UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Propuesta TFG. Curso 2024/2025

GRADO: Grado en Biotecnología

CÓDIGO DEL TFG: 251-005-2024/2025

1. DATOS BÁSICOS DEL TFG:

Título: Aplicación de la Inteligencia Artificial para el desarrollo de una herramienta informática de predicción funcional de proteínas

Descripción general (resumen y metodología):

La capacidad de predecir la función de una proteína a partir de su secuencia de aminoácidos es crucial para entender su papel biológico y su potencial aplicación en diversos campos, como la medicina, la biotecnología y la investigación farmacéutica. Dentro de este ámbito se engloban una serie de problemas de índole diversa: anotación funcional de secuencias, predicción de unión a ligandos (docking), clasificación funcional de proteínas y predicción de contactos cuaternarios e interacción proteína-proteína, son algunos de ellos.

Son todos ellos problemas no resueltos, de gran relevancia estratégica en el campo de la biocomputación, que han recibido una considerable atención a lo largo de los últimos años. La predicción funcional de proteínas es clave, de hecho, en diseño e ingeniería de proteínas, en la identificación de dianas terapéuticas y biomarcadores y en genómica funcional, tanto en la vertiente de anotación de genomas como en la búsqueda de homólogos funcionales.

En este trabajo de fin de carrera, se propone el desarrollo de un prototipo de herramienta informática basada en Inteligencia Artificial (IA) para predecir la función de proteínas. Los recientes avances en el desarrollo de algoritmos de IA cada vez más potentes y los resultados espectaculares obtenidos en los últimos años en el campo de la predicción de la estructura tridimensional de proteínas, hacen suponer que su potencial para predecir la función de una proteína a partir de su secuencia o de su estructura puede ser muy grande.

Para abordar el problema existen varios algoritmos y modelos que podrán emplearse como punto de partida: redes neuronales, árboles de decisión aleatorios (“Random Forests”) y máquinas de vector soporte (“Support Vector Machines” o “SVM”) y modelos lingüísticos basados en GPT3 y GPT4 entre otros. Asimismo, disponemos de varios repositorios y bases de datos que podrían emplearse para entrenar los algoritmos y obtener los modelos adecuados. Destacan entre ellas: Uniprot (https://www.uniprot.org/), Protein Data Bank (PDB) (https://www.rcsb.org/), Pfam-InterPro (https://www.ebi.ac.uk/interpro/) y GO (http://geneontology.org/).

El prototipo será implementado sobre alguna de las librerías de “machine learning” disponibles en Python (TensorFlow, Keras, Scikit-Learn o PyTorch) y dispondrá de una interface gráfica (“front-end”) y amigable implementada en el entorno de desarrollo rápido de aplicaciones Lazarus (basado en Free Pascal), aunque se considerarán también otras opciones como Tkinter o PyQt, disponibles como paquetes y librerías de Python.

Metodología

Recopilación y preparación de datos: Se recopilarán conjuntos de datos de proteínas con funciones conocidas y se realizará un pre-procesamiento adecuado para garantizar la calidad y la integridad de los datos.

Desarrollo del modelo de IA: Se utilizarán algoritmos de aprendizaje automático supervisado, como redes neuronales, “random forest” o “SVM” para entrenar un modelo capaz de predecir la función de las proteínas.

Implementación de la herramienta informática: Se creará una interfaz de usuario amigable y de fácil uso que permita a los usuarios ingresar la secuencia de aminoácidos de una proteína y obtener la predicción funcional correspondiente.

Evaluación y validación: Se realizarán pruebas y análisis sistemáticos para evaluar la precisión y el rendimiento de la herramienta en comparación con las funciones conocidas de las proteínas de

prueba.
Optimización y mejoras: Se buscarán formas de mejorar la precisión y eficiencia de la herramienta a través de la optimización de parámetros y la exploración de nuevas técnicas de IA.
Documentación y presentación: Se elaborará un informe detallado del trabajo realizado, incluyendo la descripción de la metodología, los resultados obtenidos y las conclusiones derivadas del estudio.

Tipología: Estudio de casos, teóricos o prácticos, relacionados con la temática del Grado.

Objetivos planteados:

1. Diseño de una estrategia completa ("workflow") para la predicción funcional de proteínas basada en técnicas de aprendizaje automático ("machine learning") e inteligencia artificial. Este objetivo incluirá la revisión de la bibliografía, la identificación de las bases de datos y los algoritmos de IA adecuados, la definición y configuración de las herramientas informáticas necesarias y la elección de los oportunos indicadores estadísticos que permitan la evaluación de los resultados obtenidos.
2. Diseño e implementación de una interfaz de usuario intuitiva y flexible para la herramienta de predicción funcional de proteínas que permita construir la matriz de características ("features"), configurar y entrenar el algoritmo de aprendizaje para crear diferentes modelos predictivos y, finalmente, probar los modelos obtenidos y evaluar la calidad de sus predicciones.

Bibliografía básica:

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235. (https://www.rcsb.org/)
The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506-D515. (https://www.uniprot.org/)
Mitchell, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research, 47(D1), D351-D360. (https://www.ebi.ac.uk/interpro/)
The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research, 47(D1), D330-D338. (http://geneontology.org/)
Kelley, L. A., et al. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols, 10(6), 845-858.
Szklarczyk, D., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(D1), D607-D613.

Recomendaciones y orientaciones para el estudiante:

Para la realización de este TFG es altamente recomendable cursar la asignatura de Ingeniería de Proteínas o disponer de conocimientos previos de programación.

Plazas: 1

2. DATOS DEL TUTOR/A:

Nombre y apellidos: FERNANDO JESÚS REYES ZURITA

Ámbito de conocimiento/Departamento: BIOQUÍMICA Y BIOLOGÍA MOLECULAR I

Correo electrónico: ferjes@ugr.es

3. COTUTOR/A DE LA UGR (en su caso):

Nombre y apellidos: Hilario Ramírez Rodrigo

Ámbito de conocimiento/Departamento: BIOQUÍMICA Y BIOLOGÍA MOLECULAR I

Correo electrónico: hilario@ugr.es

4. COTUTOR/A EXTERNO/A (en su caso):

Nombre y apellidos:

Correo electrónico:

Nombre de la empresa o institución:

Dirección postal:

Puesto del tutor en la empresa o institución:

Centro de convenio Externo:

5. DATOS DEL ESTUDIANTE:

Nombre y apellidos:

Correo electrónico: