012; Baars *et al.*, 2007) mientras que la falta de V-ATPasa provoca células con una sola vacoula en lugar de varias vacuolas más pequeñas (Baars *et al.*, 2007).

Dado que la morfología de los controles positivos y negativos es significativamente diferentes entre sí y dada que la diferencia entre ellos es la inducción de autofagia, a pesar de que la forma de la vacuola puede deberse a diversos factoes, en este caso será asociado a la macroautofagia.

Rapamincina es un conocido inductor de macroautofagia y de IP-1 se conoce su actividad como inductor de autofagia en células de mamíferos gracias a trabajos previos del grupo de investigación de Gabriel del Río Guerra del Instituto de Fisiología Celular de la Universidad Autónoma de México.

Cómo se explicó en la sección 2, durante la macroautofagia se secuestra material del citoplasma mediante autofagosomas mientras que en la microautofagia la vacuola (homólogo del lisosoma) se deforma al emplear tubos autofágicos, invaginación o escisión. Por tanto, el observar que la vacuola es firme y simétrica nos lleva a sugerir que nos encontramos frente a un efecto de macroautofagia que coincide con el tipo de autofagia que induce Rapamicina, un control con morfología de aro mayor al resto de controles y tratamientos.

Las células que no presenten la morfología de aro no se les puede definir en un estado de autofagia basal o autofagia inducida, sin embargo, dada la morfología presentada en la vacuola (la cual es deforme o segmentada), podemos inferir que esta se encuentra en microautofagia. Si esto es cierto, entonces, en su estado basal la célula de S. cerevisiae se encuentra en microautofagia.

Al ser la primera vez que el autor de esta tesis trabajó con experimentos *in vitro*, algunos de los ensayos, en especial los primeros, presentaron algunos fallos tales como la poca densidad en las muestras para las fotografías, por lo cual este experimento puede mejorarse en el futuro.

La hipótesis que se ha manejado a lo largo de este trabajo plantea que los AMCPPs serán AIPs. Para esto, primero hay que recordar Inv10, presenta señales de toxicidad, como pudo verse en la Sección 4.3.2. Lundberg y Lo Langel (2003) nos dicen que ciertos CPPs, cómo los anfipáticos, pueden causar fugas en el citoplasma dada una ruptura en la membrana, similar a algunos AMPs. Cuando se trataron las cepas, se observó que las tratadas con Inv10 presentaban residuos celulares (cell debris), lo que lleva a pensar que Inv10 pertenece a esta categoría. Ignorando a Inv10 por los motivos mencionados, los resultados obtenidos fueron que 1) dos de cuatro AMCPPs con actividad similar a IP-1, uno con actividad inferior y uno sin actividad, 2) el único AMnCPP presenta actividad similar a IP-1 y 3) un nAMCPP con actividad inferior a IP-1 y uno sin actividad.

Recordando que la macroautofagia es una respuesta al estrés, es lógico pensar que un agente extraño dentro del citoplasma con propiedades antimicrobianas generará estrés a la célula, la cual se defenderá iniciando el proceso de macroautofagia. Además la Rapamicina es un inductor de macroautofagia que actúa inhibiendo la señalización mTOR, mecanismo por el cual también la hambruna induce macroautofagia y se sabe que la limitación de nutrientes induce una fusión vacoular (Baars *et al.*, 2007), por tanto, lo que observamos como morfología de aro, podría ser señal de macroautofagia.

Cuatro de los cinco AMCPPs considerados (un control positivo y 4 tratamientos) inducen macroautofagia, lo cual apoya a la hipótesis de que AMCPPs son AIPs, sin embargo, se tendría que definir cúal es la diferencia entre péptidos como SynB y Bradikin que no siguen la norma. La razón por la cual esto sucede queda por investigarse como trabajo futuro.

5.3. Búsqueda de posibles péptidos inductores de autofagia en proteomas de mamíferos

Esta última parte pretende entregar secuencias de aminoácidos con una probabilidad alta de ser AIP. Aquí se utilizan los resultados anteriormente generados. En la prueba de concepto se observó cómo los péptidos AMPs+CPPs inducían más las morfologías atribuidas a la macroautofagia. Para encontrar las actividades se utilizaron

predictores y dado que la información sobre los predictores de CPPs era limitada, evaluar y seleccionar el mejor predictor era esencial.

La información experimental de la literatura junto con predictores y técnicas computacionales permitieron generar un conjunto reducido de posibles AIPs.

Mientras que no se observó una diferencia entre los AIPs identificados y la longevidad de los animales, se identificó que *Heterocephalus glabler* tenía un procentaje de AMCPP$_r$s mayor a cualquier otro mamífero evaluado. Esta es una especie con autofagia basal alta (Zhao *et al.*, 2014), por lo tanto resulta congruente que hasta el 70 % de sus AMCPPs hayan resultado AMCPP$_r$s cuando se tomaban proteínas con $\gamma > 1$.

Como se vio en la Sección 4.2.1, la presencia de los AMCPPs en las proteínas relacionadas a autofagia se duplica para todas las especies cuando solo se tomas aquellos péptidos de proteínas con $\gamma > 1$ y que este efecto se mitiga por completo cuando los péptidos en cuestión no son AMCPPs.

Podemos observar cómo la actividad antimicrobiana y la penetrante impactan en la detección de estas secuencias de aminoácidos. También se pudo observar que los menores porcentajes de péptidos$_r$ provenían de aquellos péptidos cuya actividad antimicrobiana estaba ausente, independientemente de la actividad penetrante, por lo que para esta sección los AMPs son dominantes sobre los CPPs.

Los péptidos que se proponen como posibles AIPs son péptidos que provienen de proteínas con $\gamma > 1$ así como los tres péptidos comerciales son subcadenas de la proteína Tat-beclin.

Capítulo 6. Conclusiones

En esta tesis se presentaron diferentes metodologías con el propósito de generar un conjunto de secuencias de aminoácidos con la mayor probabilidad de ser AIP. Durante el proceso se evaluaron y analizaron diferentes predictores y conjuntos de CPPs, se observaron las morfológicas de vacuolas en las células de *S. cerevisiae* como respuesta a diversos tratamientos y se cribó la base de datos AutophagyDatabase (Homma *et al.*, 2011) a partir de péptidos de proteomas de mamíferos de longevidad variada. Con todo lo anterior, presentamos las siguientes conclusiones.

6.1. Conclusiones

Evaluación de predictores de CPPs

- KELM, MLCPP, DCF y CPPP son los predictores más competitivos.
- Los mejores conjuntos fueron elaborados a partir de reducción de redundancia para los casos positivos y conjuntos negativos aleatorios de péptidos de origen natural con actividad biológica.
- La diferencia entre los predictores de alta calidad y los no tan buenos, recae en el conjunto de datos de entrenamiento.
- El conjunto CPP924 fue el conjunto peor valorado a pesar de ser el de mayor popularidad.
- La construcción de conjuntos de nCPPs necesita ser abordado, más nCPPs experimentalmente validados son necesarios para lograr predictores de mayor confiabilidad.
- La selección de los predictores de CPPs permitió hacer una evaluación de los péptidos producto de una proteólisis in silico como AMCPPs.

Identificación de inducción de autofagia de forma experimental

- Bac1-15, MG2d y Bradikinn sugieren ser inductores de macroautofagia similares a IP-1, sin embargo, estos datos deben se ser analizados a detalle en trabajos futuros.
- Bac15-24 y CGRDK sugieeren ser inductores de macroautofagia en un grado menor que IP-1, sin embargo, estos datos deben se ser analizados a detalle en trabajos futuros.
- SynB y Bip16 no presentan señales de ser inductores de macroautofagia.

Búsqueda de posibles péptidos inductores de autofagia en proteomas de mamíferos

- Fueron localizados 175 potenciales AIPs, provenientes de proteínas de proteomas de mamíferos seleccionados por una proteólisis in silico, una predicción de actividades AMP y CPP y una búsqueda por homología en AutophagyDatabase.
- Las actividades AMP y CPP están más presentes en las proteínas relacionadas a la autofagia que péptidos sin estas actividades, encontrando hasta 1.6 veces más péptidos que en el resto de combinaciones de actividades.
- Cuando una proteína tiene más de un AMCPP, es más probable que esta esté relacionada a autofagia. Además se observa una mayor direfencia en las actividades AMP+nCPP, nAMP+CPP y nAMP+CPP teniendo hasta 3 veces más péptidos provenientes de proteínas encontradas en AutophagyDatabase mediante búsqueda por homología.
- *Heterocephalus galber*, el cuarto mamífero de mayor coeficiente de longevidad, presentó la mayor proporción de potenciales AIPs siendo de hasta 70 % cuando se consideran péptidos de proteínas con más de un AMCPP.
- No se observó relación alguna entre la longevidad de los animales respecto a la cantidad de AMCPPs encontrados o los AMCPP$_r$ localizados.

6.2. Trabajo futuro

A continuación se presentan un total de seis proyectos propuestos que surgen a partir de los resultados de este trabajo de tesis así como para completar a los mismos.

Identificación de macroautofagia mediante aprendizaje de máquina. Durante este trabajo la identificación de células se llevó de forma manual invirtiendo horas de trabajo identificando cada una de las imágenes. Dado la limitación de tiempo no fue posible desarrollar un algoritmo de identificación de estas imágenes, por lo que se propone un modelo de aprendizaje de máquina para identificación de células en autofagia basal y en macroautofagia inducida utilizando las células clasificadas en este documento de manera individual para el entrenamiento y prueba.

Este proyecto permitiría escalar el análisis de experimentos futuros hasta incluso experimentos de alta producción.

Búsqueda de autofagia mediante RNASeq. Dado que las técnicas de RNASeq permiten cuantificar el cambio en los niveles de expresión bajo diferentes condiciones (Wang *et al.*, 2009), se podría aprovechar la tecnología para tener una evidencia cuantitativa de la presencia de autofagia.

Identificación de autofagia por espectrometría. Pese a que en este trabajo se abordó el uso de Synergy, la configuración utilizada presentó resultados indistinguibles entre los controles de autofagia basal, inducción por fármacos, inducción peptídica e incluso con el blanco. Sin embargo, se detectó que existe una configuración que presenta las diferencias entre los controles y los tratamientos. Dado que en este trabajo no se pudo abordar esta nueva configuración se propone examinar este enfoque con la configuración correcta.

Construcción de conjunto de datos de CPPs y nCPPs. Se mostró en este documento cómo los conjuntos de datos de entrenamiento para la predicción de CPPs necesita de conjuntos negativos con mayor similitud a los conjuntos positivos y que

técnicas como la generación por elección aleatoria de aminoácidos no dan lugar a buenos predictores. Por tanto, se propone la construcción de un nuevo conjunto de datos para nCPPs que utilice péptidos que por su función no tienda a penetrar la célula, por ejemplo, ligandos peptídicos, o péptidos que pierdan características esenciales de los CPPs como la relación entre la carga, la hidrofobicidad y anfipaticidad (Hansen *et al.*, 2008b; Milletti, 2012).

Prueba experimental con péptidos propuestos. En este trabajo se trató de manera experimental péptidos con actividades AMP, CPP, nAMP y nCPP, sin embargo, se propone examinar la inducción de autofagia en los 175 péptidos propuestos. Esto permitiría respaldar las técnicas aquí elaboradas y la identificación de más péptidos inductores de autofagia.

Construcción de red de similitud de AIPs propuestos. Las únicas relaciones que presentamos entre los péptidos evaluados in silico fueron sus activdades como AMP y CPP y su origen. Proponemos que utilizando herramientas como starPep toolbox (Aguilera-Mendoza *et al.*, 2019) se identifiquen las relaciones que estos péptidos tienen en cuanto a funcionalidad y posteriormente respecto a su espacio químico. Observar cómo se comportan los péptidos entre ellos, con AIPs validados experimentalmente y con otros péptidos, permitirá entender las propiedades que definen a un péptido como AIP.

Literatura citada

(2000). The proton-translocating ATPase (H +-ATPase. Reporte técnico.

Abramoff, M.D., M. P. R. S. (2004). Image Processing with ImageJ. *Biophotonics International*, **11**(7): 36–42.

Agrawal, P., Bhalla, S., Usmani, S. S., Singh, S., Chaudhary, K., Raghava, G., y Gautam, A. (2016). CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Research*, **44**(D1): D1098–D1103.

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*, p. btz260.

Ahmed, M., Quoc Nguyen, H., Seok Hwang, J., Zada, S., Huyen Lai, T., Soo Kang, S., y Ryong Kim, D. (2018). Systematic characterization of autophagy-related genes during the adipocyte differentiation using public-access data. *Oncotarget*, **9**(21): 15526.

