**PREGUNTA INICIAL DE INTERÉS BIOLÓGICO:**

¿Es posible entrenar una red neuronal para que a partir de los datos de posiciones metiladas de una muestra sea capaz de identificar a qué categoría de tejido pertenece (clasificarla en una de las 20 categorías de tejidos del TCGA)?.

Variación de la pregunta:

* Lo mismo pero considerando un algoritmo random forest.
* Lo mismo pero considerando un algoritmo árbol de decisión. Este algoritmo tiene la ventaja de ser fácilmente interpretable, por lo que en el supuesto de que tuviera una capacidad predictiva aceptable podríamos buscar la coherencia de sus criterios de clasificación con los resultados de estudios previos de metilación diferencial entre grupos.
* Lo mismo pero en lugar de utilizar datos de metilación usando los datos (picos) de histonas modificadas.

Se partiría de los datos depositados en TCGA (aproximadamente ¿? 11000 muestras existentes en TCGA) cuyas tablas de fenotipos contemplan unas ¿? 400 variables, entre las que se encuentran los tipos de tejido.

El planteamiento trata de ser algo diferente del que he observado en otros trabajos y que resumido consistiría en partir de dos grupos de muestras (dos grupos fenotípicamente distintos) y obtener las “features” (expresión de genes, locis metilados, picos de histonas) que estadísticamente poseen valores diferenciales.

En cuanto a la búsqueda de bibliografía:

Partiendo de Búsquedas en GOOGLE scholar:

* + - epigenetic methylation
    - promoter methylation
    - [histone methylation](https://scholar.google.es/scholar?hl=es&as_sdt=0,5&qsp=3&q=histone+methylation&qst=ib)
    - dna methylation

Ordenaría por fecha de publicación, puesto que las librerías R para manejar los repositorios de datos de TCGA son recientes no creo que sea preciso remontarse muchos años atrás para comprobar si existe publicado algún trabajo con este enfoque.

Este es el apartado en el que, de ser posible, querría alguna indicación para poder avanzar estos dos meses de verano: disponer de algún criterio de selección de artículos dado que su número es abrumador.

**PROBLEMÁTICA ESPERADA**

* Obtener una Conclusión del trabajo de dos palabras: “NO FUNCIONA”. Si el objetivo es la comparación de dos grupos de muestras fenotípicamente distintas siempre vamos a encontrar algún grupo de “features” con valores significativamente diferentes y por lo tanto llegaremos a algo que nos permite redactar una “Conclusión”. En esta propuesta no esta garantizado que sea capaz de diseñar y entrenar un modelo con resultados predictivos aceptables.
* Que el volumen de muestras disponible no sea ni de lejos el necesario para entrenar una red neuronal y por lo tanto la matriz de confusión final sea decepcionante (estaría en la situación descrita en el punto anterior).
* Que las capacidades de cálculo no sean suficientes y no consiga entrenar el algoritmo.
* Las habituales de todo el mundo de limitación de tiempo y conocimientos a las que hay que añadir las mías propias.

Los documentos y scripts del trabajo Fin de Máster de máster UOC se almacenan en la ubicación:

C:\TFM UOC

https://github.com/ajoven10/TFM\_UOC.git