PEC 1 – Análisis Estadístico de Interacciones Proteína - Fármaco

Gabriel Manzano Reche

MU Bioinf. y Bioest.

Diseño de Fármacos y Biología Estructural

Nombre Tutor de TFM

Gonzalo Colmenarejo Sánchez

10/10/2024

Índice

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis Estadístico de interacciones proteína - fármaco
Nombre del autor:	Gabriel Manzano Reche
Nombre del consultor/a:	Gonzalo Colmenarejo Sánchez
Titulación o programa:	Máster Bioinformática y Bioestadística
Área del Trabajo Final:	Diseño de Fármacos y Biología Estructural
Idioma del trabajo:	Castellano
Palabras clave	Diseño de fármacos, quimioinformática, interacción proteína – ligando, grupos funcionales, familias de proteínas, clases químicas

1. Contexto y Justificación del Trabajo

Comprender cómo las proteínas codifican la especificidad de unión de ligandos tiene una importancia similar a descifrar el código genético. El reconocimiento molecular es un requisito fundamental de los sistemas biológicos. Las interacciones entre proteínas y moléculas pequeñas son fundamentales para la biología, ya que permiten a las células percibir su entorno y responder de forma adecuada. De las 10^{60} moléculas pequeñas que se pueden sintetizar, solo se ha explorado una ínfima parte de las posibles interacciones entre proteínas y ligandos (Colwell, 2018).

Para el reconocimiento proteína-ligando, la combinación de una variedad casi infinita de formas y patrones de grupos químicos en las interfaces de unión hace que el problema sea especialmente difícil de racionalizar y encontrar reglas predictivas (Yamanishi et al., 2011). Este trabajo se centrará en estudiar estas interacciones que estabilizan los complejos proteína–ligando, claves en el diseño racional de fármacos.

El diseño racional de fármacos no sólo depende de las interacciones directas entre las proteínas y los ligandos, sino que también depende de la estructura y reactividad de los grupos funcionales de las moléculas involucradas, además de las moléculas de agua, las cuales juegan un papel crucial en la estabilización de estos complejos.

En el pasado, distintos trabajos han abordado este tema, directamente o de forma relacionada. Por ejemplo, Szél et al., (2024), estudiaron el papel que desempeñan las moléculas de agua en el diseño de fármacos, ya que influyen tanto en las propiedades como en los comportamientos de los complejos. Por otro lado, Mukherjee et al., (2023) se centraron en la predicción de grupos funcionales, claves durante el diseño de fármacos.

Todos estos estudios, en los que se han investigado los patrones presentes en las interacciones proteína–ligando, se centran en aspectos peculiares de las mismas, o utilizan versiones antiguas de algunas herramientas (Mukherjee et al., 2023; Raschka et al., 2018; Szél et al., 2024; Yamanishi et al., 2011). Esto muestra la gran importancia de este trabajo, ya que aborda este tema de una forma generalizada y más actualizada, situándose además en un campo de investigación con impacto directo en el descubrimiento y desarrollo de fármacos.

Además, este proyecto no se limitará al análisis de estas interacciones, sino que también tiene como objetivo desarrollar un algoritmo de Machine Learning que sea capaz de predecir los sitios de unión de un ligando a una proteína (en el caso de que ésta sea un receptor de aquél), lo que proporcionará una herramienta poderosa para el diseño racional de fármacos.

2. Descripción General

El objetivo principal de este trabajo de fin de máster (TFM) es el *análisis estadístico de las interacciones entre proteínas y fármacos orales, para identificar patrones que faciliten el diseño racional de nuevos medicamentos*. Para ello, el proyecto se llevará a cabo en las siguientes fases.

