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

Repositorio

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 Aprendizaje Automático para clasificar los AMPs según sus funciones específicas, tales como anticancerígenos, antifúngicos, antivirales, entre otros. En este informe, se presenta el análisis de datos realizado y la extracción de características, los cuales sientan 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 registros, de los cuales, 18209 son no AMP y 13448 AMP y 20 variables o columnas y la otra la base de datos de Starpep que cuenta con 45119 registros y diversas funcionalidades antimicrobianas, que en particular solo 23156 cuentan con las funcionalidades antimicrobianas necesarias para este modelo.

LIMPIEZA Y PREPARACIÓN DE LOS DATOS

El proceso comienza con el preprocesamiento de los conjuntos de datos. En la base de datos de la UNAL, se eliminaron las características fisicoquímicas, conservando únicamente las variables objetivo (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 filtrar péptidos según su capacidad antimicrobiana, solo se dejan aquellos con las funcionalidades: antifúngico, antiviral, antibacteriano, anti-Gram positivo, anti-Gram negativo,

anticancerígeno, anti-VIH, antiparasitario y antitumoral. Después, 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 tercer paso consiste en el renombramiento de las columnas en ambas bases de datos por separado, facilitando su posterior combinación.

En el cuarto paso, tras la refactorización de los nombres y el orden de las columnas 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 (que pueden inhibir el crecimiento o eliminar microorganismos, bacterias, hongos, virus, etc.), como quienes no la poseen.

Luego, ambos conjuntos de datos finales se filtran para eliminar péptidos con secuencias menores a 7 aminoácidos y mayores a 100, como resultado de esta eliminación, se encontró que no había secuencias mayores a 100, sin embargo, se eliminaron 498 péptidos menores a 7 aminoácidos, además, se eliminaron aquellos que contenían aminoácidos no naturales y péptidos cuya secuencia tiene una variación menor a tres en sus aminoácidos, 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 como la composición de aminoácidos (AAC), composición de dipéptidos (DPC), autocorrelación de Moreau-Broto normalizada (MBauto), autocorrelación de Moran (Moranauto), autocorrelación de Geary (Gearyauto), composición, transición y distribución (CTD), números de acoplamiento del orden de secuencia (SOCN), cuasi-orden de secuencia (QSO) y composición de pseudo- aminoácidos (PAAC)., y la librería 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: 22,192 filas x 1507 columnas

Df_final: 40,840 filas x 1507 columnas

Al hacer un conteo de los péptidos AMP según sus funcionalidades específicas se obtuvieron los siguientes resultados.

Tabla 1. Cantidad de péptidos por actividad antimicrobiana

Funcionalidad	Cantidad
Antimicrobiano	13859
Antifúngico	5986
Antiviral	4252
Anti-Gram +	8786
Anti-Gram -	8708
Anticancerígeno	1762
Antibacteriano	8261
Anti-VIH	953
Antiparasitario	425
Antitumoral	550

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 aprendizaje automático orientados a la clasificación de péptidos.

ANÁLISIS EXPLORATORIO DE DATOS

Tras unificar los conjuntos de datos 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.

NORMALIZACIÓN Y SELECCIÓN DE CARACTERÍSTICAS

Primero, se realizó la prueba de Shapiro para evaluar si las variables seguían una distribución normal, resultando que no la seguían. Esto determinó el tipo de normalización a aplicar. Se observó una gran variabilidad entre las escalas de las variables sin normalizar, lo que justificó la necesidad de normalización previa a la selección de características.

Se aplicó la normalización Min-Max por ser la más adecuada para los datos, descartando otras opciones como Z-score, robust scaling y log transformation, debido a su alto costo computacional o la falta de distribución normal en los datos.

Posteriormente, se analizó la varianza de las columnas normalizadas y se utilizó el método VarianceThreshold, con un umbral de 0.0025, logrando reducir de 1507 a 903 columnas, manteniendo un equilibrio entre la reducción de dimensionalidad y la retención de información relevante.

Como segundo método, se empleó la selección basada en árboles (Tree-based), utilizando la variable objetivo antimicrobiano en el primer nivel de nuestro modelo multinivel, y 50 árboles para optimizar el tiempo de ejecución. Este método eliminó 468 variables, quedando con 435.

Finalmente, se aplicó SelectKBest con el parámetro k = 160, basándonos en un análisis PCA que sugiere que con 160 componentes se puede explicar un alto porcentaje de la variabilidad (90%), sin embargo, falta evaluar los modelos con estas características y ver que tan eficientes son en el entrenamiento y la predicción.

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.