UNIVERSIDAD NACIONAL DE SAN AGUSTÍN
ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERIA DE PRODUCCIÓN Y SERVICIOS

Detección *in Silico* de Neoantígenos Utilizando *Transformers* y *Transfer Learning* en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer

Tesis presentada por el Magister:
Vicente Enrique Machaca Arceda

Para optar el Grado de:
Doctor en Ciencia de la Computación

Asesor:
Prof. Dr. Cristian Lopez Del Alamo

Arequipa - Perú
2022

Declaración de autenticidad

I, Yo Vicente Machaca Arceda, declaro que la tésis titulada, 'Detección de neo antígenos utilizando aprendizaje profundo en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del Cáncer' y el trabajo presentado en este son de mi propiedad intelectual y confirmo que:

- Este trabajo fue desarrollado durante mi candidatura a grado de doctor de esta universidad.
- Ninguna parte de esta tésis ha sido presentado para otro grado de esta universidad o cualquier otra institución.
- Cuando cito a otros autores, las fuentes has sido brindadas y con excepción de estas citas, mi trabajo es de mi autoría.
- He agradecido las principales fuentes de ayuda.
- En caso de que mi tesis haya sido desarrollado con un equipo de trabajo, yo he sido claro y he detallada la parte exacta de mi autoría.

Firma: _____

Fecha: _____

"Con fe, disciplina y desinteresada devoción al deber, no hay nada que merezca la pena que no puedas lograr."

Muhammad Ali Jinnah

Dedico este trabajo a mi esposa Pamela Laguna Laura, quien me ha acompañado durante todo este proceso, me ha motivado y sobre todo me ha dado su amor, que me ha ayudado a prevalecer y siempre seguir adelante. De igual forma, a mis padres Vicente Machaca Chino y Victoria Arceda Arenas, de ellos he aprendido el valor de la disciplina, la fuerza por emprender y la importancia de los valores sin importar las circunstancias; gracias a ellos he logrado cumplir mis objetivos.

Resumen

La detección de neo antígenos, es la fase más importante para el desarrollo de vacunas personalizadas contra el cáncer. El proceso para identificar neo antígenos, es complejo y existen varias sub fases como: secuenciamiento, alineamiento, detección de mutaciones, identificación de péptidos, *peptide-MHC binding*, *peptide-MHC presentation* y la interacción pMHC-TCR. La mayoría de publicaciones, se ha centrado en el problema de *peptide-MHC binding*, y han logrado buenos resultados, pero menos del 5 % de los péptidos identificados, llegan a la membrana de las células y logran presentarse ante las células T. En este contexto, surge un nuevo problema llamado *peptide-MHC presentaion*, enfocado en predecir que péptidos logran enlazarse a la molécula MHC y permancer unida a ellas hasta llegar a la membrana. Gracias a la tecnología de *Mass spectrometry*, se está secuenciando cada vez más muestras de compuestos pMHC de la membrana de las células; de esta forma se están construyendo nuevas bases de datos que puedan dar solución al problema de *peptide-MHC presentation*.

Las redes neuronales *Transformers* han revolucionado el campo de NLP, y se han abierto a muchas otras aplicaciones. Luego, las redes BERT, como una actualización a las *Transformer*, han sido aplicadas en problemas de interacción de proteínas. Pero, la interacción entre un péptido y la molécula, es una interacción entre proteínas; de esta forma han surgido trabajos que utilizan redes BERT para predecir la afinidad *peptide-MHC*. De esta forma, en esta tesis, se propone el uso de redes BERT para dar solución al problema de *peptide-MHC presentation*. Además en la propuesta se utilizó varias muestras de *Mass spectrometry*, recolectada de bases de datos públicas y trabajos similares. Finalmente, también se ha aplicado *transfer learning*, del modelo TAPE y ESM-1b (modelos entrenados con millones de secuencias de aminoácidos).

Esta tesis, presenta dos contribuciones: primero, se ha realizado una revisión sistemática de la literatura referente a la detección de neo antígenos y enfocada en estudiar los métodos basados en *deep learning*; segundo, se ha desarrollado un nuevo método basado en redes BERT y *transfer learning* para dar solución al problema de *peptide-MHC presentaion*.

Índice general

Índice de figuras

Índice de tablas

Abreviaciones

ANN	Artificial Neural Network
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
bp	Base pair in DNA
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DNA	Deoxyribonucleic Acid
GNN	Graph Neural Netowrk
G-BERT	Graph Bidirectional Encoder Representations from Transformers
HLA	Human Leukocyte Antigens
MCC	Matthews Correlation Coefficient
MHC-I	Major Histocompatibility Complex Class I
MHC-II	Major Histocompatibility Complex Class II
MHC-III	Major Histocompatibility Complex Class III
mRNA	Messenger Ribonucleic Acid
NLP	Natural Language Processing
pMHC	Peptide-MHC ligand
pMHC-TCR	pMHC T-cell receptor ligand
RNA	Ribonucleic Acid
RoBERTa	Optimized BERT
RSL	Revisión Sistemática de la Literatura
tRNA	Transfer Ribonucleic Acid
TCR	T-cell receptor

Capítulo 1

Introducción

1.1. Contexto y Motivación

El cáncer representa el desafío de salud global más significativo (Siegel et al., 2023). Además, según el Instituto de Investigación del Cáncer del Reino Unido, se registraron más de 18 millones de nuevos casos y 10 millones de muertes en 2020 (UK, 2023b). Además, se predice que habrá alrededor de 28 millones de nuevos casos anualmente para alrededor de 2040 si la incidencia se mantiene estable y el crecimiento de la población y el envejecimiento continúan según las tendencias recientes (UK, 2023a). Esto representa un aumento del 54.9 % desde 2020, con un aumento esperado mayor en hombres (60.6 %) que en mujeres (48.8 %).

En este contexto, se sabe que los métodos tradicionales basados en cirugía, radioterapia y quimioterapia tienen baja eficacia y efectos secundarios adversos (Peng et al., 2019). Por lo tanto, ha surgido el desarrollo de la inmunoterapia contra el cáncer, con el objetivo de estimular el sistema inmunológico del paciente (Borden et al., 2022). Existen tratamientos como vacunas personalizadas, terapias con linfocitos T adoptivos e inhibidores de puntos de control inmunológico. De entre estos, las vacunas basadas en neoantígenos han mostrado un gran potencial al potenciar las respuestas de los linfocitos T y se consideran las más propensas a tener éxito (Borden et al., 2022). Además, los neoantígenos se utilizan en la terapia de bloqueo de puntos de control inmunológico. Los neoantígenos se consideran biomarcadores predictivos y objetivos para el tratamiento sinérgico en la inmunoterapia contra el cáncer (Fang et al., 2022).

El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo que depende de la detección precisa de neoantígenos (ver Figura 1.1). Estos neoantígenos son péptidos que se encuentran exclusivamente en las células cancerosas. El objetivo de un tratamiento

basado en vacunas personalizadas es entrenar a los linfocitos (células T) del paciente para que reconozcan estos neoantígenos y activen el sistema inmunológico (Mattos et al., 2020; Peng et al., 2019). El proceso se resume en la Figura 1.1(b) y consta de las siguientes fases:

1. Obtener muestras de tejidos cancerosos y sanos. Ambos tejidos se secuencian para obtener ADN y/o ARN. Algunos enfoques incluyen información del *immunopeptidome* obtenida mediante *Mass Spectrometry* (MS).

2. En la etapa *in-silico*, se realiza el alineamiento de secuencias, se desarrolla un proceso de llamada de variantes para detectar variaciones y/o mutaciones, y se anotan estas variantes (posible detección de neoantígenos). Hay disponibles varias herramientas con buen rendimiento para esta etapa.

