upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA Y BIOCIENCIAS
NEKAZARITZAKO INGENIERITZAKO ETA BIOZIENTZIETAKO GOI MAILAKO ESKOLA TEKNIKOA

Estudio de la Decodificación de Proteínas con Fines Predictivos

GRADO EN CIENCIA DE DATOS

Presentado por Irati Sanz Delgado

Dirigido por José Antonio Moler

Septiembre 2023

Resumen

Este estudio se enfoca en investigar el potencial predictivo de diversas decodificaciones de secuencias proteicas en relación con características determinantes para la idoneidad de las proteínas en estudios clínicos. Además de analizar las secuencias proteicas, se abordan aspectos como la actividad de las proteínas mediante un problema de clasificación binaria, utilizando datos suministrados por Telum Therapeutics S.L. Se emplearon cinco decodificaciones del paquete *protp* de R Studio y se exploraron diversos modelos predictivos, incluyendo *Random Forest*, *SVM Radial* y *Gradient Boosting* como enfoques de aprendizaje automático, y Regresión Logística como un enfoque estadístico. Los resultados destacan que las decodificaciones de menor dimensionalidad demostraron un rendimiento superior, independientemente del modelo utilizado. Sin embargo, se observó que no existe un modelo universalmente efectivo para todos los problemas planteados. Se sugiere que un aumento en el tamaño de la muestra podría proporcionar un respaldo sólido para ciertos resultados, como el mayor rendimiento de la decodificación basada en la composición en la predicción de la actividad de las proteínas, en comparación con los datos experimentales proporcionados por Telum. Este enfoque podría resultar en un ahorro significativo de tiempo y recursos en términos de la cantidad de experimentos necesarios para determinar la actividad de las proteínas.

Abstract

This study focuses on investigating the predictive capability of various protein sequence decodings concerning critical features for protein suitability in clinical trials. In addition to the analysis of protein sequences, aspects related to protein activity are addressed through a binary classification problem, using data provided by Telum Therapeutics S.L. Five decodings from the protp package of R Studio were employed, and various predictive models were explored, including Random Forest, Radial SVM, and Gradient Boosting as machine learning approaches, and Logistic Regression as a statistical approach. The results highlight that lower-dimensional decodings exhibited superior performance, regardless of the model used. However, it was observed that there is no universally effective model for all the problems solved. It is suggested that an increase in the sample size could provide strong support for certain results, such as the improved performance of composition-based decoding in predicting protein activity compared to experimental data provided by Telum. This approach could result in significant time and resource savings in terms of the number of experiments needed to determine protein activity.

Índice

1 Introducción **3**

2 Conceptos Biológicos **4**

3 Proceso Telum **8**
3.1 Departamento APEXp 8
3.2 Departamento de Proteínas 10

4 Conjuntos de Datos **15**

5 Métodos de Decodificación **17**

6 Metodología **21**
6.1 Estudio de la actividad de las proteínas 21
6.1.1 Predicción de la actividad con el conjunto de variables 21
6.1.2 Predicción de la actividad con decodificaciones 29
6.2 Estudio del conjunto de proteínas activo 31
6.2.1 Predicción con decodificaciones 32
6.2.2 Predicción del effecto de las proteínas e imipenen 35

7 Resultados **38**
7.1 Resultados actividad de las proteínas 38
7.2 Resultados de proteínas activas 41

8 Conclusiones **46**

9 Lineas Futuras **47**

Referencias **49**

1 Introducción

El estudio de la decodificación de proteínas es un campo de investigación en constante evolución que busca comprender las proteínas y sus características. Esta investigación tiene aplicaciones predictivas, permitiendo anticipar la calidad de las proteínas en función de atributos cruciales. Las proteínas desempeñan un papel fundamental en diversos procesos biológicos, incluida su interacción con patógenos. Telum Therapeutics S.L., una empresa biotecnológica, se enfoca en experimentar con proteínas específicas en respuesta a patógenos.

En este contexto, el objetivo de esta investigación es analizar diferentes métodos de decodificación de proteínas y evaluar su capacidad para predecir la calidad de las proteínas según su aptitud para caracterizar variables importantes como el MIC (Concentración mínima inhibitoria), la solubilidad y la concentración de proteína muralítica. También utilizando conjuntos de datos proporcionados por departamentos de Telum, realizaremos predicciones de la actividad y el efecto de las proteínas en combinación con el antibiótico imipenem.

Emplearemos varios enfoques de modelado, incluyendo Random Forest, SVM radial, Gradient Boosting y regresión logística. Esto nos permitirá evaluar la efectividad de la predicción utilizando tanto decodificaciones como valores experimentales de los departamentos.

El modelo Random Forest es una técnica de aprendizaje automático que utiliza múltiples árboles de decisión para realizar predicciones. Este modelo tiene la capacidad de manejar grandes conjuntos de datos y puede identificar patrones complejos en los datos de entrada como pueden ser las distintas decodificaciones [1]. Esta cualidad lo hace adecuado para el estudio que se quiere realizar.

Otro algoritmo de aprendizaje automático es el modelo SVM radial, o Support Vector Machine con kernel radial, es utilizado para la clasificación y regresión. Utiliza un kernel radial para mapear los datos a un espacio de características de alta dimensión, donde busca encontrar un hiperplano óptimo que separe las clases o ajuste la regresión. Es especialmente útil cuando los datos no son linealmente separables, ya que puede modelar relaciones complejas entre las variables [2].

El último algoritmo de *machine learning* es el modelo Gradient Boosting que construye un modelo predictivo combinando múltiples modelos más simples, generalmente árboles de decisión, de manera secuencial. Un punto destacado del modelo es su capacidad para manejar problemas de conjuntos de datos desbalanceados de manera efectiva. Esto se debe a que el algoritmo prioriza la corrección de los errores, lo que significa que prestará más atención a las clases minoritarias, mejorando así su capacidad para identificar ejemplos raros o poco representados en los datos [3].

Por otro lado, el modelo estadístico de regresión logística es una técnica que se utiliza para predecir la probabilidad de que ocurra un evento binario en función de una o más variables predictoras. Utiliza una función logística para modelar la relación entre las variables predictoras y la variable de respuesta, produciendo una estimación de la probabilidad. Se interpreta en términos de odds, lo que significa que proporciona la probabilidad relativa de que ocurra el evento en comparación con que no ocurra [4].

Mediante la aplicación de estos métodos predictivos en conjunto con los experimentos y las decodificaciones, buscamos explorar y estudiar diversos aspectos de las proteínas. Además, tenemos

como objetivo evaluar la capacidad predictiva de los modelos y los datos, con el fin de determinar si podemos anticipar las características fundamentales de las proteínas.

2 Conceptos Biológicos

Debido a que este trabajo se centra en la decodificación de protéinas, resulta crucial una comprensión clara de conceptos biológicos. En particular, aquellos conceptos relacionados con proteínas y aminoácidos

Las proteínas son moléculas grandes y complejas que cumplen muchas funciones importantes en el cuerpo. Son vitales para la mayoría de los trabajos que realizan las células y son necesarias para mantener la estructura, función y regularización de los tejidos y órganos del cuerpo. Una proteína esta formada por una o más cadenas largas, plegadas de aminoácidos (polipéptidos), cuya secuencia de aminoácidos es especifica y viene determinada genéticamente [5].

En otras palabras, podemos decir que son cadenas de aminoácidos unidos en una secuencia especifica que produce una molécula funcional la cual puede plegarse y ser una enzima, o ser parte de la estructura de la célula, etc.... En total, hay miles y miles de proteínas que se producen cada día en las células y en el cuerpo. En el genoma humano, hay aproximadamente 20.000 genes que codifican para la producción de proteínas [5].

La composición de una proteína se basa en una cadena de aminoácidos. Los aminoácidos son ácidos carboxílicos que consisten en un carbono central unido a cuatro grupos funcionales diferentes: un grupo amino, un grupo carboxilo, un átomo de hidrógeno y un grupo R variable, que confiere propiedades únicas a cada aminoácido (figura 1). Las proteínas se forman a través de la síntesis por deshidratación, en la cual el nitrógeno del grupo amino de un aminoácido se enlaza con el carbono del grupo carboxílico de otro aminoácido, formando así un enlace peptídico [6].

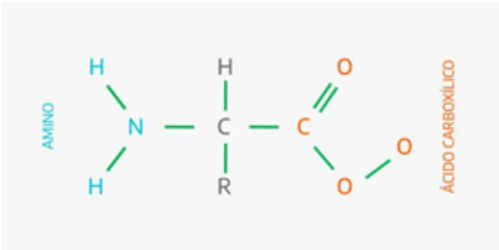


Figura 1: Estructura básica y general de un aminoácido [7]

En resumen, el aminoácido es la unidad básica que constituye la estructura fundamental de las proteínas. Aunque su función principal es esa, los aminoácidos también desempeñan diversas funciones en el organismo por sí solos. Existen 20 aminoácidos diferentes (figura 2). Algunos de ellos pueden ser sintetizados por el cuerpo, y se les llama aminoácidos no esenciales. Sin embargo, hay otros aminoácidos que el cuerpo no puede sintetizar por sí mismo y, por lo tanto, deben obtenerse a través de la alimentación [8]. Estos se conocen como aminoácidos esenciales. Son indispensables para las funciones vitales del organismo y las proteínas que contienen casi todos los aminoácidos esenciales se denominan proteínas de alto valor biológico [7].

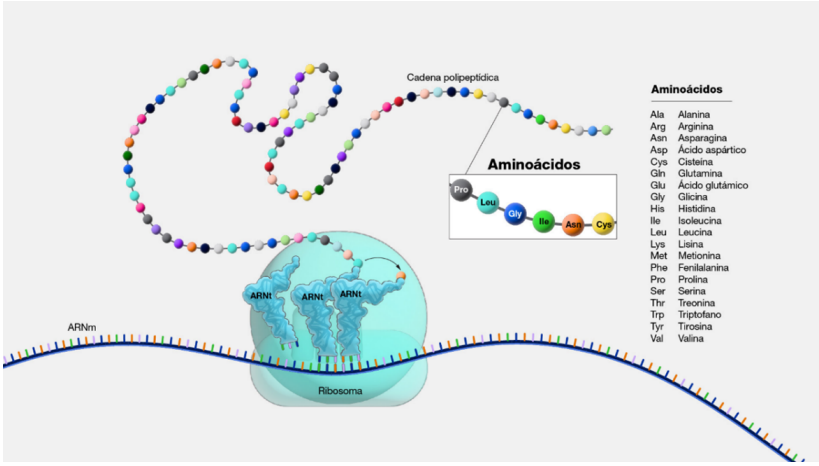


Figura 2: Cadena de aminoácidos y aminoácidos esenciales [8]

Por otro lado, la función de una proteína está determinada por su estructura, por lo que cualquier cambio en su conformación conlleva una pérdida de su capacidad funcional. Además de la estructura tridimensional, las funciones de las proteínas también dependen del tipo, la ubicación y la naturaleza de los grupos R presentes en los aminoácidos [6].

Las proteínas tienen cuatro niveles de organización (figura 3). La estructura primaria es la secuencia de una cadena de aminoácidos. Mientras que la secundaria se produce cuando la secuencia de aminoácidos se pliega y adopta una forma tridimensional en algunas regiones, dichas formas se denominan hoja plisada y hélice alfa. La estructura terciaria se da cuando una proteína madura (secundaria) se pliega sobre si misma. Por último, la estructura cuaternaria se produce gracias a la asociación de varias cadenas polipeptídicas [5].

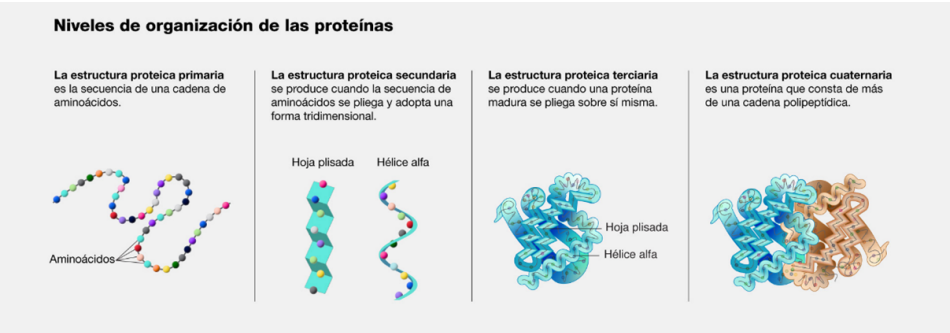


Figura 3: Niveles de organización de las proteínas [5]

También se pueden clasificar según la polaridad de sus cadenas laterales. Estas clasificaciones incluyen categorías como las cadenas laterales polares, que muestran afinidad por el agua, ejemplificadas por la asparagina, glutamina, tirosina, serina y treonina. Por otro lado, las cadenas laterales no polares, que tienen una baja afinidad por el agua, incluyen aminoácidos como la glicina, alanina, valina, leucina, isoleucina, triptófano, prolina, cisteína, metionina y fenilalanina. Además, los

aminoácidos básicos, como la arginina, lisina e histidina, presentan cargas positivas en sus cadenas laterales, mientras que los aminoácidos ácidos, como el ácido aspártico y el ácido glutámico, exhiben cargas negativas [6].

Los genes desempeñan un papel esencial en la producción de proteínas al transmitir la secuencia de aminoácidos necesaria a través del ADN. Estos genes, que son la unidad básica de la herencia, se heredan de los progenitores y contienen información que determina características físicas y biológicas específicas. La mayoría de los genes codifican proteínas individuales o segmentos de proteínas, las cuales cumplen diversas funciones en el organismo [9].

El ácido desoxirribonucleico (ADN) es la molécula encargada de transportar la información genética necesaria para el desarrollo y funcionamiento de un organismo. Está compuesto por dos cadenas complementarias que se entrelazan y forman una estructura en forma de doble hélice similar a una escalera de caracol (figura 4). Cada cadena está formada por grupos alternados de azúcar (desoxirribosa) y fosfato. Cada azúcar se une a una de las cuatro bases nitrogenadas: adenina (A), citosina (C), guanina (G) o timina (T). Las dos cadenas de ADN están unidas mediante enlaces químicos entre las bases, específicamente, adenina se une a timina y citosina se une a guanina [10].

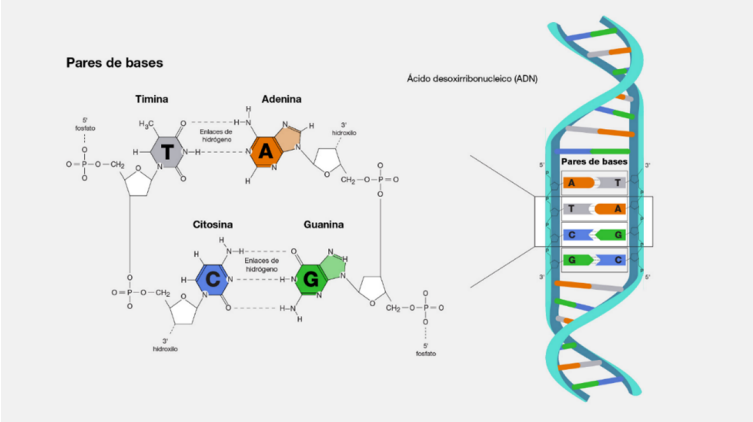


Figura 4: Pares de bases y estructura del ADN [11]

La información genética se codifica en el orden de las bases a lo largo de la estructura del ADN, la cual determina la secuencia de aminoácidos en una proteína. Los genes son segmentos de ADN capaces de determinar características físicas específicas. Sin embargo, estas características no se derivan directamente del ADN, sino a través del ARN mensajero (ARNm), que se forma a partir del ADN y lleva consigo la información necesaria para la producción de proteínas [10].

Este proceso se conoce como el dogma central de la biología molecular: los genes presentes en el ADN actúan como instrucciones para generar ARNm, y este último, a su vez, dirige la síntesis de proteínas. El ARN mensajero (ARNm) es un tipo de ARN de cadena única que participa en la síntesis proteica [10].

Durante el proceso de transcripción, el ARNm se genera a partir de una de las hebras del ADN (figura 5). En el ARNm, la base nitrogenada timina (T) se sustituye por otra llamada uracilo (U), lo que significa que la base nitrogenada complementaria a la adenina pasa a ser el uracilo. De esta

manera, se establece una cadena complementaria que lleva la información genética desde el ADN hasta la síntesis de proteínas [12].

El ARNm actúa como portador de la información genética codificada en el ADN, específicamente en relación a la secuencia de aminoácidos de una proteína. Su función principal es transportar esta información desde el núcleo de la célula, donde se encuentra el ADN, hacia el citoplasma, el entorno acuoso donde tiene lugar la síntesis de proteínas. Es importante destacar que esta sustitución de bases nitrogenadas en el ARNm, de timina por uracilo, es una característica específica del ARN en comparación con el ADN [12].

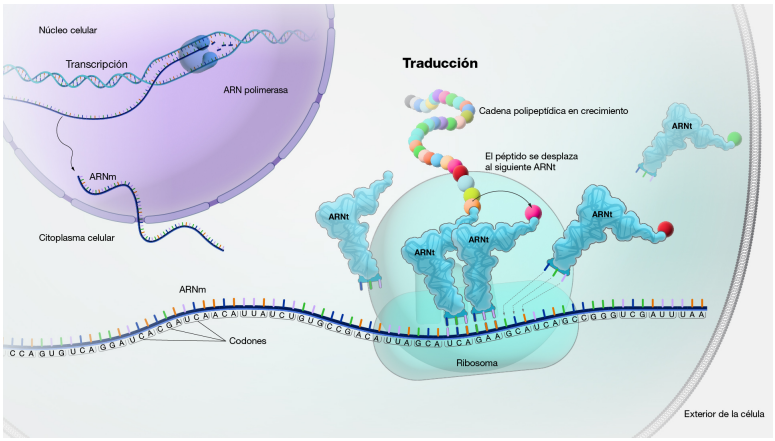


Figura 5: Transcripción y traducción del ADN [13]

La maquinaria de síntesis de proteínas lee el ARNm y decodifica la secuencia de bases del ARNm en una serie de codones, cada uno compuesto por tres bases (figura 5). Este proceso es conocido como traducción. Cada codón es reconocido por un ARN de transferencia (ARNt) específico, que porta el correspondiente aminoácido. De esta manera, los codones del ARNm se traducen en la secuencia de aminoácidos de una cadena proteica en crecimiento [13].

La traducción de los codones a aminoácidos se hace mediante lo que es conocido como el código genético. El código genético se refiere al conjunto de reglas o instrucciones que determina cómo se traducen los tripletes de bases nitrogenadas en el ARNm en la secuencia de aminoácidos durante el proceso de síntesis de proteínas. Es un sistema altamente organizado y universal que permite la traducción precisa de la información genética [14].

