PEC 3 – 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

12/12/2024

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. Descripción del avance del proyecto

El proyecto ha avanzado considerablemente desde la realización de la PEC2, logrando completar tareas esenciales como la extracción de los valores pChEMBL, búsqueda de Decoys para los ligandos y la preparación de representaciones funcionales y estructurales para ligandos y proteínas. Con estas actividades se pretende construir un dataset robusto que permita desarrollar un modelo para predictivo utilizando para ello redes neuronales.

En esta PEC, se han introducido nuevas tareas:

- **Tarea 10:** Generación de decoys para los ligandos.
- **Tarea 11:** Extracción del sitio de unión y codificación one-hot.
- **Tarea 12:** Codificación n-hot de los grupos funcionales de los ligandos.
- **Tarea 13:** Creación de un dataset final para llevar a cabo el modelo predictivo.

La introducción de estas nuevas tareas ha provocado un ligero retraso en el desarrollo del modelo predictivo. Además, debido a las limitaciones computacionales que se presentan, se ha solicitado acceso a EuropeHPC para llevar a cabo el modelo predictivo. Si no se concede este acceso, se plantea utilizar Google Colab.

2. Relación de las actividades realizadas

2.1 Grado de cumplimiento de los objetivos y resultados previstos en el plan de trabajo

- **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: Completada.
    - **Tarea 2:** Limpieza y preparación de los datos: Completada.
    - **Tarea 3:** Análisis e identificación de interacciones proteína-ligando con BINANA2: Completada.
- **Objetivo General 2:** Identificar asociaciones significativas estadísticamente entre clases químicas y familias de proteínas.
    - **Tarea 4: Uso de ClassyFire y ChEMBL/UniProt**: Completado.
    - **Tarea 5:** Pruebas estadísticas para determinar asociaciones significativas: Completado.
- **Objetivo General 3:** Desarrollar modelos de Machine Learning que permitan predecir sitios de unión en proteínas.
    - **Tarea 6:** Extracción de los grupos funcionales: Completada.
    - **Tarea 7:** Análisis estadístico para identificar asociaciones relevantes: Completada.
    - **Tarea 8:** Desarrollo del modelo de Machine Learning: En progreso.
    - **Tarea 9:** Búsqueda e incorporación de valores de pChEMBL: Completada.
    - **Tarea 10:** Generación de decoys para los ligandos.: Completada.
    - **Tarea 11:** Extracción del sitio de unión y codificación one-hot: Completada.
    - **Tarea 12:** Codificación n-hot de los grupos funcionales de los ligandos. Completada.

- **Tarea 13:** creación de un dataset final para llevar a cabo el modelo predictivo: En progreso.

2.2 Justificación de los cambios en caso necesario

Los cambios que se proponen en el cronograma representado en el punto 4, se justifican por la introducción de nuevas tareas, las cuáles son necesarias para crear un dataset completo y robusto que permita obtener un modelo predictivo del sitio de unión utilizando redes neuronales. Las nuevas tareas implementadas han permitido generar un dataset que incluye las siguientes características clave:

- **Datos de Afinidad (pChEMBL):** obtenidos de ChEMBL[1], Binding DB [2] y PDBBind [3]. Este valor indica la afinidad del ligando hacia su proteína objetivo, representado en escala logarítmica. Estos datos dan información sobre la interacción proteína – ligando.
- **Representación n-hot de los grupos funcionales:** representa la presencia de los 100 grupos funcionales más frecuentes en los ligandos en formato n-hot. Da información acerca de los grupos funcionales presentes en el ligando.
- **Representación one-hot del sitio de unión:** extracción y codificación one-hot de las secuencias aminoacídicas del sitio de unión. Da información estructural del sitio de unión entre proteína y ligando.
- **Generación de decoys de los ligandos:** se seleccionaron los decoys que presentan un menor coeficiente de Tanimoto con los ligandos para representar compuesto que no interactúan con la proteína objetivo, ampliando de esta forma la diversidad del dataset.

Con estas nuevas tareas implementadas, se ha mejorado significativamente la calidad del dataset final, y con ello, del modelo que se entre a partir de este. La inclusión de decoys permite representar interacciones negativas, permitiendo entrenar un modelo más robusto. Además, las codificaciones one-hot y n-hot dan al modelo información estructural y funcional, facilitando al modelo patrones complejos de interacción proteína-ligando. Estas tareas aseguran que el modelo predictivo entrenado a partir del dataset genere resultados precisos.