Primero, se extraerán todas las interacciones entre proteínas y fármacos orales desde el Protein Data Bank (PDB) (Berman et al., 2000), utilizando para ello la API de PDB y el módulo Bio.PDB de Biopython que permitirá limpiar y preparar los datos. Posteriormente, se utilizará la herramienta BINANA 2 (Young et al., 2022) para analizar las interacciones entre las proteínas y ligandos, identificando interacciones moleculares clave. Seguidamente, se identificarán asociaciones estadísticamente significativas entre las clases químicas de los compuestos y las familias de proteínas, identificadas aquellas mediante el uso del programa y algoritmo ClassyFire (Djoumbou Feunang et al., 2016) y éstas mediante ChEMBL/UniProt (Mendez et al., 2019).

La extracción de grupos funcionales de las moléculas se realizará mediante el algoritmo Peter Ertl (Ertl, 2017), lo que permitirá buscar asociaciones entre las interacciones proteína-ligando y los grupos funcionales, así como con las familias de proteínas y clases químicas, utilizando para ello de nuevo la herramienta BINANA 2.

Finalmente, se desarrollará un modelo de Machine Learning para predecir sitios de unión en proteínas, lo que proporcionará una herramienta para el diseño racional de fármacos.

En resumen, este TFM combinará enfoques de quimioinformática, bioinformática, estadística y aprendizaje automático para abordar un problema crítico en la investigación farmacéutica, contribuyendo así a la mejora de la eficacia y especificidad de los medicamentos.

3. Objetivos

Objetivos Generales

1. Analizar todas las interacciones entre proteínas y fármacos orales a nivel estructural para identificar patrones que faciliten el diseño racional de nuevos medicamentos.

2. Identificar asociaciones significativas estadísticamente entre clases químicas, familias de proteínas y grupos funcionales.

3. Desarrollar modelos de Machine Learning que permitan predecir sitios de unión en proteínas.

Objetivos Específicos

- Extraer de PDB todas las interacciones establecidas entre proteínas y fármacos orales
    - Utilizar la API del PDB y el módulo Bio.PDB de Biopython para descargar ficheros de estructuras de complejos entre proteínas y fármacos orales.
    - Implementar un proceso sistemático que incluya la descarga y el filtrado de los datos relevantes en un formato adecuado para su análisis posterior.
- Analizar e identificar interacciones proteína – ligando mediante BINANA 2.
- Identificar asociaciones estadísticamente significativas entre clase química de los compuestos y familias de proteínas
    - Generar clases químicas mediante el uso ClassyFire y familias de proteínas con ChEMBL y UniProt.
    - Aplicar test estadísticos para evaluar la asociación entre las clases químicas de los compuestos, y las familias de proteínas, con los distintos tipos de interacciones moleculares.
    - Interpretar los resultados para identificar patrones o tendencias en la interacción entre diferentes grupos de fármacos y familias de proteínas.
- Extraer de las moléculas los grupos funcionales presentes utilizando para ello el algoritmo de Peter Ertl.
- Identificar asociaciones estadísticamente significativas entre interacciones y grupos funcionales
    - Realizar análisis estadísticos que relacionen las interacciones proteína-ligando obtenidas con BINANA.
    - Realizar test para ver asociaciones con clases químicas y clases de proteínas.
- Desarrollar modelos de Machine Learning que permitan predecir sitios de unión para nuevas moléculas
    - Utilizar algoritmos de aprendizaje automático, como redes neuronales profundas, para construir modelos predictivos que determinen la probabilidad de unión de nuevos fármacos.

4. Enfoque y método a seguir

Para este proyecto se ha planteado un enfoque en el que se plantea el uso de numerosas herramientas quimioinformáticas, bioinformáticas y algoritmos de Machine Learning debido a que este es el más adecuado para cumplir con los objetivos planteados, ya que se requiere analizar grandes volúmenes de datos, identificar patrones complejos y realizar predicciones.

Este enfoque utiliza herramientas para la extracción y análisis de datos (API de PDB y módulo Bio.PDB de BioPython), combinadas con herramientas que permiten la identificación de interacciones (BINANA 2) y herramientas de análisis químico (ClassyFire, ChEMBL, UniProt, Algoritmo Peter Erl). Finalmente, se integran modelos de Machine Learning para predecir sitios de unión entre proteínas y ligandos.