El código genético establece que cada codón de tres bases en el ARNm corresponde a un aminoácido específico (figura 6). Sin embargo, la traducción a aminoácidos no es única debido a que algunos aminoácidos pueden ser codificados por más de un codón. En total, existen 64 codones diferentes, lo que representa los 20 aminoácidos utilizados en la formación de las proteínas [14].

Este sistema de codificación es altamente conservado a lo largo de la evolución y es compartido por la mayoría de los organismos, desde bacterias hasta seres humanos. Esta universalidad del código genético permite que la información genética se transmita y se traduzca de manera precisa a través de diferentes especies.

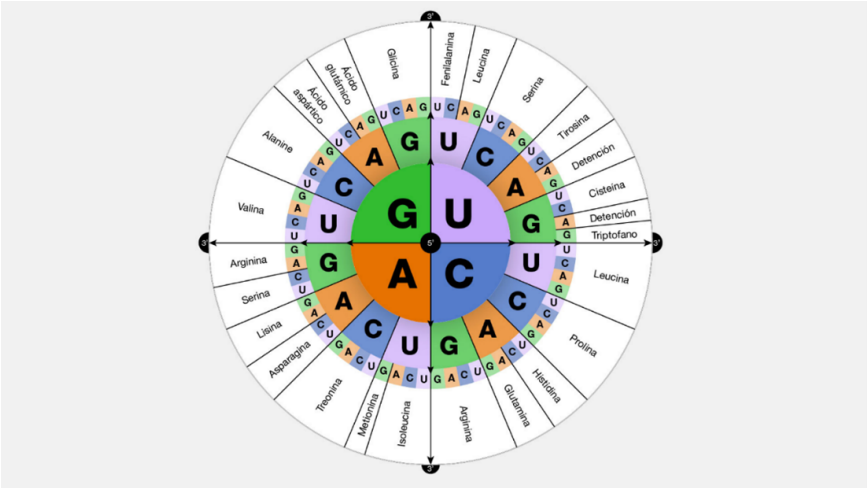


Figura 6: Código genético [14]

3 Proceso Telum

Telum Therapeutics S.L., dedicada al desarrollo de productos antimicrobiales basados en proteínas fago líticas modificadas, se encuentra inmersa en un extenso proceso de investigación para identificar las proteínas más adecuadas que puedan ser utilizadas como enzibióticos. Es decir, substancias alternativas o complementarias a los antibióticos. Las proteínas fago líticas, tal como su nombre sugiere, tienen la capacidad de llevar a cabo la lisis, es decir, la ruptura de la cápsula que resguarda el ADN de las bacterias (patógenos). Esto les permite enfrentar a estas bacterias de manera efectiva.

El proceso de investigación se divide principalmente en dos etapas. En primer lugar, las proteínas atraviesan el departamento APEXp; posteriormente, una vez seleccionadas las mejores proteínas candidatas, pasan al departamento de proteínas. Cada departamento aplica procedimientos específicos para evaluar las proteínas. Como resultado, se generan dos conjuntos de datos: uno que contiene todas las proteínas sometidas a pruebas y sus respectivos resultados, y otro que incluye únicamente aquellas proteínas que avanzan al departamento de proteínas y los procesos específicos a los que son sometidas.

3.1 Departamento APEXp

Dentro del departamento APEXp, se albergan miles de clones de bacterias que portan las proteínas de interés, las cuales serán expresadas y sometidas a pruebas. En esta etapa de la investigación, la secuencia de proteínas transportada por cada clon es desconocida. Los clones se transfieren a una placa con 96 pocillos para evaluar su actividad. Aquellos candidatos que demuestren mayor actividad serán seleccionados para el proceso de caracterización que se lleva a cabo en el departamento de proteínas.

Para realizar la selección de las candidatas, se emplean dos análisis de alto rendimiento, que se combinan con la secuenciación de los clones para examinar tanto el ADN como la secuencia de proteínas del clon de interés. Esta estrategia permite evitar la inclusión de secuencias duplicadas y

garantizar la identificación precisa de las proteínas candidatas más prometedoras.

El primer análisis de alto rendimiento utilizado es la prueba Halo, que consiste en expresar las proteínas y evaluar su actividad contra diferentes patógenos en un medio sólido. Durante esta prueba, se observan las colonias rodeadas por un halo transparente (figura 7), lo cual indica actividad antimicrobiana. Los resultados obtenidos se registran en una plantilla de Excel denominada "Template Halo Assay". En esta plantilla, se anotan los datos correspondientes a los tamaños de los halos formados por las proteínas contra cada patógeno probado. El criterio de anotación del tamaño del halo se define de la siguiente manera: 0 si no hay halo, 1 si el halo envuelve la colonia en un rango de 1 a 2 mm, 2 si el halo envuelve de 2 a 4 mm, 3 si el halo envuelve de 4 a 5 mm, y 4 si el halo es mayor de 5 mm. Esta información permite identificar las proteínas candidatas más efectivas contra los patógenos evaluados.

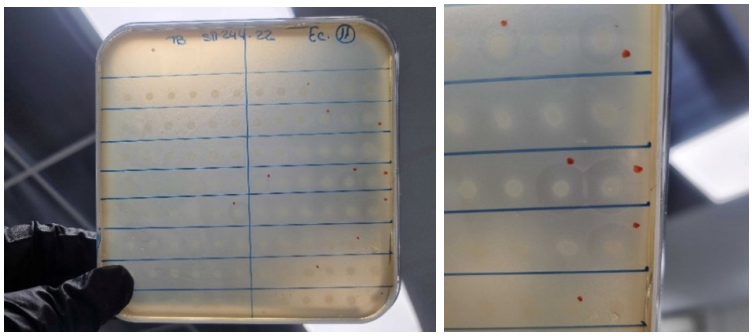


Figura 7: Prueba halo

El segundo análisis de alto rendimiento corresponde a una prueba de inhibición de crecimiento, en la que se expresan las proteínas y se evalúan en un máximo de 8 patógenos distintos en un medio líquido. Los datos resultantes de este experimento son generados mediante un espectrofotómetro y una incubadora. El equipo utilizado produce un archivo Excel que contiene los resultados de la prueba. A partir de este archivo, se calcula el área bajo la curva (AUC), lo cual permite cuantificar la eficacia de las proteínas en la inhibición del crecimiento de los patógenos.

El departamento APEXp culmina su proceso con la secuenciación de las muestras. El plato de 96 pocillos se envía a LGC-Genomics para que realicen la secuenciación. Los resultados obtenidos de la secuenciación vienen con un sistema de código basado en el orden numérico, el cual no debe ser modificado para garantizar el seguimiento preciso de cada muestra. Una vez recibidos los resultados, se cargan en la plataforma Geneious Prime, donde las secuencias son alineadas con la secuencia de ADN de referencia (figura 8). Los resultados del alineamiento son guardados automáticamente, y además se genera un informe sobre el ensamblaje de la secuencia.

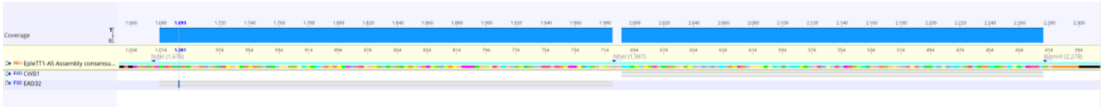


Figura 8: Alineamiento de secuencias protéicas

El ensamblaje de la secuencia implica la tarea de alinear y unir fragmentos de secuencias de ADN para reconstruir la secuencia original del ADN de referencia. Durante este proceso, se deben tener en cuenta varios aspectos visuales, como la presencia del codón de inicio, una secuencia de ADN correcta y sin mutaciones, la secuencia de proteína correcta en el marco de lectura, la presencia de nucleótidos correctos en cada posición (especialmente si las proteínas provienen de clonación o han sido generadas mediante técnicas de mutagénesis o ingeniería de proteínas), la presencia de la etiqueta 6xHisTag (etiqueta de histidina que se utiliza facilitar la purificación y detección de proteínas) y la presencia del codón de parada.

Una vez que la secuencia se considera correcta, los resultados se registran en un archivo Excel llamado "Summary sequencing results", donde el identificador de la secuencia es el código proporcionado por LGC, facilitando así su seguimiento y análisis posterior.

Una vez concluidas las pruebas llevadas a cabo por el departamento APEXp, se procede a seleccionar cuidadosamente las proteínas que cumplen con los criterios de calidad para someterlas a una caracterización más exhaustiva en el departamento de proteínas. La decisión de selección se basa en los resultados obtenidos en las pruebas anteriores y se hace de manera manual. Los criterios son: la presencia de halo, que el AUC sea por lo menos 4 veces el obtenido en el pocillo de control que contiene antibiótico y las proteínas que poseen una secuencia de ADN precisa y única en el plásmido utilizado para la expresión de la proteína. Una vez completada la selección, se procede a copiar manualmente las proteínas elegidas junto con sus características en la base de datos.

3.2 Departamento de Proteínas

En el departamento de proteínas, se llevan a cabo una serie de experimentos con el objetivo de seleccionar las proteínas más activas, que serán consideradas como candidatas para las pruebas preclínicas. Estos experimentos se enfocan en evaluar y comparar la actividad de las proteínas seleccionadas anteriormente en el departamento APEXp.

En el departamento de proteínas, el primer experimento consiste en llevar a cabo la sobreexpresión y purificación de las proteínas de interés. Este proceso se divide en dos tipos: .ºn the benchƶ .ºn an AKTA". Para el enfoque "on the Bench" se realiza la sobreexpresión y purificación de las proteínas en el laboratorio. Después de la purificación, se realiza una técnica llamada SDS-PAGE para separar las proteínas en función de su peso molecular. La imagen resultante de la SDS-PAGE permite calcular el peso molecular de la proteína de interés (figura 9).

Para el segundo enfoque, además de la sobreexpresión y purificación de las proteínas, se emplea un sistema automatizado llamado AKTA. La salida de este proceso también proporciona la información necesaria para calcular el peso molecular de la proteína. Sin embargo, en este caso, hay tres posibles tipos de purificaciones a realizar: cromatografía de afinidad de metal (IMAC), cromatografía de intercambio iónico (IEX) y cromatografía de exclusión por tamaño (SEC) (figura 9). Estas purificaciones se pueden realizar de forma independiente o en una combinación secuencial para obtener una purificación más completa y específica.

Dentro del primer paso del experimento de sobreexpresión y purificación de proteínas, también se realiza la cuantificación de proteínas para determinar su concentración. Este proceso se lleva a cabo utilizando un colorímetro y el método de Bradford, el cual se basa en la afinidad de un tinte conocido como .ᴬzul de Coomassie"para unirse selectivamente a las proteínas en una muestra y modificar su

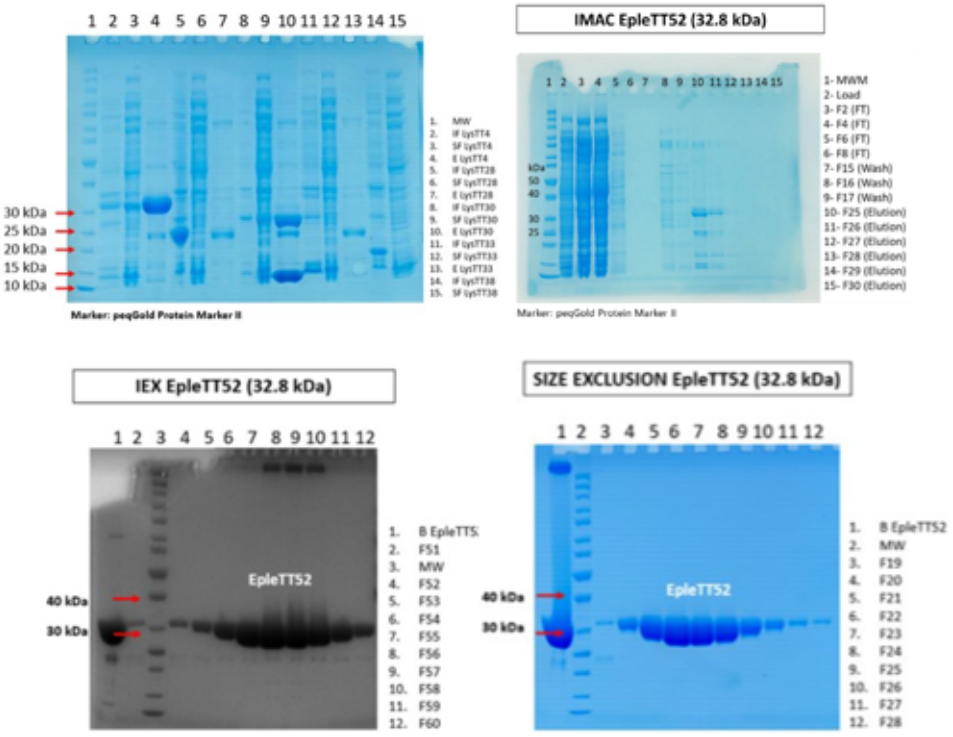


Figura 9: Procesos de expresión y purificación de proteínas *on the bench* (esquina superior izquierda) y *on AKTA* para purificaciones IMAC, IEX y SEC

color en relación a la concentración de proteínas presente. Cuando este tinte se incorpora a una solución que contiene proteínas, estas se unen al mismo, lo que provoca un cambio en la absorbancia [15]. El cambio en la absorbancia permite determinar la concentración de proteínas en la muestra utilizando una curva estándar.

Para la construcción de la curva estándar se preparan soluciones de una proteína de referencia con distintas concentraciones conocidas. Luego, se aplica el método de Bradford a cada una de estas soluciones y se registra el cambio en la absorbancia resultante. Estos valores obtenidos se utilizan para construir la curva estándar, que servirá como referencia para cuantificar la concentración de proteínas en muestras desconocidas [15].

Los datos de la prueba se obtienen a través de la medición de la absorción de las muestras utilizando los equipos EQ18 y EQ19. Estos resultados son luego utilizados para calcular la concentración de proteínas en mg/ml, basándose en la curva estándar previamente establecida. Una vez calculada la concentración de proteína, esta información es registrada manualmente en la base de datos, al igual que el peso molecular previamente calculado para cada proteína.

El segundo paso en este departamento es la caracterización de la actividad antimicrobial de las proteínas in vitro. Los experimentos realizados son los siguientes “Checkerboard”, MIC (Concentra-

ción Mínima Inhibitoria) y MBC (Concentración Mínima Bactericida) en un medio; y el ensayo de reducción de turbidez (TRA).

Los experimentos Checkerboard, MIC y MBC se hacen en placas de 96 pocillos con intención de ver la cantidad de proteína expresada en µg/ml necesario para obtener la mínima concentración inhibitoria (inhibe el crecimiento), la mínima concentración bactericida (mata el microorganismo) y el efecto de la combinación con antibiótico para eliminar la bacteria. Son técnicas para evaluar la actividad antimicrobiana de compuestos y determinar su eficacia frente a microorganismos (patógenos) específicos en medios de cultivo.

En el contexto del experimento de Checkerboard y MIC, los resultados se evalúan visualmente y se registran manualmente según la presencia de un pocillo transparente en la placa. Si el resultado es incierto o ambiguo, se utiliza una sustancia llamada *pesto blue* para teñir la placa. Cuando las células están muertas, la placa adquiere un color azul. Por otro lado, si las células están vivas, la placa muestra un color rosáceo (figura 10).

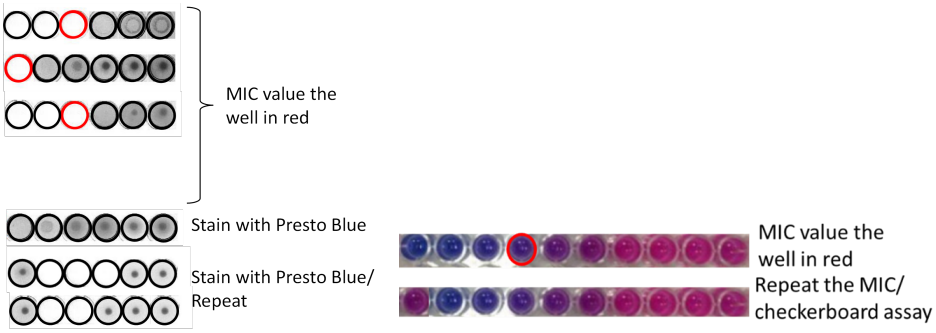


Figura 10: Prueba para calcular el valor MIC

El valor MIC se apunta en la base de datos directamente, mientras que el valor del checkboard requiere de cálculos que dependen de los valores MIC. Los cálculos son realizados manualmente obteniendo el valor del Índice de Concentración Fraccional (FICI).

$$FICI = \sum FIC = FIC_A + FIC_B \quad \text{donde } FIC \text{ se calcula como:}$$

$$FIC_A = \frac{A}{MIC_A} \quad y \quad FIC_B = \frac{B}{MIC_B}$$

A y B son el MIC del antibiótico y de la proteína, respectivamente cuando son combinadas en un solo pocillo por fila de la placa. MIC_A y MIC_B son los MIC del antibiótico o de la proteína, respectivamente cuando se usan individualmente sin combinarlos en el mismo pocillo.

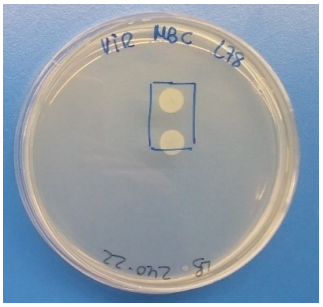
En el análisis de los resultados, se pueden obtener tres posibles conclusiones: sinergismo, indiferencia o antagonismo, y esto dependerá de los valores del FICI. Los resultados obtenidos son introducidos manualmente en cada correspondiente experimento.

- FICI $\leq 0,5$: Sinergismo
- 0,5 $<$ FICI $<$ 4: Indiferencia

- FICI ≥ 4: Antagonismo

Cuando el valor del FICI es menor o igual a 0,5, se considera un resultado de sinergismo, lo cual indica que el antibiótico y la proteína son complementarios al enfrentar al microorganismo, potenciando su actividad de manera conjunta. En caso de que el valor del FICI esté en el rango de 0,5 a 4, se clasifica como resultado de indiferencia. Esto significa que la combinación del antibiótico y la proteína no potencia ni inhibe significativamente la actividad frente al patógeno, mostrando una interacción neutra. Por último, si el valor del FICI es mayor o igual a 4, se considera un resultado de antagonismo. Esto indica que la combinación de ambos tiene un efecto contrario y no se complementan, lo que resulta en una reducción en su actividad frente al microorganismo.

Para obtener los valores de MBC (Concentración Mínima Bactericida), se requiere un paso adicional en el proceso. Este paso consiste en transferir los pocillos que mostraron el MIC a una nueva placa de cultivo (Figura 11a), con el objetivo de observar el número de colonias bacterianas que crecen.



(a) Placa de cultivo.