Además, se prevén limitaciones computacionales para el desarrollo del modelo predictivo debido a el entrenamiento con redes neuronales. Para abordar este desafío se plantea:

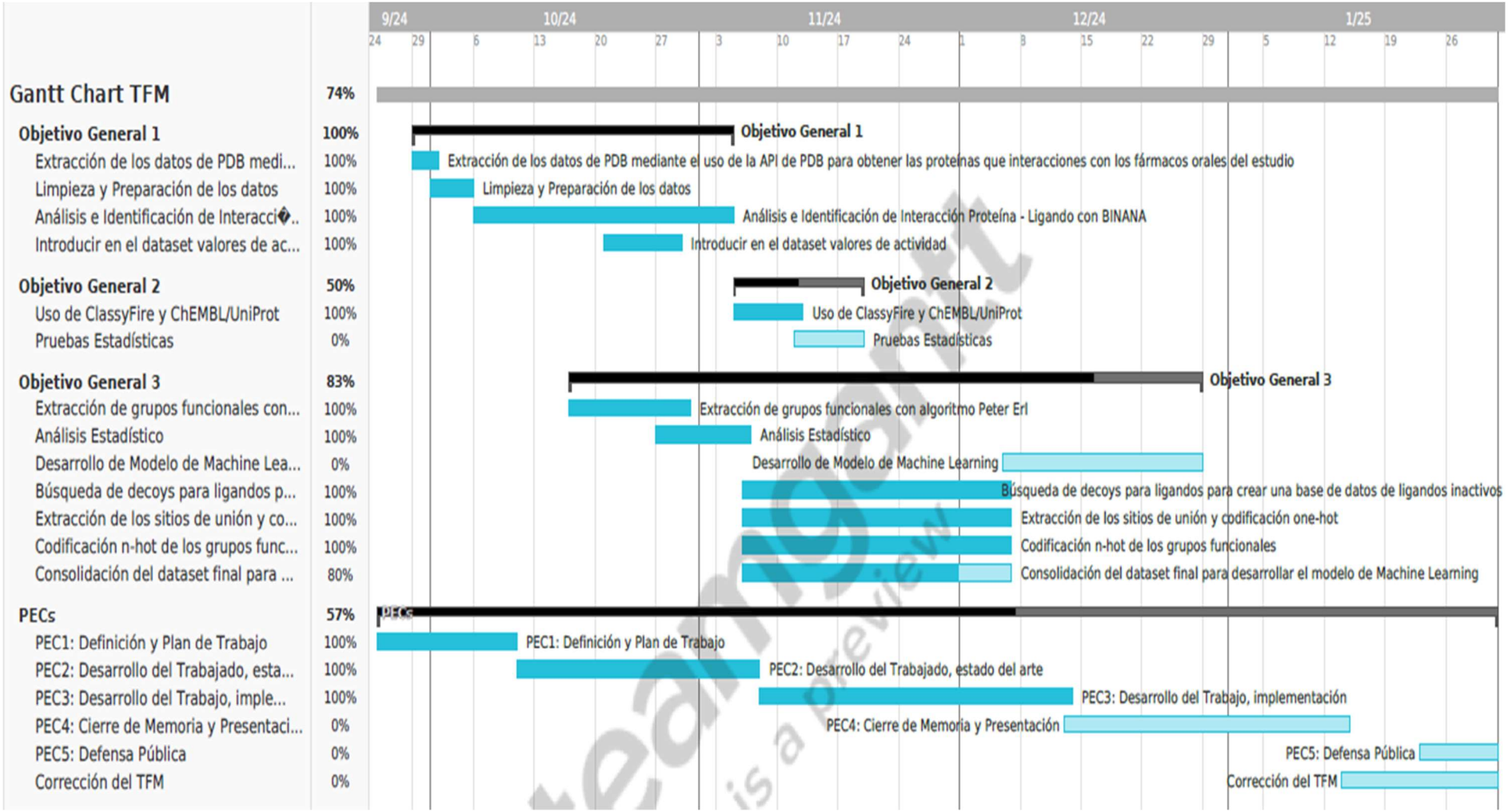
- **Solicitud EuropeHPC:** se ha solicitado el acceso a recursos computacionales avanzados en esta plataforma que permitirán entrenar el modelo con el volumen y la complejidad de los datos actuales.
- **Alternativa en Google Colab:** en caso de no obtener el acceso a EuropeHPC, se usará Google Colab como alternativa.

3. Relación de las desviaciones en la temporización y acciones de mitigación

- **Aparición de nuevas tareas:** tareas asociadas a la creación de un dataset sólido y robusto que permita obtener un modelo predictivo del sitio de unión utilizando redes neuronales.

    - Estas tareas se han completado en el periodo comprendido entre 3 de noviembre y el 5 de diciembre.
- **Retraso en la creación del modelo predictivo:** debido a la introducción de tareas adicionales y a la consolidación de los datos. Se usarán plataformas externas de computación avanzada (EuropeHPC o Google Colab) para crear este modelo.
    - Se prevé trabajar en esta tarea entre el 6 y el 29 de diciembre.

