PEC 2 – 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

07/11/2024

Índice

1. Descripción del avance del proyecto 3
2. Relación de las actividades realizadas 3
2.1 Grado de cumplimiento de los objetivos y resultados previstos en el plan de trabajo 3
2.2 Justificación de los cambios en caso necesario 4
3. Relación de las desviaciones en la temporización y acciones de mitigación 4
4. Diagrama de Gantt.....5
5. Listado de los resultados parciales obtenidos hasta el momento 6

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, con algunos objetivos y tareas finalizados antes de lo programado mientras que otras han experimentado ciertos retrasos. Entre las tareas completadas o cerca de ser completadas se encuentran:

- **Tarea 1:** Extracción de los datos de PDB mediante el uso de la API de PDB para obtener las proteínas que interaccionan con los fármacos orales del estudio
- **Tarea 2:** Limpieza y Preparación de los Datos
- **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.

Como se puede ver, se ha adelantado la extracción de los grupos funcionales usando el algoritmo de Peter Erl y el análisis estadístico para identificar asociaciones relevantes entre interacciones y grupos funcionales. Sin embargo, el análisis e identificación de interacciones proteína ligando de BINANA ha sufrido un retraso debido a falta de recursos computacionales.

Además, han aparecido nuevas tareas.

- **Tarea 9:** búsqueda e incorporación en el dataset final de valores de actividad, como el caso de pChEML. Estos valores serán utilizados para entrenar el modelo de Machine Learning.

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:** Limpieza y Preparación de los Datos
  - Las tareas destinadas a la limpieza y preparación de los datos han sido completadas con éxito, cumpliendo lo establecido en el plan inicial de trabajo.
  - La tarea de análisis e identificación de interacciones proteína-ligando con BINANA2 ha experimento un leve retraso debido a la falta de recursos computacionales, sin embargo, gracias a recursos externos ofrecidos por el tutor del TFM, esta tarea se espera completar pronto.
- **Objetivo General 2:** Uso de Classyfire y ChEMBL/UniProt
  - La ejecución de las tareas destinadas a completar este objetivo general 2 dependen de análisis e identificación de interacciones proteína-ligando con BINANA2. Por esta razón, el plazo de cumplimiento de este objetivo se extenderá en el tiempo.
  - A pesar de ello, se ha avanzado en la estructuración para esta etapa, lo que va a facilitar su inicio una vez se completen las fases previas.
- **Objetivo General 3:** Extracción de Grupos Funcionales y Análisis Estadístico
  - Como consecuencia del retraso de las otras tareas, se ha adelantado y completado algunas de las tareas de este objetivo.

- Se ha completado la extracción de grupos funcionales utilizando el algoritmo de Peter Erl y, además, se ha realizado un análisis estadístico de estos grupos funcionales, permitiendo una caracterización avanzada de los complejos del dataset.

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 las limitaciones de recursos computacionales, necesarias para el análisis e identificación de interacciones con BINANA2. Para solventar este problema, el código se ejecutará usando los servidores del grupo de investigación del tutor del TFM. Viendo el retraso que ha sufrido esta actividad, se ha adelantado la realización de estas tareas.

Además, se ha propuesto una nueva tarea:

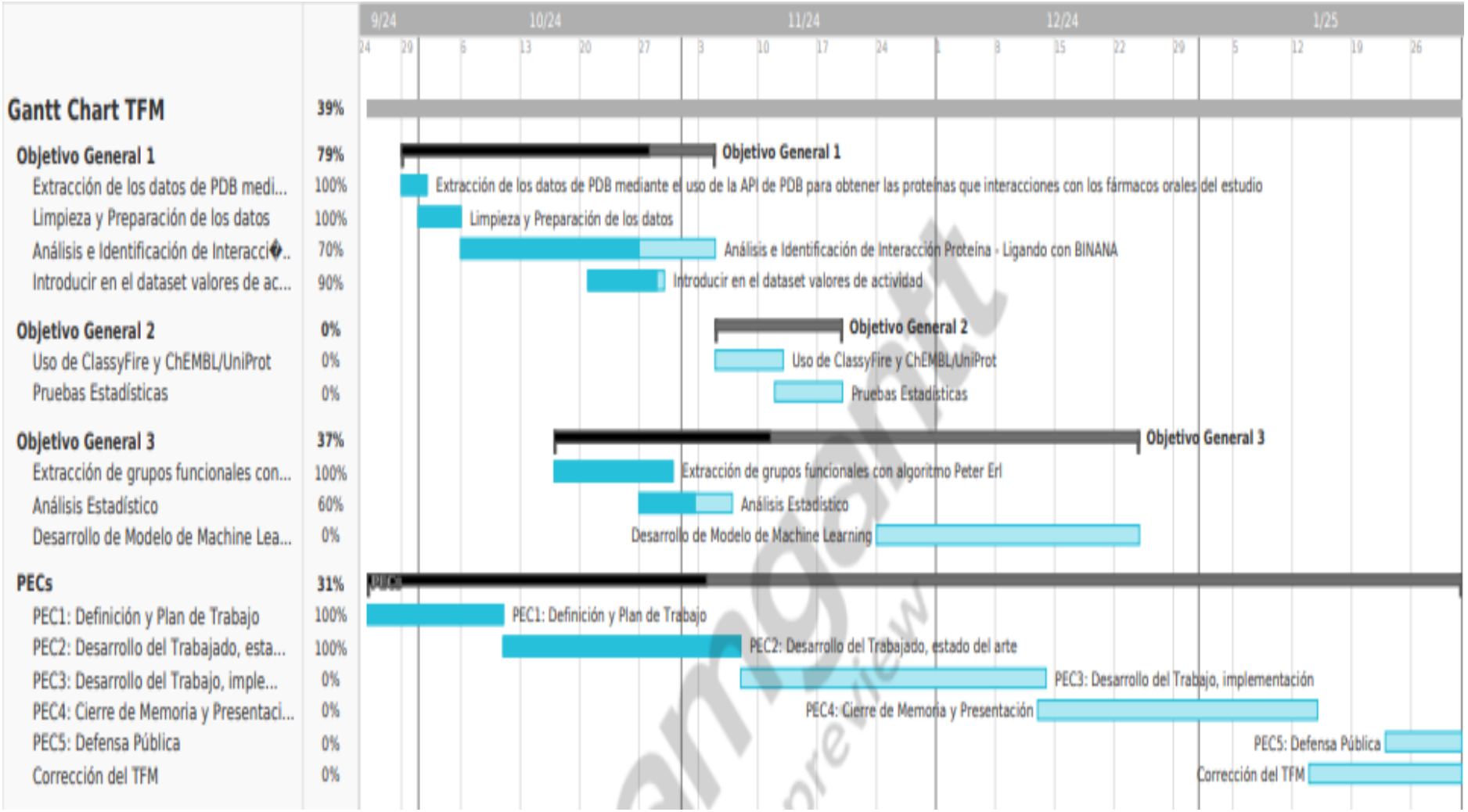
- **Tarea 9:** búsqueda e incorporación en el dataset final de valores de actividad, como el caso de pChEML.

Estos nuevos datos que se implementarán en el dataset final se usarán para entrenar el modelo de Machine Learning. El tiempo de incorporación de esta tarea se ha planificado para no afectar al cronograma. Se plantea realizarlo en paralelo con las tareas y ejecución de BINANA2 en los servidores externos.

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

- **Retraso en el análisis de interacciones proteína-ligando con BINANA2**
  - **Causa:** falta de recursos computacionales
  - **Solución:** el código se ejecutará en los servidores del grupo de investigación del tutor del TFM, lo que permitirá completar esta tarea en un menor tiempo. Frente al retraso de esta tarea se ha adelantado en la realización de otras.
- **Aparición de una nueva tarea:** búsqueda e incorporación en el dataset final de valores de actividad, como el caso de pChEML.
  - Ante esto, se reorganizará el cronograma para completar esta nueva tarea sin que afecte al resto del proyecto.
- **Extensión del tiempo propuesto para completar el Objetico General 2:**
  - Causa: se ha retrasado el cumplimiento de este objetivo debido al retraso de análisis de interacciones proteína-ligando con BINANA2
  - Frente al retraso de esta tarea se ha adelantado en la realización de otras.