Example

Well	Amount of agent (µg/mL)	CFU/plate	Total CFU	≥99.9% kill
1	128	2, 0	2	Yes
2	64	3, 8	11	Yes
3	32	10,8	18	Yes
4	16	15, 11	33	No
5	8	TNCT	TNCT	No
MBC = 32 µg/mL				

(b) Tabla para determinar MBC.

Figura 11: Prueba MBC.

Basándose en el conteo de colonias, se realiza una interpretación utilizando la tabla de la figura 11b. En esta tabla se registran los números identificativos de los pocillos, el número de agentes utilizados, las unidades formadoras de colonias (colony-forming unit, CFU) por placa, el total de CFU y si se ha conseguido eliminar más del $99,9\,\%$ de los microorganismos. Finalmente, el valor de MBC se determina como el número mínimo de agentes necesarios para lograr la eliminación del porcentaje establecido de microorganismos. Los valores de MBC se introducen manualmente en los experimentos correspondientes. Posteriormente, estos datos son ingresados en la base de datos de Telum, donde se recopilan los resultados de los tres experimentos realizados en esta fase del departamento de proteínas.

Para concluir la primera fase de los experimentos in vitro, se realiza un ensayo de reducción de turbidez (TRA). Este ensayo se lleva a cabo en una placa de 96 pocillos con el objetivo de obtener la actividad específica de la proteína expresada como AUC/µg de proteína. En el ensayo, se prueban diferentes concentraciones de proteínas y se registran los valores de absorción utilizando los equipos EQ18 y EQ19. Estos resultados obtenidos de los equipos se utilizan para calcular el AUC/µg de proteína. El valor calculado se introduce manualmente en el experimento correspondiente y en la base de datos de Telum.

Los candidatos más activos avanzan a la siguiente fase, donde se someten a una caracterización más detallada. Estos experimentos adicionales permiten una evaluación completa de los candidatos y ayudan a determinar su potencial como agentes antimicrobianos en diferentes condiciones. En esta etapa, se llevan a cabo varios experimentos para determinar el valor MIC y MBC en presencia del 50 % de suero humano y otro en presencia de 50 % suero de ratón. Además, se realiza el ensayo de tiempo de eliminación (*Time Killing Assay*) para evaluar la capacidad del candidato para matar microorganismos a lo largo del tiempo. Asimismo, se realizan experimentos de prevención de formación de biofilm para evaluar la capacidad del candidato para evitar la adhesión y formación de biofilms en superficies. El último experimento en esta fase es la eliminación de biofilm para evaluar si el candidato puede eliminar biofilms ya formados.

La obtención de los valores MIC y MBC en diferentes medios sigue el mismo procedimiento mencionado en la fase anterior del departamento de proteínas. Se realizan los ensayos correspondientes utilizando suero humano y suero de ratón como medio de cultivo. Los resultados de los ensayos se registran manualmente en la base de datos.

El ensayo de tiempo de eliminación se realiza para determinar la carga bacteriana en presencia de proteínas durante períodos de 3 y 6 horas. En este ensayo, se cuenta el número de bacterias presentes en las placas (figura 12) donde se lleva a cabo el experimento. Los resultados son obtenidos manualmente y se calcula la CFU/ml. Durante el ensayo, se evalúa cómo la presencia de las proteínas afecta la cantidad de bacterias a lo largo del tiempo, lo que proporciona información importante sobre la capacidad de las proteínas para eliminar o reducir la carga bacteriana en el entorno específico de estudio.

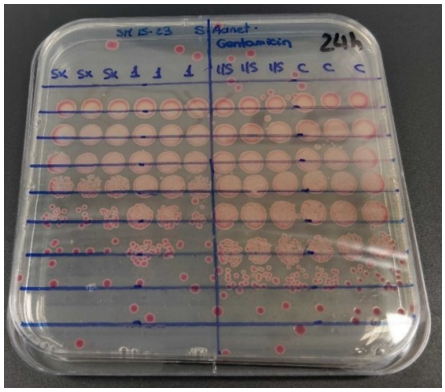


Figura 12: Placa donde lleva a cabo el experimento de tiempo de eliminación.

En cuanto a los ensayos relacionados con el biofilm, la prevención de su formación se realiza utilizando una placa de 96 pocillos, a partir de la cual se determina el porcentaje de biofilm formado. Un biofilm es un complejo ecosistema microbiano que involucra una o varias especies de microorganismos que se adhieren a una superficie, ya sea viva o inerte. En este ensayo, se prueban diferentes concentraciones de proteínas y se mide el valor de absorción utilizando el equipo EQ19. Los resultados obtenidos permiten calcular el porcentaje de inhibición del biofilm. La medida del porcentaje de inhibición proporciona información sobre la efectividad de las proteínas en evitar la adhesión y formación de biofilms en diferentes superficies.

El segundo ensayo relacionado con el biofilm es la prueba de eliminación de biofilm, también llevada a cabo en una placa de 96 pocillos. Este ensayo sigue el mismo procedimiento que el de prevención de biofilm, con la diferencia de que en este caso se mide el porcentaje de biofilm eliminado por los candidatos que ya tenían biofilm formado. Todos los resultados obtenidos en esta fase son introducidos manualmente en la base de datos de Telum.

Una vez completada la ejecución de los experimentos mencionados, los candidatos más prometedores avanzan a una tercera fase en la que se realizan procedimientos tanto ex vivo como in vivo.

Los ensayos ex vivo abarcan la prevención y eliminación de biofilms en tres modelos de piel distintos: piel sin heridas, piel con heridas y piel con heridas por quemaduras. Los resultados de estos experimentos se obtienen en dos intervalos de tiempo distintos, y se calcula la reducción logarítmica en comparación con los datos de control. Los datos recopilados provienen del recuento de colonias en las placas (Imagen X, la placa de tiempo de eliminación) utilizadas en el experimento.

En términos de la microbiología y de pruebas antimicrobianas, la reducción logarítmica implica la disminución del número de microorganismos (como bacterias) en una muestra o superficie, expresada en una escala logarítmica. Por ejemplo, una reducción logarítmica de 1 significa que la cantidad de microorganismos se ha reducido en 10 veces. Este enfoque permite evaluar de manera cuantitativa el impacto de las intervenciones antimicrobianas en la reducción de la carga microbiana en los diferentes modelos de piel utilizados en el estudio.

El último experimento llevado a cabo en la fase de generación de la base de datos de Telum es el ensayo in vivo. Para realizar este experimento, se colabora con una empresa externa que proporciona los resultados en formato de presentación de PowerPoint, el cual no siempre sigue un formato uniforme. Los datos numéricos relativos al porcentaje de supervivencia, la reducción logarítmica en el bazo y la reducción logarítmica en los pulmones se ingresan manualmente en la base de datos, junto con los resultados de los ensayos ex vivo. La realización de este ensayo in vivo proporciona información valiosa sobre el rendimiento de los candidatos en un entorno más cercano a las condiciones reales. Por ahora no hay datos de dichos ensayos.

4 Conjuntos de Datos

Gracias al proceso llevado a cabo por Telum, ahora contamos con una base de datos que nos proporciona un punto de partida para nuestro trabajo. Durante el proceso de investigación, hemos observado que, en cada fase o etapa, solo las proteínas candidatas más activas son sometidas a todos los experimentos. Esto ha dado lugar a un conjunto de datos que muestra cierta dispersión y una presencia significativa de valores faltantes.

Para lograr un éxito en la aplicación de técnicas estadísticas y de aprendizaje automático, es esencial contar con un conjunto de datos lo más completo posible. Con el fin de abordar esta necesidad, Telum ha facilitado una base de datos más enfocada, reduciendo la cantidad de variables y centrándose en los experimentos realizados en todas las proteínas de dicho departamento. Esta estrategia nos permite obtener un conjunto de datos más coherente y abarcador según nuestras necesidades. Además, se ha provisto un historial que incluye todas las proteínas evaluadas en el departamento APEXp.

Iniciamos al describir el conjunto de datos que abarca el historial de las proteínas evaluadas en el departamento APEXp, junto con los resultados de sus experimentos. En esta colección de datos, encontramos tanto las proteínas que han avanzado al departamento de proteínas como aquellas que no lo han hecho. Con la perspectiva de realizar una predicción binaria en etapas posteriores, hemos optado por seleccionar las variables con las que abordaremos nuestros objeticos. Este conjunto esta formado por 231 instancias y 12 variables.

- **Name**: una variable categórica que identifica a la proteína.
- **Number of amino acids**: una variable numérica que refleja la longitud de la secuencia proteica, es decir, la cantidad de aminoácidos presentes.
- **Protein sequence**: secuencia de aminoácidos que conforman la proteína.
- **Molecular weight**: una variable numérica que indica el peso molecular de la proteína.
- **pI (punto isoeléctrico)**: una variable numérica que representa el pH en el cual la carga neta de la proteína es cero, lo que significa que las cargas positivas y negativas están equilibradas. Debido a la neutralidad de la carga, su movilidad electroforética es nula [16].
- **Charge at pH 7**: la carga de la molécula en un entorno con pH neutro, expresada como un valor numérico.
- **Extinction coefficient**: en el contexto biológico, este coeficiente mide cuánta luz absorbe una sustancia a una longitud de onda específica. Es una propiedad intrínseca de la sustancia y puede variar según su naturaleza química y la longitud de onda de la luz utilizada para medir la absorción. Es una herramienta esencial para cuantificar y caracterizar la interacción de las biomoléculas con la luz [17].
- **Frequency of acidic groups n (%)**: la proporción de grupos ácidos en la secuencia de la proteína.
- **Frequency of basic groups n (%)**: la proporción de grupos básicos en la secuencia.
- **Frequency of charged groups n (%)**: la proporción de grupos con carga en la secuencia
- **Frequency of polar uncharged groups n (%)**: la proporción de grupos polares sin carga en la secuencia.
- **Frequency of polar hydrophobic groups n (%)**: la proporción de grupos polares hidrófobos en la secuencia.

Por otro lado, emplearemos un segundo conjunto de datos que incluye exclusivamente las proteínas más activas que han sido procesadas en el departamento homónimo. Al igual que hicimos con el conjunto anterior, seleccionaremos las variables más relevantes para nuestro estudio. Las dimensiones de este conjunto son más reducidas que las del anterior estando formado por 94 instancias y 13 variables.

- **Protein**: variable que contiene el identificador de la proteina.
- **Protein sequence**: secuencia de aminoácidos

- **Protein length**: longitud de la secuencia de la proteina, es decir, número de aminoácidos.
- **Molecular weight**: peso molecular.
- **pI**: punto isoeléctrico
- **Pathogen**: patógeno frente al que se prueba la proteina lítica.
- **Protein purification solubility (%)**: porcentaje de solubilidad durante la purificación de la proteina.
- **Muralytic (TRA)**: resultados del ensayo de reducción de turbidez.
- **MIC (µM)**: resultados del ensayo MIC.
- **Checkerboard with imipenem**: valores calculados en el ensayo checkerboard con antibiótico.
  - **Effect**: el efecto de la proteina en conjunto con el antibiótico. Los resultados pueden ser sinergismo, indiferencia o antagonismo.
  - **Concentration imipenem for synergism (µM)**: cuando el efecto es sinergético se calcula la concentración de antibiótico necesaria.
  - **Concentration protein for synergism (µM)**: cuando el efecto es sinergético se calcula la concentración de proteína necesaria.

Vamos a emplear ambos conjuntos de datos con el propósito de investigar la decodificación de proteínas con un enfoque predictivo. Asimismo, aprovecharemos la disponibilidad de los datos experimentales para utilizar estas variables en nuestras predicciones. De esta manera, podremos comparar las predicciones resultantes de ambas fuentes y evaluar si las decodificaciones por sí solas son suficientes para predecir las variables de interés.

5 Métodos de Decodificación

Uno de los objetivos de este trabajo es explorar las diversas formas de decodificar las secuencias de proteínas y determinar si, solo con estas decodificaciones, podemos predecir aspectos significativos del proceso llevado a cabo por Telum. Nuestra principal herramienta de trabajo será el paquete *protr*, ampliamente utilizado en cuestiones de bioinformática. Este paquete cuenta con múltiples tipos de decodificaciones, que van desde las más simples hasta las más complejas. Sin embargo, para aprovechar estas implementaciones, primero debemos formatear las secuencias proteicas en un formato específico.

El formato requerido para las secuencias es el formato FASTA, en el cual las secuencias deben estar identificadas como ">SecuenciaX", donde "X.^{es} el número identificador asignado a la proteína. Después del identificador, debe haber un salto de línea antes de comenzar a escribir la secuencia de aminoácidos. Cada 80 aminoácidos, se debe agregar otro salto de línea. Todas las proteínas deben estar escritas en este formato, una detrás de la otra, sin espacios entre ellas.

Entendemos que este proceso puede volverse tedioso, especialmente si se trabaja con un conjunto de datos grande. Por esta razón, hemos desarrollado un código que automatiza la conversión de las secuencias al formato FASTA. Este código genera dos archivos FASTA: uno que contiene todas las secuencias del conjunto de datos históricos, en el cual es necesario eliminar la fila 230, ya que está

vacía y podría causar errores al intentar decodificarla. El segundo archivo contiene únicamente las secuencias de proteínas del conjunto formado por las más activas.

Después de leer el archivo FASTA, es importante verificar que las secuencias solo contienen los 20 aminoácidos esenciales(figura 13). Si encontramos una proteína que contiene un aminoácido que no pertenece a los 20 mencionados, como en el caso de la proteína 178 del conjunto histórico, esa secuencia se considera no apta para el estudio de las decodificaciones y debe ser eliminada del análisis.

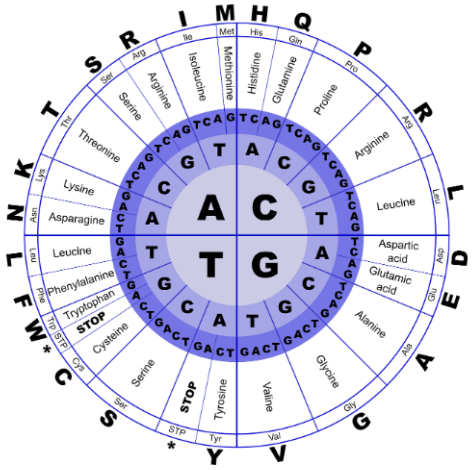


Figura 13: Siglas en las que se traduce cada aminoácido [18]

Hemos seleccionado cinco métodos diferentes de decodificación de secuencias de proteínas. Los dos primeros métodos son los más simples y se basan en las proporciones de aminoácidos. El primero de ellos decodifica las secuencias según la proporción de cada uno de los 20 aminoácidos, lo que nos proporciona 20 variables predictoras. El segundo método es similar, pero calcula las proporciones de todos los pares posibles de aminoácidos, es decir, los dipéptidos. Con esta decodificación, tenemos 20 aminoácidos que se combinan con otros 20, lo que nos da un total de 400 dipéptidos como variables predictoras [18].

El tercer método se basa en la composición, transición y distribución de aminoácidos dentro de las secuencias proteicas. En total, se evalúan siete características clave de las proteínas:

- **Hidrofobicidad:** Esta propiedad refleja la tendencia de las sustancias a evitar o repeler el agua debido a la presencia de propiedades no polares o regiones no polares en su estructura molecular.
- **Volumen normalizado de van der Waals:** Se trata de una medida esencial para comprender el espacio ocupado por las moléculas y cómo interactúan con otras en contextos químicos y biológicos [19].
- **Polaridad:** Aquí se analiza la existencia y la distribución de regiones cargadas eléctricamente en la estructura tridimensional de las proteínas [20].

- **Polarizabilidad:** Esta característica mide la capacidad de una sustancia para responder a un campo eléctrico externo, lo que implica un cambio en su distribución electrónica y la creación de una polarización eléctrica temporal [20].
- **Carga:** La carga se calcula sumando las cargas eléctricas de los aminoácidos que conforman la secuencia proteica. En este contexto, la carga se clasifica en tres tipos distintos: positiva, neutra o negativa.
- **Estructura secundaria:** Se refiere a la organización tridimensional de las regiones locales en la cadena polipeptídica de la proteína. Describe cómo interactúan entre sí los aminoácidos que componen la proteína a nivel local, formando patrones regulares de plegamiento.
- **Accesibilidad solvente:** Esta propiedad está relacionada con la capacidad de la proteína para interactuar con otras moléculas. Mide el grado en que ciertos átomos de la proteína están expuestos y accesibles al solvente circundante, generalmente agua [21].

	Group 1	**Group 2**	**Group 3**
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals Volume	0-2.78	2.95-4.0	4.03-8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	4.9-6.2	8.0-9.2	10.4-13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	0-1.08	0.128-0.186	0.219-0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary Structure	Helix	Strand	Coil
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent Accessibility	Buried	Exposed	Intermediate
	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, S, P, T, H, Y

Cuadro 1: Tabla de clasificación de aminoácidos para las propiedades de composición, transición y distribución [18]

Cada una de estas características se organiza en tres grupos, lo que da como resultado un total de 21 variables predictoras. La asignación de un valor a un grupo específico se basa en los aminoácidos presentes en la secuencia proteica, como se especifica en la tabla 1. Entre los atributos disponibles, se pueden calcular tres descriptores: composición, transición y distribución. En este contexto, hemos optado por utilizar el descriptor de composición, que proporciona un porcentaje global para cada clase codificada a lo largo de la proteína, tal como se detalla en la tabla mencionada. Tomemos, por ejemplo, la secuencia "MTEITAAMVKELRESTGAGA". Si la codificamos según su hidrofobicidad, obtendremos la siguiente representación: "32132223311311222222". Aquí, hemos asignado 5, 10 y 5 individuos a cada clase correspondiente [18]. Luego, sustituimos estos valores en la fórmula (1) y calculamos los resultados.

$$C_r = \frac{n_r}{n} \quad r = 1, 2, 3 \tag{1}$$

La cuarta decodificación se refiere a la composición anfifílica de pseudo-aminoácidos, que se basa en la composición de aminoácidos y se centra en la clasificación de estos en dos grupos distintos: aminoácidos anfipáticos y no anfipáticos. Esta decodificación resulta útil para caracterizar la tendencia de una secuencia de proteína a tener regiones anfipáticas, las cuales son zonas que presentan una parte hidrofóbica y otra hidrofílica. Estas regiones anfipáticas desempeñan un papel crucial en muchas proteínas debido a su participación en la interacción con membranas celulares y otras proteínas.

Con esta decodificación, generamos 20 variables constantes que representan la frecuencia de cada aminoácido con respecto a la frecuencia total y la media de las características calculadas multiplicadas por un vector de pesos. Estas características están relacionadas con la correlación en términos de hidrofobicidad e hidrofilicidad. Se parte de los valores conocidos de hidrofobicidad e hidrofilicidad para cada aminoácido y se calculan las correlaciones con respecto al aminoácido "j"de la secuencia. Para aclarar, si estamos en el aminoácido ïz tomamos "jïgual a uno, la primera característica será la media de las correlaciones entre aminoácidos contiguos. Esta característica se calcula tanto para la hidrofobicidad como para la hidrofilicidad, lo que resulta en 2 variables por cada característica. Por defecto, el número de características es 30, lo que nos da un total de 80 variables [18].