4. Diagrama de Gantt



5. Listado de los resultados parciales obtenidos hasta el momento

Los resultados parciales obtenidos hasta ahora incluyen desde la búsqueda de proteínas que formen complejos proteicos con los fármacos orales propuestos para este estudio, hasta la creación de un dataset final que se usará para el modelo de redes neurales que prediga la unión entre ligando y proteína. Estos resultados se encuentran organizados y estructurados en un repositorio de Github para facilitar la accesibilidad al proyecto.

En este repositorio se pueden encontrar los datos, scripts y notebooks que permiten llevar a cabo el procesamiento y análisis de las interacciones entre proteínas y fármacos. A continuación, se explicará en mayor detalle los elementos añadidos al repositorio desde la PEC2.

pChEMBL.py

Este script ha sido actualizado. Procesa un conjunto de datos de complejos proteína-ligando para obtener valores de actividad biológica (pChEMBL) a partir de las bases de datos ChEMBL[1], Binding DB [2] y PDBBind [3]. Además, para aquellos valores en los que no se han encontrado valores de afinidad en estas bases de datos, se generan valores aleatorios entre el segundo y tercer cuartil.

- **Pasos Principales:**
  - Consulta de valores pChEMBL en las bases de datos ChEMBL, BindingDB y PDBBind.
  - Asignación de valores aleatorios entre segundo y tercer cuartil para los ligandos sin afinidades conocidas.
- **Input:**
  - Filtered_Extract_Ligands_Uniprot.csv
  - BindingDB_All.tsv (base de datos extraída de Binding DB)
  - pdb_bind_data.xlsx (base de datos extraída de PDBBind)
- **Output:**
  - dataset_with_pchembl.csv: este dataset contiene los valores de pChEMBL.

binding_site.py

Este script extrae el sitio de unión de proteínas con sus ligandos y lo codifica en one-hot para usarlo en el modelo predictivo.

- **Pasos Principales:**
  - Identificación de los sitios de unión usando para ello la estructura de las proteína en formato .cif.
  - Codificación de las secuencias de aminoácidos del sitio de unión en formato one-hot.
- **Input:**
  - filtered_extract_ligands_uniprot.csv
  - Estructuras de las proteínas en formato .cif.
- **Output:**
  - binding_site_one_hot.csv: contiene la codificación one-hot de los sitios de unión.

ligand_decoy.py

Este script genera y procesa decoys de ligandos, seleccionando aquellas que presentan una menor similitud de Tanimoto para enriquecer el dataset con datos inactivos, es decir, de ligandos que no se unen a la proteína objetivo.

- **Pasos Principales:**
  - Creación de archivo txt con los SMILES de los ligandos.
  - Generación de decoys de los ligandos utilizando el software DUD-E [5].
  - Procesamiento de los archivo .picked donde se encuentran los decoy.
  - Cálculo de similitudes de Tanimoto.
  - Selección y filtrado de decoys menos similares
- **Input:**
  - Valid_smiles.txt
  - DUD-E_results (carpeta con los decoys generados)
- **Output:**
  - ligand_decoy.csv: contiene los decoys generados.

n_hot_fgs.py

Este script codifica en formato n-hot los 100 grupos funcionales más frecuentes entre los ligandos, los cuales fueron extraídos previamente en el script fgs. Estos datos enriquecen el dataset con características funcionales.

- **Pasos Principales:**
  - Identificación de los 100 grupos funcionales más frecuentes.
  - Codificación n-hot de los grupos funcionales para cada ligando.
- **Input:**
  - fgs_pdb.csv
- **Output:**
  - fgs_n_hot.csv: Codificación n-hot de los grupos funcionales

dataset_final.py

Este script une los datos estructurales, funcionales y de afinidad obtenidos a partir de los scripts anteriores en un único dataset, listo para entrenar el modelo predictivo.

- Pasos Principales:
  - Conversión de los InChI a SMILES para comparar con los decoys.
  - Fusión de los datos funcionales, estructurales y de afinidad.
  - Procesamiento y ajuste de decoys.
  - Eliminación de filas con valores nulos.
- Input
  - dataset_with_pchembl.csv: Datos de afinidad.
  - binding_site_one_hot.csv: Datos estructurales.
  - fgs_n_hot.csv: Codificaciones funcionales.
  - ligand_decoy.csv: Decoys procesados.
- Output
  - dataset_final: Dataset final consolidado y listo para el modelo.

6. Bibliografía

1. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*. 2019;47(D1):D930-D940. doi:10.1093/NAR/GKY1075
2. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016;44(D1):D1045-D1053. doi:10.1093/NAR/GKV1072
3. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*. 2015;31(3):405-412. doi:10.1093/BIOINFORMATICS/BTU626
4. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;55(14):6582-6594. doi:10.1021/JM300687E