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 el análisis de los grupos funcionales presentes en estos compuestos. 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. Un componente clave en este repositorio es el archivo main.py, que actúa como punto de entrada para la ejecución automática de todos los scripts del proceso. Gracias a esta automatización del proceso, se optimiza el tiempo de procesamiento y además se minimizan el riesgo de errores. A continuación, se explicará en mayor detalle cada uno de los elementos que se encuentran en el repositorio.

interacciones_proteina.py

Este script analiza un conjunto de fármacos orales obtenidos en DrugBank para identificar sus complejos proteína – ligando en la base de datos de PDB. A continuación, se muestran los pasos principales:

- **Pasos Principales:**
  - Carga y limpieza de datos.
  - Estandarización y filtrado de estructuras moleculares.
  - Búsqueda de identificadores PDB para los fármacos.
  - Obtención de complejos proteína-ligando asociados a los fármacos.
  - Exportación de los resultados.
- **Input:**
  - drugs.csv: contiene los fármacos orales obtenidos de la base de datos Drug Bank.
- **Output:**
  - **result_drugbank.csv:** contiene los fármacos filtrados y estandarizados con su drugbank ID correspondiente.
  - **result_pdb.csv:** contiene las proteínas presenten en PDB que forman complejos proteicos con los fármacos orales.
  - **complejos_PDB.txt:** contiene el PDB ID de los complejos proteicos encontrados.

descarga_estructuras_pdb.py

Este script automatiza la descarga de las estructuras de los complejos proteína – fármaco en formato CIF desde la API de RCSB PDB usando para ello la lista de identificadores obtenida en el script anterior. A continuación, se muestran los pasos principales:

- **Pasos Principales:**
  - Lectura de los Identificadores de PDB
  - Descarga Archivos CIF
  - Registro de los resultados obtenidos para llevar un control del proceso
- **Input:**

  - Complejos_PDB.txt
- **Output:**
  - Estructura PDB en formato CIF
  - **Salida_terminal.txt:** contiene un resumen de la descarga
  - **Errores.txt:** lista de estructuras que no se descargaron correctamente

extract_ligands_uniprot.py

Este script asocia cada uno de los complejos proteicos con su correspondiente identificador de UniProt. Esto permite ver las relaciones entre PDB_ID, UniProt_ID y los ligandos en las estructuras de proteínas. Durante el proceso, se filtran los ligandos no deseados usando la lista de exclusión de ligandos (blacklist.txt).

- **Pasos Principales:**
  - Obtención del UniProt_ID.
  - Extracción de los ligandos.
  - Creación de archivo que contiene la asociación entre Complejos PDB – Ligandos – UniProt_ID.
  - Filtrado de los ligandos usando la blacklist
- **Input:**
  - Complejos_pdb.txt
  - Archivos en formato CIF
  - blacklist.txt: contiene los fármacos que se van a excluir.
- **Output:**
  - filtered_extract_ligands_uniprot.csv: de archivo que contiene la asociación entre Complejos PDB – Ligandos – UniProt_ID.

fgs.py

Este script aplica el Algoritmo de Peter Erl para extraer los grupos funcionales de los complejos proteicos y representarlos como pseudo-SMILES.

- **Pasos Principales:**
  - Conversión de Inchi a objeto Mol de RDKit
  - Obtención de grupos funcionales aplicando Algoritmo Peter Erl
  - Creación y Almacenamiento de archivo de salida
- **Input:**
  - filtered_extract_ligands_uniprot.csv
  - result_drugbank.csv
- **Output:**
  - fgs_pdb.csv
  - fgs_drugbank.csv

analisis_fgs.py

Este script se utiliza para realizar un análisis estadístico de los grupos funcionales. A continuación, se presentan las funciones presentes en este archivo:

- **calcular_estadisticas_fgs:**
  - Calcula estadísticas generales de FGs, incluyendo el número total de fragmentos, fragmentos únicos, promedio de fragmentos por molécula, y otras métricas.
- **plot_top_20_fgs:**

  - Genera un gráfico de barras de los 20 fragmentos funcionales más comunes.
- **calcular_estadisticas_aromaticos_heteroatomos:**
  - Calcula estadísticas sobre fragmentos aromáticos y con heteroátomos, incluyendo el número binario y único de cada tipo de FG.
- **calcular_estadisticas_heteroatomos:**
  - Realiza un análisis similar al anterior pero enfocado en fragmentos que contienen oxígeno, nitrógeno, azufre, fósforo y halógenos.
- **visualizar_coocurrencia_top_20_fgs:**
  - Crea y visualiza la matriz de co-ocurrencia de los 20 fragmentos funcionales más comunes.

