

Índice general

AGRADECIMIENTOS	I
1. Introducción	1
1.1. Marco teórico	5
1.1.1. Leucemia.	5
1.1.2. Antígenos y anticuerpos	6
1.1.3. Auto antígenos y auto reactividad.	11
1.2. Hipótesis	12
1.3. Objetivo general	12
1.3.1. Objetivos específicos.	12
2. Modelamiento predictivo aplicado a sistemas de interacción antígeno anticuerpo	14
2.1. Marco teórico.	16
2.1.1. Data mining	16
2.1.2. Machine learning	17
2.1.2.1. Supervised learning	18
2.1.2.1.1. K-Nearest Neighbors (KNN)	18
2.1.2.1.2. Decision Tree	19
2.1.2.1.3. Naive Bayes	19
2.1.2.1.4. Métodos de ensamble	20
2.1.2.1.5. Métodos de Regresión	20
2.1.2.2. Unsupervised learning	21
2.1.2.3. Problemas asociados al desarrollo de modelos	22
2.1.2.3.1. Evaluación de conjuntos de datos	22
2.1.2.3.2. Dimensionalidad de conjuntos de datos	23
2.1.2.3.3. Sobreajuste y validacion de modelos.	24
2.1.2.4. Medidas de desempeño.	27
2.1.3. Neural networks o redes neuronales	29
2.1.4. Codificación de secuencias	35
2.1.5. Embeddings y auto encoders.	36
2.1.6. Herramientas de auto encoding	39
2.1.7. Natural Language Process (NLP)	41
2.2. Metodología y estrategias de desarrollo.	43
2.2.1. Limpieza y preparación del set de datos.	44

2.2.2. Codificación de secuencias	44
2.2.3. Estrategias de entrenamiento y selección de modelos predictivos.	45
2.2.4. Comparar y evaluar estrategias de codificación.	49
2.2.5. Corroborar y estimar escalabilidad del modelo predictivo. .	50
2.2.6. Herramientas.	51
2.3. Resultados y discusión.	53
2.4. Conclusiones	71
3. Bases de datos	74
3.1. Bases de datos previamente reportadas	76
3.1.1. Bases de datos de interacción Ab-Ag.	77
3.1.2. Bases de datos de anticuerpos.	80
3.1.3. Bases de datos de antígenos.	87
3.1.4. Bases de datos de epítopes.	90
3.2. Metodología.	93
3.2.1. Descarga de set de datos.	93
3.2.2. Procesamiento de set de datos.	93
3.2.3. Unión y depuración set de datos.	94
3.2.4. Diseño e implementación base de datos.	98
3.2.5. Caracterización de secuencias.	99
3.3. Resultados y discusión.	102
3.4. Conclusiones	114
4. Proyecciones y trabajo a futuro	117
4.1. Metodología	119
4.1.1. Alineamiento de secuencias auto antígenas	119
4.1.2. Distribución de distancias	119
4.1.3. Categorización con respecto a distribución	120
4.1.4. Alineamiento contra secuencias experimentales	120
4.1.5. Comparación y categorización	120
4.2. Resultados parciales	121
4.2.1. Métodos de clasificación binaria basados en propiedades filogenéticas	127
4.3. Discusión	130
5. Discusión y conclusiones generales.	132
Referencias	141
Material suplementario	162
A. Material suplementario	162

Índice de cuadros

2.2.1.Tabla de algoritmos e iteraciones	46
2.2.1.Tabla de algoritmos e iteraciones	47
2.2.2.Tabla de herramientas.	53
2.3.1.Tabla de modelos por algoritmo	63
3.1.1.Tabla resumen bases de datos de interacción antígeno - anticuerpo.	79
3.1.1.Tabla resumen bases de datos de interacción antígeno - anticuerpo.	80
3.1.2.Tabla resumen bases de datos de anticuerpos.	81
3.1.2.Tabla resumen bases de datos de anticuerpos.	82
3.1.2.Tabla resumen bases de datos de anticuerpos.	83
3.1.2.Tabla resumen bases de datos de anticuerpos.	84
3.1.2.Tabla resumen bases de datos de anticuerpos.	85
3.1.2.Tabla resumen bases de datos de anticuerpos.	86
3.1.2.Tabla resumen bases de datos de anticuerpos.	87
3.1.3.Tabla resumen bases de datos de antígenos.	88
3.1.3.Tabla resumen bases de datos de antígenos.	89
3.1.3.Tabla resumen bases de datos de antígenos.	90
3.1.4.Tabla resumen bases de datos de epítopes.	91
3.1.4.Tabla resumen bases de datos de epítopes.	92
3.2.1.Tabla resumen propiedades caracterizadas y predichas.	99
3.2.1.Tabla resumen propiedades caracterizadas y predichas.	100
3.2.1.Tabla resumen propiedades caracterizadas y predichas.	101
A.0.1Tabla modelos generados por set de datos utilizado en proceso exploratorio	162
A.0.1Tabla modelos generados por set de datos utilizado en proceso exploratorio	163
A.0.1Tabla modelos generados por set de datos utilizado en proceso exploratorio	164
A.0.1Tabla modelos generados por set de datos utilizado en proceso exploratorio	165
A.0.1Tabla modelos generados por set de datos utilizado en proceso exploratorio	166