Akman, H. O., Raghavan, A., y Craigen, W. J. (2011). Animal Models of Glycogen Storage Disorders. *Progress in Molecular Biology and Translational Science*, **100**: 369–388.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P., Wilson, J., y Hunt, T. (2015). *Molecular Biology of the Cell*. Garland Science.

Angulo Bahón, C. (2001). Aprendizaje con máquinas núcleo en entornos de multiclasificación. *Inteligencia Artificial*, **6**(17): 72–82.

Aufschnaiter, A. y Büttner, S. (2019). The vacuolar shapes of ageing: From function to morphology. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1866**(5): 957–970.

Baars, T. L., Petri, S., Peters, C., y Mayer, A. (2007). Role of the V-ATPase in Regulation of the Vacuolar Fission–Fusion Equilibrium. *Molecular Biology of the Cell*, **18**(10): 3873–3882.

Beltran, J. A., Aguilera-Mendoza, L., y Brizuela, C. A. (2017). Feature weighting for antimicrobial peptides classification: a multi-objective evolutionary approach. En: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE International Conference on Bioinformatics and Biomedicine, pp. 276–283.

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**(S7): 672.

Boland, B., Kumar, A., Lee, S., Platt, F. M., Wegiel, J., Yu, W. H., y Nixon, R. A. (2008). Autophagy Induction and Autophagosome Clearance in Neurons: Relationship to Autophagic Pathology in Alzheimer's Disease. *Journal of Neuroscience*, **28**(27): 6926–6937.

Boman, H. G. (1995). Peptide antibiotics and their role in innate immunity. *Annual review of immunology*, **13**(1): 61–92.

Cai, Z. y Yan, L.-J. (2013). Rapamycin, Autophagy, and Alzheimer's Disease. *Journal of biochemical and pharmacological research*, **1**(2): 84–90.

Chen, L., Chu, C., Huang, T., Kong, X., y Cai, Y. D. (2015). Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino Acids*, **47**(7): 2485–1493.

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.

Cuervo, A. M., Bergamini, E., Brunk, U. T., Dröge, W., Ffrench, M., y Terman, A. (2005). Autophagy and aging: the importance of maintaining çleançells. *Autophagy*, **1**(3): 131–140.

De Magalhães, J. P. y Costa, J. (2009). A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*, **22**(8): 1770–1774.

de Magalhães, J. P., Costa, J., y Church, G. M. (2007). An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *The journals of gerontology. Series A, Biological sciences and medical sciences*, **62**(2): 149–60.

Diener, C., Garza Ramos Martínez, G., Moreno Blas, D., Castillo González, D. A., Corzo, G., Castro-Obregon, S., y Del Rio, G. (2016). Effective Design of Multifunctional Peptides by Combining Compatible Functions. *PLoS Computational Biology*, **12**(4): 1004786.

Dobchev, D. A., Mäger, I., Tulp, I., Karelson, G., Tamm, T., Tämm, K., Jänes, J., Langel, [U+FFFD], y Karelson, M. (2010). Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks. Reporte técnico, Bentham Science Publishers.

Dowaidar, M., Gestin, M., Cerrato, C. P., Jafferali, M. H., Margus, H., Kivistik, P. A., Ezzat, K., Hallberg, E., Pooga, M., Hällbrink, M., y Langel, [U+FFFD] (2017). Role of autophagy in cell-penetrating peptide transfection model. *Scientific Reports*, **7**(1): 12635.

Duda, R. O., Hart, P. E. P. E., y Stork, D. G. (1973). *Pattern classification*. Jonh Wiley & Sons, INC., segunda edición.

Eastwood, M., Cheung, S., Lee, K., Moffat, J., y Meneghini, M. (2012). Developmentally Programmed Nuclear Destruction during Yeast Gametogenesis. *Developmental Cell*, **23**(1): 35–44.

Ferea, T. L. y Bowman, B. J. (1996). The Vacuolar ATPase of Neurospora crmsa Is Indispensable: Inactivation of the uma-1 Gene by Repeat-Induced Point Mutation. *Genetics*, **143**(1): 147–154.

Friedman, J., Hastie, T., y Tibshirani, R. (2001). *The elements of statistical learning*, Vol. 1. Springer series in statistics Springer, Berlin.

Gallegos, J. C. P., Soto, A. T., Aguilera, F. S. Q., Sprock, A. S., Flor, E. U. M., Casali, A., Scheihing, E., Valdivia, Y. J. T., Soto, M. D. T., y Zapata, F. J. O. (2014). *Inteligencia Artificial*. Iniciativa Latinoamericana de Libros de Texto Abiertos.

Galluzzi, L., Baehrecke, E. H., Ballabio, A., Boya, P., Manuel, J., y Pedro, B.-S. (2017). Molecular definitions of autophagy and related processes. *Nektarios Tavernarakis*, **9**(11): 1811–1836.

Garcia-Valtanen, P., Del Mar Ortega-Villaizan, M., Martinez-Lopez, A., Medina-Gali, R., Perez, L., Mackenzie, S., Figueras, A., Coll, J. M., y Estepa, A. (2014). Autophagy-inducing peptides from mammalian VSV and fish VHSV rhabdoviral G glycoproteins (G) as models for the development of new therapeutic molecules. *Autophagy*, **10**(9): 1666–1680.

Gattass, R. R. y Mazur, E. (2018). Nanostructure fabrication; (160.2540) Fluorescent and luminescent materials. References and links 1. *Appl. Phys. Lett*, **12**(3): 10.

Gautam, A., Singh, H., Tyagi, A., Chaudhary, K., Kumar, R., Kapoor, P., y Raghava, G. P. S. (2012). CPPsite: a curated database of cell penetrating peptides. *Database*, **2012**(0): bas015.

Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., y Raghava, G. P. (2013). In silico approaches for designing highly effective cell penetrating peptides. *Journal of Translational Medicine*, **11**(1): 74.

Geng, J. y Klionsky, D. J. (2008). The Atg8 and Atg12 ubiquitin-like conjugation systems in macroautophagy 'Protein Modifications: Beyond the Usual Suspects' Review Series. *EMBO reports*, **9**: 859–864.

Godballe, T., Nilsson, L. L., Petersen, P. D., y Jenssen, H. (2011). Antimicrobial *β*-Peptides and *α*-Peptoids. *Chemical Biology and Drug Design*, **77**(2): 107–116.