main.py

Este script automatiza el proceso para analizar estadísticamente las interacciones proteína – ligando. Primero, proceso y filtra los datos de fármacos orales obtenidos de DrugBank, luego descarga las estructuras de los complejos proteicos en formato CIF, y los asocia con su identificador UniProt. Finalmente, se extraen los grupos funcionales de cada complejo utilizando una aproximación del algoritmo de Peter Erl.

analisis_fgs.ipynb

Este notebook utiliza las funciones definidas en el script análisis_fgs.py para analizar los grupos funcionales obtenidos para los datos de PDB y para los datos de DrugBank. A continuación, se muestran los resultados obtenidos:

Tabla 1: La tabla muestra el análisis de fragmentos funcionales (FGs) para los conjuntos de datos PDB y DrugBank. La columna "N" representa el número de moléculas únicas analizadas. "Total FGs" es el número total de fragmentos encontrados, mientras que "FGs /mol" muestra el promedio de fragmentos por molécula. "FGs Binarios" representa la cantidad de fragmentos considerando solo su presencia o ausencia en cada molécula, y "FGs binarios/mol" indica el promedio de estos fragmentos binarios por molécula. "FGs Únicos" muestra el número total de fragmentos únicos, y "FGs Únicos/mol" refleja el promedio de fragmentos únicos por molécula.

	N	Total FGs	FGs /mol	FGs Binarios	FGs binarios/mol	FGs Únicos	FGs Únicos/mol
PDB	483	2749	5,69	1881	3,89	128	0,27
DrugBank	1306	6841	5,24	4691	3,59	234	0,18

Tabla 2: detalla el análisis de fragmentos funcionales aromáticos (Ar FGs) y fragmentos que contienen heteroátomos (Het FGs) en los conjuntos de datos PDB y DrugBank. "Ar FGs (bin) / mol" indica el promedio de fragmentos aromáticos binarios por molécula, mientras que "Ar FGs (un)" muestra el total de fragmentos aromáticos únicos en cada conjunto. "Ar FGs (un) / mol" proporciona el promedio de estos fragmentos únicos por molécula. De manera similar, "Het FGs (bin) / mol" representa el promedio de fragmentos con heteroátomos por molécula, "Het FGs (un)" es el total de

fragmentos únicos con heteroátomos, y "Het FGs (un) / mol" indica el promedio de estos fragmentos únicos por molécula en cada conjunto de datos.

	Ar FGs (bin) / mol	Ar FGs (un)	Ar FGs (un) / mol	Het FGs (bin) / mol	Het FGs(un)	Het FGs (un) / mol
PDB	1,743271	483	1	5,515528	1808	3,743271
DrugBank	1,37	1049	0,80	5,08	4527	3,47

Tabla 3: Esta tabla presenta el análisis detallado de la presencia de grupos funcionales específicos (FGs) que contienen heteroátomos en los datos de PDB y DrugBank. Las columnas indican el promedio de cada tipo de grupo funcional por molécula y la fracción de fragmentos únicos en cada conjunto de datos. Por ejemplo, "O FGs (bin) / mol" muestra el promedio de fragmentos que contienen oxígeno por molécula, mientras que "Frac O FGs (un)" refleja la fracción de fragmentos de oxígeno únicos. Esto se aplica también a grupos funcionales que contienen nitrógeno (N), azufre (S), fósforo (P) y halógenos (X), con sus respectivos valores de fracción única (un) y promedio por molécula (bin).

	O FGs (bin) / mol	Frac O FGs (un)	N FGs (bin) / mol	Frac N FGs (un)	S FGs (bin) / mol	Frac S FGs (un)	P FGs (bin) / mol	Frac P FGs (un)	X FGs (bin) / mol	Frac X FGs (un)
PDB	2,67	2,03	2,61	1,77	0,21	0,21	0,02	0,02	0,73	0,42
DrugBank	2,70	1,96	2,39	1,64	0,26	0,24	0,03	0,02	0,59	0,33