La última técnica de decodificación que hemos elegido emplear es la autocorrelación de Moreau-Broto normalizada (2). Esta técnica implica la definición de un número específico de *lags* (retrasos) y el cálculo de las autocorrelaciones de Moreau-Broto considerando estos lags en las secuencias de proteínas (3). Este cálculo de autocorrelación se realiza para cada una de las ocho propiedades predefinidas y se ejecuta para cada lag, que por defecto es de 30. En consecuencia, obtenemos un conjunto de 240 variables distintas para nuestro análisis [18].

$$ATS(d) = \frac{AC(d)}{N - d} \quad d = 1, 2, \ldots, \text{nlag} \tag{2}$$

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad d = 1, 2, \ldots, \text{nlag} \tag{3}$$

Las ocho propiedades abarcan una amplia gama de características relacionadas con la química y la física de las secuencias proteicas:

- **Normalized Average Hydrophobicity Scales** (CIDH920105): Escala de hidrofobicidad normalizada que mide la tendencia de una sustancia a repeler el agua debido a sus regiones no polares.
- **Average Flexibility Indices** (BHAR880101): Índices de flexibilidad promedio que cuantifican la flexibilidad de una secuencia proteica en términos de la variabilidad en la distancia entre sus átomos [22].
- **Polarizability Parameter** (CHAM820101): Parámetro de polarizabilidad.
- **Free Energy of Solution in Water, kcal/mole** (CHAM820102): Energía libre de solución en agua, medida en kilocalorías por molécula, que indica la capacidad de una sustancia para disolverse en agua [23].

- **Residue Accessible Surface Area in Tripeptide** (CHOC760101): Área superficial accesible de residuos en tripeptidos, que se refiere a la medida en que los átomos de una proteína están expuestos al solvente [24].
- **Residue Volume** (BIGC670101): Volumen de residuo, que cuantifica el espacio ocupado por un aminoácido o un residuo en una proteína. Residuo se denomina a la parte única entre cada uno de los 20 aminoácidos [25].
- **Steric Parameter** (CHAM810101): Parámetro estérico que describe la influencia del tamaño y la forma de un grupo funcional en las interacciones moleculares [26].
- **Relative Mutability** (DAYM780201): Mutabilidad relativa, que evalúa la probabilidad de que un aminoácido en una secuencia proteica mute o cambie a lo largo del tiempo.

6 Metodología

En este apartado, seguiremos procedimientos similares para abordar los desafíos planteados tanto en el conjunto histórico como en el conjunto de proteínas activas. Se centrará en la adecuada lectura de datos, el preprocesamiento, un análisis exploratorio exhaustivo y la fase predictiva como los principales elementos a abordar.

6.1 Estudio de la actividad de las proteínas

En este contexto particular, nuestro objetivo es determinar si podemos prever si una proteína es lo bastante activa para avanzar al departamento de proteínas. Para abordar esta tarea, vamos a establecer un problema de clasificación binaria. Para ello, necesitamos introducir una nueva columna en el registro histórico de proteínas. Esta columna tendrá un valor binario que indicará si la proteína ha avanzado al siguiente departamento o no.

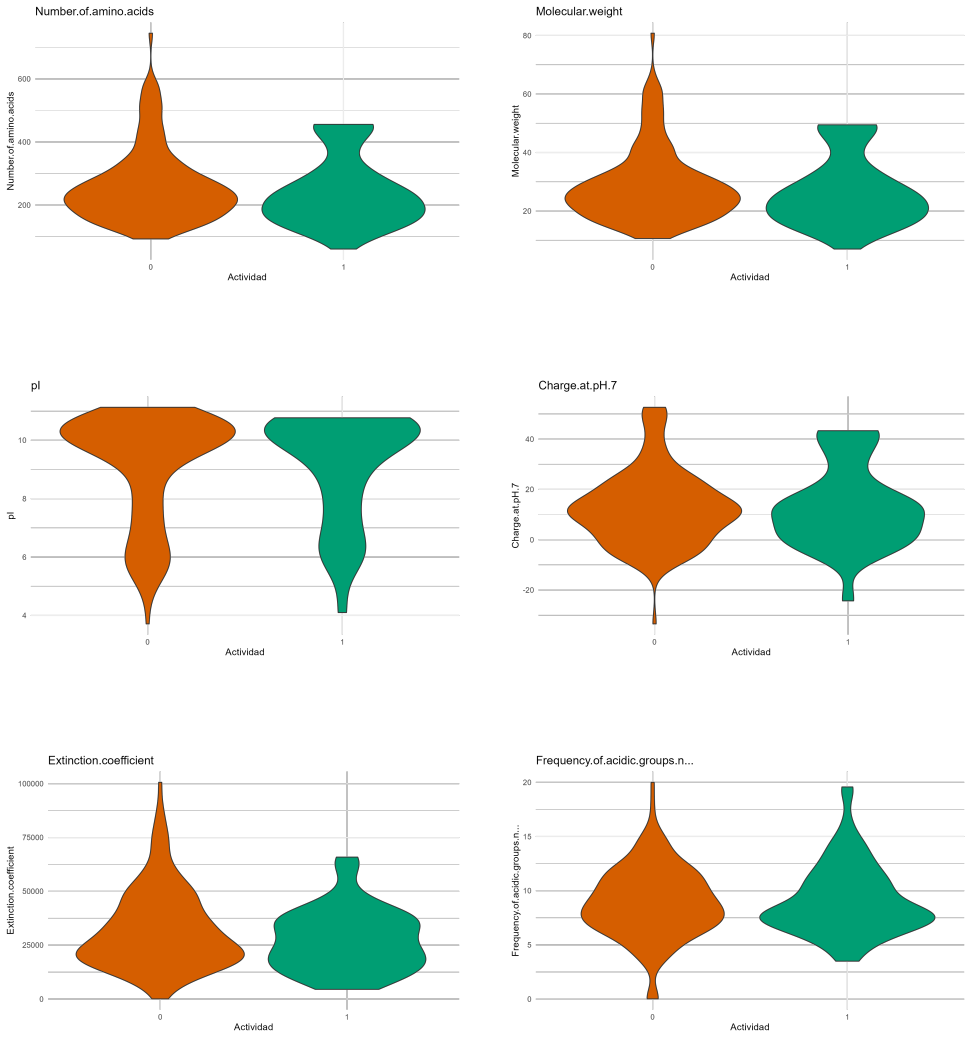
Iniciamos el proceso con la lectura de los datos, utilizando los dos conjuntos de datos proporcionados. Aunque en esta primera fase nos enfocamos exclusivamente en los datos históricos, es necesario utilizar el conjunto de datos del departamento de proteínas para generar la columna binaria. Seleccionaremos la columna "Protein", que contiene los nombres de las proteínas, y la compararemos con la columna "Name.[en] el conjunto de datos históricos. Aquellos nombres que coincidan recibirán un valor de 1 en la nueva columna, mientras que los que difieran obtendrán un valor de 0. A través de esta columna binaria, logramos identificar las proteínas más y menos activas de manera efectiva.

6.1.1 Predicción de la actividad con el conjunto de variables

En una primera aproximación al problema, nos centraremos exclusivamente en el uso de las variables proporcionadas por los experimentos, excluyendo la variable de secuencia proteica. Esta última la emplearemos en una etapa posterior de nuestro análisis.

Una vez que hemos creado la columna binaria, procedemos a realizar un análisis exploratorio de nuestro conjunto de datos. En primer lugar, excluimos la columna que contiene los nombres que identifican las proteínas, ya que esta es una variable puramente identificativa. A continuación, examinamos los valores faltantes en el conjunto de datos y notamos que una proteína presenta valores faltantes en casi todas las columnas. Dado este hecho, optamos por eliminar la fila correspondiente (fila 230), ya que carece de información significativa.

Siguiendo con nuestro análisis exploratorio, procedemos a examinar las distribuciones de las variables en función de las dos clases previamente definidas (figura 14). En términos generales, no se observan variaciones significativas en el comportamiento de las variables en relación con las clases. Esto se verifica mediante el test de Wilcoxon, que evalúa si los valores medianos de dos grupos dependientes difieren de manera significativa [27]. Notablemente, el p-valor más pequeño obtenido corresponde al coeficiente de extinción, con un valor de 0,09625, que no alcanza significación estadística.



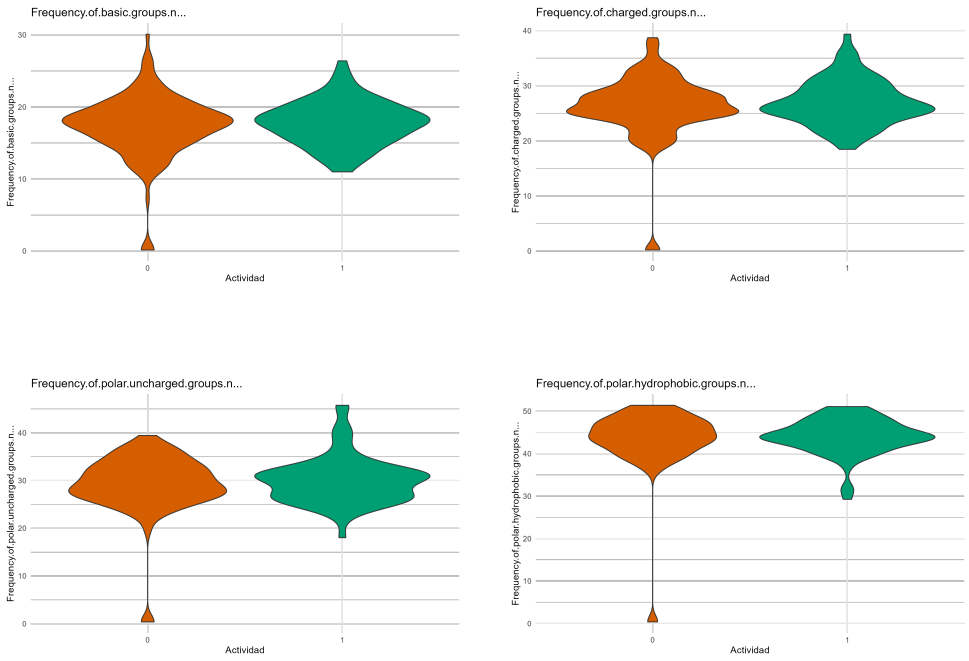


Figura 14: Gráficos de violin para observar la distribución de las variables

En la mayoría de los gráficos, se pueden identificar ciertos valores que resultan inusuales a simple vista, ya que se encuentran notablemente alejados de las zonas donde se concentran la mayoría de los ejemplos analizados. Estos valores anómalos, también conocidos como outliers, a menudo tienen el efecto de distorsionar las distribuciones de las variables y pueden impactar la eficacia de las predicciones. Estos valores atípicos pueden deberse a errores en la entrada de datos o a problemas en los procesos de obtención de los mismos. En tales casos, lo ideal es abordar y corregir los errores subyacentes. En otros casos, estos valores atípicos pueden surgir debido a la variabilidad aleatoria en los datos.

Antes de abordar la gestión de valores atípicos, llevamos a cabo una evaluación de la normalidad en cada una de las variables. Para este propósito, aplicamos la prueba de Lilliefors y observamos que todos los p-valores obtenidos son notablemente pequeños, todos ellos por debajo de 0,05. Esta evidencia sugiere que ninguna de las variables sigue una distribución normal, lo que implica que no podemos aplicar técnicas de imputación de valores atípicos basadas en la suposición de que los datos se distribuyen normalmente, como podría ser el método intercuartílico.

Implementamos la distancia de Mahalanobis para identificar valores atípicos. Una característica esencial de este método radica en su capacidad para considerar todas las variables de manera conjunta al calcular la distancia. determinamos una distancia específica para cada individuo y luego evaluamos si, en un contexto global, dicho individuo se aparta significativamente de la norma [28]. El cálculo de esta distancia toma en consideración tanto las medias como la matriz de covarianzas de todas las variables. Posteriormente, calculamos el estadístico chi cuadrado de las distancias de Mahalanobis con k-1 grados de libertad y para ver la significancia del estadístico, calculamos la densidad acumulada. Cuando el p-valor obtenido es menor a 0,001 clasificamos el individuo como

outlier. En este caso, nos encontramos con que ninguno es menor y por lo tanto, concluimos que no hay instancias atípicas.

Hemos completado la evaluación de la distribución individual de cada variable y avanzaremos al siguiente paso, que consiste en explorar las relaciones entre las variables. Para ello, llevaremos a cabo un análisis de correlación y correlación parcial. Nos enfocamos en identificar pares de variables que muestren una correlación superior a 0,75 en valor absoluto, además de una correlación parcial que supere el mismo valor. Es crucial considerar ambos valores, ya que las correlaciones tradicionales solo reflejan la relación entre el par de variables que se está analizando, sin tener en cuenta las demás en el conjunto.

Por otro lado, correlaciones parciales nos permiten evaluar la relación entre dos variables controlando el efecto de las demás [29]. El propósito fundamental es eliminar la influencia de las otras variables en la relación entre las dos bajo estudio. Esta aproximación nos brinda la capacidad de identificar cuáles variables adicionales pueden estar influyendo en la correlación entre las dos variables principales. Cuando observamos una correlación parcial significativa, indica que incluso después de remover la influencia de las demás variables, las dos variables en cuestión seguirán manteniendo una correlación sustancial. En contraste, una correlación alta pero una correlación parcial baja sugiere que la correlación es principalmente influenciada por otras variables en el conjunto.

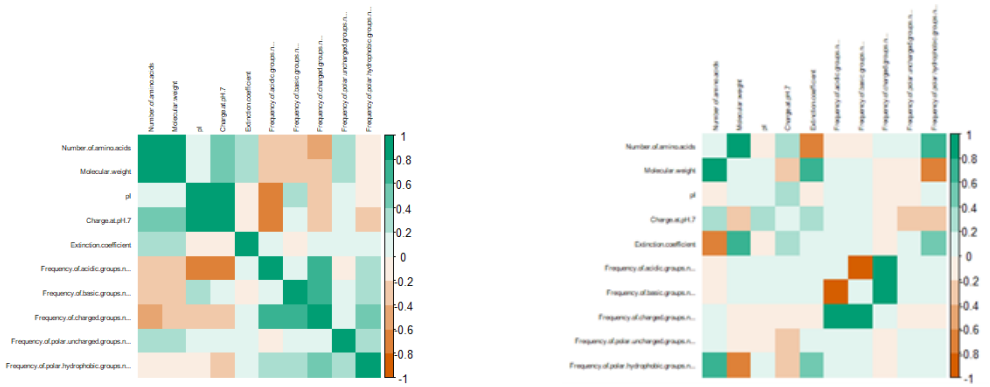


Figura 15: Matriz de correlaciones y correlaciones parciales del historial de proteínas

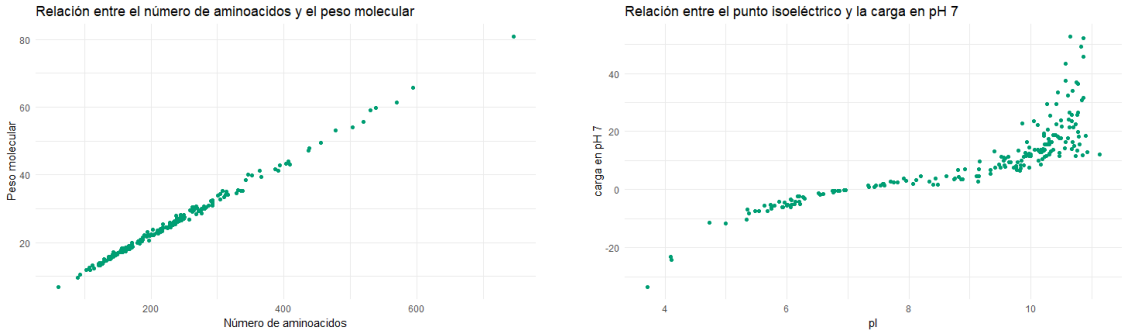
Variable 1	Variable 2	Correlación	Correlación Parcial
Número de aminoácidos	Peso molecular	0.9984	0.9986
Punto isoeléctrico (pI)	Carga en pH 7	0.8048	0.3053
Punto isoeléctrico (pI)	Frecuencia de grupos ácidos	-0.7942	0.0642
Carga en pH7	Frecuencia de grupos ácidos	-0.7384	0.07365
Frecuencia de grupos básicos	Frecuencia de grupos con carga	0.7872	0.9998
Frecuencia de grupos básicos	Frecuencia de grupos ácidos	0.04852	-0.9992
Frecuencia de grupos ácidos	Frecuencia de grupos con carga	0.6541	0.9996

Cuadro 2: Tabla con los valores numéricos de las correlaciones y correlaciones parciales

Continuando con el análisis, notamos una correlación significativamente alta entre el número de

aminoácidos y el peso molecular. Esta correlación es muy positiva, alcanzando un valor de 0,9984 (tabla 2). En esta relación, a medida que el número de aminoácidos aumenta, también se incrementa proporcionalmente el peso molecular de la proteína (figura 16a). Esta correlación es coherente, ya que la longitud de la proteína está directamente relacionada con su tamaño y peso molecular.

Otra correlación altamente positiva se observa entre el punto isoeléctrico (pI) y la carga a pH 7, con un valor de 0,8048 (tabla 2). Existe una asociación directa entre las variables, se debe a que el pI representa el pH en el cual la carga neta de la proteína está compensada [30]. Cuando el pH es superior al pI, la proteína presenta una carga neta negativa, mientras que cuando el pH es inferior al pI, la carga neta es positiva. En este caso, estamos comparando el pH 7 con el valor de pI. Por lo tanto, a la izquierda del valor de pI menor a 7, observamos valores de carga negativa, mientras que a la derecha se presentan valores de carga positiva (16b).



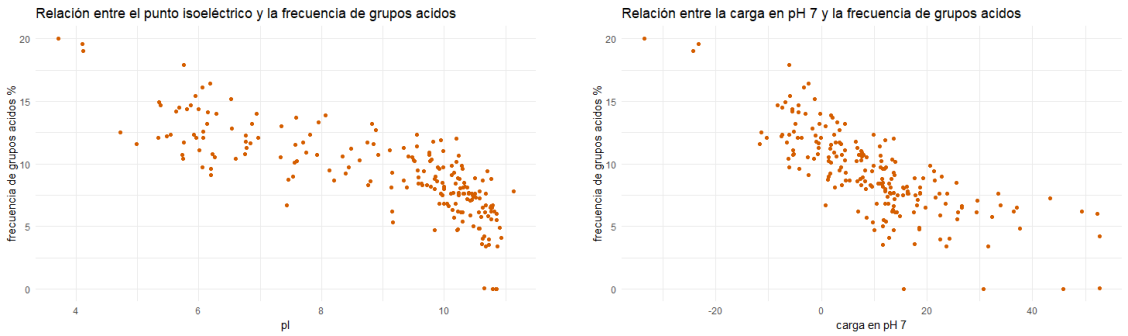
(a) Relación entre peso molecular y número de aminoácidos.