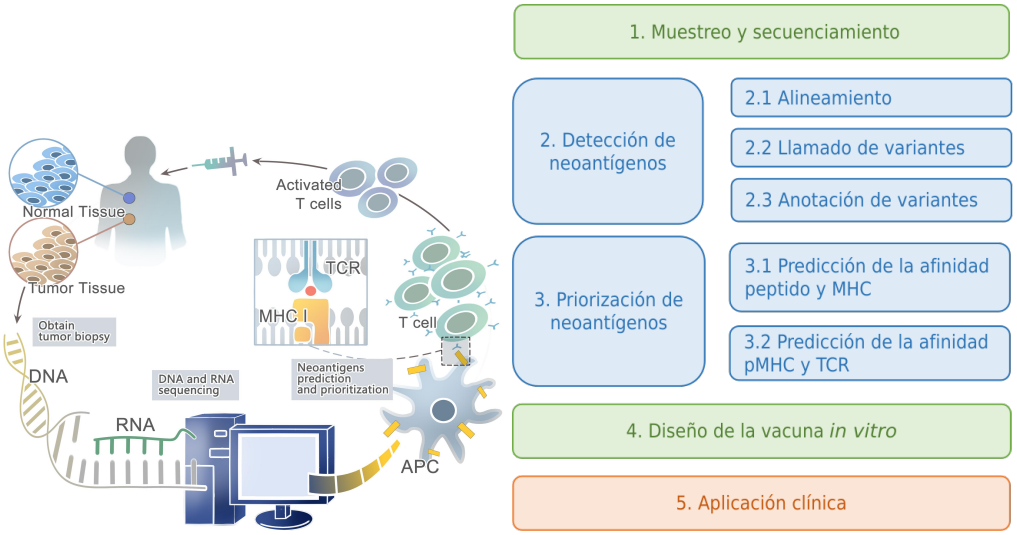
3. En esta etapa *in-silico*, se priorizan los neoantígenos. Este paso es crucial y ha recibido una atención significativa en la investigación en los últimos años debido a su complejidad y la baja efectividad de los enfoques actuales. Aquí, se evalúa la afinidad de los candidatos neoantígenos (péptidos) de la etapa anterior con el *Major Histocompatibility Complex* (MHC), conocido como la unión pMHC. Luego, se evalúa la afinidad de pMHC para unirse al *T-cell Receptor* (TCR). Al final de esta etapa, se obtienen los neoantígenos.

4. En la etapa *in-vitro*, en el laboratorio se inducen las células T del paciente para que reconozcan los neoantígenos. En este punto, se desarrollan las vacunas. Esta etapa la llevan a cabo biotecnólogos y biólogos.

5. Finalmente, el oncólogo realiza una evaluación clínica de la vacuna.

La detección *in-silico* de neoantígenos se basa en las etapas segunda y tercera representadas en la Figura 1.1(a). En este contexto, debido a la complejidad del proceso y la variedad de métodos disponibles, se han desarrollado herramientas de software y flujos de trabajo para agilizar el uso de estas herramientas. Además, los Transformers han marcado el comienzo de una nueva era en la inteligencia artificial, demostrando logros destacados en una variedad de tareas de procesamiento del lenguaje natural (Patwardhan et al., 2023). Estos modelos también han encontrado aplicación en la detección de neoantígenos, especialmente en la tercera etapa de la Figura 1.1(b). Se han propuesto modelos BERT y redes de aprendizaje profundo con mecanismos de atención para predecir la unión péptido-MHC y pMHC-TCR obteniendo resultados prometedores. Sin embargo, aún existe mucho camino por recorrer y con el incremento constante de muestras de ADN/proteínas, sumado a los nuevos mecanismos para entrenar modelos

Transformers con billones de parametros, se espera lograr avances significativos en este campo de estudio.



(a) Proceso de desarrollo (b) Pipeline

FIGURA 1.1: Marco de desarrollo para la creación de vacunas personalizadas contra el cáncer basadas en neoantígenos. (a) proporciona una visión general de cada etapa (Han et al., 2020). (b) una visión general de cada fase con un énfasis en el desarrollo *in-silico*.

1.2. Problema

Los neoantígenos son peptidos mutados específicos de tumores y son considerados los principales causantes de una respuesta inmune (Borden et al., 2022; Chen et al., 2021b; Gopanenko et al., 2020). Es así que surgen varios esfuerzos e investigación en la Inmunoterapia del cáncer, concentradas en el estudio y detección de neoantígenos. Es así, que el desarrollo de vacunas personalizadas basadas en neoantígenos es considerado uno de los métodos con mayor probabilidad de éxito (Borden et al., 2022). Incluso varias compañías como BioNTech, Genocea Biosciences, Neon Therapeutics y Gritstone Oncology realizan investigación y ofrecen el servicio de generar vacunas personalizadas a pacientes de cáncer.

Además, el MHC representa un factor clave, en la detección de neoantígenos, al ser el encargado de unirse al neoantígeno y presentarlo a la superficie de la célula. Debido a esto, este trabajo en enfoca en el desarrollo de un método para la predicción del enlace entre neoantígenos y MHC (*pMHC binding*), esto corresponde a la Fase 3.1 de la Figura 1.1(b) dentro del *pipeline* de detección de neoantígenos. Existen dos tipos: el MHC clase I (MHC-I) y MHC clase II (MHC-II), ambos presentan péptidos en la superficie celular a las células T CD8+ y CD4+, respectivamente (Janeway Jr, 1997;

Abualrous et al., 2021). En detalle, el ciclo de vida de los neoantígenos que se unen a MHC-I se puede resumir de la siguiente manera. Primero, una proteína cancerígena se degrada en péptidos en el citoplasma. Luego, los péptidos se unen al MHC (*pMHC binding*). Después, este compuesto sigue un camino hasta llegar a la membrana celular (*pMHC presentation*). Finalmente, el pMHC es reconocido por el TCR, desencadenando el sistema inmunológico (Janeway Jr, 1997; Wieczorek et al., 2017; Gasser et al., 2021). Por lo tanto, el *pMHC binding* es un paso muy importante para la inmunidad celular, y la predicción y comprensión de esta unión tienen un valioso potencial. Lamentablemente, la mayoría de las ligandos *pMHC* no llegan a la membrana celular (Mattos et al., 2020).

Adicionalmente, las proteínas MHC están codificadas por genes altamente polimórficos, llamados *Antígenos Leucocitarios Humanos* (HLA); la considerable naturaleza polimórfica de los genes MHC proporciona una variación sustancial en la unión con los neoantígenos, lo que influye en el conjunto de neoantígenos presentados a las células T (Abualrous et al., 2021). En consecuencia, los métodos propuestos se categorizan como *pan-specific* y *allele-specific*. Los métodos *allele-specific* (Rammensee et al., 1999; Reche et al., 2002; Kim et al., 2009; Nielsen and Andreatta, 2016; Vang and Xie, 2017; Shao et al., 2020; Bravi et al., 2021) entrenan un modelo para cada *allele* del MHC; mientras que los métodos *pan-specific* (Hu et al., 2019; Liu et al., 2019b; Wu et al., 2019; Phloyphisut et al., 2019; O'Donnell et al., 2018, 2020; Reynisson et al., 2020a; Venkatesh et al., 2020; Ye et al., 2021; Mei et al., 2021; Chu et al., 2022; Zhang et al., 2022b; Mei et al., 2021; Hu et al., 2019; Gfeller et al., 2023) entrenan un modelo global que toma péptidos (neoantígenos) y MHC como entradas. Además, la naturaleza polimórfica del MHC eleva bastante la complejidad de este problema, se cree que existen las 10000 diferentes MHC *alleles* (Abelin et al., 2017), esto complica mucho la detección de neo antígenos. Por lo tanto, los métodos *pan-specific* surgen con una alta posibilidad de futuras aplicaciones.

Lamentablemente, a pesar de varios esfuerzos en el desarrollo de métodos para la detección de neoantígenos, menos del 5 % de neoantígenos detectados activan el sistema immune (Mattos et al., 2020; Mill et al., 2022; Bulik-Sullivan et al., 2019; Bassani-Sternberg et al., 2015; Yadav et al., 2014). Según los autores de los métodos, las razones son:

1. La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de *Mass Spectrometry* (MS) (Kim et al., 2018). Por ejemplo, la mayoria de propuestas no utiliza datos de MS; en la actualidad, existe una creciente información de estos datos y se estan aplicado a varios campos de la Bioinformática.

2. Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC) (etapa 3.1 de la Figura 1.1(b)). La mayoria de aplicaciones, se basa en el

uso de MHCFlurry (O'Donnell et al., 2020) y NetMHCpan4.1 (Reynisson et al., 2020a). Sin embargo, actualmente, se cuenta con herramientas de mejor desempeño basado en *transformers* (Arceda, 2023). Esta tesis, se enfoca en resolver este problema.

