

Tabla de Contenido

1.1. Antecedentes	1
1.1.1. Antecedentes Generales	1
1.1.2 Métodos de Ingeniería de Proteínas	1
1.2. Motivación	2
1.4. Objetivos	3
1.4.1. Objetivo General	3
1.4.2. Objetivos Específicos	3
1.6. Resultados Esperados	3
2. Marco Teórico	4
2.1. Codificación	4
2.1.1. One-hot encoder	4
2.1.2. Composición de aminoácido (AAC)	5
2.1.4. Composición de dipéptidos y tripéptidos	5
2.1.5. Composición terminal	6
2.1.3. Composición, transición y distribución (CTD)	6
2.1.6. Composición de pseudo-aminoácidos (PseAAC)	7
2.1.7. Alfabeto reducido de aminoácidos	9
2.1.8. Modelo N-gram	9
2.1.9. Propiedades fisicoquímicas	9
2.1.10. Autocorrelación	9
2.1.11. Matriz de puntaje de posición específica (<i>Position Specific Scoring Matrix, PSSM</i>)	10
2.1.12. Transformada de Fourier (digitalización de propiedades fisicoquímicas)	10
2.2. Preprocesamiento de los datos	11
2.3. Selección de características	13
2.6. Algoritmos de aprendizaje supervisado	14
2.6.1. K-nearest neighbor (KNN)	14
2.6.3. Support vector machine (SVM)	14
2.6.4. Decision Trees	15
2.6.5. Random Forest (RF) y Gradient Boosting Decision Trees (GBDT)	16
2.6.6. Artificial Neural Network (ANN)	16
2.6.6. Convolutional Neural Network (CNN)	18

2.7. Validación cruzada	19
2.8. Indicadores de desempeño	21
2.8.1. Problemas de clasificación binaria	21
2.8.2. Problemas de clasificación de múltiples clases	22
2.8.3. Problemas de regresión	23
2.9. Alineamiento	23
3. Metodología	25
3.1. Conjunto de Datos	25
3.2. Alineamiento	25
3.3. Caracterización de los espectros:	26
4. Conjuntos de datos y caracterización	28
4.1. Recolección de los conjuntos de datos	28
4.2. Caracterización de largo de secuencias	31
4.3. Caracterización de las respuestas	37
4.4. Composición aminoacídica de los conjuntos	38
5. Alineamiento	44
6. Caracterización de espectros	48
6.1. VaxinPad:	48
6.2. DBP	49
6.3. iAMP-2L_multiclass	51
6.4. Enantioselectivity:	54
6.5. T50	55
7. Modelos predictivos	57
8. Conclusiones	66
9. Bibliografía	68
Anexos	74
Anexo A: Descripción detallada de los conjuntos de datos	74
AntiTb_primary	74
AntiTb_secondary	74
ACP-DL	74
iACP	75
QSP	75
iAMP-2L_multiclass	75
iAMP-2L_binary	76
DBP	76

Pop	77
VaxinPad	77
Solub	77
Enantioselectivity	77
Localization	77
T50	78
RT	78
Anexo B: Histogramas de largo de secuencias para los conjuntos fuera de los casos de estudio	79
Anexo C: Histogramas de respuestas para los conjuntos de regresión fuera de los casos de estudio	86
Anexo D: Hiperparámetros estimados para cada modelo	88
Random Forest	88
Support Vector Machine	89
K-Nearest Neighbor	90
Selección del algoritmo	90
Anexo E: Código	92
Formas de codificación	92
Modelos predictivos	97