Este enfoque permite el análisis de grandes volúmenes de datos de manera eficiente, permitiendo realizar predicciones de sitios de unión en proteínas para nuevas moléculas, aportando gran valor significativo al diseño racional de fármacos. Además, la automatización permite una escalabilidad que sería imposible de alcanzar con otros métodos. Sin embargo, este enfoque presenta la desventaja de que requiere grandes recursos computacionales y elevada experiencia en programación para el desarrollo de los modelos de Machine Learning.

Otro enfoque que se podría haber utilizado consistiría en realizar estudios en el laboratorio para analizar las interacciones proteína – ligando. Sin embargo, este enfoque sería más costoso, requeriría más tiempo y no permitiría el análisis de grandes volúmenes de datos como los que se plantean como en el otro enfoque propuesto.

5. Planificación

5.1 Tareas

Objetivo General 1: Analizar interacciones entre proteínas y fármacos orales para identificar patrones que faciliten el diseño racional de nuevos fármacos.

- **Tarea 1:** Extracción de los datos de PDB mediante el uso de la API de PDB para obtener las proteínas que interacciones con los fármacos orales del estudio
- **Tarea 2:** Limpieza y Preparación de los datos
- **Tarea 3:** Análisis e identificación de interacciones proteína – ligando, utilizando BINANA 2

Objetivo General 2: Identificar asociaciones significativas estadísticamente entre clases químicas y familias de proteínas.

- **Tarea 4:** Uso de ClassyFire para clasificar las clases químicas de los compuestos y uso de ChEMBL y UniProt para identificar las familias de proteínas.

- **Tarea 5:** Realizar pruebas estadísticas para determinar asociaciones significativas entre clases químicas y familias de proteínas.

Objetivo General 3: Desarrollar modelos de Machine Learning que permitan predecir sitios de unión en proteínas.

- **Tarea 6:** Extracción de los grupos funcionales con el Algoritmo de Peter Ertl.

- **Tarea 7:** Análisis estadístico para identificar asociaciones relevantes entre interacciones y grupos funcionales.

- **Tarea 8:** Desarrollo del modelo de Machine Learning que permita predecir el sitio de unión.

PECs: representan las etapas claves del desarrollo y evaluación del TFM. Van desde la introducción del proyecto hasta su defensa.

- **PEC1:** Definición y Plan de Trabajo (Introducción, establecimiento de objetivos, cronograma, enfoque del proyecto, análisis de riesgo)

- **PEC2:** Informe que detalla los avances en el trabajo, cumplimiento de objetivos, actividades realizadas, desviaciones, junto con resultados parciales.

- **PEC3:** Informe que actualiza los cambios del proyecto con respecto a la PEC2. Se inicia la redacción de la memoria final.

- **PEC4:** Redacción de la memoria final, revisión del proyecto y preparación de la presentación virtual.