3. Para la etapa 3.2 de la Figura 1.1(b), los autores no consideran la predicción del enlace pMHC al TCR (pMHC-TCR), varios autores consideran incluir esta tarea en trabajos futuros (Rubinsteyn et al., 2018).

4. Finalmente, no utilizar información de eventos de *alternative splicing*, variaciones estructurales en el ADN y las mutaciones de fusión de genes, está información esta fuertemente relacionada con varios tipos de cancer (Wood et al., 2020).

En conclusión, la detección de neoantígenos es un desafío que consta de múltiples etapas, y las herramientas actuales en el estado del arte presentan un rendimiento insuficiente. Uno de los factores clave detrás de este bajo rendimiento está relacionado con la predicción del enlace pMHC. Por esta razón, esta tesis se centra en abordar este problema mediante la propuesta de un método basado en Transformers para la predicción del enlace pMHC.

1.2.1. Formulación del problema

El presente estudio se centra en el problema de predicción del enlace pMHC-I (*pMHC binding prediction*). Esto representa un problema de clasificación binaria que toma como entrada la secuencia de aminoacidos de un péptido y el MHC. Un péptido podría representarse como: $p = \{A, ..., Q\}$ y una representación similar para el MHC sería: $q = \{A, N, ..., G\}$. Finalmente, necesitamos conocer la probabilidad de afinidad entre p y q. Si esta probabilidad es lo suficientemente alta, es posible que el péptido se enlace al MHC y por lo tanto, el péptido p en cuestión, sería un excelente candidato a neoantígeno.

1.3. Objetivos

1.3.1. Objetivo General

Proponer un método *in Silico* basado en *Transformers* y *Transfer Learning* para la detección de neo antígenos, enfocados en la predicción del enlace pMHC.

1.3.2. Objetivos específicos

(a) Analizar los métodos que utilizan Transformers para la predicción del enlace pMHC en el contexto de detección de neoantígenos.

(b) Analizar los modelos basados en Transformers TAPE, ProtBert-BFD, y EMS2 pre-entredados para diversas tareas en Proteómica y de los cuáles se puede aplicar *Transfer Learning*.

(c) Analizar técnicas como *Gradient Accumulation Steps* (GAS) y métodos de *layer-freezing* para entrenar modelos Transformers con millones de parámetros.

(d) Implementar *fine-tuning* a los modelos TAPE, ProtBert-BFD, y EMS2 para la tarea de predicción del enlace pMHC, aplicando GAS y una metodología de *layer-freezing*.

(e) Comparar los modelos de mejor desempeño con las herramientas del estado del arte como: NetMHCpan4.1, MHCFlurry2.0, Anthem, ACME y MixMHCpred2.2.

1.4. Contribuciones

Las principales contribuciones de este trabajo son:

(a) Se ha desarrollado una revisión sistemática de la literatura referente a los métodos basados en *Transformers* para la detección de neoantígenos. Esto ha generado dos publicaciones tituladas: *"Deep Learning and Transformers in MHC-Peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review"* (Machaca et al., 2023) y *"Transformers Meets Neoantigen Detection: A Systematic Literature Review"*.

(b) Se ha implementado *fine-tuning* a seis modelos de Transformers para la predicción del enlace pMHC; además, se ha evaluado el uso de GAS y una metodología de *layer-freezing*. Los resultados fueron publicados en: *"Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context"* (Arceda, 2023) y *"Fine-tuning Transformers for Peptide-MHC Class I Binding Prediction"*.

(c) Finalmente, se comparó los métodos propuestos con herramientas del estado del arte como: NetMHCpan4.1, MHCFlurry2.0, Anthem, ACME y MixMHCpred2.2. Los métodos propuestos obtuvieron los mejores resultados en *accuracy*, *Area Under the Curve* (AUC), *recall*, *f1-score* y *Matthews Correlation Coefficient* (MCC).

1.5. Organización del Trabajo

En el Capítulo 2 se presentan los conceptos básicos sobre Bioinformática e inmunoterapia del Cáncer, también son abordados los temas sobre *deep learning* y redes neuronales Transformers.

Luego, en el Capítulo 3 se describen los trabajos relacionados a la presente tesis. Debido a la gran cantidad de publicaciones, solo se ha considerado trabajos desde el 2018 y que hacen uso de Transformers o redes reuronales con mecanismos de atención.

El Capitulo 4, presenta la propuesta de la tesis. Esta se basa en un método para desarrollar *fine-tuning* a Transformers pre-entrenadas para diversas tareas de Proteómica.

Luego, en el Capítulo 5, se presentan los resultados de la investigación. Además, se presenta una comparación con los métodos del estado del arte.

Finalmente, en el Capítulo 6 son expuestos las conclusiones del presente trabajo así como también las direcciones para continuar con el mismo en la sección de trabajos futuros.

Capítulo 2

Marco Conceptual

El proyecto pertenece al área de Bioinformática y específicamente a la Inmunoinformática, en este contexto el marco teórico detalla conceptos de Biología Molecular (ADN, ARN y proteínas), Inmunología y Ciencias de la Computación.

2.1. Bioinformática y Biología Molecular

En esta sección, describiremos los principales conceptos referentes a Biología Molecular que serán considerados en la propuesta de la tesis.

2.1.1. Bioinformática

Según Luscombe et al. (2001), la Bioinformática involucra la tecnología que utiliza las computadoras para el almacenamiento, manipulación y distribución de información relacionada a la Biología Molecular como DNA, RNA y proteínas. También podemos considerar que la Bioinformática se enfoca al análisis de secuencias, estructuras y funciones de los genes y proteínas; algunas veces también puede ser llamado Computación Molecular Biológica (Xiong, 2006).

2.1.1.1. DNA, RNA y Proteínas

Deoxyribonucleic Acid (DNA) es una molécula dentro de las células que contiene información genética responsable del desarrollo y función del organismo (NCI, 2022). Gran parte del DNA se sitúa dentro del núcleo de las células (en organismos Eucariotes). Por ejemplo en la Figura 2.1, vemos como el DNA, forma parte de los cromosomas y estos

a su vez están en el núcleo. Luego, podemos notar, que los genes representan segmentos del DNA. Finalmente, en la Figura 2.1, notamos las bases nitrogenadas que componen el DNA: *Guanine*, *Cytosine*, *Adenine* y *Thymine*; normalmente, estas bases serán representadas por las letras: G, C, A, T respectivamente.

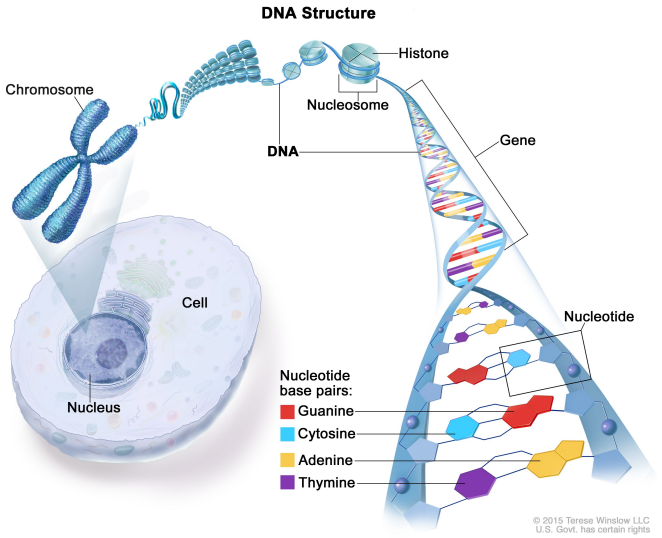


FIGURA 2.1: Localización y estructura del DNA. Fuente: NCI (2022).