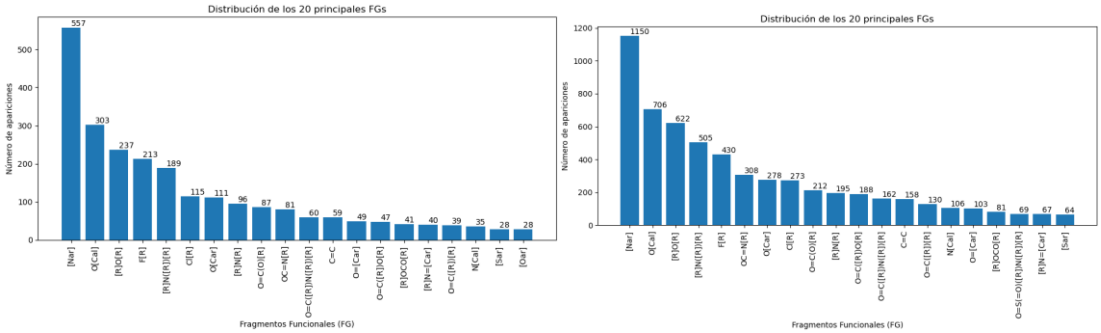


Figura 1: Estas figuras muestran la distribución de los 20 grupos funcionales más comunes en PDB por un lado(gráfico de la izquierda) y DrugBank por otro (gráfico de la derecha).

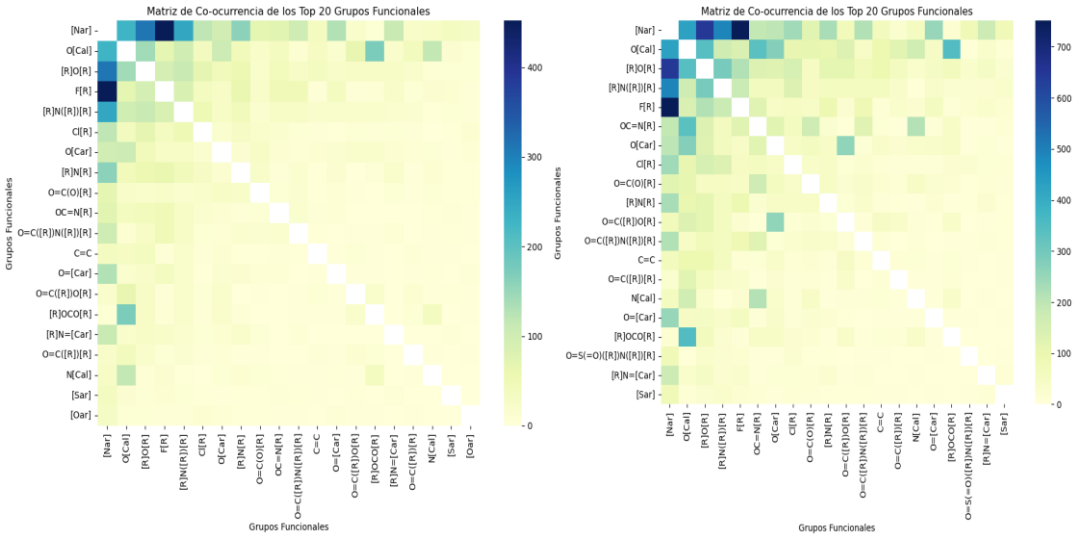


Figura 2: Heat map que muestra la co-ocurrencia de los 20 grupos funcionales más comunes en el dataset por un lado de PDB (izquierda) y por otro de DrugBank (derecha).

pChEMBL.py

Este script procesa un conjunto de datos de complejos proteicos, donde se busca obtener valores específicos de actividad biológica llamados *pChEMBL* usando para ello varias bases de datos.

- **Pasos Principales:**
  - Consulta de datos de actividad biológica en la base de datos de pChEMBL, tomando como identificador de molécula el InChi Key y el identificador de proteína UniProt ID.
  - Integración de los datos con datos de actividad biológica de PDB Bind.
- **Input:**
  - Filtered_Extract_Ligands_Uniprot.csv
  - pdb_bind_data.xlsx
- **Output:**
  - dataset_with_pchembl.csv

De los más de 5000 complejos iniciales de los que se parten, solo se ha podido obtener valores de afinidad para aproximadamente 1400. Debido a ello, se plantea la búsqueda de nuevos datos de afinidad en la base de datos de **Binding DB** para completar este conjunto de datos.