Gomez, J. A., Chen, J., Ngo, J., Hajkova, D., Yeh, I.-J., Gama, V., Miyagi, M., y Matsuyama, S. (2010). Cell-Penetrating Penta-Peptides (CPP5s): Measurement of Cell Entry and Protein-Transduction Activity. *Pharmaceuticals (Basel, Switzerland)*, **3**(12): 3594–3613.

Griffin, R. W. (1981). *Química orgánica moderna*. reverté.

Gutierrez, D. D. (2015). *Machine Learning and Data Science: An introduction to statistical larning methods with R*. Technics Publications.

Hällbrink, M. y Karelson, M. (2005). Prediction of cell-penetrating peptides. *Cell-Penetrating Peptides: Methods and Protocols*, **1324**.

HAN, K., KIM, J., y CHOI, M. (2014). Computer simulations unveil the dynamics of autophagy and its implications for the cellular quality control. *Journal of Biological Systems*, **22**(4): 659–675.

Hansen, M., Chandra, A., Mitic, L. L., Onken, B., Driscoll, M., y Kenyon, C. (2008a). A Role for Autophagy in the Extension of Lifespan by Dietary Restriction in C. elegans. *PLoS Genetics*, **4**(2): e24.

Hansen, M., Kilk, K., y Langel, [U+FFFD] (2008b). Predicting cell-penetrating peptides. *Advanced Drug Delivery Reviews*, **60**(4-5): 572–579.

Harrington, P. (2012). *Machine learning in action*, Vol. 5. Manning Greenwich, CT.

Hernandez, D., Torres, C., Setlik, W., Cebrián, C., Mosharov, E., Tang, G., Cheng, H.-C., Kholodilov, N., Yarygina, O., Burke, R., Gershon, M., y Sulzer, D. (2012). Regulation of Presynaptic Neurotransmission by Macroautophagy. *Neuron*, **74**(2): 277–284.

Holton, T. A., Pollastri, G., Shields, D. C., y Mooney, C. (2013). CPPpred: Prediction of cell penetrating peptides. *Bioinformatics*, **29**(23): 3094–3096.

Homma, K., Suzuki, K., y Sugawara, H. (2011). The autophagy database: An all-inclusive information resource on autophagy that provides nourishment for research. *Nucleic Acids Research*, **39**(1): D986–D990.

Huang, P.-S., Boyken, S. E., y Baker, D. (2016). The coming of age of de novo protein design. *Nature*, **537**(7620): 320–327.

Jenssen, H. (2011). Descriptors for antimicrobial peptides. *Expert Opinion on Drug Discovery*, **6**(2): 171–184.

Jeong, J. K., Moon, M. H., Lee, Y. J., Seol, J. W., y Park, S. Y. (2013). Autophagy induced by the class III histone deacetylase Sirt1 prevents prion peptide neurotoxicity. *Neurobiology of Aging*, **34**(1): 146–156.

K. Schmidt-Nielsen y Knut, S. N. (1986). *Scaling. Why is animal size so important?*, Vol. 69. New York: Cambridge University. pp. 129–130.

Kanehisa, M. y Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1): 27–30.

Kanehisa, M., Sato, Y., y Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, **428**(4): 726–731.

Kelleher, J. D., Mac Namee, B., y D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Kerstens, W. y Van Dijck, P. (2018). A Cinderella story: how the vacuolar proteases Pep4 and Prb1 do more than cleaning up the cell's mass degradation processes. *Microbial cell (Graz, Austria)*, **5**(10): 438–443.

Kowalska, K., Carr, D. B., y Lipkowski, A. W. (2002). Direct antimicrobial properties of substance P. *Life Sciences*, **71**(7): 747–750.

Kubat, M. (2015). *An introduction to machine learning*. Springer.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Lantz, B. (2015). *Machine learning with R*. Packt Publishing Ltd.

Levine, B. y Kroemer, G. (2008). Autophagy in the Pathogenesis of Disease.

Li, W. y Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13): 1658–1659.

Lu, S., Tager, L. A., Chitale, S., y Riley, L. W. (2006). A cell-penetrating peptide derived from mammalian cell uptake protein of Mycobacterium tuberculosis. *Analytical biochemistry*, **353**(1): 7–14.

Lundberg, P. y Lo Langel, U. [U+FFFD] (2003). A brief introduction to cell-penetrating peptides. *Journal of molecular recognition J. Mol. Recognit*, **16**: 227–233.

Maciel-Herrerías, M. y Cabrera-Beníez, S. (2016). El papel de la autofagia en enfermedades pulmonares. *Revista del Instituto Nacional de Enfermedades Respiratorias*, **75**(3): 227–236.

Madeo, F., Zimmermann, A., Maiuri, M. C., y Kroemer, G. (2015). Essential role for autophagy in life span extension. *Journal of Clinical Investigation*, **125**(1): 85–93.

Maglogiannis, I. G. (2007). *Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. IOS Press. p. 407.

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., y Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, **8**(44): 77121.

Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., y Lee, G. (2018). Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *Journal of Proteome Research*, **17**(8): 2715–2726.

McKee, T., McKee, J. R. T., y McKee, J. R. (2003). *Bioquímica: la base molecular de la vida*. McGraw-Hill/Interamericana,.

Meher, P. K., Sahu, T. K., Saini, V., y Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, **7**.

Melendrez C., G. (2018). *Análisis de transcriptomas para el descubrimiento de péptidos antimicrobianos*. Tesis de doctorado, Centro de Investigación Científica y de Educación Superior de Ensenada.

Michael, M. y Lin, W.-C. (1973). Experimental study of information measure and inter-intra class distance ratios on feature selection and orderings. *IEEE Transactions on Systems, Man, and Cybernetics*, (2): 172–181.

Milletti, F. (2012). Cell-penetrating peptides: Classes, origin, and current landscape. *Drug Discovery Today*, **17**(15-16): 850–860.

Minkiewicz, P., Darewicz, M., Iwaniak, A., Sokołowska, J., Starowicz, P., Bucholska, J., Hrynkiewicz, M., Minkiewicz, P., Darewicz, M., Iwaniak, A., Sokołowska, J., Starowicz, P., Bucholska, J., y Hrynkiewicz, M. (2015). Common Amino Acid Subsequences in a Universal Proteome—Relevance for Food Science. *International Journal of Molecular Sciences*, **16**(9): 20748–20773.

Mitchell, T. (1997). *Machine Learning*, Vol. 1. McGraw Hill.

Montaño-Pérez, K. (2002). Péptidos antimicrobianos: un mecanismo de defensa ancestral con mucho futuro. *Interciencia*, **27**(1): 21–29.

Murray, R. K., Granner, D. K., Rodwell, V. W., Retana, M. P., y Moreno, M. M. (2007). *Harper bioquímica ilustrada*. Mc Graw Hill.

Murty, M. N. y Devi, V. S. (2015). *Introduction to pattern recognition and machine learning*, Vol. 5. World Scientific.

Nakamura, S. y Yoshimori, T. (2018). Autophagy and Longevity. *Mol. Cells*, **41**(1): 65–72.

Nakatogawa, H., Ichimura, Y., y Ohsumi, Y. (2007). Atg8, a Ubiquitin-like Protein Required for Autophagosome Formation, Mediates Membrane Tethering and Hemifusion. *Cell*, **130**(1): 165–178.

Oku, M. y Sakai, Y. (2018). Three Distinct Types of Microautophagy Based on Membrane Dynamics and Molecular Machineries. *BioEssays*, **40**(6): 1800008.

Pandey, P., Patel, V., George, N. V., y Mallajosyula, S. S. (2018). KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. *Journal of Proteome Research*, **17**(9): 3214–3222.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, [U+FFFD] (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**(Oct): 2825–2830.

Peña, A. (1988). *Bioquímica*. Limusa.

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).

Porto, W. F., Pires, [U+FFFD] 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.

Qi, Y., Bar-Joseph, Z., y Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, **63**(3): 490–500.

Qiang, X., Zhou, C., Ye, X., Du, P.-f., Su, R., y Wei, L. (2018). CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Briefings in Bioinformatics*, p. bby091.

Reggiori, F. y Klionsky, D. J. (2013). Autophagic Processes in Yeast: Mechanism, Machinery and Regulation. *Genetics*, **194**(2): 341–361.

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature*, **572**(7767): 27–29.

Rodriguez Plaza, J. G., Morales-Nava, R., Diener, C., Schreiber, G., Gonzalez, Z. D., Ortiz, M. T. L., Blake, I. O., Pantoja, O., Volkmer, R., Klipp, E., Herrmann, A., y Del Rio, G. (2014). Cell penetrating peptides and cationic antibacterial peptides: Two sides of the same coin. *Journal of Biological Chemistry*, p. bby091.

Sanders, W. S., Johnston, C. I., Bridges, S. M., Burgess, S. C., y Willeford, K. O. (2011). Prediction of Cell Penetrating Peptides by Support Vector Machines. *PLoS Computational Biology*, **7**(7): e1002101.

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J. B., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler-Bauer, A., Phan, L., Schneider, V. A., Schoch, C. L., Pruitt, K. D., y Ostell, J. (2019). Database resources of the National Center for Biotechnology Information.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., y Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, **9**(7): 676–682.

Shoji-Kawata, S., Sumpter, R., Leveno, M., Campbell, G. R., Zou, Z., Kinch, L., Wilkins, A. D., Sun, Q., Pallauf, K., MacDuff, D., Huerta, C., Virgin, H. W., Bernd Helms, J., Eerland, R., Tooze, S. A., Xavier, R., Lenschow, D. J., Yamamoto, A., King, D., Lichtarge, O., Grishin, N. V., Spector, S. A., Kaloyanova, D. V., y Levine, B. (2013). Identification of a candidate therapeutic autophagy-inducing peptide. *Nature*.

Takeshige, K. (1992). Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. *The Journal of Cell Biology*, **119**(2): 301–311.

Tang, H., Su, Z. D., Wei, H. H., Chen, W., y Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochemical and Biophysical Research Communications*, **477**(1): 150–154.

Temerin, L. A. (1985). Size, function, and life history. By W.A. Calder III. Cambridge: Harvard University Press. 1984. xii + 431 pp., figures, tables, appendices, index. $32.50 (cloth). *American Journal of Physical Anthropology*, **66**(3): 340–342.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., y Idicula-Thomas, S. (2010). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, **38**(suppl_1): D774–D780.

Tsukada, M. y Ohsumi, Y. (1993). Isofation and charactization of autophagy-defective mutants of. Reporte técnico 1.

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

Wang, Z. y Wang, G. (2004). APD: the antimicrobial peptide database. *Nucleic acids research*, **32**(suppl 1): D590–D592.

Wang, Z., Gerstein, M., y Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1): 57–63.

Wei, L., Tang, J., y Zou, Q. (2017a). SkipCPP-Pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics*.

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., y Zou, Q. (2017b). CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *Journal of Proteome Research*, **16**(5): 2044–2053.

Wei, L., Zhou, C., Chen, H., Song, J., y Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, **34**(23): 4007–4016.

Wei, T., Liu, J., Ma, H., Cheng, Q., Huang, Y., Zhao, J., Huo, S., Xue, X., Liang, Z., y Liang, X.-J. (2013). Functionalized nanoscale micelles improve drug delivery for cancer therapy in vitro and in vivo. *Nano letters*, **13**(6): 2528–34.

Welter, E., Thumm, M., y Krick, R. (2010). Quantification of nonselective bulk autophagy in S. cerevisiae using Pgk1-GFP. *Autophagy*, **6**(6): 794–797.

Xie, Z., Nair, U., y Klionsky, D. J. (2008). Atg8 Controls Phagophore Expansion during Autophagosome Formation. *Molecular Biology of the Cell*, **19**(8): 3290–3298.

Yasuhiro, A., Naoyuki, U., Ryogo, H., y Yoh, W. (1989). Structure and Function of the Yeast Vacuolar Membrane Proton ATPase. *Journal of Bioenergetics and Biomembrane*, **21**(5): 589–603.

Zhao, S., Lin, L., Kan, G., Xu, C., Tang, Q., Yu, C., Sun, W., Cai, L., Xu, C., y Cui, S. (2014). High Autophagy in the Naked Mole Rat may Play a Significant Role in Maintaining Good Health. *Cellular Physiology and Biochemistry*, **33**(2): 321–332.

Anexo

División de secuencias

Tabla 27. Cantidad de secuencias de aminoácidos obtenidos en las distintas especies a determinada longitud.