Durante el ciclo de vida de la célula, ocurre un proceso llamado Transcripción (ver Figura 2.2), en este proceso se generan cadenas de *Ribonucleic Acid* (RNA) a partir de la cadena de DNA (NCI, 2022). Durante este proceso la base nitrogenada *Thymine* (T) es reemplazada por *Uracil* (U). El proceso mencionado, ocurre dentro del núcleo de la célula y en esta etapa el RNA es llamado *messenger RNA* (mRNA). Una vez el mRNA sale del núcleo, es transportado por *transfer RNA* (tRNA) hacia los Ribosomas (ver Figura 2.2). En está, última etapa ocurre la Traducción, cada grupo de tres bases nitrogenadas (codones) se convierten en un aminoácido diferente, luego estos aminoácidos forman cadenas polipeptídicas y estas a su vez forman las proteínas; normalmente, cada gen genera una proteína (Xiong, 2006; NCI, 2022).

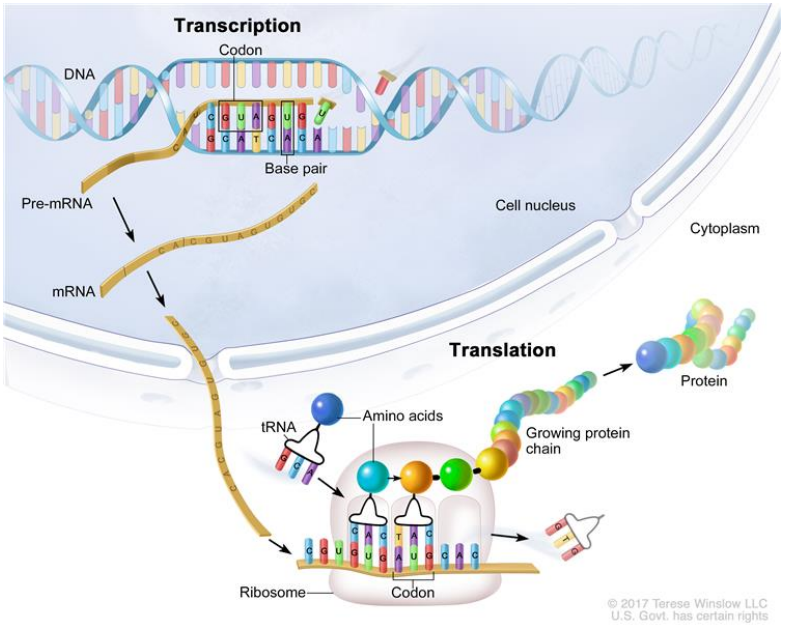


FIGURA 2.2: Transcripción y traducción. Fuente: NCI (2020).

Durante el proceso de Traducción, puede ocurrir un fenómeno llamado *Alternative Splicing*. Por ejemplo , en la Figura 2.3, notamos como un gen puede generar tres proteínas distintas, cada una con funciones distintas. Este fenómenos, complica bastante el análisis de DNA.

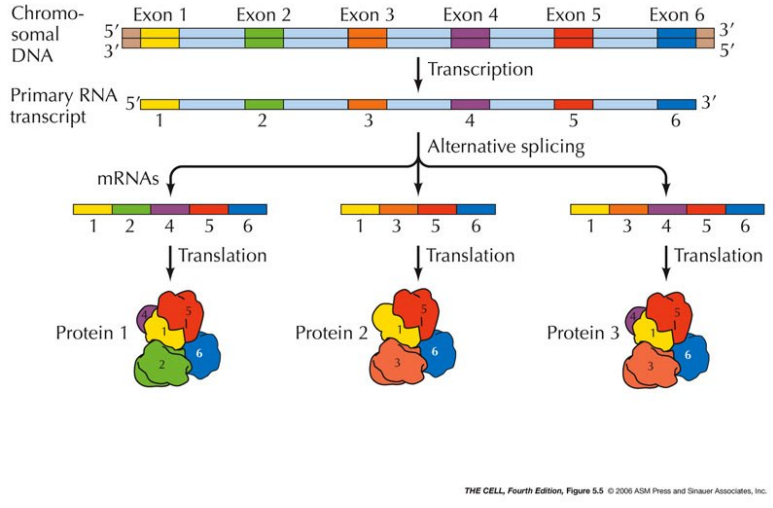


FIGURA 2.3: *Alternative Splicing*. Fuente: NCI (2020).

2.1.2. Mutaciones

Las mutaciones también llamadas variaciones, representan cualquier cambio en la secuencia de DNA, estos pueden ocurrir durante la división celular o por la exposición a agentes químicos o radioactivos. Estas mutaciones pueden ser beneficiosas, dañinas (cuando afectan la generación de proteínas) o no tener algún efecto (NCI, 2022). Varios tipos de Cáncer son ocasionados por estas mutaciones (Borden et al., 2022; Chen et al., 2021b; Mattos et al., 2020).

Según el tipo de célula afectada, tenemos: mutaciones somáticas y mutaciones *germline* (una mutación en estas células puede ser heredada a la descendencia) (Clancy, 2008). Según (Xu, 2018), las variaciones genómicas pueden clasificarse en tres grupos: *Single-Nucleotide Variant* (SNV), inserciones y eliminaciones (INDELS) y *Structural Variation* (SV). Una mutación se considera SNV cuando las variaciones afectan a menos de 10 bases.

En la Figura 2.4, presentamos ejemplos de SNV. Por ejemplo, las sustituciones pueden afectar la generación de un aminoácido, pero las inserciones o eliminaciones pueden afectar en cadena la generación de varios aminoácidos, a este tipo de fenómeno se le conoce como *frameshit mutation* (Xu, 2018).

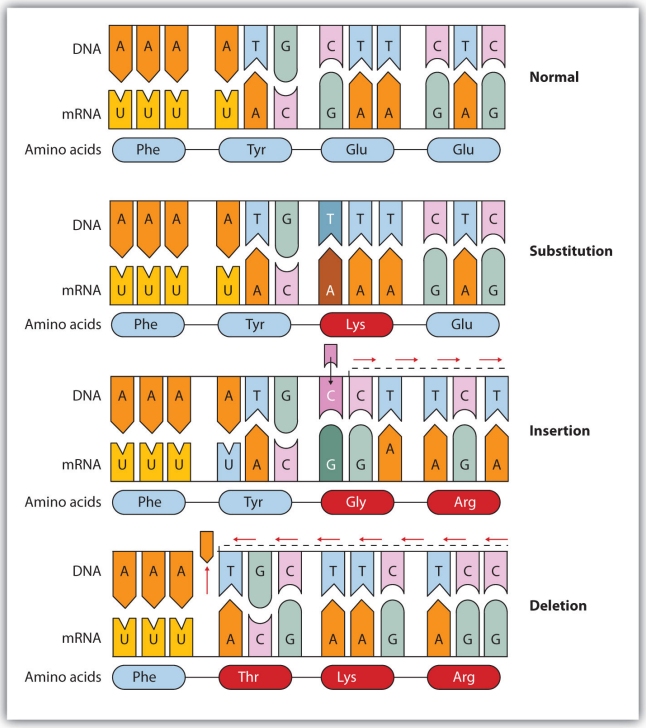


FIGURA 2.4: Ejemplos de SNV en el DNA. Fuente: Socratic.org (2022)

En la Figura 2.5, mostramos algunos tipos de SV. En este caso, también se pueden presentar INDELS, *Tanden duplication*, inversiones, traslocaciones y *Copy Number Variants* (CNV). Los CNVs, representan fuertes candidatos para ser biomarcadores de varios tipos de Cáncer (Pan et al., 2019; Lucito et al., 2007). Otra mutación importante, es referente a la fusión de genes, en estos casos dos o más genes se fusionan y forman una proteína completamente diferente, este tipo de mutación también está fuertemente relacionado a varios tipos de Cáncer (Kerbs et al., 2022; Kim and Zhou, 2019; Heyer and Blackburn, 2020).

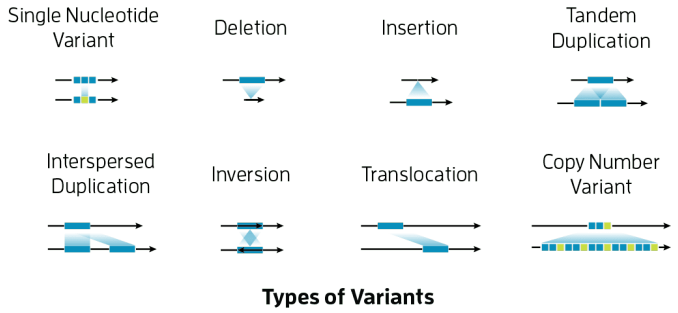


FIGURA 2.5: Ejemplos de variaciones en el DNA. Fuente: PacBio (2021)

2.2. Sistema inmunitario

El sistema inmunitario hace referencia al conjunto de células y procesos químicos que tiene como función protegernos de agentes extraños como: microbios, bacterias, células de Cáncer, toxinas, etc. Marshall et al. (2018). En esta sección, se explicará de forma breve el comportamiento del sistema inmunitario frente cuando un agente extraño (antígeno) ingresa al cuerpo humano.

2.2.1. Células T y APC

Las células T también llamadas linfocitos T, se forman a partir de la médula ósea y son los encargados de eliminar agentes extraños (antígenos) NCI (2022). Estas células están compuestas por un T-cell Receptor (TCR), que es el encargado de reconocer y enlazar a los antígenos. Luego, algunas células T, requieren de la acción de los *Antigen Presenting Cells* (APC), estás células APC son: células dentríticas, macrofagos, células B, fibroblastos y células epiteliales. Normalmente, los APC devoran los antígenos y luego los presentan a las células T para su eliminación (Marshall et al., 2018).

2.2.2. MHC I y II

Major Histocompatibility Complex (MHC) I y II, son proteínas que desempeñan un rol importante en el sistema inmunitario. Ambas proteínas tienen la función de presentar péptidos (antígenos) en la superficie de las células, para que sean reconocidas por la células T (Abualrous et al., 2021). MHC-I se encarga de la presentación de las células con núcleo, mientras que MHC-II, de las células APC.

El proceso de presentación de los antígenos por MHC-I es el siguiente (Figura 2.6): la proteína foránea es degradado por el proteasoma y se producen péptidos (posibles antígenos), luego estos péptidos son transportados al Endoplasmic Reticulum (ER) con la ayuda de *Transporter associated Antigen Processing* (TAP), luego es migrado al aparato de Golgi para ser presentado en la superficie de la célula y es enlazado a la proteína MHC-I, una vez en la superficie, el antígeno puede ser reconocido por las células CD8+T (Zhang et al., 2019).

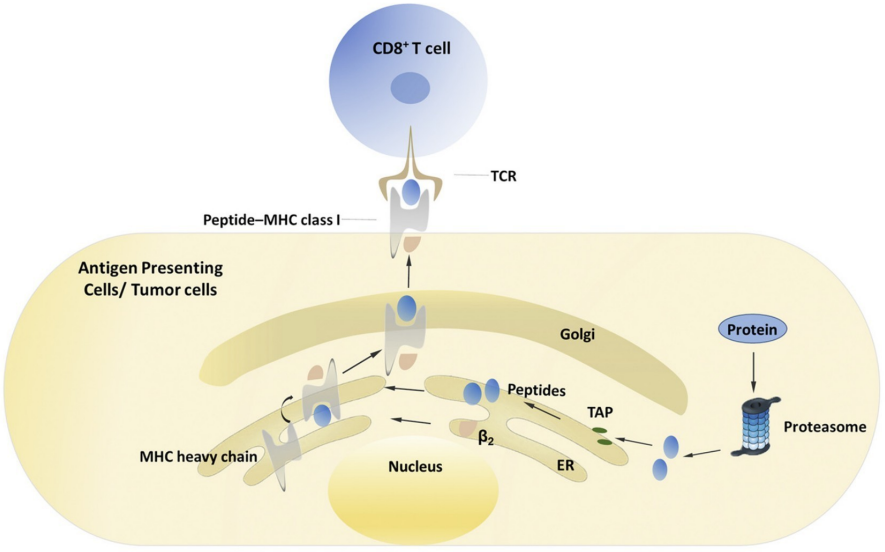


FIGURA 2.6: Presentación de antígenos por MHC-I. Fuente: Zhang et al. (2019)

Para el caso de MHC-II, es un proceso similar (Figura 2.7): primero, los patógenos son devorados por fagocitosis, los péptidos asociados a MHC-II son producidos en el Endoplasmic Reticulum (ER), para luego ser trasladados al aparato de Golgi, y luego ser transportados a la superficie de las células una vez enlazadas con MHC-II, finalmente, son reconocidas por las células CD4+T (Zhang et al., 2019).

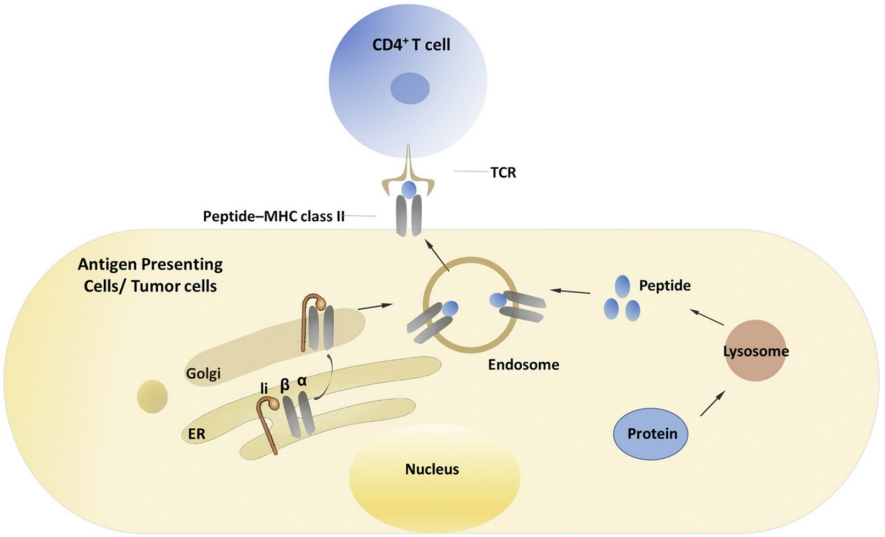


FIGURA 2.7: Presentación de antígenos por MHC-II. Fuente: Zhang et al. (2019)

2.2.3. Neo antígenos

Es una proteína que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA. Los neo antígenos cumplen un rol importante al estimular una respuesta inmune en contra de células de Cáncer. En la actualiadad, se estudia su uso en el desarrollo de vacunas contra el Cáncer NCI (2022). Una característica importante de los neo antígenos, es que solo están presentes en células tumorales y no en células sanas, debido a eso son considerados factores clave en la inmunoterapia del Cáncer Borden et al. (2022). En la actualidad hay varios métodos para detectar a predecir neo antígenos, pero solo una pequeña porción de ellos logran estimular al sistema inmune Chen et al. (2021b); Hao et al. (2021).

Este proceso para la detección de neo antígenos, generalmente consiste en: (1) extracción del tejido tumoral, (2) identificación de mutaciones, (3) detección de neo antígenos y predicción de inmunogenicidad, (4) desarrollo de experimentos in vitro y (5) desarrollo de la vacuna (Mattos et al., 2020; Peng et al., 2019) (ver Figura 2.8).

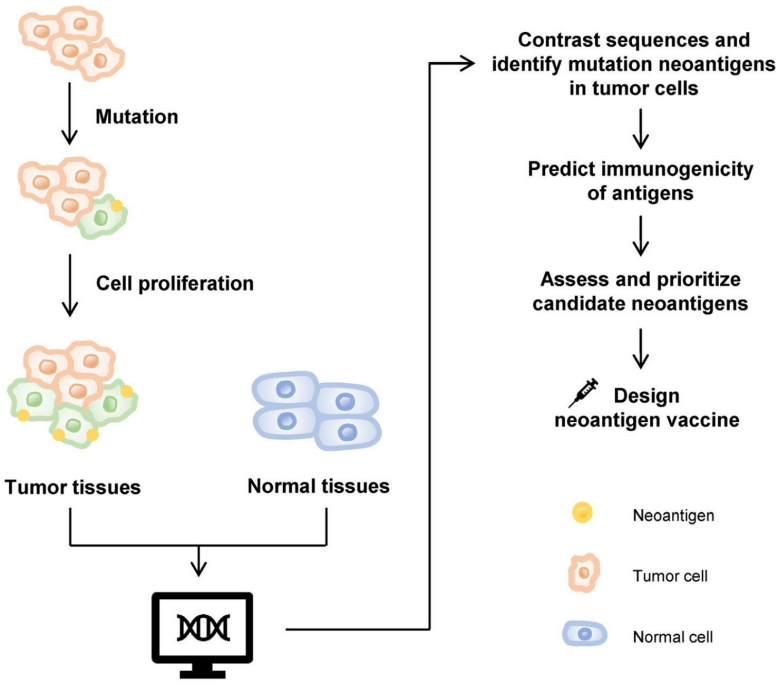


FIGURA 2.8: Proceso para la detección de neo antígenos y generación de vacunas personalizadas. Fuente: (Mattos et al., 2020)

2.3. *Machine Learning*

Machine Learning (ML) es una categoría de algoritmos computacionales capaces de emular algunas acciones inteligentes. Es el resultado de varias disciplinas como: inteligencia artificial, probabilidad, estadística, ciencia de la computación, teoría de la computación, psicología y filosofía (El Naqa and Murphy, 2022). *Machine Learning* tiene varias definiciones, pero una de las mas acertadas, según Samuel (1967): "Campo de estudio que brinda a las computadoras la habilidad de aprender sin haber sido explicitamente programado".

