UNIVERSIDAD DE BURGOS
ESCUELA DE DOCTORADO

TESIS DOCTORALES

TÍTULO: MACHINE LEARNING APPROACHES IN BIOINFORMATICS; ADVANCES IN TRANSCRIPTION AND PROTEIN FITNESS PREDICTION.

AUTOR: BARBERO APARICIO, JOSÉ ANTONIO

PROGRAMA DE DOCTORADO: INGENIERÍA Y TECNOLOGÍAS INDUSTRIAL, INFORMÁTICA Y CIVIL

ACTO Y FECHA DE LECTURA: EL ACTO PÚBLICO DE DEFENSA DE TESIS SE DESARROLLARÁ, EL DÍA 01 DE DICIEMBRE DE 2023, A LAS 10:30 HORAS, PRESENCIALMENTE EN LA SALA DE JUNTAS DE LA ESCUELA POLITÉCNICA SUPERIOR (RÍO VENA), DE LA UNIVERSIDAD DE BURGOS, Y DE MANERA TELEMÁTICA, A TRAVÉS DE LA APLICACIÓN MICROSOFT TEAMS.

DIRECTORES: D. CÉSAR IGNACIO GARCÍA OSORIO
D. JOSÉ FRANCISCO DÍEZ PASTOR

TRIBUNAL: D. GONZALO CERRUELA GARCÍA
D. ÁLVAR ARNAIZ GONZÁLEZ
D. MARCIN BLACHNIK
DÑA. AIDA DE HARO GARCÍA
D. MARIO JUEZ GIL

RESUMEN: A medida que nos seguimos adentrando en la era de la información, la bioinformática está pasando a ser cada vez más importante en la biología moderna, en gran parte debido a su papel crítico en el procesamiento y análisis de la gran cantidad de datos complejos generados en el campo. A menudo los métodos tradicionales encuentran dificultades debidas al gran volumen y complejidad de estos datos, posicionando a las técnicas de aprendizaje automático como una solución más óptima. La exploración de la intersección entre aprendizaje automático y la bioinformática ofrece numerosas oportunidades para el desarrollo y la mejora de herramientas computacionales diseñadas para manejar y obtener información crítica sobre estos grandes conjuntos de datos.

El objetivo principal de esta tesis es el desarrollo de una exploración exhaustiva de las posibilidades del aprendizaje automático en bioinformática, con un enfoque específico en problemas como la predicción del inicio de la transcripción y la predicción del fitness en las proteínas. Además, dadas las similitudes entre las secuencias bioinformáticas y el campo del procesamiento del lenguaje natural, la investigación tiene un claro énfasis en el uso de métodos basados en secuencias.

Este trabajo ha resultado en la producción de varias contribuciones al campo en forma de tres artículos científicos. Los dos primeros se centran en la predicción del inicio de la transcripción. En el primero de ellos descubrimos que la integración de simulaciones biofísicas en conjunto con la secuencia de ADN

puede mejorar los resultados de los métodos de aprendizaje automático. En el segundo, además, llegamos a la conclusión de que, mientras que las máquinas de soporte vectorial han sido una opción muy establecida en el campo de la predicción del inicio de la transcripción, nuestra investigación sugiere que los métodos de aprendizaje profundo los superan, marcando un cambio de paradigma en el área. Además, presentamos conjuntos de datos personalizados a partir de datos de Ensembl, proporcionando un recurso valioso para futuros estudios. El tercer artículo aborda la predicción del fitness en proteínas, específicamente en escenarios con conjuntos de datos escasos y concluye que los métodos de deep transfer learning se establecen como la mejor alternativa ante otras estrategias bien adaptadas a tales situaciones, como los métodos de aprendizaje semi-supervisado.

Palabras clave: Bioinformática, Aprendizaje Automático, Sitio de Inicio de la Transcripción, Aprendizaje Profundo, Aptitud de la Proteína.

Keywords: Bioinformatics, Machine Learning, Transcription Start Site, Deep Learning, Protein Fitness.