

# Tabla de Contenido

<b>1. Introducción y contexto general</b>	<b>1</b>
1.1. Ingeniería de proteínas y estrategias de diseño de variantes . . . . .	3
1.2. Métodos computacionales aplicados en ingeniería de proteínas . . . . .	6
1.2.1. Métodos de análisis filogenéticos . . . . .	6
1.2.2. Métodos de análisis de estructuras . . . . .	7
1.2.3. Métodos de estudio de mutaciones . . . . .	8
1.3. Inteligencia artificial y machine learning . . . . .	8
1.4. Machine learning y tipos de aprendizaje . . . . .	9
1.4.1. Aprendizaje supervisado . . . . .	11
1.4.2. Algoritmos clásicos de aprendizaje supervisado . . . . .	11
1.4.3. Neuronas artificiales, redes neuronales y deep learning . . . . .	12
1.4.4. Desempeño y evaluación de los algoritmos . . . . .	15
1.4.5. Principales problemas en el entrenamiento de modelos . . . . .	16
1.4.6. Generando un modelo predictivo . . . . .	16
1.4.7. Algoritmos clásicos de aprendizaje no supervisado . . . . .	17
1.4.8. Evaluación y desempeño . . . . .	19
1.5. Aplicaciones del machine learning en ingeniería de proteínas . . . . .	20
1.5.1. Machine learning y deep learning como soporte de biología estructural	21
1.5.2. Machine learning y deep learning como soporte para el diseño de nuevas secuencias . . . . .	21
1.5.3. Aplicaciones del <i>deep learning</i> y el <i>machine learning</i> para resolver ta- reas específicas de ingeniería de proteínas . . . . .	22

1.6.	Los mayores desafíos de la ingeniería de proteínas en las últimas décadas . . .	22
1.6.1.	Representaciones numéricas, cuál es la mejor alternativa? . . . . .	22
1.6.2.	Estudio de mutaciones puntuales, cómo podemos caracterizar las mutaciones? . . . . .	23
1.6.3.	Diseñar mutaciones, proponiendo nuevas estrategias de diseño . . . .	24
1.7.	Hipótesis . . . . .	25
1.8.	Objetivos . . . . .	25
1.8.1.	Objetivo general . . . . .	25
1.8.2.	Objetivos específicos . . . . .	25
<b>2.</b>	<b>Representaciones numéricas</b>	<b>27</b>
2.1.	Estrategias y metodologías de codificación . . . . .	28
2.1.1.	One hot y ordinal encoder . . . . .	28
2.1.2.	Frecuencias de residuos . . . . .	30
2.1.3.	Uso de propiedades fisicoquímicas . . . . .	30
2.1.4.	Natural language processing . . . . .	31
2.2.	Imágenes y estructuras de grafos como estrategias de representación de proteínas	33
2.2.1.	Estructuras de grafos aplicadas a representación de proteínas . . . . .	33
2.2.2.	Aplicaciones de imágenes a representaciones de proteínas . . . . .	34
2.3.	Aplicaciones de digital signal processing . . . . .	35
2.4.	Transformaciones de Fourier . . . . .	36
2.4.1.	Transformada rápida de Fourier (FFT) . . . . .	37
2.4.2.	Aplicaciones de las transformadas de Fourier en ingeniería de proteínas	38
2.5.	Principales problemáticas asociadas a la representación de proteínas . . . . .	39
2.6.	Metodología . . . . .	40
2.6.1.	Selección de propiedades desde AAIndex . . . . .	40
2.6.2.	Selección de casos de estudio y preparación de los conjuntos de datos	42
2.6.3.	Comparaciones de rendimiento para tareas de ingeniería de proteínas	43

2.6.4.	Implementaciones y comentarios generales . . . . .	43
2.7.	Resultados y discusiones . . . . .	43
2.7.1.	Identificación de grupos semánticos de propiedades fisicoquímicas . . . . .	43
2.7.2.	La combinación de FFT y codificadores de propiedades fisicoquímicas semánticas mejoran el rendimiento de modelos predictivos . . . . .	46
2.7.3.	Reconocimiento de patrones visuales en plegamientos y funciones enzimáticas . . . . .	50
2.8.	Conclusiones y comentarios generales . . . . .	51
<b>3.</b>	<b>Métodos de ensamble para mejorar el rendimiento de modelos predictivos</b>	<b>53</b>
3.1.	Estrategias de entrenamiento de modelos predictivos . . . . .	54
3.2.	Metodología . . . . .	56
3.2.1.	Descripción del pipeline . . . . .	56
3.2.2.	Codificación y procesamiento de las secuencias . . . . .	57
3.2.3.	Casos de uso y conjuntos de datos de prueba . . . . .	59
3.3.	Resultados y discusiones . . . . .	60
3.4.	Conclusiones y principales comentarios . . . . .	64
<b>4.</b>	<b>Estrategias de diseño de secuencias con propiedades deseables</b>	<b>66</b>
4.1.	Metodología . . . . .	67
4.1.1.	Diseño e implementación de herramientas de exploración de landscapes . . . . .	68
4.1.2.	Diseño e implementación de estrategias de exploración de secuencias de péptidos para evaluar actividades biológicas deseables . . . . .	68
4.1.3.	Diseñando modelos en conjuntos de secuencias poco informativos y explorando sistemas probabilísticos de predicción . . . . .	69
4.2.	Resultados y discusiones . . . . .	76
4.2.1.	Exploración de landscapes . . . . .	76
4.2.2.	Exploración de actividades biológicas para secuencias de péptidos . . . . .	79
4.2.3.	Diseño e implementación de modelos productivos para elaboración de péptidos con propiedades deseables . . . . .	81
4.3.	Conclusiones y comentarios generales . . . . .	84

<b>5. Análisis de mutaciones puntuales e identificación de sitios relevantes para mutagénesis</b>	<b>86</b>
5.1. Metodología . . . . .	88
5.1.1. Caracterización de mutaciones en conjuntos de proteínas . . . . .	88
5.1.2. Diseño y entrenamiento de modelos predictivos . . . . .	89
5.1.3. Identificación de sitios relevantes . . . . .	89
5.1.4. Validación y conjuntos de datos de prueba . . . . .	90
5.2. Resultados y discusiones . . . . .	90
5.2.1. Entrenamiento de modelos predictivos para sistemas de mutaciones puntuales . . . . .	90
5.2.2. Etapas y consideraciones al identificar sitios de interés . . . . .	92
5.2.3. Identificación de sitios relevantes en Epoxide Hydrolase en estudios de enantioselectividad . . . . .	93
5.3. Conclusiones y comentarios generales . . . . .	94
<b>6. Conclusiones, trabajos en desarrollo y perspectivas a futuro</b>	<b>96</b>
6.1. Conclusiones generales . . . . .	96
6.2. Trabajo en desarrollo y resultados preliminares . . . . .	97
6.3. Perspectivas y proyectos a futuro . . . . .	99
<b>Bibliografía</b>	<b>115</b>