(b) Relación entre pI y carga en pH 7.

Figura 16: Gráficos de dispersión.

La última correlación altamente positiva que analizaremos es aquella entre la frecuencia de grupos básicos y grupos con carga, con valor 0,7872 (tabla 2). Esta relación química entre moléculas se debe a la presencia de grupos funcionales particulares. Los grupos básicos son componentes moleculares capaces de aceptar electrones, lo que les permite comportarse como bases en reacciones químicas. Por otro lado, los grupos con carga son porciones de la molécula que pueden llevar una carga eléctrica [31]. Su naturaleza iónica puede indicar la presencia de grupos básicos que han aceptado protones o grupos ácidos que han cedido protones. En resumen, los grupos básicos pueden contribuir a la formación de grupos con carga dentro de una molécula, dependiendo de las condiciones químicas y del entorno en el que se encuentren.

Además de las correlaciones positivas entre las variables del conjunto de datos, también se presentan correlaciones negativas, como en el caso de la relación entre el pI y la frecuencia de grupos ácidos (figura 17a). Esta asociación, con un valor de -0,7942 (tabla 2), es previsible debido a que bioquímicamente los aminoácidos con cadenas ácidas tienen un punto isoeléctrico menor a 7. Esto se debe a que la presencia de aminoácidos con cadenas laterales acidas presentan cargas negativas y necesitan de pH más bajo para ser neutralizadas [32].



(a) Relación entre pI y la frecuencia de grupos ácidos. (b) Relación carga en pH 7 y frecuencia de grupos ácidos.

Figura 17: Gráficos de dispersión.

La siguiente relación que analizaremos es la existente entre la carga en pH 7 y la frecuencia de grupos ácidos. Esta correlación se caracteriza por ser negativa, con un valor de -0,7384 (tabla 2). Al igual que en las relaciones previas, en este caso también entra en juego la bioquímica. Cuando la frecuencia de cadenas laterales ácidas es mayor, aumenta la probabilidad de que la proteína presente carga negativa en un entorno de pH 7 (figura 17b). Esto se debe a que los grupos ácidos liberan protones, y en un entorno neutro con un pH de 7 donde la concentración de protones es considerable, los grupos ácidos se ionizan, resultando en una mayor carga negativa en la proteína [32].

Los gráficos de dispersión por pares no resultan efectivos para identificar correlaciones parciales, dado que no permiten visualizar la influencia de las otras variables involucradas. En nuestro análisis, hemos identificado dos pares de variables que exhiben una alta correlación y, además, una correlación parcial que supera el umbral establecido (tabla 2). Como respuesta a esto, hemos optado por tomar la decisión de eliminar una variable de cada par. Específicamente, hemos eliminado la variable número de aminoácidos por un lado y la variable frecuencia de grupos con carga por el otro. Esta acción se realiza con el propósito de simplificar el modelo y evitar el sobre entrenamiento.

Optamos por no normalizar los datos debido a que, en ciertos contextos, las disparidades en las escalas pueden brindar información relevante al modelo. Normalizar todos los datos a la misma escala podría potencialmente eliminar información valiosa contenida en variables específicas.

En esta etapa de nuestro trabajo, nos centraremos en evaluar la capacidad de predecir la actividad de las proteínas. Dado que estamos abordando un problema de clasificación binaria, hemos optado por emplear el modelo de regresión logística como nuestro enfoque principal. Con el propósito de definir el modelo de manera óptima, aprovecharemos la función *step*, que nos permitirá examinar la relevancia de la inclusión de variables en el modelo.

La función *step* utiliza el criterio de Información de Akaike (AIC) para guiar la selección del modelo. Este enfoque busca lograr un equilibrio entre la calidad del ajuste y la simplicidad del modelo. En consecuencia, nuestro objetivo es minimizar el valor del AIC, ya que ello indica un ajuste más preciso de los datos y una menor complejidad del modelo.

Hemos empleado la función *step* y hemos encontrado que el mejor equilibrio entre ajuste y

simplicidad se logra mediante un modelo que incluye las siguientes variables: peso molecular, carga en pH 7, coeficiente de extinción, frecuencia de grupos ácidos, frecuencia de grupos básicos y frecuencia de grupos polares sin carga. Para ver si la reducción del número de variables es significativa hacemos ANOVA del modelo nulo, el modelo proporcionado por *step* y el modelo completo, que incluye todas las variables.

La prueba ANOVA proporciona un p-valor para cada modelo, permitiéndonos realizar una inferencia estadística para determinar si existen diferencias significativas entre ellos. En este contexto, el modelo generado mediante la función *step* muestra significación estadística, ya que su p-valor es de 0,009 en comparación con el modelo nulo. Por otro lado, el modelo completo no alcanza significación en comparación con el "modelo step".

Nuestro interés radica en que todas las variables incluidas en el modelo final sean estadísticamente significativas. Sin embargo, observamos que no todas las variables presentes en el modelo generado por el método *step* alcanzan esta significancia. En particular, el p-valor para la variable coeficiente de extinción es de 0,1419, mientras que la variable frecuencia de grupos básicos se encuentra en el umbral de la significancia con un p-valor de 0,0562. En respuesta a esta situación, hemos creado dos modelos adicionales. El primero excluye tanto el coeficiente de extinción como la frecuencia de grupos básicos, mientras que el segundo modelo únicamente excluye el coeficiente de extinción.

Procedemos a realizar nuevamente la prueba ANOVA para evaluar la significancia de la exclusión de las variables mencionadas anteriormente. Los resultados muestran que ambos modelos son estadísticamente significativos en comparación con el modelo nulo. Además, el segundo modelo propuesto presenta significancia con relación al primero, con un p-valor de 0,0274. En consecuencia, determinamos que las variables que serán empleadas para la predicción son las siguientes: peso molecular, carga en pH 7, frecuencia de grupos ácidos, frecuencia de grupos básicos y frecuencia de grupos polares sin carga. En otras palabras, hemos reducido la dimensionalidad del conjunto de datos original, manteniendo únicamente 5 de las 10 variables iniciales.

Para evaluar el rendimiento de los modelos, es fundamental dividir el conjunto de datos en dos partes: el conjunto de entrenamiento y el conjunto de prueba. En este caso, destinaremos el 70 % de las instancias al conjunto de entrenamiento y el 30 % restante al conjunto de prueba. El proceso consistirá en entrenar los modelos utilizando el conjunto de entrenamiento y luego evaluar su capacidad predictiva utilizando el conjunto de prueba. Esta división nos permitirá estimar cómo se desempeñan los modelos en datos no vistos y nos proporcionará una medida objetiva de su rendimiento.

Antes de proceder con la implementación de los modelos, es de gran importancia establecer un umbral que determine cuándo una predicción se asignará a la clase 0 o a la clase 1. En este sentido, emplearemos la curva ROC (Receiver Operating Characteristic) para obtener el umbral que optimiza la proporción entre verdaderos positivos y verdaderos negativos, es decir, el umbral que maximiza el área bajo la curva (AUC). Tras analizar la curva ROC (figura 18), hemos determinado que el umbral óptimo es 0,332 para el conjunto de datos.

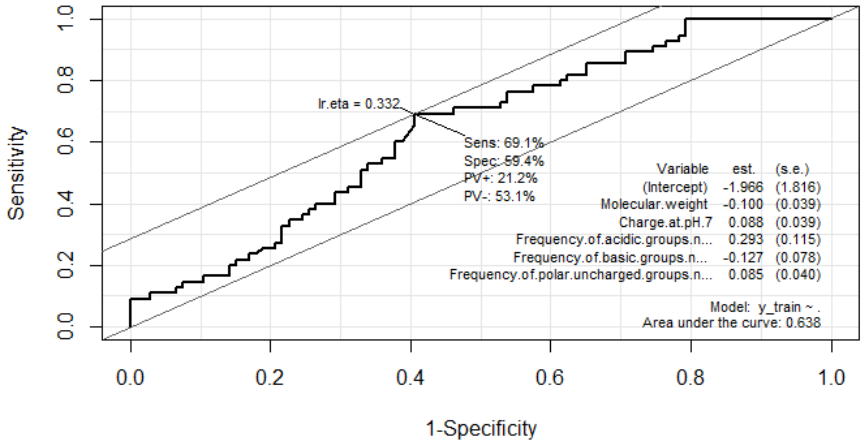


Figura 18: Gráficos de la curva ROC del modelo de regresión logística con los datos historicos

Una vez que hemos determinado las variables que serán incluidas en los modelos, establecido el umbral para las predicciones y dividido el conjunto de datos en entrenamiento y prueba, estamos listos para entrenar los modelos seleccionados. Además de la regresión logística mencionada previamente, nuestra selección de modelos incluye: el clasificador Random Forest, SVM radial, y Gradient Boosting. Cabe destacar que algunos de estos modelos están enfocados en técnicas de aprendizaje automático, mientras que otros se basan en fundamentos estadísticos.

Para evaluar el desempeño de los modelos mencionados, debido al desequilibrio en el número de muestras entre las clases, es más relevante y esclarecedor considerar medidas como la especificidad y la sensibilidad del modelo. Estas métricas nos permiten evaluar el rendimiento específico de cada clase, lo que resulta crucial para identificar el modelo que mejor se adapta a nuestros objetivos. En este contexto, nuestro objetivo es predecir de manera óptima la clase 1, que representa las proteínas más activas que avanzarán al departamento de proteínas.

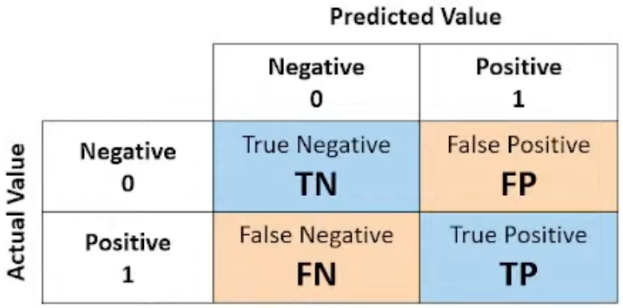


Figura 19: Matriz de confusión

Calcularemos estas métricas mediante la elaboración de matrices de confusión (figura 19), que nos permiten comparar las predicciones con las clases reales de cada instancia. Nos centraremos especialmente en dos métricas clave: la sensibilidad, que indica el porcentaje de instancias de la clase minoritaria que se predijeron correctamente, y la precisión, que representa cuántas de las instancias clasificadas como clase 1 realmente pertenecen a esa clase. Dado que estamos interesados en lograr un equilibrio entre falsos positivos y falsos negativos, utilizaremos la métrica F1-score. Esta métrica se calcula considerando tanto la sensibilidad como la precisión, lo que nos proporcionará una evaluación más completa del rendimiento del modelo. Las métricas mencionadas se calculan de la siguiente manera:

- Sensibilidad o Recall = $\frac{TP}{TP+FN}$
- Precisión o positive predicted values (ppv) = $\frac{TP}{TP+FP}$
- f1-score = $\frac{2*\text{Sensibilidad}*\text{Precisión}}{\text{Sensibilidad}+\text{Precisión}}$

6.1.2 Predicción de la actividad con decodificaciones

En esta segunda aproximación del problema, utilizaremos las decodificaciones previamente explicadas. Dado que en la primera aproximación ya hemos creado la columna binaria que determina si una proteína es activa o no, podemos avanzar. Sin embargo, durante la creación de los archivos FASTA, hemos identificado un aminoácido adicional en el individuo 178 que no puede ser analizado con el paquete de decodificación que estamos utilizando. Por lo tanto, es necesario excluir este individuo de la columna binaria.

Una vez que tenemos los conjuntos x (decodificaciones) e y (actividad), procederemos a realizar un análisis exploratorio similar al del problema anterior. En primer lugar, examinaremos cada conjunto para determinar si existen variables constantes que no aportan información. La única variable constante identificada es el dipéptido CH, cuyo valor es constante e igual a 0 para todos los individuos. Antes de continuar, eliminaremos esta variable del conjunto de dipéptidos.

Calculamos la correlación entre las variables de cada decodificación de forma individual. Observamos la presencia de variables altamente correlacionadas en cuatro de las cinco decodificaciones, en concreto en los dipéptidos, la composición, los aminoácidos anfifílicos y la autocorrelación de Moreau-Broto. A continuación, intentamos calcular las correlaciones parciales para determinar si estas correlaciones se deben a la influencia de otras variables, pero nos encontramos con que la matriz de datos es singular y, por lo tanto, no es invertible.

Decidimos aplicar otro método para evaluar la colinealidad entre las variables, que es la función alias. Esta función nos indica si una variable presenta colinealidad con respecto a las demás[33]. En caso de que sea así, consideramos que esa variable es redundante, ya que no proporciona información adicional al modelo, dado que su información puede ser deducida a partir de otras variables.

Aunque la decodificación de las proporciones de aminoácidos no muestra correlación, hemos observado que la proporción del aminoácido V exhibe una colinealidad negativa perfecta con respecto a todos los demás aminoácidos. Este fenómeno se debe a que las proporciones se calculan en función de la longitud de cada secuencia de proteínas, y, por lo tanto, la suma de todas las proporciones debe ser igual a 1. En consecuencia, si la proporción de V aumenta, las proporciones de algunos de los demás aminoácidos disminuirán y viceversa.

En el caso anterior, es sencillo analizar la colinealidad, pero en conjuntos más extensos como es el caso de los dipéptidos, el análisis se complica. Nos encontramos con que 234 variables de las 399 son colineales con algunas de las demás, algunas debido a la misma razón que con las proporciones y otras por razones que desconocemos. Sucede algo similar con la decodificación de la autocorrelación de Moreau-Broto, en este caso, son 15 variables las que muestran colinealidad de las 240. Todas las colinealidades en este conjunto se dan entre los retardos de la variable mutabilidad relativa. La visualización de la matriz que presenta los coeficientes de colinealidad resulta inviable debido a la cantidad de variables con las que las otras podrían mostrar colinealidad.

Por otro lado, tenemos el conjunto con la decodificación anfifílica, la cual no presenta ninguna colinealidad. Sin embargo, en el conjunto de composición, de las 21 variables iniciales, 9 de ellas son combinaciones lineales de otras. Por ejemplo, tanto la hidrofobicidad, el volumen normalizado de van der Waals, la polaridad, la carga y la estructura secundaria correspondientes al grupo 3, presentan colinealidad negativa perfecta respecto a los grupos 1 y 2 de los respectivos atributos.

Como colinealidades perfectas curiosas, observamos que la polarizabilidad del grupo 2 es positivamente colineal tanto con el volumen van der Waals del grupo 1 como del grupo 2, además de ser negativamente colineal con la polarizabilidad del grupo 1. También, la polarizabilidad del grupo 3 es negativamente colineal con los valores de van der Waals 1 y 2, y no presenta relación con la polarizabilidad del grupo 1. Las dos últimas colinealidades las presentan la accesibilidad solvente del grupo 2 y 3. La del grupo 2 corresponde a una colinealidad perfecta respecto a la hidrofobicidad del grupo 1. Mientras que la del grupo 3 presenta colinealidad negativa perfecta respecto a la hidrofobicidad del grupo 1 y la accesibilidad solvente del grupo 1.

Dado que estamos empleando un modelo de regresión logística para predecir la actividad, es fundamental abordar la colinealidad, ya que puede impactar negativamente en la precisión de nuestras predicciones. Por esta razón, hemos decidido eliminar las variables que exhiben colinealidad perfecta, dado que su información ya se encuentra contenida en otras variables del conjunto. Además, al eliminar estas variables, podremos evaluar si la colinealidad influye en las altas correlaciones que previamente hemos observado en nuestros datos. Cabe destacar que, incluso después de eliminar las variables redundantes, algunas correlaciones continúan siendo notables, lo que indica que otros factores pueden estar contribuyendo a estas asociaciones.

Para tomar decisiones sobre qué variables correlacionadas eliminar y cuáles conservar, utilizaremos el Factor de Inflación de la Varianza (VIF), que cuantifica la correlación y su intensidad entre las variables predictoras en un modelo de regresión, considerando el impacto del resto de las variables [34]. Sin embargo, es importante mencionar que no hemos podido aplicar este método a los conjuntos de dipéptidos y de autocorrelación de Moreau-Broto. A pesar de nuestros esfuerzos para eliminar la colinealidad, aún encontramos que el modelo presenta variables con una alta colinealidad. Esta situación podría deberse a la elevada dimensionalidad de estos conjuntos de datos. Por lo tanto, hemos optado por eliminar una variable de cada par que tenía una correlación superior a 0,8 en ambos conjuntos.

Adicionalmente, logramos aplicar con éxito el Factor de Inflación de la Varianza (VIF) en los conjuntos de composición y aminoácidos anfifílicos. Utilizando este índice, hemos realizado una selección de variables a eliminar en estos conjuntos debido a las altas correlaciones detectadas. Establecimos un umbral de 30 como indicador de una correlación potencialmente significativa entre

una variable dada y otras variables predictoras del modelo. En el caso de la composición, inicialmente identificamos 4 variables para su eliminación basándonos en la correlación, y luego, con la asistencia del índice VIF (figura 20), eliminamos tres de ellas: polaridad del grupo 2, volumen normalizado de van der Waals del grupo 1 e hidrofobicidad del grupo 2. La única variable que retenemos es la estructura secundaria del grupo 2, ya que no muestra una correlación severa con otras variables.

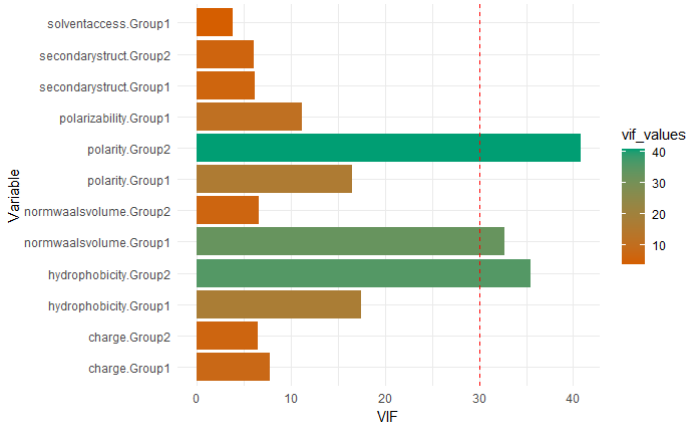


Figura 20: ïndices VIF de las variables correspondientes a la decodificación basada en la composición

En cuanto al conjunto relacionado con la anfifilia, hemos seguido un procedimiento similar. Inicialmente, consideramos la eliminación de 9 variables debido a sus correlaciones elevadas. Sin embargo, tras aplicar el índice VIF y establecer nuestro umbral, finalmente eliminamos solo 5 de estas variables. Este proceso demuestra la utilidad del índice VIF para filtrar de manera efectiva las variables que deben ser eliminadas. Es importante destacar que tanto las correlaciones parciales como los índices VIF tienen en cuenta la influencia del resto de las variables en el conjunto, lo que contribuye a una selección más precisa de las variables redundantes.

En esta nueva fase del proyecto, tras completar el análisis exploratorio de datos, nos preparamos para entrenar modelos de clasificación y realizar predicciones sobre la actividad de las proteínas utilizando las decodificaciones. En esta ocasión, hemos optado por no llevar a cabo una selección de variables durante el proceso de entrenamiento de los diversos modelos. Nuestra razón para ello radica en el interés de analizar en profundidad las variables que desempeñan un papel más relevante en la clasificación de las proteínas como activas o inactivas.

6.2 Estudio del conjunto de proteínas activo

Comprendemos que las proteínas más activas avanzan al departamento especializado en pruebas proteicas, donde se someten a una serie de evaluaciones adicionales. El propósito de estas evaluaciones es determinar cuáles de estas proteínas muestran un desempeño lo suficientemente destacado como para avanzar hacia ensayos clínicos in vivo. Los expertos de Telum nos han compartido las variables de mayor relevancia en las que centran su atención al momento de evaluar si una proteína arroja resultados satisfactorios o no.

6.2.1 Predicción con decodificaciones

Utilizaremos el conjunto de proteínas activas y las pautas proporcionadas por Telum para explorar si podemos predecir los atributos cruciales utilizando las decodificaciones que se explicaron previamente. Nuestro primer paso consiste en importar el archivo FASTA que alberga las secuencias provenientes del departamento de proteínas. En total, este estudio incluye a 95 individuos en análisis.

Asimismo, procedemos a la selección de las variables de interés en el conjunto mencionado. Estas variables incluyen la solubilidad, el MIC (Concentración Mínima Inhibitoria) y el valor muralítico. Además, incorporaremos la variable patógenos al conjunto de datos, dado que el comportamiento de una proteína puede variar en función del patógeno al que se enfrente. También consideraremos la variable efecto, la cual, a pesar de tener tres valores posibles, en este conjunto de datos específico solo presenta casos con efecto sinérgico o indiferente.

Etiquetamos las tres primeras variables según los criterios que los expertos de Telum consideran como indicativos de una proteína buena o mala. Realizamos pruebas utilizando restricciones tanto más rigurosas como más flexibles. Estas restricciones representan las condiciones óptimas para que una proteína sea considerada excelente en cuanto a esa característica, así como condiciones que, aunque no son ideales, tampoco son consideradas como malas.

Utilizamos histogramas para determinar los puntos de corte en la etiqueta de las variables (figura 21). En el caso de la solubilidad, una mayor solubilidad se considera mejor, ya que una solubilidad baja puede aumentar el coste de producción y afectar la actividad de la proteína. Telum sugiere que una solubilidad superior al 80 % es lo ideal. Sin embargo, consideramos que las proteínas con una solubilidad superior al 50 % también son adecuadas. Estos dos puntos de corte nos permiten abordar un problema con un desequilibrio en las clases y otro donde más o menos la mitad de los individuos pertenecen a cada clase.

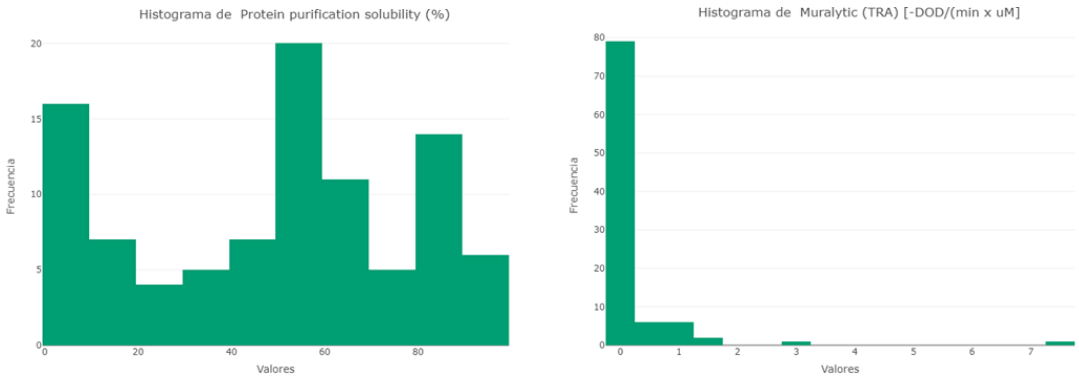


Figura 21: Histograma de los valores asociados a la solubilidad y el valor muralítico.

En lo que respecta al valor muralítico, este indicador evalúa la actividad de la proteína en células muertas al medir su capacidad degradadora del peptidoglicano. El peptidoglicano es un componente exclusivo de las bacterias que otorga rigidez y consistencia a la pared celular bacteriana. La degradación de este compuesto conduce a la lisis celular, lo que a su vez permite que los antibióticos ejerzan su acción [35]. En consecuencia, a medida que el valor del experimento del valor muralítico aumenta,

se refleja una mayor actividad degradadora de la proteína. En este contexto, hemos establecido dos puntos de corte: 1 y 0,24.

La tercera de las variables mencionadas es el valor MIC, que como hemos mencionado previamente, representa la concentración mínima de proteínas requerida para inhibir el crecimiento bacteriano. En este caso, cuanto menor sea la cantidad de proteína necesaria, mejor será la proteína en términos de su capacidad para inhibir el crecimiento bacteriano. Sin embargo, es importante tener en cuenta que algunos de los valores de MIC están precedidos por el carácter ">", lo cual indica que la proteína no presenta un valor de MIC o que este valor es considerablemente mayor que la concentración probada. Esto puede deberse a limitaciones en el laboratorio, donde no es posible producir una mayor cantidad de proteína para realizar pruebas a concentraciones más altas.

El carácter ">"que precede a algunos valores de MIC dificulta considerar la variable como numérica y, por lo tanto, calcular estadísticas con ella. Sin embargo, dado que un valor de MIC bajo indica una proteína más efectiva, hemos decidido asignar un valor numérico muy alto, como 150 (en comparación con los demás valores), a los individuos que presenten este carácter. Esta imputación nos permite tener valores numéricos para todos los casos y, así, asignar las etiquetas correspondientes. En este contexto, consideraremos un punto de corte ideal de 25 para la variable MIC, mientras que 50 será el punto de corte menos restrictivo (figura 22).

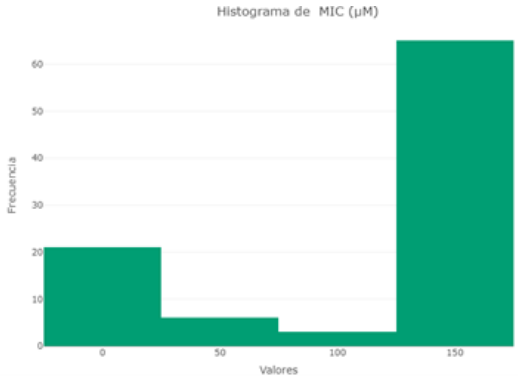


Figura 22: Historama de los valores MIC

Desde Telum, se hace especial hincapié en la importancia de la variable .efecto sinérgico o indiferente". Esto es especialmente relevante en el caso del MIC, ya que para Telum, tanto un valor alto del MIC como la ausencia de MIC, pero con un efecto sinérgico, se consideran resultados positivos. Esto se debe a que, en el tratamiento en humanos, siempre se opta por una combinación de la proteína y un antibiótico, por lo que la existencia de un efecto sinérgico entre la proteína y el antibiótico resulta fundamental.

Teniendo en cuenta lo mencionado, en lugar de utilizar la variable MIC etiquetada, vamos a crear dos nuevas variables a partir de ella, teniendo en consideración los valores del efecto. Las condiciones para estas nuevas columnas son las siguientes: si el valor de MIC es bueno, entonces la nueva variable tendrá un valor bueno; y si el valor de MIC es malo, pero hay un efecto sinérgico, entonces la nueva variable también tendrá un valor bueno.

Por último, antes de empezar con las decodificaciones de este problema, vamos a crear otras dos variables a partir de la combinación del resto de etiquetas. De esta manera, consideraremos todos los atributos importantes a la hora de decidir si una proteína es buena. Definiremos una proteína como buena cuando al menos dos de las tres variables se consideren buenas. La primera combinación estará formada por las etiquetas ideales, y la segunda por las etiquetas menos restrictivas.

El proceso que llevamos a cabo con las decodificaciones es muy similar al realizado en el problema anterior. Examinamos variables redundantes y encontramos que en el conjunto de dipéptidos, 9 variables son constantes y siempre toman el valor 0. De estas, una es la columna CH, al igual que en la decodificación del conjunto histórico. Esto nos lleva a pensar que las proteínas activas no contienen los 8 dipéptidos constantes que hemos identificado. Sin embargo, no podemos confirmarlo completamente, ya que el conjunto de datos es pequeño y este descubrimiento podría deberse a la falta de individuos activos en la muestra.

Continuamos con la evaluación de la colinealidad de las variables en cada conjunto de datos. En el conjunto de proporciones de aminoácidos, observamos una colinealidad similar con la variable V. Sin embargo, en el caso de los dipéptidos, identificamos la presencia de 325 colinealidades. Aunque planeamos eliminarlas del conjunto, la falta de visualización debido a la alta dimensión del conjunto nos impide determinar si son colinealidades perfectas o no. En cuanto al conjunto de decodificación por composición, encontramos las mismas colinealidades perfectas que en el problema anterior.

A diferencia del problema anterior, en el conjunto de decodificación anfifílica encontramos 7 colinealidades. Sin embargo, debido al gran número de variables, no podemos analizar los coeficientes de colinealidad en detalle. Pero podemos observar que se presentan en las características 28, 29 y 30 de hidrofobicidad e hidrofilicidad, así como en la característica 27 de hidrofilicidad. Por otro lado, en el conjunto de autocorrelación de Moreau-Broto, esta vez nos encontramos con 164 colinealidades, que, al igual que en el caso de los dipéptidos, no podemos visualizar debido a la alta dimensionalidad del conjunto.

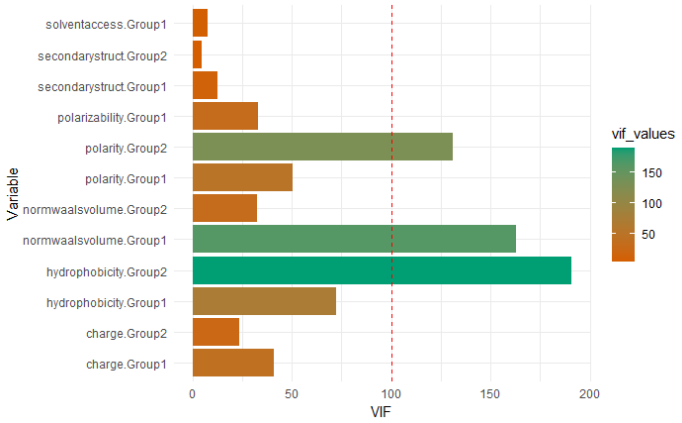


Figura 23: ïndices VIF de las variables correspondientes a la decodificación basada en la composición del conjunto activo.

El siguiente paso implica utilizar los valores VIF para filtrar las variables que deben ser eliminadas debido a la alta correlación. En los conjuntos de proporciones y dipéptidos, no encontramos

correlaciones significativamente altas. Sin embargo, en el conjunto de composición, aunque identificamos 5 correlaciones elevadas, decidimos eliminar solo 3 de ellas debido a que sus valores VIF son extremadamente altos (figura 23). Estas variables son la polaridad y la hidrofobicidad del grupo 2, así como el volumen normalizado de van der Waals del grupo 1. Estas variables coinciden con las eliminadas en el problema anterior.

En los conjuntos de decodificaciones anfifílicas y Moreau-Broto, no hemos podido aplicar los valores VIF para seleccionar las variables a eliminar. El problema de las variables aliadas persiste incluso después de eliminar algunas de ellas. En el caso de los pseudo-aminoácidos anfifílicos, eliminamos 8 variables, y en el caso de Moreau-Broto, eliminamos 2. Las dos últimas corresponden a la relación entre los lags 1 de las variables volumen de residuos y parámetro de polarizabilidad. Esta relación es previsible ya que la polarizabilidad depende del residuo, es decir, de la cadena específica del aminoácido. La otra correlación que eliminamos es entre los lags 8 del parámetro de polarizabilidad y el área superficial accesible de residuos, que también es una relación esperada.

Hemos mencionado que, además de las decodificaciones como variables, vamos a hacer uso de la información de los patógenos. Esto se debe a que podemos tener secuencias de proteínas repetidas debido a que se experimenta frente a más de un patógeno. Los patógenos los hemos codificado utilizando el método one-hot encoding, el cual asigna el valor 1 a la clase analizada y 0 al resto de las clases, para todas las clases que forman la variable patógenos. En total, contamos con 8 patógenos, lo que significa que tenemos 8 columnas binarias que nos indican el patógeno asociado a cada instancia.

Analizando estos nuevos datos, hemos observado que el patógeno S. epidermidis ATCC 35984 presenta una colinealidad perfectamente negativa con respecto al resto de las variables. Esto se debe a que es el último patógeno en la lista y, por lo tanto, ocupa las filas restantes. En nuestro análisis predictivo, no tendremos en cuenta esta variable. Por otro lado, hemos notado que hay una correlación muy baja entre el resto de los patógenos, lo que indica que son independientes entre sí.

Finalmente, procedemos con la predicción de las variables etiquetadas. Hemos seguido el mismo procedimiento de predicción que se explicó en el problema anterior. La única diferencia es que utilizamos dos conjuntos distintos para entrenar los modelos. Uno de ellos contiene solo las decodificaciones, mientras que el otro incorpora también la información de los patógenos.

6.2.2 Predicción del effecto de las proteínas e imipenen

Además del análisis de las decodificaciones, Telum destaca la importancia de ciertas variables, como el efecto de la proteína y el antibiótico. Con el objetivo de profundizar en esta variable, iniciamos un análisis exploratorio del conjunto de datos del departamento de proteínas.

Siguiendo el mismo enfoque que en los problemas anteriores, comenzamos por examinar la redundancia de las variables. Encontramos que ninguna variable es constante, pero identificamos una correlación entre dos variables: la longitud de la proteína y el peso molecular. Esta relación entre variables ya ha sido analizada previamente. Sin embargo, es importante destacar que la matriz es singular, lo que nos impide calcular las correlaciones parciales.

A pesar de no identificar relaciones significativas entre las variables, Telum ha señalado su interés en ciertas relaciones, como la posible influencia del peso molecular y el punto isoeléctrico en la

solubilidad. Sin embargo, al examinar la muestra de proteínas activas, no hemos encontrado un patrón claro que indique una relación definitiva entre estas variables (figura 24).

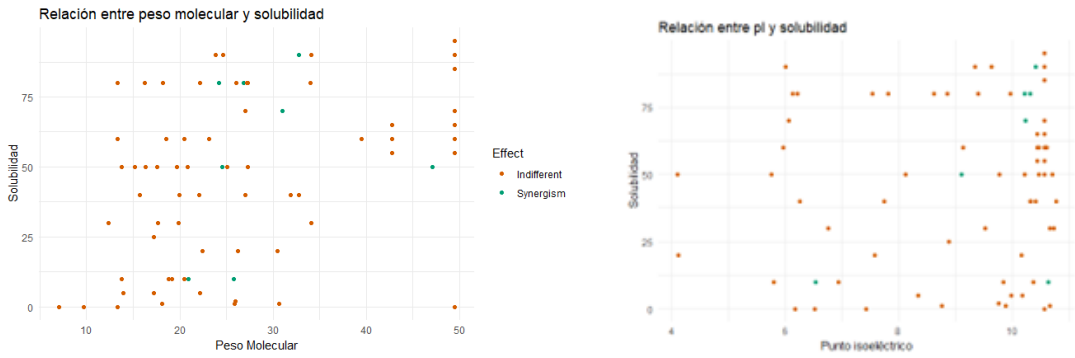


Figura 24: Gráficos de dispersión de las variables peso molecular y punto isoeléctrico frente solubilidad y la variable efecto.

También consideran relevante relacionar las secuencias de proteínas con el valor MIC, como se ha explorado en secciones anteriores, así como examinar la relación entre el valor MIC y el punto isoeléctrico (pI). Sin embargo, al igual que en el caso anterior, no hemos identificado un patrón definido que relacione estas dos variables (figura 25). Se ha sugerido que existe una tendencia en la que un mayor valor de pI se asocia con un menor valor de MIC. Excluyendo los valores de MIC de 150, que indican experimentos inválidos, podemos observar que en general se cumple esta suposición, aunque hay ciertos individuos que no siguen esta tendencia esperada.

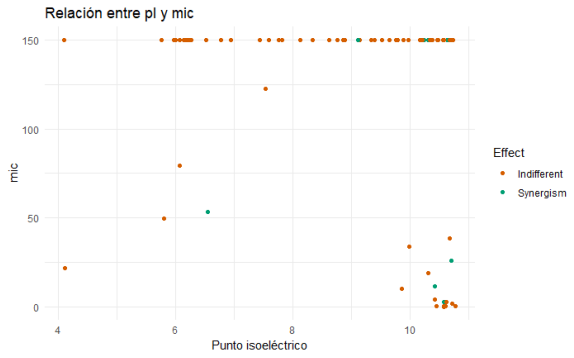


Figura 25: Gráfico de dispersión de las variables MIC frente al punto isoeléctrico y la variable efecto.

Adicionalmente, al observar los gráficos, notamos que no hay valores predefinidos para los individuos con efecto sinergético y efecto indiferente. Similar a los valores correspondientes a las proteínas indiferentes, los valores para las proteínas sinergéticas también parecen dispersarse ampliamente en todo el rango de valores (figuras 24 y 25).

Con la ayuda de la prueba de Wilcoxon, llegamos a la conclusión de que no existe evidencia significativa para afirmar que las clases de la variable efecto son iguales para las siguientes variables:

concentración de imipenem en sinergismo y concentración de proteína en sinergismo. Obtenemos p-valores de $2,2 \times 10^{-16}$ y $9,866 \times 10^{-6}$, respectivamente, ambos inferiores al valor de significancia de 0.05. Este resultado era esperado, ya que estas pruebas solo se realizan en las proteínas que muestran sinergismo con el antibiótico.

También obtenemos valores significativos para los patógenos A. baumannii, E. coli 70081223 y K. pneumoniae. En cuanto a los patógenos, hemos observado que algunos de ellos no presentan ningún caso con efecto sinergético. Estos son E. coli ATCC 25922, E. faecium, P. aeruginosa, S. aureus y S. epidermidis. Al profundizar un poco más, notamos que no todos los patógenos han sido evaluados la misma cantidad de veces.

