Druggable Proteins Prediction

Mauricio Sandoval Cuenca

Enero 2021

1 Introducción

En los últimos años, el avance de la tecnología ha permitido recolectar una gran cantidad de datos biológicos de gran relevancia, sin embargo, a pesar de que el espacio de estas colecciones ha crecido exponencialmente, el impacto de esta información sigue siendo bajo en aplicaciones biomédicas como el desarrollo de fármacos [1].

Los métodos computacionales tienen la ventaja de poder tomar en cuenta las propiedades globales de una molécula, no limitándose a estudiar únicamente su secuencia, estructura o función. Las técnicas del aprendizaje de máquina han resultado ser herramientas poderosas para extraer información relevante de conjuntos de datos masivos y ruidosos.

El proceso de desarrollo de fármacos implica una gran cantidad de etapas costosas y laboriosas en las que es necesario un alto nivel de seguridad. Típicamente el punto de partida es la identificación de moléculas objetivo (usualmente proteínas) sobre las cuales los fármacos puedan tener un efecto favorable. A esta propiedad se le conoce como *druggability* [2].

A pesar de que se han usado diversos algoritmos basados en: Support Vector Machines, árboles de decisión, ensemble of classifiers, regresión logística, redes bayesianas, etc., para identificar la *drogabilidad* de una molécula; las redes neuronales aún no han sido ampliamente exploradas para predecir cuando una proteína objetivo puede emplearse efectivamente para diseñar fármacos.

En este proyecto trabajaremos con el conjunto de datos ubicado en el siguiente repositorio

`https://github.com/mauriciosandovl/druggable_proteins_prediction.git`

La base fue construida manualmente a partir de datos extraído mediante técnicas de dinámica molecular. Nuestro conjunto de datos está formado por 39 características biológicas de un total de 5597 proteínas. Dadas estas características el objetivo es predecir si una proteína es es candidata o no para emplearse en el diseño de fármacos (*druggable protein*).

Las ideas originales de este proyecto fueron planteadas con ayuda del Dr. Marcelino Arciniega Castro[1], que actualmente se encuentra asesorando mi proyecto de tesis en el área de biología molecular computacional.

2 Procedimiento

Para la visualización y el pre procesamiento de los datos, usaremos las librerías `pandas` y `sklearn`. Una vez terminado el pre procesamiento guardaremos los datos en un archivo nuevo que será el que ingresará a nuestra red neuronal.

Empezaremos el pre procesamiento normalizando los datos sustrayendo la media y dividiendo entre la desviación estándar para que el rango de valores sea uniforme. Ahora, como nuestra base de datos original consta de 39 características biológicas de proteínas, en principio podríamos pensar que quizá no sean necesarias todas para hacer la clasificación, en efecto, realizando una matriz de correlación (Figura 1) podemos

[1] `http://www.ifc.unam.mx/investigadores/marcelino-arciniega`

observar que hay características que guardan una correlación alta y concluimos que hay información que no es indispensable.

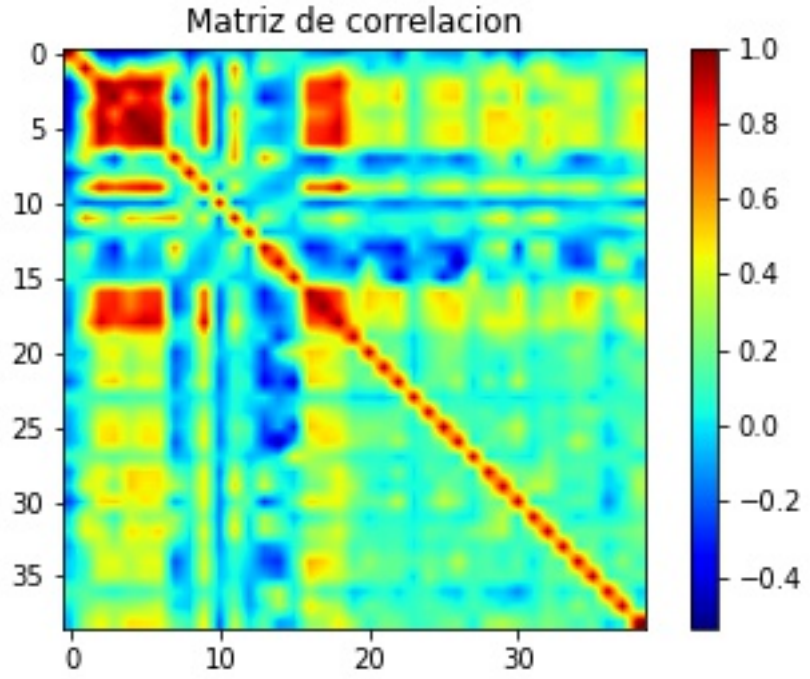

Figure 1: Correlación de las características que conforman nuestros datos

Para disminuir la dimensión haremos uso del *Principal Component Analysis* (PCA) sobre nuestros datos normalizados para disminuir la dimensión y optimizar el costo computacional.

Posteriormente dividiremos los datos en un conjunto de entrenamiento con un 80% de los datos y conjunto de prueba con el resto. Aplicaremos los datos pre procesados en *minibatches* de tamaño 10 a una red de aprendizaje profundo que en principio, tendrá arquitectura conformada por:

- 1 capa de entrada con tamaño igual a la dimensión obtenida después de aplicar PCA
- 2 capas ocultas completamente conectadas mediante ReLU ambas de tamaño 30
- 1 capa de salida con dos nodos seguida de la función log softmax

Para el entrenamiento tomaremos un total de 50 tiempos, usaremos entropía cruzada como función de pérdida, como optimizador usaremos descenso por el gradiente estocástico y graficaremos la función de pérdida respecto al conjunto de entrenamiento y de prueba simultáneamente. Finalmente, con el conjunto de entrenamiento realizaremos una matriz de confusión para evaluar el rendimiento de nuestra red.

References

[1] Kandoi G, Acencio ML and Lemke N (2015) *Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review.* Front. Physiol. 6:366. doi: 10.3389/fphys.2015.00366

[2] `https://en.wikipedia.org/wiki/Druggability`. Consultado el 26 de noviembre de 2020 a las 16:00 hrs.