2.3.1. Algoritmos de aprendizaje

Un algoritmo de aprendizaje o *machine learning algorithm*, es aquel algoritmo que no debe ser programado explícitamente, este aprende de la experiencia, a partir de datos (Goodfellow et al., 2016). Según Mitchell (1997): "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". La traducción a español indicaría: "Un programa de computadora puede aprender de una experiencia E, para una tarea T y con una métrica de desempeño P, si el desempeño de

la tarea T, medido con P, mejorar con la experiencia E". Esto, nos da a entender que un programa de computadora puede aprender si mejora su desempeño según aumente su experiencia o datos.

2.3.1.1. La tarea, T

La tarea T de ML, puede ser descrito como de la forma en que el sistema de ML procesa una muestra o ejemplo. Según Goodfellow et al. (2016) las tareas más comunes de ML son:

- **Clasificación**. En este caso, el algoritmo de ML debe predecir la clase a la que pertenece la muestra. Entonces, al algoritmo debe producir una función: $f : \mathbb{R}^n \rightarrow \{1, ..., k\}$. También puede escribirse como: $y = f(x)$, aquí x representa la entrada y la función f determinará la clase a la que pertenece.
- **Regresión**. El algoritmo debe producir una función: $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Es decir, dada como entrada un vector x de reales, el algoritmo de ML debe predecir un valor en los números reales.
- **Transcripción**. En este caso, dada como entrada datos no estructurados, el algoritmo de ML debe generar información de forma textual. Por ejemplo: dada una imagen como entrada, la salida sería el texto encontrado en la imagen.
- **Maquinas de traducción**. Como el nombre indica, la entrada es un texto en un lenguaje y la salida es un texto en otro lenguaje.
- **Salida estructurada**. En este caso la salida es un vector o alguna estructura de datos de varios valores. El procesamiento natural de lenguaje es un buen ejemplo, la entrada es un texto y la salida es un árbol que denota la estructura gramatical y semántica de la entrada.
- **Detección de anomalías**. En este tipo de problemas el algoritmo de ML, busca detectar eventos anómalos, es decir muestras que no corresponden a la distribución normal de los datos. Un ejemplo, es la detección de transacciones fraudulentas.
- **Síntesis y muestreo**. En este caso, el algoritmo de ML debe generar nuevas muestras a partir de un conjunto de entrenamiento. Esto se aplica en los videojuegos, para la generación automática de texturas para objetos de gran tamaño.

2.3.1.2. El desempeño, *P*

Es muy importante medir el desempeño de un algoritmo de ML, usualmente la métrica utilizada puede variar según la tarea *T*. Para tareas de clasificación, usualmente se suele aplicar *Precision* y *Recall*, estos estan detallados en las Ecuaciones 2.1 y 2.2 respectivamente (Dalianis, 2018).

$$Precision : P = \frac{tp}{tp + fp} \tag{2.1}$$

$$Recall : R = \frac{tp}{tp + fn} \tag{2.2}$$

tp, hace referencia a la cantidad de muestras que eran verdaderas y han sido reconocidas como verdaderas; *fp*, son las muestras que eran falsas, pero fueron reconocidas como verdaderas; *fn*, son las muestras que eran negativas y fueron reconocidas como negativas. Otra métrica importante es el *F-score*, este puede ser definido como el peso promedio de *Precision* y *Recall* (Dalianis, 2018). En la Ecuación 2.3, presentamos la definición.

$$F - score : F_\beta = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \tag{2.3}$$

Cuando $\beta = 1$:

$$F - score : F_1 = 2 * \frac{P * R}{P + R} \tag{2.4}$$

Finalmente otra métrica, aunque no muy recomendada para datos no balanceados es el *accuracy*. Este representa el porcentaje de muestras reconocidas correctamente.

$$Accuracy : acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.5}$$

Para otro tipo de problemas, como regresión se puede aplicar el *error rate*, esta es una medida en los números reales y nos indica que tan diferente es la predicción realizada por un algoritmo de ML Goodfellow et al. (2016).

2.3.1.3. La experiencia, *E*

Según el tipo de experiencia que realizan los algoritmos de ML, se pueden clasificar en: Aprendizaje supervisado y Aprendizaje no supervisado Goodfellow et al. (2016).

- **Aprendizaje supervisado**. En este caso, cada muestra par el entrenamiento tiene los datos de entrada x y una etiqueta l. La idea es que el algoritmo de ML, pueda aprender de estos datos y luego realizar predicción de la etiqueta j tomando como entrada sólo los datos x.
- **Aprendizaje no supervisado**. En este caso, solo se cuenta con muestras no etiquetadas. Entonces el algoritmo de ML, debe agrupar los datos en *clusters*. Un ejemplo de estos problemas es la segmentación de clientes, segmentación de noticias, etc.

2.3.2. Redes neuronales

Uno de los modelos mas representativos de ML son la redes neuronales. Estas se basan en unidades llamadas neuronas (perceptron). En la Figura 2.9, se muestra esta representación, donde x_i, representa un atributo, w_i es el peso que se asigna al atributo x_i, de esta forma la neurona representa el resultado de multiplicar un peso a un atributo: $\sum_{i=1}^{d} x_i \cdot w_i$, una representación vectorial sería: $\mathbf{x}^T\mathbf{w}$ (Nielsen, 2015). Luego, a dicho resultado se aplica una función de activación, la función mas utilizada es la función sigmoidea (Equación 2.6 y 2.7).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.6}$$

, donde $z = \sum_i w_i \cdot x_i - b$.

$$\frac{1}{1 + e^{-\sum_i w_i \cdot x_i - b}} \tag{2.7}$$

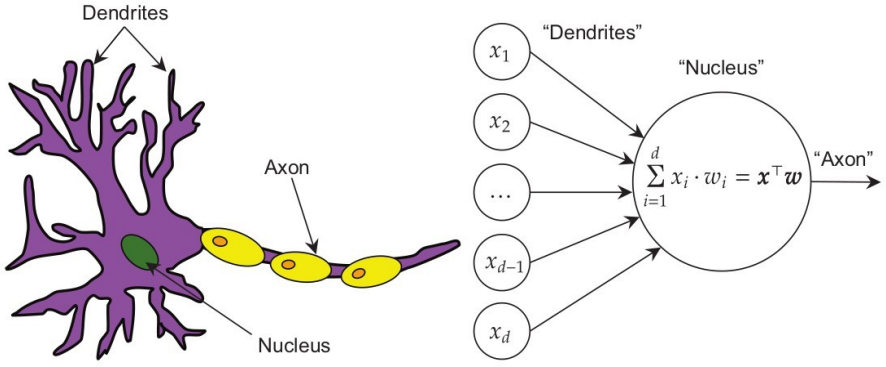


FIGURA 2.9: Representación de una neurona. Fuente: Raff (2022).

El perceptron, es capaz de solucionar varios problemas, pero para casos complejos puede formar una red, como se presenta en la Figura 2.10.

