Clasificación de péptidos antimicrobianos con funcionalidades específicas

Integrantes:
Santiago Rivera Montoya (santiago.riveram@udea.edu.co)
Emanuel López Higuita (emanuel.lopezh@udea.edu.co)
Tutor:
Carlos Andrés Mera (carlos.mera@udea.edu.co)

Proyecto Integrador I - 2508700 - G17

INTRODUCCIÓN

En la literatura actual, se han desarrollado diversos clasificadores para determinar si un péptido posee capacidad antimicrobiana, e incluso se han propuesto métodos para su generación artificial mediante técnicas de inteligencia artificial. Sin embargo, son pocos los trabajos que se han enfocado en clasificar los Péptidos Antimicrobianos (AMPs) según sus funciones específicas, utilizando las bases de datos ya existentes creadas por expertos.

Este proyecto tiene como objetivo abordar esta necesidad, tomando dichas bases de datos y aplicando técnicas de Machine Learning para clasificar los AMPs según sus funciones específicas, tales como anticancerígenos, antifúngicos, antivirales, entre otros. En este primer informe, se presenta el análisis de datos realizado, el cual sienta las bases para la construcción de modelos de clasificación que permitirán identificar las funciones específicas de estos péptidos.

DESCRIPCIÓN DE LOS DATOS

Se cuenta con dos bases de datos, una proveniente de la Universidad Nacional que cuenta con 31659 filas y 20 variables o columnas y la otra la base de datos de Starpep que cuenta con 45119 filas y diversas características fisicoquímicas.

LIMPIEZA Y PREPARACIÓN DE LOS DATOS

El proceso comienza con el preprocesamiento de los datasets. En la base de datos de la UNAL, se eliminaron las características fisicoquímicas, conservando únicamente las variables objetivo (como anticáncer, antifúngico, antiviral, entre otras) y los identificadores de cada péptido.

En el caso de la base de datos de Starpep, se extrajo la metadata y se agrupó utilizando variables dummy, lo que permitió conservar solo las variables objetivo y el identificador de cada péptido.

El segundo paso implica el filtrado de péptidos según su capacidad antimicrobiana. Posteriormente, se combinaron ambas bases de datos, resultando en una base de datos exclusivamente de péptidos AMP y otra que incluye tanto péptidos con capacidad antimicrobiana como aquellos que no la poseen.

En el tercer paso, se verifica la existencia de secuencias idénticas en ambas bases de datos. En caso de encontrar péptidos

duplicados, se eliminan para evitar duplicidades al combinar las bases de datos.

El cuarto paso consiste en el renombramiento de las columnas en ambas bases de datos por separado, facilitando su posterior combinación.

Tras la refactorización de los nombres y el orden de las columnas, se procede a unir las bases de datos, obteniendo un dataframe de péptidos exclusivamente AMP y otro dataframe que incluye tanto péptidos AMP como NO-AMP.

Luego, ambos dataframes finales se filtran para eliminar péptidos con secuencias mayores a 100 aminoácidos y aquellos que contengan aminoácidos no válidos, en preparación para la futura extracción de características.

Finalmente, se exportan los archivos en formato .fasta para llevar a cabo la extracción de características fisicoquímicas. Estas características se obtienen mediante las librerías propy3 en Python3 para extraer descriptores de la composición de aminoácidos, y Peptides en R para calcular la longitud, índice de Boman, carga, punto isoeléctrico e hidrofobicidad.

Al concluir este proceso, se obtienen dos dataframes finales:
Df_final_AMP: 23,105 filas x 37 columnas
Df_final: 65,304 filas x 37 columnas

Este proceso de preprocesamiento y extracción de características fisicoquímicas ha permitido generar dos conjuntos de datos depurados y estructurados que son fundamentales para realizar el análisis exploratorio de los datos y para los futuros modelos de machine learning orientados a la clasificación de péptidos.