Índice de figuras

1.1.1.Esquema de un anticuerpo.	9
1.1.2.Esquema de la recombinación VDJ.	10
2.1.1.Ejemplo matriz de confusión.	30
2.1.2.Esquema de una red neuronal con arquitectura Feed-forward	33
2.1.3.Esquema de una red neuronal con arquitectura Recurrent	34
2.1.4.Esquema de autoencoder	38
2.2.1.Esquema codificación de secuencias	45
2.2.2.Esquema de entrenamiento de modelos	46
2.2.3.Esquema construcción del modelo ensamblado	48
2.2.4.Esquema predicción en modelo ensamblado.	49
2.2.5.Esquema de evaluación de estrategias de codificación.	51
2.2.6.Ejemplos de codificaciones utilizadas.	52
2.2.7.Esquema proceso Leave One Antibody Out	52
2.3.1.Número de ejemplos de clase por anticuerpo	54
2.3.2.Logos conjunto de anticuerpos analizado.	55
2.3.3.Logos conjunto completo de anticuerpos.	56
2.3.4.Docking anticuerpo A021 y antígeno uORF:IOH38079.	58
2.3.5.Docking anticuerpo A120 y antígeno uORF:IOH38079.	60
2.3.6.Largo vectores por codificación	62
2.3.7.Performance promedio algoritmos	64
2.3.8.Performance validación cruzada	66
2.3.9.Performance Leave One Antibody Out	68
2.3.10.Performance modelo ensamblado	69
2.3.11.Predicción de clases modelo ensamblado	70
3.1.1.Esquema de bases de datos	78
3.2.1.Esquema descarga de datos	94
3.2.2.Esquema procesamiento de datos.	95
3.2.3.Esquema unión y depuración de datos.	97
3.2.4.Ejemplo set de datos.	98
3.3.1.Resumen bases de datos de anticuerpos.	103
3.3.2.Resumen bases de datos de antígenos.	105
3.3.3.Resumen bases de datos de epítopes.	106
3.3.4.Aporte realizado a colección de anticuerpos.	107

3.3.5.Aporte realizado a colección de antígenos.	108
3.3.6.Aporte realizado a colección de epítopes.	108
3.3.7.Redundancia en los set de datos	109
3.3.8.Esquema representativo del sistema IMDb.	111
3.3.9.Ejemplo visualización interacción Ag-Ab.	113
4.2.1.Histograma términos GO función molecular auto antígenos.	121
4.2.2.Histograma términos GO proceso biológico auto antígenos.	122
4.2.3.Histograma términos GO componente celular auto antígenos.	123
4.2.4.Histograma dominios Pfam auto antígenos.	124
4.2.5.Histograma términos GO función molecular antígenos IMDb.	124
4.2.6.Histograma términos GO proceso biológico antígenos IMDb.	125
4.2.7.Histograma términos GO componente celular antígenos IMDb.	126
4.2.8.Histograma dominios Pfam antígenos IMDb.	127
4.2.9.Esquema representativo para sistema de clasificación propuesto	129
A.0.1Gráfico Ramachandran modelo A021	167
A.0.2Gráfico Ramachandran modelo A120	168
A.0.3Gráfico Ramachandran modelo uORF:IOH38079	169
A.0.4Performance entrenamiento promedio algoritmos	170
A.0.5Performance testeo promedio algoritmos	170
A.0.6Performance entrenamiento promedio datasets	171
A.0.7Performance testeo promedio datasets	171