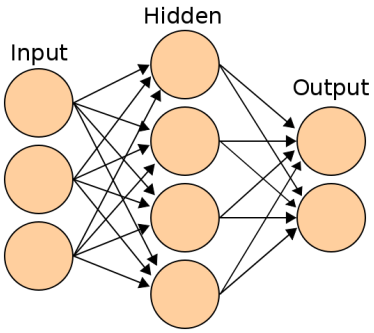


FIGURA 2.10: Representación de una red neuronal.

2.4. *Deep learning*

Deep learning (DL) es una subcategoría de *Machine Learning*, a diferencia de los algoritmos tradicionales de ML, usualmente DL trata con señales sin pre-procesamiento, los modelos (basados en redes neuronales) son mucho mas complejos tanto en dimensión como en el método de aprendizaje (El Naqa and Murphy, 2022). Por ejemplo, en la Figura 2.11, presentamos la relación entre inteligencia arficial, ML y DL, de ahí podemos concluir que ML es parte de la IA y DL es parte de ML (El Naqa and Murphy, 2022).

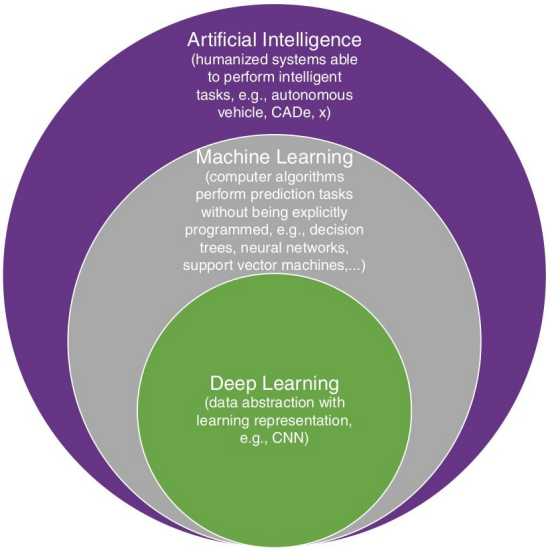


FIGURA 2.11: Relación entre Inteligencia Artificial, *Machine Learning* y *Deep Learning*. Fuente: El Naqa and Murphy (2022).

2.4.1. *Deep Feedforward networks*

Deep Feedforward networks son perceptrones multicapa o *multilayer perceptrons*(MLP). Su objetivo es aproximar una función f^*, para el caso de clasificación, podría modelarse como $y = f^*(x)$. Luego, un *feedforward network*, define un mapeo $y = f(x;\theta)$ y aprende los valores de los parametros θ Goodfellow et al. (2016). Entonces un *Deep Feedforward networks*, es una red neuronal tradiconal pero con un número grande de neuronas y capas (Figura 2.12). Existen muchos tipos de *Deep Feedforward networks*, estas serán detalladas en los siguientes apartados.

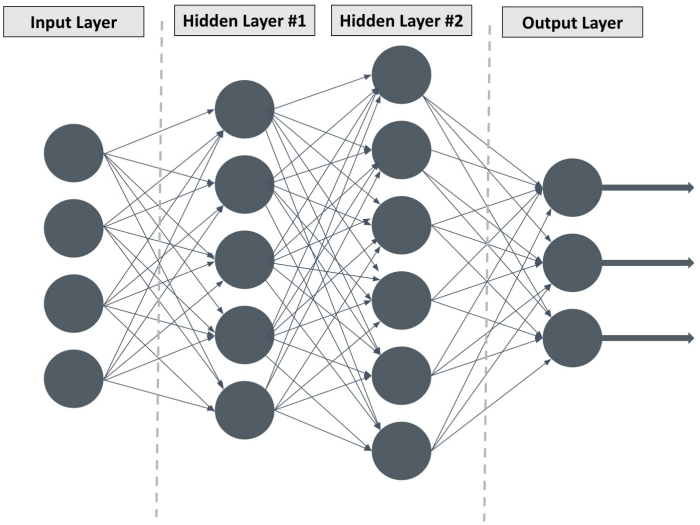


FIGURA 2.12: Representación de un *Deep Feedforward Network*. Fuente: El Naqa and Murphy (2022).

2.4.2. *Convolutional Neural Networks*

Una *Convolutional Neural Networks* (CNN), es una red neuronal basada en la operación de convoluciones (utilizada en procesamiento de imágenes). Generalmente estas redes neuronales se aplican a problemas de visión computacional (Zhang et al., 2021). La operación básica es la convolución, esta se presenta en la Figura 2.13. Se toman pequeñas ventanas de una imagen y se realiza el producto punto con un *kernel* ya establecido. Según los diferentes valores del *kernel*, se pueden obtener diferentes resultados en la imagen de salida como: detección de bordes, suavizados, dilatación, etc.

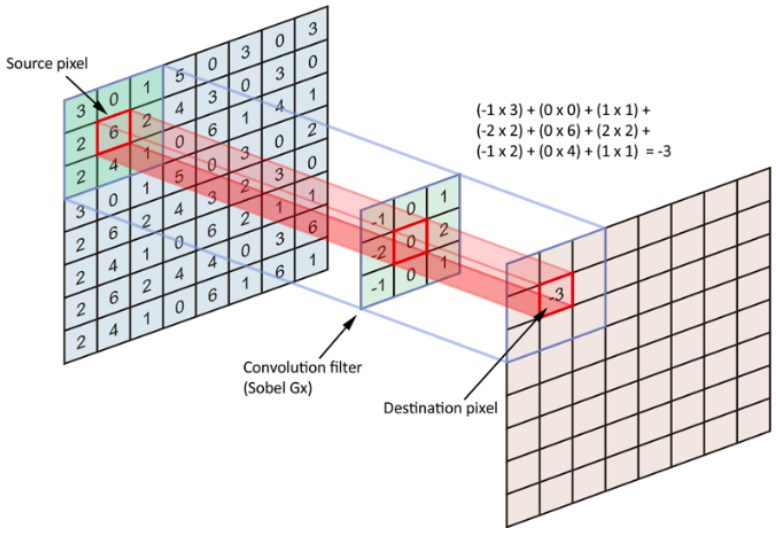


FIGURA 2.13: Ejemplo de una convolución en procesamiento de imágenes. Fuente: Shuchen (2022).

Con inspiración en la operación de convolución, se plantean las CNN por primera vez por LeCun et al. (1998). En la Figura 2.14, se presenta la LeNet-5, planteado por los autores. Luego, surgen diversa propuestas como AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015) y ResNet (He et al., 2016).

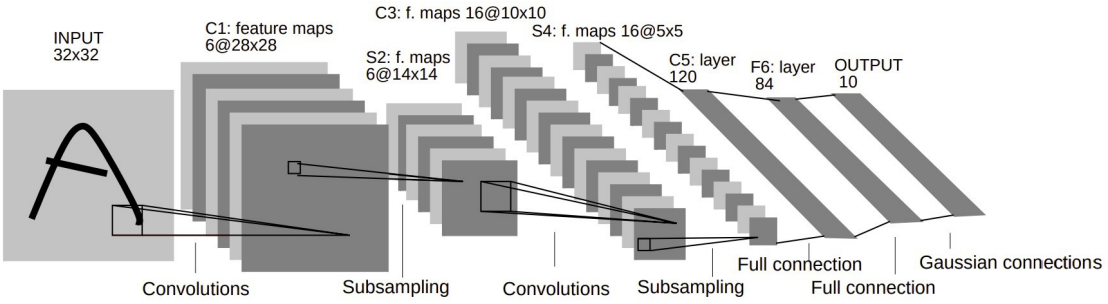


FIGURA 2.14: Arquitectura de LeNet-5, una CNN para el reconocimiento de digitos. Fuente: LeCun et al. (1998).

2.4.3. *Recurrent Neural Networks*

Mientras que las CNN están especializadas para manejar información espacial, las *Recurrent Neural Networks* (RNN), se especializan en información secuencial (Zhang et al., 2021). En este campo, se habla del tiempo como una variable y se tratan problemas de series temporales por ejemplo.

El término RNN, aparece por primera vez en los trabajos de Rumelhart et al. (1985) y Jordan (1997). Algunos autores, comentan también que el inicio de las RNN fue con las

redes de Hopfield (Hopfield, 1982). En general estas RNN, tienen dos entradas: estado actual y estado anterior; luego la RNN predice el siguiente estado. El problema de estas redes neuronales surgen por una falta de memoria, es decir cuando tenemos varios estados, el estado inicial va a influenciar cada vez menos a los estados futuros.

