Máster Oficial en Big Data Science
Propuesta Trabajo Fin de Máster (TFM)
Año Académico 2022-2023

Título del proyecto: Diseño de un sistema inteligente de predicción de interacciones genómicas *in silico.* (Código D3)

Empresa / Departamento: Universidad de Navarra / Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI)

Tutor de Empresa: Rubén Armañanzas Arnedillo
Contacto/Email: rarmananzas@unav.es
Cotutor de Empresa: Aitor Oviedo Madrid
Contacto/Email: aoviedomadr@unav.es

Descripción

La validación de interacciones moleculares entre los compuestos derivados de la activación de dos genes es una labor tediosa y realizada principalmente en base a experimentos de laboratorio. Este proyecto enmarcado dentro de la temática de *Systems Biology* busca probar la hipótesis de que un modelo de caja negra entrenado con interacciones ya validadas es capaz de predecir nuevas interacciones entre pares de genes de manera sintética, *i.e., in silico.*

Detalles del proyecto: objetivos, metodología, fases, herramientas recomendadas, etc.

Los objetivos principales del proyecto son:

- Desarrollar un sistema de información capaz de almacenar de manera estructurada toda la base de conocimiento disponible en repositorios públicos.
- Diseñar una representación de *embedding* para una secuencia genómica determinada.
- Desarrollar una red neuronal profunda capaz de predecir interacciones entre dos genes.

El trabajo consistiría en las siguientes fases de trabajo:

1. Obtener los ficheros de datos en bruto desde ftp.ncbi.nlm.nih.gov, entre otros:

   Gene RIFs, interaciones ya descritas, secuenciación completa de cada gen, información de contexto del gen, publicaciones de referencia, etc.

   Anotación automática de tags/características relevantes basados en análisis del lenguaje natural de las anotaciones individuales.

2. Implementar un sistema de información capaz de almacenar y devolver información estructurada de los elementos obtenidos en 1, así como las interacciones ya descritas (p. ej., una base de datos noSQL).

3. Diseñar una estructura de *embedding* que sea capaz de codificar eficientemente la secuencia genómica de genes humanos (*homo sapiens*).

4. Diseñar, entrenar, y validar un modelo basado en arquitecturas de aprendizaje profundo (*deep learners)* capaz de predecir con cierto grado de certeza la posible interacción entre los

transcritos de dos genes determinados (en Python). Entre otras tareas requerirá:

Codificación de la capa de entrada, número y tipo de capas internas, funciones de activación, tipo de salida (determinista *vs.* probabilística).

Prueba de concepto con particiones *train+test* de los datos y estimación del rendimiento.