Trabajo de Final de Máster - PEC 0: Definición de los conceptos del trabajo

---

Alumno: Juan Vigueras Díaz

Tutor: Jorge Valencia Delgadillo

Título: Comparación entre modelos de aprendizaje profundo basados en estructuras y enfoques tradicionales de aprendizaje automático basados en descriptores para la predicción de la bioactividad en el marco del desarrollo de nuevos de fármacos contra las bacterias Escherichia coli, Staphylococcus aureus y Pseudomonas aeruginosa.

Palabras clave: Deep Learning, Machine Learning, Predicción In-Silico de Bioactividad, Desarrollo de Fármacos, Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa

Temática escogida: El desarrollo de fármacos constituye una serie de procesos de larga durabilidad en los que además se invierte una gran cantidad de recursos económicos. Ser capaces de detectar de manera precoz la inviabilidad de una molécula como posible fármaco implicaría el ahorro de estos recursos en favor de enfocar los estudios hacia propuestas con un potencial de éxito mayor. Bajo este contexto, el trabajo se enfoca en la aplicación de algoritmos de inteligencia artificial (*deep learning*, *machine learning*) para la predicción de bioactividad en tres bacterias comúnmente conocidas por ser las causantes de diversas patologías: E. coli, S. aureus y P. aeruginosa; así como la evaluación de viabilidad de esta tecnología como herramienta complementaria en la etapa inicial de investigación pre-clínica de una o varias moléculas candidatas.

Problemática a resolver: Aplicar inteligencia artificial para la predicción de bioactividad no es una tarea directa o trivial. Escoger adecuadamente el algoritmo, la tipología de los datos (estructurales, basados en la configuración espacial de la molécula, o indicativos, basados en descriptores moleculares), así como el (o los) indicativo(s) empleado(s), requiere de la comparación de múltiples experimentos hasta hallar, si existe, el modelo que presente mayor nivel de rendimiento para cada una de las especies de bacterias seleccionadas.

Objetivos:

La meta de este trabajo es estudiar la viabilidad y comparar el rendimiento de algoritmos de inteligencia artificial en la predicción de bioactividad para tres bacterias. Para ello, se define la siguiente lista de objetivos principales:

En relación al pre-tratamiento de datos obtenidos en la Base de Bioactividad 'ChEMBL':

- Seleccionar, para cada bacteria, una enzima de estudio que actúe como diana terapéutica y que contenga el suficiente número de registros para poder trabajar con modelos de IA. De no haber suficientes registros, se trabajaría con las bacterias al completo, a coste de perder información relativa al mecanismo de acción.
- Procesar las moléculas propuestas como fármacos sobre las que se tiene registro de bioactividad. Con la librería RdKit de Python, calcular ciertos descriptores moleculares a utilizar en los modelos de predicción de *machine learning*. Con la librería Pytorch Geometric de Python, calcular los grafos correspondientes a cada molécula a utilizar en los modelos de predicción de *deep learning*.

En relación a la implementación, entrenamiento y evaluación de distintos modelos de inteligencia artificial:

- Implementar una arquitectura de aprendizaje profundo (*deep learning*) con redes neuronales basada en capas gráficas convolucionales y perceptrones multicapa. Implementar los siguientes algoritmos de aprendizaje automático (*machine learning*): *x-gradient boost*, *random forest* y *support vector machine*.
- Entrenar estos modelos para cada una de las bacterias de estudio y para cada uno de los indicadores de bioactividad escogidos - IC50 y MIC[1] (en caso de trabajar con bacterias) o solamente IC50 (en caso de trabajar con enzimas). Optimizar los hiper-parámetros de manera progresiva en los sucesivos experimentos.
- Evaluar y analizar el rendimiento de los modelos para cada uno de los modelos con métricas adecuadas para una tarea de regresión. Con esto, inducir la eficacia del mejor modelo y su posible aplicación en el desarrollo de fármacos a la hora evaluar el potencial de eficacia de una nueva molécula de estudio.

