

Índice general

1. Introducción	1
1.1. Marco Teórico	1
1.1.1. Sistema Inmune	1
1.1.1.1. Sistema inmunológico innato	2
1.1.1.2. Sistema inmunológico adaptativo	2
1.1.2. Interacción antígeno-anticuerpo	3
1.1.2.1. Anticuerpo	4
1.1.2.2. Antígeno	5
1.1.3. Autoinmunidad	7
1.1.4. Leucemia	8
1.1.5. Machine Learning y Deep Learning aplicado a Ingeniería de Proteínas	8
1.1.5.1. Machine Learning e Inteligencia Artificial	9
1.1.5.2. Estrategias de representación de proteínas	15
1.1.5.3. Grafos	18
1.1.5.4. Deep Learning	19
1.2. Estado del Arte	23
1.3. Justificación del problema	25
2. Hipótesis y objetivos	27
2.1. Hipótesis	27
2.2. Objetivo General	27
2.3. Objetivos Específicos	27
3. Metodología	28
3.1. Conjunto de datos	29
3.2. Enfoques de trabajo	29
3.2.1. Enfoque: Nodos	29
3.2.2. Enfoque: Grafos	31
3.2.3. Comparación entre enfoques	33
3.3. Entrenamiento de modelos	33
3.4. Ajuste de hiperparámetros	36
4. Resultados y Discusiones	37
4.1. Conjunto de datos	37

4.2. Predictor para complejos Antígeno-Anticuerpo	38
4.2.1. Rosetta	38
4.2.2. AlphaFold	41
4.3. Codificaciones	41
4.4. Cálculo de distancias	42
4.5. Generación de grafos	42
4.6. Resultados del entrenamiento	43
4.6.1. Enfoque en nodos	43
4.6.2. Enfoque en grafos	46
5. Conclusiones y futuros trabajos	51
Referencias	55

Índice de cuadros

3.2.1.Comparación entre ambos enfoques descritos para la generación de grafos.	33
4.6.1.Mejor resultado para el enfoque en nodos proveniente de la codificación fisicoquímica Alpha Structure.	44
4.6.2.Mejor resultado para el enfoque en nodos proveniente de la codificación fisicoquímica Alpha Structure.	47

Índice de figuras

- 1.1.1. **Comparación entre sistemas inmunes.** El sistema inmune innato esta compuesto por células del cuerpo que actúan como primera defensa ante una enfermedad, mientras que el sistema inmune adaptativo se encuentra conformado principalmente por los anticuerpos, donde su interacción con los antígenos permite inhibir el patógeno y generar memoria inmunológica. 3
- 1.1.2. **Representación estructural del anticuerpo.** Los anticuerpos son proteínas en forma de "Y" conformadas por dos cadenas ligeras idénticas (L) y dos cadenas pesadas idénticas (H), además de una región constante (C) y una región variable (V) por cada cadena. La región constante revela la clase o tipo de anticuerpo (IgA, IgD, IgE, IgG e IgM), mientras que la región variable otorga la especificidad al momento de interactuar con un antígeno en particular. Las dos cadenas se emparejan mediante enlaces disulfuro para así conformar un hetero dímero idéntico correspondiente al anticuerpo. 5
- 1.1.3. **Representación estructural de los epítomos.** Los antígenos poseen una o varias regiones proteicas denominadas epítomos, las cuales son reconocidas específicamente por los anticuerpos (región hipervariable). Hay dos tipos de epítomos, los epítomos continuos o lineales y los epítomos discontinuos o conformacionales. Estos últimos son los más comunes y se diferencian por la disposición que adoptan los residuos a lo largo de la cadena, en donde la posición de los aminoácidos varía a lo largo de la cadena a diferencia de los lineales. 6
- 1.1.4. **Representación de la interacción antígeno-anticuerpo.** Los antígenos tienen regiones proteicas denominadas epítomos, los cuales interactúan de forma *específica* con la región hipervariable de un anticuerpo. En el caso de la Figura 1.1.4, el antígeno posee 5 epítomos (Ep), en donde dos de ellos están interactuando con un anticuerpo en particular. 7