- **PEC5:** Defensa Pública del TFM

5.2 Calendario / Hitos

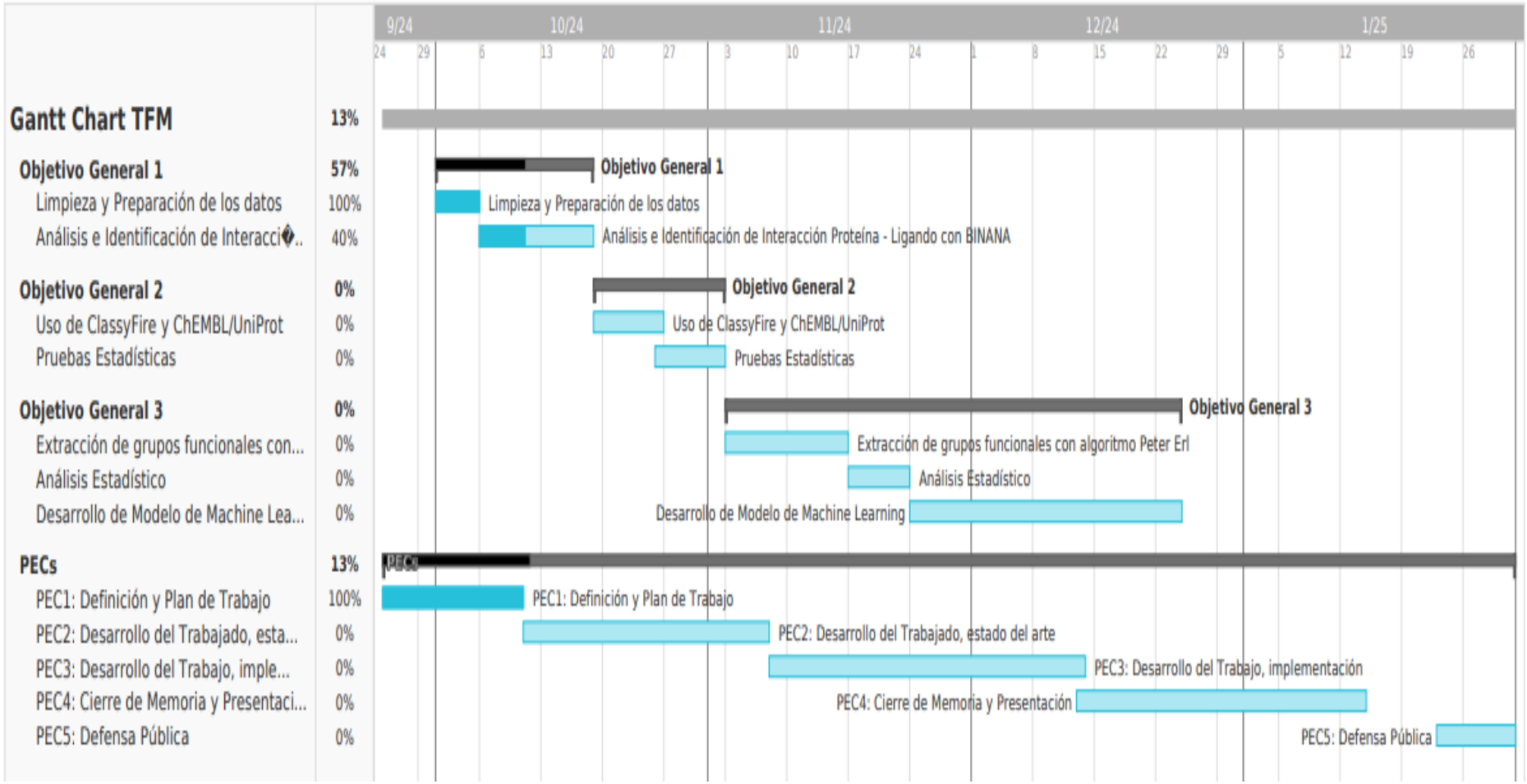

Ilustración 1: Diagrama de Gantt

5.3 Análisis de riesgos

- **Riesgo asociado al alcance del proyecto y tiempo:**
    - Es posible que, dada la magnitud del trabajo planteado, en el tiempo previsto no de tiempo al desarrollo de un modelo de Machine Learning completamente optimizado. En ese caso, se aproximará con una serie de reglas de predicción de unión en función de la clase química y de diana.
- **Riesgo de no disponer de los suficientes recursos computacionales:**
    - Algunos de los cálculos requeridos en este trabajo son altamente intensivos y es posible que no se puedan realizar de forma realista en el ordenador personal del estudiante. Para mitigar este problema se buscarán recursos de computación alternativos como los recursos de supercomputación que ofrece la Unión Europea o alternativamente se ejecutarán en los servidores del grupo del director del TFM.
- **Riesgo de errores en herramientas:**
    - Es posible que una o más de estas herramientas (e.g. BINANA 2, Bio.PDB), den errores al ejecutarse debido a la presencia de “bugs” de programación. En esos casos se buscarán herramientas alternativas, ya que las usadas no representan las únicas opciones disponibles.