EDA

Luego de haber unificado los datasets con el proceso realizado anteriormente, se procedió con el análisis mediante gráficos; el de cajas y bigotes arrojó que parece haber una gran cantidad de datos "atípicos" pero se conoce que las propiedades fisicoquímicas de los péptidos tienen una gran variabilidad según la cantidad y variedad de aminoácidos que posean, así que con el gráfico no hay suficiente evidencia para eliminarlos. Además, se sacaron los datos que parecían ser atípicos, sin embargo, dado que, tras analizar cada columna, se concluye que estos datos atípicos no se deben a errores de medición, cálculo o datos incorrectos. Por el contrario, son fuentes únicas de información que enriquecerán los modelos, aumentando la robustez de estos.

Con el mapa de calor, se analizó la correlación entre columnas, lo que arrojó que las correlaciones entre aminoácidos (A, R, N, etc.) suelen ser bajas, lo que indica que la presencia de un aminoácido específico no afecta significativamente a otro en este conjunto de datos. Sin embargo, algunas combinaciones de aminoácidos tienen correlaciones negativas o positivas moderadas, lo que podría reflejar patrones específicos de interés en la secuencia de aminoácidos.

Por otro lado, las propiedades fisicoquímicas como Hidrofobicidad, Índice de Boman, y Carga muestran correlaciones diversas con los diferentes aminoácidos. Por ejemplo, la correlación negativa entre Hidrofobicidad y el aminoácido R sugiere que cuando R es alto, la Hidrofobicidad tiende a ser baja.

Al agrupar los péptidos según su funcionalidad específica y calcular la media de cada aminoácido en cada categoría, se observó que la L (Leucina) y la K (Lisina) destacan en todas las

categorías funcionales, siendo los aminoácidos que aparecen en mayor cantidad promedio en cada funcionalidad. Esto sugiere que tanto la leucina como la lisina podrían desempeñar roles críticos en la eficacia de los péptidos en sus respectivas funciones biológicas.

Por último, al agrupar los péptidos según su función específica, y analizarlos con las propiedades de: carga, longitud, índice de boman, e hidrofobicidad, mediante el gráfico de cajas y bigotes se encontró que por índice de Boman, carga e hidrofobicidad, los tres gráficos arrojan que no hay mucha variación según sus características, no obstante, por longitud Los AMP antifúngicos y antibacterianos tienden a ser más largos y mostrar mayor variabilidad en comparación con otros y por punto isoeléctrico el anti VIH tiene un menor punto isoeléctrico en comparación con los demás.

PRÓXIMOS PASOS

Una vez hecha la selección y extracción de características, se exploraron los datos y analizaron las correlaciones, además de analizar las características según las propiedades fisicoquímicas, en el proyecto se buscarán los modelos que mejor se adapten a los datos para compararlos y evaluarlos.

BIBLIOGRAFÍA

1. Organización Mundial de la Salud. Plan de acción mundial sobre la resistencia a los antimicrobianos. Organización Mundial de la Salud. 2016. https://iris.who.int/handle/10665/255204.
2. Vélez A, Mera C, Orduz S , Branch JW. Generación de péptidos antimicrobianos mediante redes neuronales recurrentes. Revista DYNA. 88(216), pp. 210-219. 2021.
3. Wu, Q., Ke, H., Li, D., Wang, Q., Fang, J., Zhou, J.: Recent progress in machine learning-based prediction of peptide activity for drug discovery. Current topics in medicinal chemistry 19(1), 4–16 (2019). https://doi.org/10.2174/1568026619666190122151634
4. Waghu, F.H., Barai, R.S., Gurung, P., Idicula-Thomas, S.: CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Research 44(D1), D1094–D1097 (2016). https://doi.org/10.1093/NAR/GKV1051
5. Szymczak, Paulina & Szczurek, Ewa. (2023). Artificial intelligence-driven antimicrobial peptide discovery. Current opinion in structural biology. 83. 102733. 10.1016/j.sbi.2023.102733.