Observamos que, en general, los patógenos que muestran significancia con respecto a las clases de la variable efecto son los más evaluados, especialmente A. baumannii, que representa 40 de las 95 instancias. Sin embargo, no podemos afirmar que el número de experimentos realizados sea la única razón de esta significancia, ya que los patógenos S. aureus y S. epidermidis han sido evaluados un número similar de veces que E. coli 70081223 y K. pneumoniae, pero no muestran la misma significancia en relación con la variable efecto.

En el proceso de selección de variables para nuestro modelo, aplicaremos el mismo enfoque que utilizamos en el primer problema. Emplearemos tanto la función step como la prueba ANOVA para determinar qué variables son significativas para el modelo. Antes de comenzar, eliminaremos las variables concentración de proteína e imipenem para sinergismo, ya que como mencionamos previamente, estas dos variables se calculan solo en casos donde la prueba checkerboard arroja un efecto sinergético. Por lo tanto, no las tendremos en cuenta ni en el proceso de selección ni en el entrenamiento de modelos. Además, excluiremos el patógeno S. epidermidis debido a su multicolinealidad con otras variables.

Después de aplicar la función step, se sugiere que el modelo esté formado solo por tres variables: valor muralítico y los patógenos E. coli 70081223 y K. pneumoniae. Para confirmar esta selección, aplicamos la prueba ANOVA como un paso adicional. Observamos que el modelo sugerido es significativo en comparación con el modelo nulo, con un p-valor de $1,634 \times 10^{-5}$. Además, el uso de todas las variables disponibles para formar el modelo no es significativo en comparación con el modelo sugerido por la función step.

En el modelo sugerido, encontramos que el valor muralítico no resulta significativo. Sin embargo, al aplicar la prueba ANOVA entre un modelo que excluye esta variable y el modelo sugerido por la función step, observamos que el modelo con el valor muralítico está al límite de ser significativo, con un p-valor de 0,0633. La eliminación la variable del modelo significaría trabajar únicamente con dos patógenos, los cuales toman valores binarios debido a la codificación one-hot. Por lo tanto, hemos decidido mantener la variable en el modelo debido a su importancia para los expertos y para evitar trabajar únicamente con columnas codificadas.

Finalizamos esta sección de metodología con el entrenamiento de modelos, que sigue el mismo procedimiento que en los problemas anteriores. Sin embargo, aprovechamos que este modelo está formado por un número reducido de variables. Esta simplicidad nos permite realizar un análisis detallado de la curva ROC y definir el umbral óptimo para el modelo de regresión logística, que en este caso es 0,049 (figura 26). También nos da la oportunidad de interpretar los coeficientes de cada variable, lo cual abordaremos en la siguiente sección de resultados.

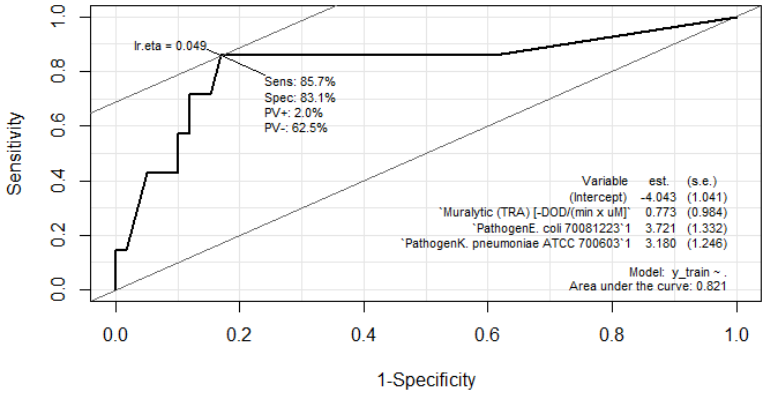


Figura 26: Curva ROC del modelo de regresión logística.

7 Resultados

En esta sección, analizamos los resultados obtenidos del entrenamiento de los modelos y su capacidad para predecir instancias que no formaron parte del proceso de entrenamiento. Como mencionamos previamente, entrenamos cuatro modelos diferentes y utilizamos el f1-score como métrica de evaluación, que proporciona un equilibrio entre la sensibilidad y la precisión.

Para todos los problemas abordados, dividimos la población en dos conjuntos: el 70 % de los individuos se destina al conjunto de entrenamiento, mientras que el 30 % restante se reserva para el conjunto de prueba. Es importante destacar que, en el conjunto histórico inicial, contábamos con 331 individuos, de los cuales solo podemos utilizar 229. En el conjunto activo, disponemos de 95 individuos, todos ellos válidos para el análisis.

7.1 Resultados actividad de las proteínas

Empezamos analizando los resultados de la actividad de las proteínas teniendo en cuenta los experimentos realizados en el departamento APEXp.

Modelo	f1-score	Sensibilidad	Precisión
Reg. Logística	0.5172	0.3846	0.7895
RFC	0.5263	0.5263	0.5263
SVM Radial	---	---	---
Gradeint Boosting	0.5306	0.4333	0.6842

Cuadro 3: Tabla con los rendimientos para cada modelo de la predicción de la actividad de las proteínas.

Los resultados obtenidos hasta el momento no son particularmente favorables, ya que un rendimiento del 50 % no cumple con las expectativas óptimas. Por lo general, se busca alcanzar al menos un 80 % de rendimiento en los modelos. En el caso de la predicción de la actividad de las proteínas en el conjunto histórico, el modelo que muestra el mejor rendimiento es el de Gradient Boosting, con un 53 % (Tabla 3). Sin embargo, es importante señalar que, aunque consideramos que este modelo es el más efectivo, los demás modelos también tienen un rendimiento bastante similar, con la excepción de SVM Radial, que se desempeña menos satisfactoriamente.

Cuando las diferencias entre los rendimientos de los modelos son tan pequeñas, hemos optado por seleccionar el modelo de Regresión Logística. Esto se debe a que la Regresión Logística es un modelo estadístico que ofrece una mayor interpretabilidad en comparación con los modelos de *Machine Learning*. La Regresión Logística nos permite realizar interpretaciones en términos de *odds* (razón de probabilidades), que indican la probabilidad relativa de que ocurra un evento en comparación con la probabilidad de que no ocurra [36]. En este contexto, estamos evaluando la probabilidad de que una proteína sea activa en comparación con la probabilidad de que no lo sea.

Al analizar los coeficientes obtenidos en la regresión logística, podemos interpretar cómo un cambio unitario en una variable predictora afecta a los *odds*. Observamos que a medida que aumenta la frecuencia de grupos ácidos y grupos polares sin carga, así como la carga en pH 7, la probabilidad de que la proteína sea activa también aumenta (figura 27). Por otro lado, cuando la frecuencia de grupos básicos y el peso molecular aumentan, la probabilidad de que la proteína sea activa disminuye. De estas cinco variables, la frecuencia de grupos ácidos y grupos básicos son las que tienen el mayor impacto en la probabilidad de actividad de la proteína.

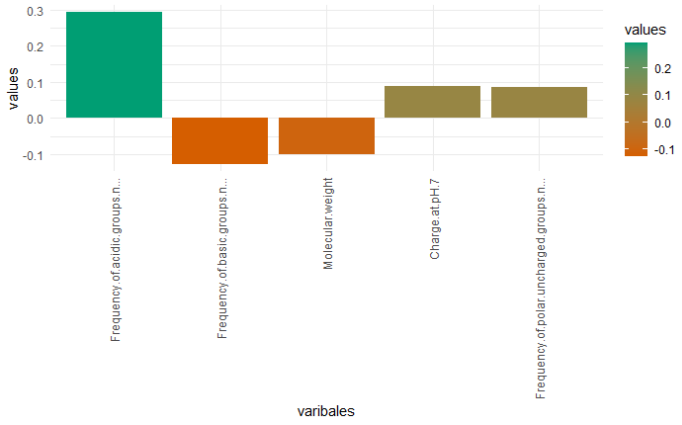


Figura 27: Coeficientes del modelo de regresión logística en la predicción de la actividad de las proteínas.

Además, podemos examinar la importancia de las variables, que indica cuán influyente es cada variable en términos generales en la capacidad predictiva. En otras palabras, muestra la contribución de cada variable al rendimiento del modelo. Un valor más alto en esta métrica implica una mayor influencia de la variable en la predicción. En este caso, hemos encontrado que las variables más influyentes para este modelo son el peso molecular y la frecuencia de grupos ácidos (figura 28). Es importante destacar que los valores de importancia asociados a cada variable son muy similares en magnitud, por lo que más o menos contribuyen de igual manera al modelo.

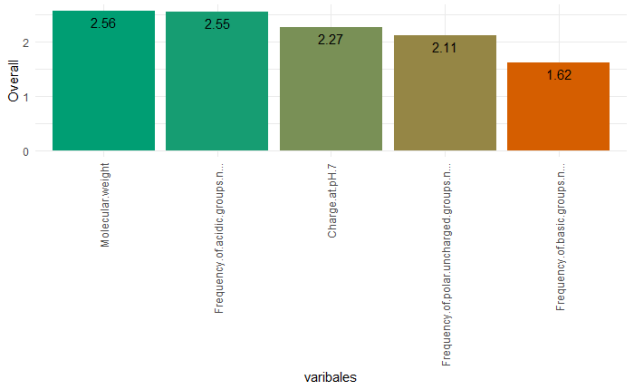


Figura 28: Valores de importancia en términos generales de las variables respecto al modelo de regresión logística en la predicción de la actividad de las proteínas.

Proseguimos con los resultados derivados de la predicción de la actividad de las proteínas, centrándonos exclusivamente en el uso de las decodificaciones de las secuencias proteicas. Al emplear la decodificación basada en la composición junto con el modelo Gradient Boosting, logramos alcanzar un rendimiento del 65 % (tabla 4). Aunque aún no es un resultado óptimo, representa una mejora con respecto al rendimiento obtenido al utilizar solamente los datos de los experimentos realizados en el departamento.

Modelo	Decodificación	f1-score	Sensibilidad	Precisión
Reg. Logística	CTD	0.5091	0.3684	0.8235
RFC	Dipeptidos	0.5625	0.6	0.5294
SVM Radial	Proporcion	0.5	0.6364	0.4118
Gradeint Boosting	CTD	0.65	0.5652	0.7647

Cuadro 4: Tabla con el mejor rendimiento y decodificación para cada modelo de la predicción de la actividad de las proteínas.

Cuando se construye un árbol en Gradient Boosting, se elige una variable para dividir el conjunto de datos en dos subconjuntos. El valor *Gain* mide la reducción de impureza resultante de la división, esencialmente muestra cuánto mejora la pureza de los nodos del árbol al dividirlo. Un valor mayor indica que una variable es más importante para el modelo y contribuye más a la capacidad del modelo para hacer predicciones precisas.

Al analizar estos valores, podemos identificar las variables que más contribuyen a una predicción precisa o a una separación más precisa de las clases de proteínas activas y no activas. En este caso, observamos que la hidrofobicidad del grupo 1, el volumen normalizado de van der Waals del grupo 2 y la estructura secundaria del grupo 2 son las variables más influyentes (figura 29).

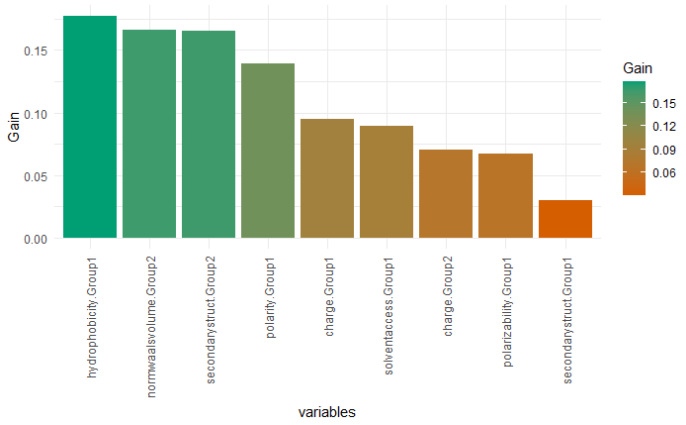


Figura 29: Ganancia del modelo al dividir los nodos por la variable correspondiente en la decodificación CTD.

Estos grupos corresponden a aminoácidos polares específicos en el caso de la variable hidrofobicidad, a un rango de valores de volumen de van der Waals entre 2,95 y 4, y finalmente, a proteínas con una estructura secundaria en forma de hebra o cadena (tabla 1). Estas variables desempeñan un papel significativo en la capacidad del modelo para realizar predicciones precisas sobre la actividad de las proteínas.

7.2 Resultados de proteínas activas

Una vez concluido el análisis del conjunto histórico, procedemos a examinar los resultados obtenidos para el conjunto de proteínas activas. Los resultados que encontramos para las distintas etiquetaciones de variables.

Variable Etiquetada	Mejor modelo	Decodificación	f1-score
Effect	Reg. Logística	ctd	0.6667
Solubility 50	SVM Radial	apaac	0.8571
Solubility 80	Reg. Logística	mb	0.3077
Muralytic 0.24	RFC	ctd	0.7273
Muralytic 1	RFC	ctd	0.4
Nuevo MIC 25	SVM Radial	ctd	0.7059
Nuevo MIC 50	SVM Radial	ctd	0.7368
Combinación Simple	RFC	mb	0.8333
Combinación Ideal	Reg. Logística	apaac	0.5

Cuadro 5: Tabla con el mejor rendimiento, decodificación y modelo para cada variable etiquetada del conjunto activo.

En la predicción del efecto de la proteína y el antibiótico, obtenemos un f1-score de casi el 67 % con la decodificación basada en la composición y un modelo de regresión logística (tabla 5). Por otro

lado, observamos que las variables más importantes resultan ser la carga en el grupo 1 (aminoácidos con carga positiva), el patógeno K. pneumoniae y el acceso solvente del grupo 1 (figura 30), que corresponde a tener los aminoácidos correspondientes a este grupo escondidos (buried, tabla 1).

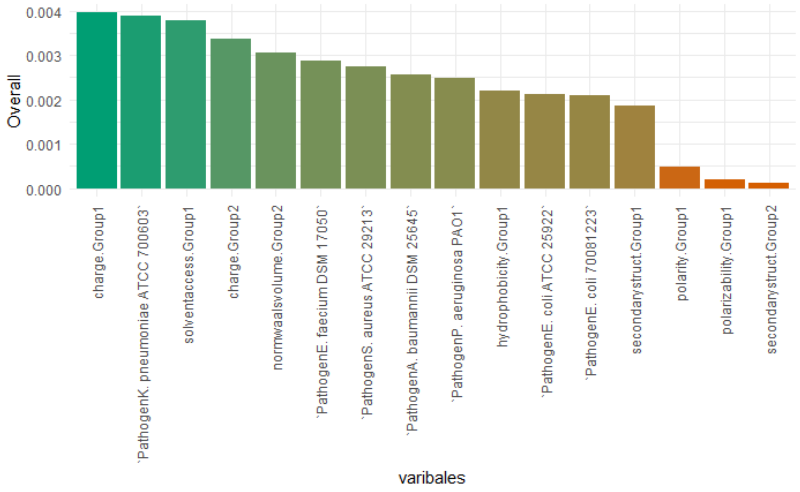


Figura 30: Importancia de las variables de la decodificación CTD para la predicción del efecto de la proteína e imipenem.

Los resultados relacionados con la solubilidad varían significativamente según el punto de corte considerado. Se observa que el rendimiento es considerablemente más bajo para la solubilidad ideal debido al desequilibrio en los datos, mientras que el rendimiento mejora sustancialmente cuando el 50 % de los individuos pertenecen a cada clase, alcanzando un 85 % (tabla 5).

Además, se destaca que las decodificaciones más efectivas para predecir la solubilidad son las anfifílicas y las de Moreau-Broto. Esta elección tiene una justificación sólida, ya que la decodificación anfifílica se basa en la hidrofobicidad e hidrofilicidad de los aminoácidos, propiedades que están estrechamente relacionadas con la solubilidad. Además, varios atributos de la decodificación de Moreau-Broto están relacionados con aspectos que influyen en la solubilidad, como la escala de hidrofobicidad normalizada, la polarizabilidad y, especialmente, la energía libre en solución acuosa, que indica la capacidad de una sustancia para disolverse en agua.

Continuando con la siguiente variable etiquetada, nos encontramos con el valor muralítico. Al igual que con la variable anterior, observamos que el rendimiento mejora en problemas balanceados en comparación con los desbalanceados (tabla 5). En este caso, la mejor decodificación y modelo predictivo coinciden en ambos puntos de corte. Al examinar la importancia de las variables, notamos que en ambos casos los patógenos K. pneumoniae, E. coli ATCC, E. faecium, S. aeruginosa y S. aureus no aportan información significativa al modelo. Por otro lado, observamos que variables como la polaridad y la carga del grupo 1 son muy importantes para el problema balanceado, mientras que la estructura secundaria del grupo 2 y la polaridad del grupo 1 son las más relevantes para el desbalanceado.

Para la variable "nuevo MIC,"se observa un buen rendimiento en las diferentes etiquetaciones. Este éxito se debe en gran parte a que es una combinación de la variable MIC y la variable efecto,

lo que resulta en un aumento en el número de individuos clasificados como buenas proteínas. En ambos puntos de corte, se logra un rendimiento superior al 70 % utilizando el modelo SVM radial junto con la decodificación basada en la composición (tabla 5).

En cuanto a la combinación de columnas, se observa que el rendimiento de las restricciones flexibles es significativamente mayor que el de las restricciones ideales. En la combinación simple, se logra un rendimiento del 83 % (tabla 5). Las variables más importantes en este caso están relacionadas con la hidrofobicidad, el índice de flexibilidad medio y la polarizabilidad (figura 31).

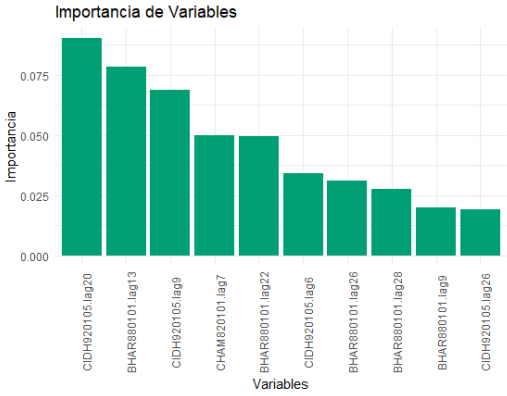


Figura 31: Las 10 variables más importantes de la decodificación Moreau-Broto para la predicción de la combinación simple.