Como productos finales del proyecto:

- Crear un repositorio de código en GitHub con el código implementado y los experimentos realizados.

---

[1] Mitad de la concentración inhibitoria máxima, IC50%; Concentración mínima inhibitoria, MIC

- En caso de obtener modelos con resultados de evaluación razonables para su puesta en producción, crear una aplicación para cargar una o varias moléculas en formato SMILES y devolver una predicción del indicativo de bioactividad correspondiente.

Bibliografía:

A continuación se muestra y se comenta con una breve descripción la bibliografía destacable que ha servido de inspiración para definir los conceptos y objetivos de este trabajo.

Consultas principales:

- Base de Datos ChEMBL. https://www.ebi.ac.uk/chembl/. Recuperado por última vez el 18/02/2022.

Base de datos de bioactividad que se utilizará en este trabajo. Contiene registros para un gran número de moléculas y bacterias, incluyendo diversos indicadores (MIC, IC50, % de Actividad, etc.).

- RdKit. https://www.rdkit.org. Recuperado por última vez el 18/02/2022.

Librería implementada en lenguaje Python para el tratamiento de formato SMILES y el cálculo de descriptores moleculares.

- Pytorch Geometric. https://pytorch-geometric.readthedocs.io/. Recuperado por última vez el 18/02/2022.

Extensión de la biblioteca Pytorch para el tratamiento de estructuras de datos de tipo nube, malla, red, etc. Incluye implementación de capa gráfica convolucional.

Artículos principales:

- Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. CoRR, abs/1510.02855. Opgehaal van http://arxiv.org/abs/1510.02855

Punto de partida propuesto por el tutor del trabajo en cuanto al uso de modelos convolucionales de inteligencia artificial para predecir bioactividad en el marco de desarrollo de fármacos.

- Zhang, H., Liao, L., Saravanan, K. M., Yin, P., & Wei, Y. (2019). DeepBindRG: a deep learning based method for estimating effective protein-ligand affinity. PeerJ, 7, e7362. https://doi.org/10.7717/peerj.7362

Alternativa al artículo de partida. En este caso se propone predecir el grado de afinidad para evitar así trabajar con estructuras moleculares tridimensionales, lo cual consumiría mucho tiempo y esfuerzo.

- Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. CoRR, abs/1609.02907. Opgehaal van http://arxiv.org/abs/1609.02907

Las redes gráficas convolucionales se presentan como una solución para abordar el tratamiento de estructuras de datos correspondientes a nubes de puntos categorizados unidos por enlaces (bordes).

- Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241–1250. doi:10.1016/j.drudis.2018.01.039

Se expone el auge y la utilidad de modelos de inteligencia artificial (en concreto, de *deep learning*) en marco de desarrollo de fármacos.

- Allel K, García P, Labarca J, Munita JM, Rendic M; Grupo Colaborativo de Resistencia Bacteriana; et al. Socioeconomic factors associated with antimicrobial resistance of Pseudomonas aeruginosa, Staphylococcus aureus, and Escherichia coli in Chilean hospitals (2008–2017). Rev Panam Salud Publica. 2020;44:e30. https://doi.org/10.26633/RPSP.2020.30

Estudio en el que se pone de manifiesto la reciente multi-resistencia desarrollada por este trío de bacterias que, además, son las causantes de diversas patologías comunes en humanos.

- Mirani ZA, Fatima A, Urooj S, Aziz M, Khan MN, Abbas T. Relationship of cell surface hydrophobicity with biofilm formation and growth rate: A study on Pseudomonas aeruginosa, Staphylococcus aureus, and Escherichia coli. Iran J Basic Med Sci. 2018;21(7):760-769. doi:10.22038/IJBMS.2018.28525.6917

Análisis sobre las asociaciones que este trío de bacterias pueden llegar a presentar en caso de infectar de manera simultánea.