	Cantidad de péptidos						
Longitud	**Myotis lucifugus**	**Heterocephalus glaber**	**Desmodus rotundus**	**Eptesicus fuscus**	**Cavia porcellus**	**Rattus norvegicus**	**Condylura cristata**
11	2012	2102	1962	2096	1962	1386	1901
12	1604	1688	1491	1631	1620	1105	1591
13	1553	1568	1419	1506	1461	950	1421
14	1389	1472	1337	1431	1336	923	1323
15	1285	1298	1208	1284	1199	817	1160
16	1214	1213	1097	1259	1174	785	1139
17	1153	1208	1089	1181	1116	784	1088
18	1092	1142	1001	1077	1093	717	1050
19	1020	1087	940	1055	989	682	966
20	1093	1055	1004	1048	1000	687	907
21	1014	998	920	994	958	630	911
22	928	927	885	915	875	569	868
23	895	960	830	905	866	564	764
24	870	936	798	865	853	531	802
25	775	853	740	792	796	553	749
26	765	770	715	710	739	467	716
27	759	770	672	776	744	546	707
28	761	780	680	758	742	477	734
29	721	697	628	702	684	500	588
30	693	697	601	704	631	405	601
Total	21,596	22,221	20,017	21,689	20,838	14,078	19,986

Contabilizaciones de células tomadas en el microscopio NanoImager

Tabla 28. Células contabilizadas en cada tratamiento señalando el ensayo del que fue obtenido.

Ensayo	Tratamiento	Cantidad de aros	Cantidad de palomita
1	Bac1-15	52	54
1	Bac1-15	71	58
2	Bac1-15	56	64
2	Bac1-15	52	93
1	Bac15-24	36	33
1	Bac15-24	52	52
2	Bac15-24	3	8
2	Bac15-24	1	9
2	Bac15-24	2	13
2	Bac15-24	82	75
2	Bac15-24	84	77
1	Bip16	29	36
1	Bip16	53	53
1	Bip16	53	67
2	Bip16	0	9
2	Bip16	3	12
2	Bip16	2	13
2	Bip16	32	32
2	Bip16	37	57
2	Bip16	21	78
2	Bip16	16	45
1	Bradikin	17	18
1	Bradikin	25	31
1	Bradikin	38	41
2	Bradikin	7	18
2	Bradikin	13	44
2	Bradikin	57	70
2	Bradikin	88	79
2	Bradikin	60	87
1	Control	31	41
2	Control	1	8
2	Control	10	14
2	Control	4	14
2	Control	5	23
2	Control	5	23
2	Control	5	23
2	Control	1	25
2	Control	39	84
2	Control	39	85
2	Control	40	93
	Control	5	23
2	CRGDK	17	28
2	CRGDK	11	30
2	CRGDK	13	33

Tabla 29. Células contabilizadas en cada tratamiento señalando el ensayo del que fue obtenido (continuación).

Ensayo	Tratamiento	Cantidad de aros	Cantidad de palomita
2	CRGDK	15	34
2	CRGDK	2	33
2	CRGDK	1	63
1	GRCDK	11	9
1	GRCDK	45	32
1	GRCDK	51	42
1	GRCDK	61	57
1	Inv10	26	17
1	Inv10	38	21
1	Inv10	31	33
2	Inv10	3	40
2	Inv10	4	43
2	Inv10	4	49
2	Inv10	4	63
2	Inv10	49	62
2	Inv10	54	52
2	Inv10	54	51
2	Inv10	50	55
1	IP-1	47	44
1	IP-1	49	57
1	IP-1	49	46
1	IP-1	80	41
2	IP-1	89	32
2	IP-1	77	42
2	IP-1	14	37
2	IP-1	27	39
2	IP-1	27	39
2	IP-1	16	42
2	IP-1	16	61
2	IP-1	42	44
2	IP-1	41	41
	IP-1	27	39
1	MG2d	26	30
1	MG2d	14	34
1	MG2d	35	29
2	MG2d	25	25
2	MG2d	11	27
2	MG2d	19	29
2	MG2d	29	39
2	MG2d	48	61
2	MG2d	63	65

Tabla 30. Células contabilizadas en cada tratamiento señalando el ensayo del que fue obtenido (continuación).

Ensayo	Tratamiento	Cantidad de aros	Cantidad de palomita
2	MG2d	13	75
1	Rapamicina	70	58
1	Rapamicina	119	67
1	Rapamicina	143	51
2	Rapamicina	55	22
2	Rapamicina	42	27
2	Rapamicina	38	27
2	SynB	0	5
2	SynB	0	5
2	SynB	1	6
2	SynB	7	7
2	SynB	7	12
2	SynB	6	13
2	SynB	27	46
2	SynB	49	66
2	SynB	38	68
2	SynB	0	2

Manual de usuario de herramienta CPPEval

Introducción

Bienvenido y gracias por usar estas aplicaciones.

El CPPEvalApp es una herramienta que ayudará a los usuarios que necesitan evaluar una gran cantidad de péptidos, permitiendo una evaluación sencilla aunque estos se encuentren en uno o varios archivos sin importar tu tamaño mientras estos se encuentren formato fasta. El usuario podrá evaluar con cualquier combinación de ocho de los nueve predictores disponibles en la actualidad.

Va a facilitar el proceso de evaluación dando la oportunidad de ignorar las restricciones de los predictores y seguir aprovechándolos.

La aplicación hará las peticiones a los servidores correspondientes y dará formato a las salidas para obtener un archivo de salida csv limpio.

Disfrute.

Ventajas

Esta herramienta permite a los usuarios predecir múltiples archivos fasta sin importar si estas secuencias se encuentran en una sola línea o divididas en renglones. Los archivos pueden contener secuencias no válidas y simplemente se elegirán las secuencias adecuadas para cada predictor en lugar de rechazar todo el archivo y la salida de todos lo predictores será entregada en un formato uniforme. Los archivos pueden ser tan largos como se requieran y el programa se encargará de separarlo según las capacidades de los predictores.

La salida será guardada en un archivo csv con todas las evaluaciones de los predictores en lugar de tener que descargar cada uno o incluso de tener que copiar y pegar.

Desventajas

Existen un par de atributos de los predictores que se pierden durante el proceso, como lo es evaluar la eficiencia de los predictores, diseñar nuevos péptidos o buscar subsecuencias de los péptidos con actividad penetrante.

Predictores

Esta aplicación llama a los predictores disponibles en línea Cell-PPD (Gautam *et al.*, 2013), CPPPred (Holton *et al.*, 2013), C2Pred (Tang *et al.*, 2016), DCF (Diener *et al.*, 2016), CPPPred-RF (Wei *et al.*, 2017b), Skip-CPPPred (Wei *et al.*, 2017a), KELM-CPPPred (Pandey *et al.*, 2018), y MLCPP (Manavalan *et al.*, 2017). La forma en que CPPred-Fl (Qiang *et al.*, 2018) evalúa las secuencias es lo suficientemente diferente como para no estar en esta aplicación. En el futuro esto podría ser agregado. Todos los enlaces a los predictores mencionados se encuentran en la Tabla 31. Es importante destacar que no todas las características de todos los descriptores se utilizan para esta aplicación,

CPPD es principalmente para diseñar nuevos CPP y tanto SkipCPP como MLCPP pueden asignar una eficiencia de penetración a los CPP predichos.