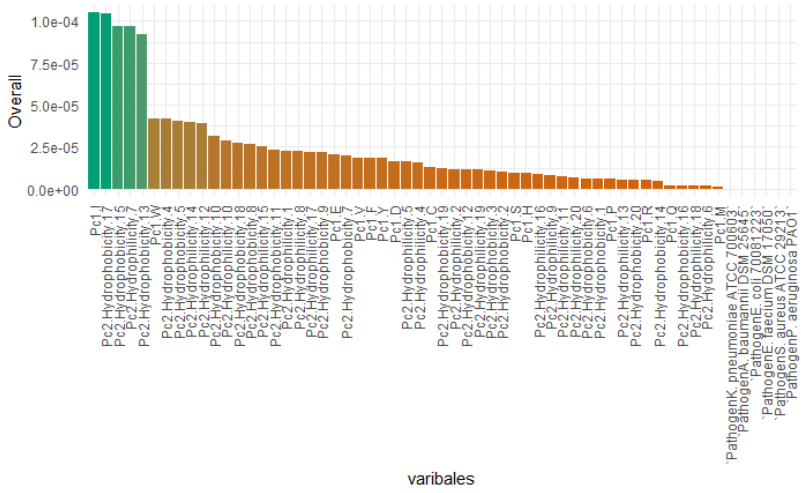


Figura 32: Importancia de las variable de la decodificación anfifílica para la predicción de la combinación ideal.

Por otro lado, en la combinación ideal, las variables más importantes son la proporción de aminoácidos I, la hidrofobicidad para las características 17 y 15, y la hidrofilicidad para las carac-

terísticas 7 y 13 (figura 32). Es importante destacar que en ambos modelos, las variables relacionadas con los patógenos no parecen tener un peso importante en las predicciones.

Hacemos una pequeña mención acerca de la contribución de las variables relacionadas con los patógenos a los modelos predictivos. Observamos que los resultados obtenidos sin la incorporación de las variables binarias correspondientes a los patógenos apenas varían. Es cierto que han resultado esclarecedores para el rendimiento de la variable efecto, mientras que han tenido un efecto perjudicial en el rendimiento de la combinación ideal. Sin embargo, para el resto de las variables, los valores son bastante similares (tablas 5 y 6).

Variable Etiquetada	Mejor modelo	Decodificación	f1-score
Effect	Reg. Logística	ctd	0.3333
Solubility 50	SVM Radial	ctd	0.8571
Solubility 80	RFC	ctd	0.3077
Muralytic 0.24	SVM Radial	appac	0.8333
Muralytic 1	Reg. Logística	mb	0.4
Nuevo MIC 25	Reg. Logística	ctd	0.7368
Nuevo MIC 50	Reg. Logística	ctd	0.7619
Combinación Simple	Reg. Logística	proporción	0.8
Combinación Ideal	RFC	ctd	0.6667

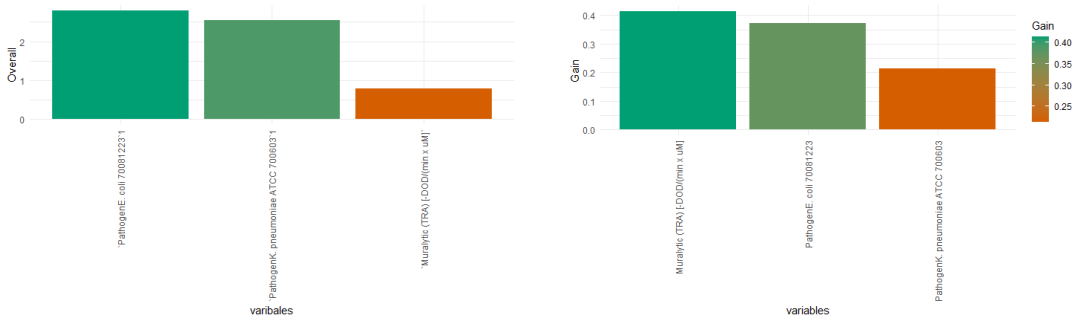
Cuadro 6: Tabla con el mejor rendimiento, decodificación y modelo para cada variable etiquetada del conjunto activo sin tener en cuenta los patógenos.

Por último, analizamos los resultados obtenidos para la predicción del efecto de la proteína e imipenem. Observamos que no se obtienen resultados para la predicción con el modelo SVM radial. Además, notamos que tanto la regresión logística como el modelo de gradient boosting logran el mismo rendimiento, un 66 % (tabla 7). Este valor es idéntico al que se obtiene al predecir la variable efecto utilizando la decodificación basada en composición y los patógenos. Sin embargo, es importante tener en cuenta que el modelo que proporciona estos resultados está formado por tan solo tres variables, en contraste con las 16 que componen el modelo analizado previamente.

Modelo	f1-score	Sensibilidad	Precisión
Reg. Logística	0.6667	0.6	0.75
RFC	0.4	1	0.25
SVM Radial	---	---	---
Gradeint Boosting	0.6667	0.6	0.75

Cuadro 7: Tabla con el mejor rendimiento para cada modelo en la predicción del efecto de la proteina e imipenem.

Como mencionamos previamente, en estos casos optamos por el modelo de regresión logística. Además de permitirnos ver la magnitud del aporte de cada variable al modelo, también nos brinda la capacidad de interpretar cómo la probabilidad de éxito aumenta o disminuye en función de los coeficientes obtenidos. Observamos que las variables más influyentes en el modelo son los dos patógenos incluidos, E. coli 70081223 y K. pneumoniae (figura 33a). Como era de esperar, la importancia del valor muralítico es bastante menor en comparación con las otras dos variables, ya que su significancia estaba al límite.



(a) Valores de importancia.

(b) Coeficientes.

Figura 33: Valores para interpretar la predicción de la variable efecto mediante la regresión lógistica

Por otro lado, podemos observar que todos los coeficientes de la regresión son positivos (figura 33b). Esto significa que cuando cualquiera de las tres variables aumenta en una unidad, la probabilidad de que el efecto sea sinérgico también aumenta. Es cierto que los coeficientes de los patógenos son mayores, lo que indica que cuando la proteína se prueba contra uno de estos dos patógenos, la probabilidad de sinergismo aumenta más que cuando se aumenta una unidad en el valor muralítico.

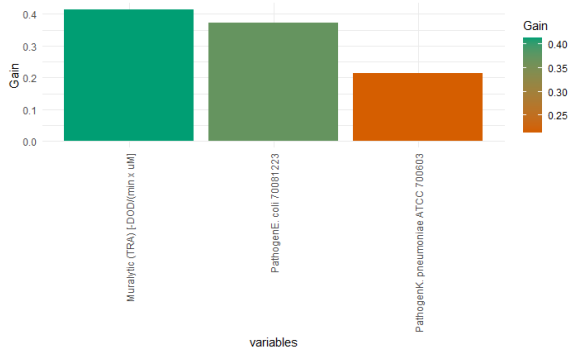


Figura 34: Ganancia del modelo gradeint boosting al dividir los nodos por la variable correspondiente en la predicción de la vriable efecto.

A pesar de obtener el mismo rendimiento en el modelo de Gradient Boosting, es interesante analizar qué variables han contribuido más a este modelo. Observamos que, a diferencia del modelo logístico, el valor muralítico es la variable que aporta la mayor ganancia al modelo de Gradient

Boosting (figura 34). Esto significa que es la variable que permite una separación más precisa entre las clases en este modelo.

8 Conclusiones

Luego de llevar a cabo un exhaustivo análisis y predicciones, podemos extraer diversas conclusiones. Durante el análisis exploratorio, hemos identificado ocho dipéptidos en el conjunto activo que constantemente presentan un valor de cero en todas sus instancias, lo que significa que no se encuentran presentes en ninguna de las 95 instancias. No obstante, en el conjunto histórico, estos dipéptidos muestran proporciones detectables. Este hallazgo nos lleva a considerar que la constante ausencia de estos dipéptidos podría ser un indicativo de la actividad de la proteína. Es importante tener en cuenta que esta hipótesis no puede ser afirmada con total certeza debido al limitado tamaño de la muestra.

Continuando con la cuestión del tamaño de la muestra, nos hemos encontrado con desafíos relacionados con las decodificaciones que generaban más variables que instancias disponibles en nuestra muestra. Esto normalmente conducía a ajustes de probabilidades extremadamente cercanas a uno y cero en los modelos predictivos, lo que indica un efecto de sobreajuste del modelo. Como resultado, observamos un rendimiento deficiente al intentar predecir el conjunto de prueba.

Por lo tanto, podemos concluir que disponer de una muestra más amplia podría contribuir significativamente a mejorar el rendimiento predictivo de la clase minoritaria en nuestros modelos. Una muestra más grande aumentaría el número de instancias con información adicional, lo que podría resultar fundamental para lograr clasificaciones más precisas y evitar el sobreajuste.

A pesar de los desafíos encontrados, hemos logrado obtener rendimientos que oscilan entre el 50 % y el 70 % para la mayoría de los problemas planteados. No hemos identificado un modelo que funcione de manera superior en todos los problemas a resolver. En lugar de eso, hemos observado que la eficacia de los modelos varía según la característica que se desea clasificar en cada ocasión.

Es importante destacar que cuando las diferencias de rendimiento eran pequeñas o se presentaban empates entre los modelos, hemos optado por utilizar de manera preferente la regresión logística. Como hemos mencionado anteriormente, además de proporcionar información sobre la contribución de cada variable al modelo predictivo, la regresión logística nos permite comprender cómo cada variable influye en el aumento o disminución de la probabilidad de éxito

Las decodificaciones han proporcionado resultados bastante sólidos, ya que con solo la secuencia de proteínas hemos logrado obtener rendimientos comparables a los obtenidos utilizando conjuntos de datos con valores experimentales. Por ejemplo, en el caso de predecir la actividad de las proteínas, utilizando la decodificación basada en la composición y el modelo Gradient Boosting, alcanzamos un rendimiento del 65 %, en contraste con el 53 % obtenido con los valores del departamento APEXp.

Este resultado merece una investigación más exhaustiva con una muestra más grande. La capacidad de predecir si una proteína es lo suficientemente activa como para avanzar al siguiente departamento podría ofrecer a Telum Therapeutics S.L. un ahorro significativo de tiempo y recursos. Dado que la investigación es costosa, la capacidad de obtener predicciones utilizando solo la

secuencia proteica podría potencialmente reducir la inversión de tiempo y dinero en los experimentos adicionales que la empresa realiza actualmente.

Las decodificaciones tienen aspectos negativos, como los que hemos mencionado. Por lo tanto, la selección de decodificaciones con una dimensionalidad razonable es fundamental para poder interpretar las variables de manera efectiva. Hemos observado que las decodificaciones basadas en la proporción de un aminoácido, la composición y la anfifilia de los pseudo-aminoácidos son las más interpretables.

Hemos alcanzado un rendimiento del 66 % en la predicción del efecto entre la proteína y el antibiótico utilizando el conjunto de datos del departamento de proteínas. Este logro se obtiene en dos modelos: Gradient Boosting y regresión logística. Es importante destacar que consideramos este rendimiento bastante satisfactorio, considerando que el modelo esta compuesto por solo tres variables.

En nuestro análisis, hemos observado que la presencia o ausencia de ciertos patógenos, como E. coli 7008122 y K. pneumoniae, desempeñan un papel fundamental en la predicción del efecto sinérgico o indiferente. Sin embargo, es importante destacar que al menos la mitad de los patógenos no mostraba ningún caso de sinergismo, lo que ha resultado en la depreciación de esos patógenos a la hora de seleccionar las variables más relevantes para la predicción.

Finalmente, es importante destacar que consideramos la decodificación basada en la composición como la mejor opción. Esto se debe a que, en la mayoría de los problemas en los que se aplicaron distintas decodificaciones, está en particular ha demostrado ofrecer los mejores resultados. Además, los rendimientos obtenidos son buenos, especialmente en relación con el número de variables con las que estábamos trabajando.

9 Lineas Futuras

El enfoque principal de este trabajo se centra en la decodificación de proteínas, habiendo seleccionado cinco métodos específicos para este propósito, aunque existen numerosos métodos adicionales disponibles. Sería beneficioso explorar diferentes técnicas de decodificación, teniendo en cuenta que la elección adecuada de la decodificación es esencial para evitar problemas de dimensionalidad.

En la misma línea, hemos implementado decodificaciones utilizando el paquete *protp* de R Studio, pero también existen otras bibliotecas con las que se pueden realizar decodificaciones de proteínas. Una de estas alternativas es *seqinr*, que podríamos utilizar para comparar el rendimiento entre las decodificaciones de distintos paquetes y explorar posibles hallazgos adicionales.

Lo mismo se aplica a los modelos predictivos utilizados, ya que otros modelos diseñados específicamente para abordar problemas desbalanceados podrían ofrecer un mejor rendimiento. La exploración de una gama más amplia de modelos podría enriquecer el estudio. Por ejemplo, la implementación de un modelo discriminante podría contribuir a identificar valores o atributos particulares que diferencien el comportamiento de cada patógeno, lo que podría conducir a una caracterización más precisa.

Por último, una dirección futura relevante sería trabajar con dominios de proteínas en lugar de secuencias de proteínas completas. Es conocido que la distribución de aminoácidos a lo largo de la secuencia y la estructura de la proteína son críticos para determinar su función. Por lo tanto, trabajar con regiones específicas de la proteína, definidas por expertos, podría resultar muy interesante. El proceso sería similar al llevado a cabo en este estudio, pero se enfocaría en decodificar las secuencias en dominios especificados.

Referencias

[1] Team, D. (2022, 1 agosto). Random Forest: Bosque aleatorio. Definición y funcionamiento. Formation Data Science — DataScientest.com. https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento

[2] Máquinas de vector soporte (SVM) con Python. (s.f.). https://cienciadedatos.net/documentos/py24-svm-python.html

[3] Remolino. (2023). ¿Qué es el gradient boosting? una técnica de ML cada vez más popular. remolinator.com. https://remolinator.com/que-es-el-gradient-boosting/

[4] Estadística, P. Y. (2023). Regresión logística. Probabilidad y Estadística. https://www.probabilidadyestadistica.net/regresion-logistica/#google_vignette

[5] Proteína. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Proteina

[6] Editorial Grudemi. (2022, 1 agosto). Proteínas - ¿Qué son?, función, importancia, ejemplos y más. Enciclopedia de Biología. https://enciclopediadebiologia.com/proteinas/

[7] Editorial Grudemi (2019). Aminoácidos. Recuperado de Enciclopedia de Biología. https://enciclopediadebiologia.com/aminoacidos/. Última actualización: agosto 2022.

[8] Aminoácido. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Aminoacido

[9] Gen. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Gen

[10] Ácido desoxirribonucleico (ADN). (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/%C3%81cido-desoxirribonucleico

[11] Par de bases. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Par-de-bases

[12] ARN mensajero (ARNM). (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/ARN-mensajero

[13] Traducción. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Traduccion

[14] Código genético. (s.f.). Genome.gov. https://www.genome.gov/es/genetics-glossary/Codigo-genetico

[15] Cellculture. (2022). The Bradford method to quantify protein (indirect cell quantitation). Cultivo de células. https://cellculture.altervista.org/the-bradford-method-to-quantify-cell-protein/?lang=es&doing_wp_cron=1693836348.4308450222015380859375

[16] Punto isoeléctrico. Diccionario médico. Clínica Universidad de Navarra. (s.f.). https://www.cun.es.https://www.cun.es/diccionario-medico/terminos/punto-isoelectrico

[17] ¿Qué es un coeficiente de extinción? (s.f.). https://es.411answers.com/a/que-es-un-coeficiente-de-extincion.html

[18] Nan Xiao . (s.f.). PROTR: R package for generating various numerical representation schemes of protein sequences. https://cran.r-project.org/web/packages/protr/vignettes/protr.html#CompositionTransitionDistribution

[19] Estructura de proteínas, fuerzas basicas. (s.f.). https://sitios.quimica.unam.mx/departamento/proteinas/estructura/EPpran2.html

[20] Admin. (2020). Diferencia entre polaridad y polarizabilidad. Pi Productora. https://piproductora.com/diferencia-entre-polaridad-y-polarizabilidad/

[21] Solvent accessibility. (s. f.). https://swift.cmbi.umcn.nl/teach/aainfo/access.shtml

[22] Barletta, P. (2020). Dinámica de cavidades proteicas: flexibilidad y cambios conformacionales. https://ridaa.unq.edu.ar/handle/20.500.11807/2183

[23] Matubayasi, N. (2017). Free-energy analysis of protein solvation with all-atom molecular dynamics simulation combined with a theory of solutions. Current Opinion in Structural Biology, 43, 45-54. https://doi.org/10.1016/j.sbi.2016.10.005

[24] Relative accessible surface area. (s.f.). https://es.abcdef.wiki/wiki/Relative_accessible_surface_area#:~:text=El%20%C3%A1rea%20de%20superficie%20relativamente%20accesible%20o%20la,del%20residuo.%20Se%20puede%20calcular%20mediante%20la%20f%C3%B3rmula%3A

[25] Cruzito. (2020, 16 septiembre). ¿Qué es el residuo de aminoácidos? — Estudyando. Estudyando. https://estudyando.com/que-es-el-residuo-de-aminoacidos/

[26] Guevara, S. R. (2022). ¿Qué es el número estérico? YuBrain. https://www.yubrain.com/ciencia/quimica/que-es-el-numero-esterico/

[27] 3.2 Wilcoxon. (s.f.). https://www.uv.es/webgid/Inferencial/32_wilcoxon.html

[28] Greyrat, R. (2022, 5 julio). ¿Cómo calcular la distancia de mahalanobis en R? – Barcelona Geeks. https://barcelonageeks.com/como-calcular-la-distancia-de-mahalanobis-en-r/

[29] Benites, L. (2022). Correlación parcial y semiparcial: definición y ejemplo. Statologos. https://statologos.com/correlacion-parcial/

[30] Peinado, J. P., & Meléndez-Valdés, F. T. (s.f). 7. Medida del pH: Disoluciones reguladoras. Precipitación isoeléctrica de la caseína. Uco.es. https://www.uco.es/organiza/departamentos/bioquimica-biol-mol/pdfs/07%20MEDIDA%20pH.pdf

[31] De Laboratorio, A. (2022). Clasificación de aminoácidos. Ciencia y Datos. https://cienciaydatos.org/quimica/bioquimica/clasificacion-de-aminoacidos/?expand_article=1

[32] Gonzalez Núñez, V. (s.f). Usal.es. https://diarium.usal.es/vgnunez/files/2012/11/3.-Amino%C3%A1cidos.-Propiedades-%C3%A1cido-base.pdf

[33] Alias Function - RDocumentation. (s.f.). https://rdocumentation.org/packages/stats/versions/3.6.2/topics/alias

[34] Statologos. (2021). Cómo calcular el factor de inflación de la varianza (VIF) en R. Statologos. https://statologos.com/factor-de-inflacion-de-la-varianza-r/#google_vignette

[35] Christian. (2023). El Peptidoglicano. Microbiología. https://microbiologia.net/bacterias/peptidoglicano/

[36] Ollé, J. (2021). Qué es y cómo interpretar una regresión logística. Conceptos Claros. https://conceptosclaros.com/que-es-regresion-logistica/