1.1.5.	Matriz de confusión general. La matriz de confusión es una herramienta muy útil para determinar la cantidad de clases que fueron predichas correctamente. Las filas de la matriz corresponden a los valores observables, mientras que las columnas a los valores predichos por el modelo resultante. La intersección de filas y columnas genera cuatro estimadores: Verdaderos Positivos (VP), Falsos Negativos (FN), Falsos Positivos (FP) y Verdaderos Negativos (VN). A partir de estos valores, se pueden determinar cinco variables o métricas de la matriz de confusión, como lo son: Exactitud (Accuracy), Precisión, Sensibilidad (Recall), Especificidad y el Coeficiente de Correlación de Matthews (MCC). Cada uno de estos estimadores permite evaluar que tan bien predice un modelo de predicción.	12
1.1.6.	Representación de la validación cruzada (Cross-validation). El conjunto de datos se divide en k grupos, donde un gran parte de ese conjunto se ocupa para entrenar, mientras que la restante de prueba. Este proceso se repite k veces, para que así cada grupo individual sea utilizado de prueba.	15
1.1.7.	Principales estrategias de representación de proteínas. La representación de proteínas es de vital importancia a la hora de desarrollar modelos predictivos o identificar patrones. Es posible dividirlos en tres grandes grupos. Numéricas, empleando imágenes, y aplicando estructuras de grafos. Dentro de las numéricas se encuentran las conocidas como One Hot Encoder, el uso de propiedades fisicoquímicas, la combinación de Digital signal processing y recientemente, el uso de transformadores basados en procesamiento de lenguaje natural. Por otro lado, el uso de imágenes puede provenir desde la estructura 3D o empleando técnicas de fingerprints. Finalmente, las estrategias de grafos facilitan la aplicación de modelamiento matemático discreto con el fin de identificar patrones o características similares.	17
1.1.8.	Ejemplo de un grafo. Ejemplo básico de un grafo, en donde los nodos corresponden a ciudades de España y las aristas a la distancia (Km) que las separa. Se puede apreciar que hay nodos que solo poseen una relación (simple), como la ciudad de Cádiz, mientras que hay otros nodos que mantienen más de una relación (múltiple), como la ciudad de Granada y Sevilla (Francisco J. Gil Gala, 2015).	19
1.1.9.	Esquema de una red neuronal. La red neuronal más básica se compone por tres capas: una capa de entrada con los datos sin procesar. Una capa oculta, la cual procesa, modifica y transfiere la información de una capa a otra. Este proceso se conoce como aprendizaje. Y por último, una capa de salida, la cual corresponde al modelo resultado del proceso de aprendizaje por parte de las capas ocultas.	20

1.1.10	Arquitectura base para una Convolutional Neural Network (Mayank Mishra, 2020).	21
1.1.11	Arquitectura base para una Graph Neural Network (George Mohler, 2018).	23
3.0.1	Metodología. Actividades a realizar para ambos enfoques a la hora de trabajar con grafos y deep learning.	28
3.2.1	Metodología. El enfoque de los nodos consiste en armar un grafo que represente toda la red de proteínas, en donde los nodos corresponden a la secuencia codificada de las proteínas, mientras que las aristas a los niveles de interacción entre ellas.	30
3.2.2	Metodología. El enfoque de los grafos consiste en armar un grafo por cada interacción antígeno-anticuerpo, en donde los nodos corresponde a los residuos de las proteínas, mientras que las aristas a las distancias euclidianas que los separan.	32
3.3.1	Generación de batchs. PyTorch apila las matrices de adyacencias en forma diagonal (se crea un grafo gigante que contiene múltiples subgrafos), mientras que los nodos y los targets se concatenan en la dimensión del nodo.	34
3.3.2	Arquitectura para una red neuronal. Los grafos en su forma matricial ingresan a la capa input de la red neuronal. Luego, en las capas ocultas se generan embedding, los cuales resumen las características de la información traspasada entre nodos o las estructuras, para posteriormente pasar a una función de activación. Una vez realizada la predicción, se calcula la función de pérdida para estimar el error de predicción. Finalmente, se calculan los gradientes para actualizar los pesos de la red en un proceso denominado Backpropagation. Este proceso se repite hasta que la función de pérdida es reducida.	35
4.2.1	Pipeline de Rosetta para complejos Antígeno-Anticuerpo.	39
4.2.2	Pipeline actualizado para la predicción antígeno-anticuerpo.	40
4.6.1	Arquitectura Enfoque en Nodos. Arquitectura aplicada para el entrenamiento de GNN con un enfoque en nodos, la cual se encuentra conformada por capas de convolución, funciones de activación ReLU, capas de Dropout para evaluar sobreajuste y una capa de Softmax final para generar el clasificador.	43
4.6.2	Mejor rendimiento para el accuracy a lo largo de las 100 epochs, tomando en consideración una arquitectura basada en capas de convolución, funciones ReLU, capas de Dropout, una capa Softmax y una codificación fisicoquímica Alpha Structure.	45

4.6.3. Arquitectura Enfoque en Grafos. Arquitectura para el enfoque en grafos basada en capas de convolución, ya sea GNN o GCN, funciones de activación ReLU, una capa global mean pool para obtener un valor promedio de los embeddings generados, una capa de Dropout y una capa Linear para generar la predicción.	47
4.6.4. Rendimiento del accuracy por parte del enfoque de los grafos a lo largo de cada epoch.	48
4.6.5. Gráfico resultante de Prosa para un complejo predicho por AlphaFold.	49
4.6.6. Gráfico resultante de Prosa para un complejo predicho por Rosetta.	50
5.0.1. Resultado para un modelo de AlphaFold. Resultado de Prosa para un complejo de AlphaFold, alcanzando un puntaje de -5.63. .	53
5.0.2. Resultado para un modelo de Rosetta. Resultado de Prosa para un complejo de Rosetta, alcanzando un puntaje de -11.39. . .	54