Tabla 31. Referencia de predictores, nombre, enlace y tipo de acceso.

Autores	Predictor	Enlace	Tipo
Gautam et al, 2013	CPPD	http://crdd.osdd.net/raghava/cellppd/	Online (para diseño)
Holton et al, 2013	CPPP	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/cpppred.php	Online
Tang et al, 2016	C2Pred	http://lin-group.cn/server/C2Pred	Online
Diener et al, 2016	DCF	https://github.com/cdiener/dcf	Application
Wei et al, 2017	CPPred-RF	https://omictools.com/cppred-rf-tool	Application
Wei et al, 2018	SkipCPP	http://server.malab.cn/SkipCPP-Pred/Index.html	Online
Pandey et al, 2018	KELMCPP	http://sairam.people.iitgn.ac.in/KELM-CPPpred.html	Online
Manavalan et al, 2018	MLCPP	http://www.thegleelab.org/MLCPP/	Online
Quiang et al, 2018	CPPPredFL	http://server.malab.cn/CPPred-FL/	Online

Interface

La interfaz está dividida en ocho partes marcadas de A a H en la Figura 18.

A, B y E representan el botón para agregar una carpeta de entrada, una carpeta de salida y para iniciar el procedimiento respectivamente.

C es un texto de entrada, aquí puede seleccionar un prefijo para los archivos de salida si lo necesita, así como ignorarlo.

D representa el panel con los predictores disponibles. Cada casilla de verificación representa un predictor. En el caso de CPPD, si está marcado, se mostrará otro panel como el que se muestra en la Figura 19 que representa los parámetros que solicita ese predictor (Figura 20).

Para KELM, los seis modelos aparecen como casillas de verificación adicionales, como se puede ver en la Figura 21. Si la casilla de verificación KELM no está marcada, ninguno de los modelos será considerado.

F y G son barras de progreso, F mostrará el progreso del archivo que se está evaluando actualmente y el progreso de toda la petición.

Finalmente, H es el log del proceso. Puedes ver lo que está haciendo y con cuántas

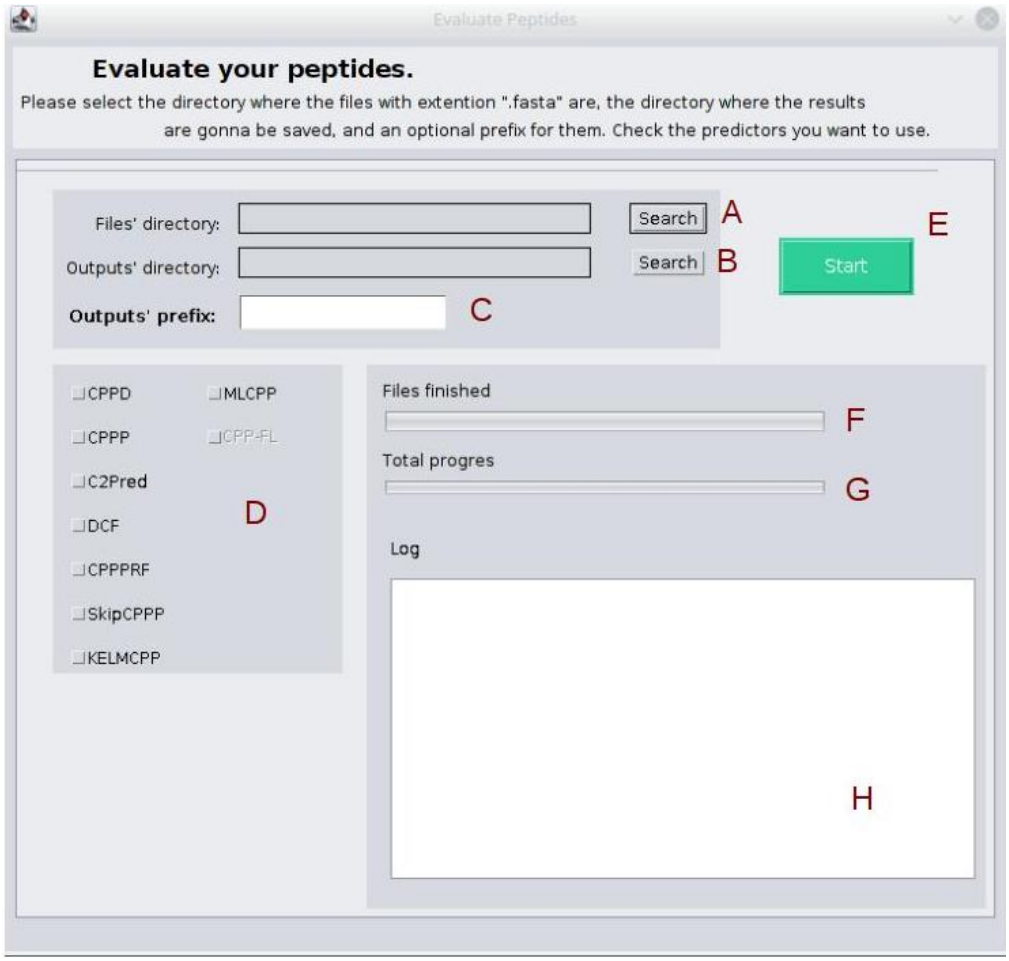

Figura 18. Interfaz de la aplicación y componentes principales marcados de A a H.

Figura 19. opciones adicionales de CPPD.