6. Resultados esperados

Al finalizar este TFM se esperan obtener los siguientes resultados:

- **Plan de Trabajo:** documento escrito que contiene descripción del proyecto, objetivos, tareas, cronograma y análisis de riesgos. Este documento servirá como guía para el desarrollo correcto del TFM.
- **Memoria:** informe que documenta todo el proyecto. Esta memoria contendrá:
    - Introducción del proyecto
    - Materiales y metodología utilizada. Descripción del uso de las herramientas bioinformáticas utilizadas (BINANA, ClassyFire, API, ChEMBL, UniProt) y del modelo de Machine Learning.
    - Resultados obtenidos
    - Conclusiones y futuras líneas de investigación
- **Análisis Estadístico y Asociaciones Identificadas**: resultado de los análisis estadísticos en los que se mostrarán las asociaciones

significativas entre las clases químicas, familias de proteínas, grupos funcionales e interacciones proteína – ligando.

- **Scripts y Modelos de Machine Learning:** scripts de Python en los que se desarrollarán los procesos de extracción de los datos de PDB, análisis estadísticos, uso de las distintas herramientas bioinformáticas y los modelos de Machine Learning para la predicción de sitios de unión de proteínas. Todos estos scripts se subirán a un repositorio de GitHub.

- **Presentación virtual:** una presentación visual que resuma los aspectos más destacados del proyecto en alrededor de 20 minutos.

- **Artículo Científico:** se explorará la posibilidad de producir un artículo científico que exponga la metodología utilizada y los resultados obtenidos, con vistas de ser enviado a una revista científica.

7. Bibliografía

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/NAR/28.1.235

Colwell, L. J. (2018). Statistical and machine learning approaches to predicting protein–ligand interactions. *Current Opinion in Structural Biology*, *49*, 123–128. https://doi.org/10.1016/J.SBI.2018.01.006

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R., & Wishart, D. S. (2016). ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, *8*(1), 1–20. https://doi.org/10.1186/S13321-016-0174-Y/FIGURES/9

Ertl, P. (2017). An algorithm to identify functional groups in organic molecules. *Journal of Cheminformatics*, *9*(1), 1–7. https://doi.org/10.1186/S13321-017-0225-Z/TABLES/1

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940. https://doi.org/10.1093/NAR/GKY1075

Mukherjee, G., Braka, A., & Wu, S. (2023). Quantifying Functional-Group-like Structural Fragments in Molecules and Its Applications in Drug Design. *Journal of Chemical Information and Modeling*, *63*(7), 2073–2083. https://doi.org/10.1021/ACS.JCIM.3C00050

Raschka, S., Wolf, A. J., Bemister-Buffington, J., & Kuhn, L. A. (2018). Protein-ligand interfaces are polarized: discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes. *Journal of Computer-Aided Molecular Design*, *32*(4), 511–528. https://doi.org/10.1007/S10822-018-0105-2

Szél, V., Zsidó, B. Z., & Hetényi, C. (2024). Enthalpic Classification of Water Molecules in Target-Ligand Binding. *Journal of Chemical Information and Modeling*, *64*(16), 6583–6595. https://doi.org/10.1021/ACS.JCIM.4C00794/ASSET/IMAGES/LARGE/CI4C00794_0007.JPEG

Yamanishi, Y., Pauwels, E., Saigo, H., & Stoven, V. (2011). Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of Chemical Information and Modeling*, *51*(5), 1183–1194. https://doi.org/10.1021/CI100476Q/SUPPL_FILE/CI100476Q_SI_004.PDF

Young, J., Garikipati, N., & Durrant, J. D. (2022). BINANA 2: Characterizing Receptor/Ligand Interactions in Python and JavaScript. *Journal of Chemical Information and Modeling*, *62*(4), 753–760. https://doi.org/10.1021/ACS.JCIM.1C01461