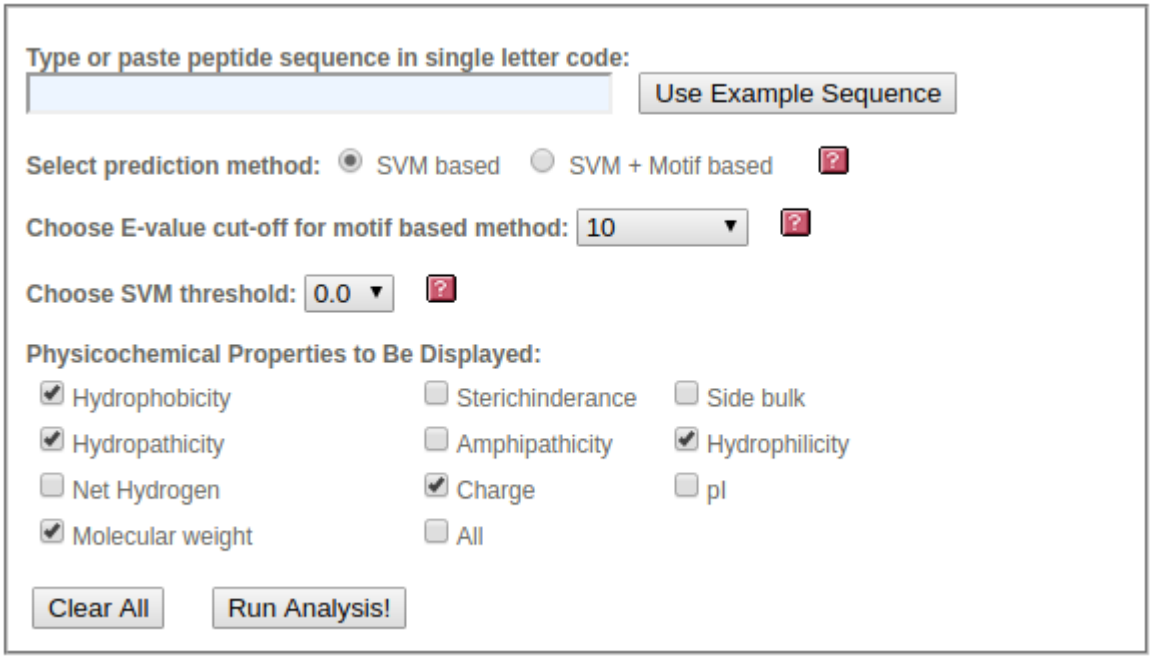

Figura 20. Formulario original del sitio web de CPPD.

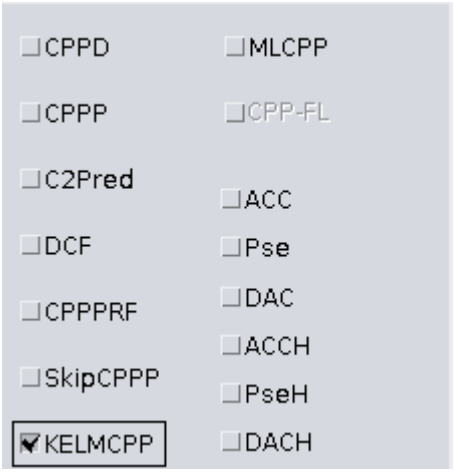

Figura 21. Los modelos de Kelm se muestran.

secuencias.

Entrada

En el botón A (ver sección) se le preguntará un directorio. La aplicación obtendrá todos los archivos .fasta y los evaluará. La cantidad de archivos se mostrará después de seleccionar la carpeta. Estos archivos pueden ser multifasta y pueden ser una sola línea por secuencia o multilínea. No hay restricción en las secuencias, pero dados los filtros de los predictores seleccionados no se evaluarán todas las secuencias.

Si algún archivo .fasta no tiene el formato de un archivo fasta, se ignorará.

En el botón B (ver sección) se le pedirá una carpeta de destino, allí se creará una carpeta llamada Evaluaciones y dentro de ella, las predicciones. Así que no tengas miedo de usar la misma carpeta si así lo deseas.

Un texto de entrada C (consulte la sección) solicitará un prefijo opcional para las salidas. Si no agrega uno, la salida se nombrará como el archivo original.

En el panel D (consulte la sección), la casilla de verificación representa los predictores que serán considerados en una ejecución en particular. Elija tantas como necesite y para el caso particular de CPPD, aparecerán configuraciones adicionales como en el original sitio web. KELM mostrará los seis modelos diferentes que utiliza, una vez más,

puede elegir tantos como necesite.

Salidas

Cada predictor tiene una manera diferente de presentar los resultados que se muestran en la Tabla 32. Algunos son binarios y otros continuos. Esta aplicación busca las predicciones y las asigna a un valor dentro del rango [0,1]. Si por alguna razón el predictor correspondiente no pudiera evaluar una secuencia, se marcará con un -50.

Tabla 32. Salidas de los predictores

Predictor	Rango de valores	CPPvalue
CPPD	(-, 1]	> 0
CPPP	[0,1]	> 0.5
C2Pred	[0,1]	> 0.5
DCF	[0,1]	> 0.5
CPPred-RF	[0,1]	> 0.5
SkipCPP	{0,1 }	1
KELMCPP	{0,1 }	1
MLCPP	[0,1]	> 0.5
CPPPredFL	[0,1]	> 0.5

Todos los péptidos evaluados se presentarán en un archivo cvs donde cada columna representa un predictor. Este archivo se guardará en el destino marcado con el prefijo seleccionado y el nombre del archivo fasta original.

Proceso

Una vez que haga clic en el botón E, inicie, la aplicación leerá cada archivo uno a tiempo, filtrará según los predictores y los enviará a todos en una o para las secciones según las capacidades empíricas de los predictores. En la barra F (ver la sección) se mostrará el progreso del archivo real y en la barra G (ver la sección) el progreso total.

A medida que avanza el progreso, en el panel H (ver sección) se iniciará el registro general. Cuando finalice el proceso, ambas barras estarán llenas y el panel H escribirá "Finish".

Recomendaciones

Si tiene muchos péptidos, puede evaluarlos con varios predictores, luego elija los que desea trabajar.

Si tiene muchas proteínas, puede utilizar CPPred-Fl (Qiang *et al.*, 2018) que no está incluido en esta aplicación y usar su resultado como entrada para esta aplicación.

CPPD y KELM son los predictores empíricos más lentos y DCF los más rápidos. KELM y MLCPP habían mostrado las mejores actuaciones. CPPD y KELM son los predictores con más errores de servidor. No siempre es así, pero generalmente las salidas se marcarán como -50 (consulte la sección).

Puede ejecutar la aplicación con **java -jar ÇPPsEvaluatorApp.jar"** para obtener más información de registro como el DCF, una información adicional de CPPRF o el proceso de cada petición paralela.

¿Tu trabajo murió inesperadamente? No te preocupes. En la carpeta donde se ubica el contenedor hay una carpeta llamada Predictores, puede ver sus evaluaciones por predictor.

Si los modelos KELM no aparecen, desmarque y vuelva a marcar la casilla de verificación.