Como alternativa de solución al problema mencionado anteriormente, surgen Long Short-Term Memory, propuesta por Hochreiter and Schmidhuber (1997). Una red neuronal LSTM, es capaz de recordar un dato relevante de una secuencia y almacenarlo varios instantes de tiempo. En la Figura 2.15, explicamos brevemente el funcionamiento de LSTM, los datos que ingresan a una compuerta (*gate*), son los datos de entrada en un tiempo específico y el estado oculto anterior. Luego, es procesado por tres capas totalmente conectadas: *input gate*, *forget gate* y *output gate* (Zhang et al., 2021).

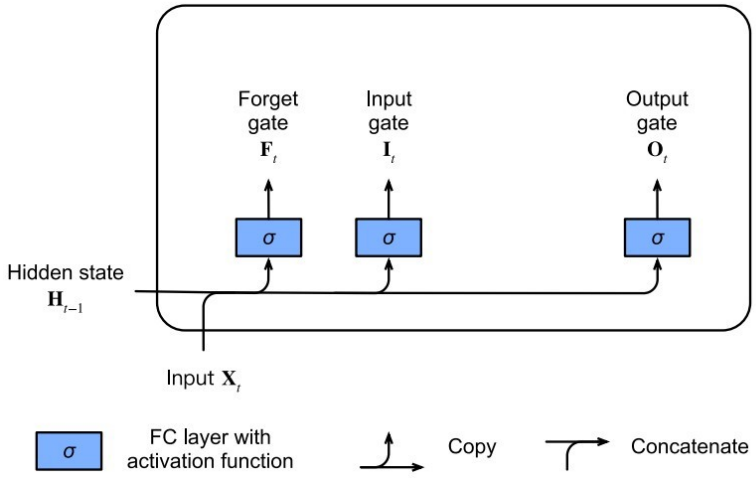


FIGURA 2.15: Ejemplo del procesamiento del *input gate*, *forget gate* y *output gate* de LSTM. Fuente: Zhang et al. (2021).

2.4.4. Transformers

El concepto del mecanismo de atención fue introducido inicialmente por Bahdanau en 2014 (Bahdanau et al., 2014) para abordar las limitaciones asociadas con vectores de codificación de longitud fija. Este enfoque novedoso produjo resultados comparables a los estados del arte en la traducción de inglés a francés. Posteriormente, el mecanismo de atención encontró aplicación en la inferencia de lenguaje natural (Parikh et al., 2016), lo que llevó a la propuesta de una red de atención estructurada (Kim et al., 2017). Sin embargo, es importante señalar que estos módulos de atención se utilizaban típicamente en conjunto con redes recurrentes. Ocurrió un cambio significativo en 2017 con la publicación del innovador artículo "Attention Is All You Need" propuesta por Vaswani et al. (2017), que presentó una nueva arquitectura de red conocida como Transformer. Esta arquitectura se basó exclusivamente en mecanismos de atención y representó una

partida fundamental de los enfoques tradicionales. En 2018, Devlin et al. (2018) introdujo el modelo bidireccional de *Transformer Bidirectional Encoder Representations from Transformers* (BERT). Desde entonces, se ha convertido en uno de los modelos de Transformer más reconocidos e influyentes en el campo. El Transformer se basa en el concepto de *self-attention*, que se refiere a cuánta atención presta una palabra a otras palabras. Por ejemplo, en la siguiente oración: "El animal no cruzó la calle porque estaba muy cansado", *self-attention* permite asociar "estaba" con "animal" (Prince, 2023).

En este contexto, el bloque principal es la autoatención $sa[\bullet]$, que toma N entradas x_n, cada una de dimensión $D \times 1$, y devuelve N vectores de salida del mismo tamaño. En el procesamiento del lenguaje natural, cada entrada x_n representa una palabra; mientras que en secuencias de proteínas, representa un aminoácido. Luego, se calculan un conjunto de valores mediante $v_n = \beta_v + \Omega_v x_n$, donde β_v y Ω_v son los sesgos y pesos, respectivamente. Así, el bloque de autoatención se calcula mediante la Ecuación **??**. El peso $a[x_m, x_m]$ es la atención que la salida x_n presta a x_m.

$$a[x_m, x_n] = \text{softmax}[k^T \cdot q_n] \tag{2.8}$$

$$Sa[X] = V \cdot \text{softmax}[k^T \cdot q_n] \tag{2.9}$$

$$\text{Sa}[X] = V \cdot \text{softmax}\left[\frac{k^T \cdot q_n}{\sqrt{D_q}}\right] \tag{2.10}$$

Además, aplicar varios *multi-head self-attention* lográ mejores resultados. Entonces, la concatenación de varios *head attentions* se presenta en la Equación 2.11.

$$\text{MhSa}[X] = \Omega_c[Sa_1[X]; Sa_2[X]; ...; Sa_H[X];] \tag{2.11}$$

2.4.5. BERT

Bidirectional Encoder Representations from Transformers (BERT), propuesta por Devlin et al. (2018), está inspirada por la red *Tranformer* y su mecanismo de atención, la cuál entiende la relación contextual entre diferentes palabras. A diferencia de una RNN, BERT no tiene dirección, es decir lee la secuencia entera. Esta característica, le permite al modelo aprender información contextual de una palabra con respecto a las otras (Kelvin, 2022).

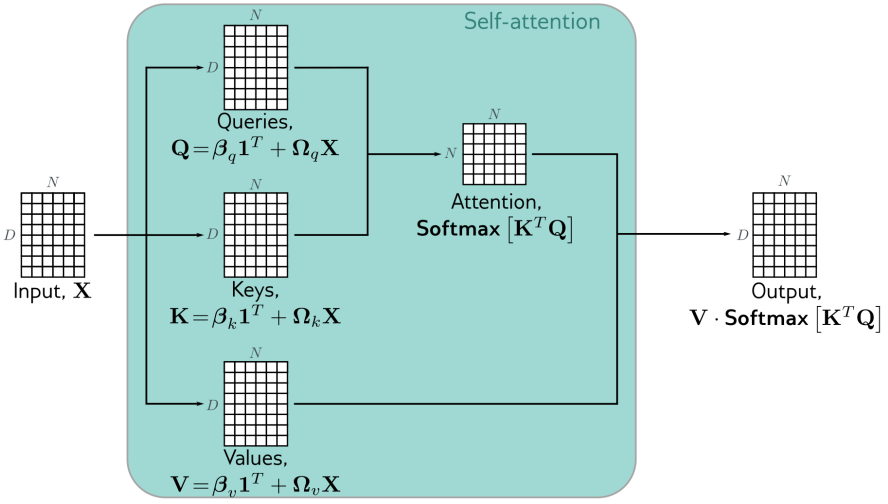


FIGURA 2.16: How to compute attention weights. Source: Prince (2023)

Capítulo 3

Estado del Arte

En este capítulo detallamos la metodología utilizada para realizar una Revisión Sistemática de la Literatura (RSL) referente a los métodos basados en Transformer y redes neuronales que utilicen mecanismos de atención para la detección de neoantígenos.

3.1. Revisión Sistemática de la Literatura (RSL)

Nuestro enfoque principal se centra en la priorización de neoantígenos (ver Figura 3.1), ya que esta área ha sido objeto de una cantidad significativa de investigaciones que utilizan Transformers. Además, integramos análisis de pipelines y estudios de ensayos clínicos para obtener información sobre los hallazgos más recientes en cuanto a la aplicación de la detección de neoantígenos en vacunas personalizadas contra el cáncer.

Según las cadenas de búsqueda de la Tabla 3.1 y considerando solo trabajos publicados a partir de 2018, se analizaron los títulos de los artículos, obteniendo un total de 151 artículos. Luego, se seleccionó un subconjunto en función de los criterios de inclusión: artículos con una categoría ERA (A o B) o artículos de revistas en los cuartiles Q1/Q2. Al final de esta etapa, se obtuvieron 79 artículos.

3.2. Resultados de la RSL

3.2.1. Detección de neoantígenos

La detección de neoantígenos se basa en la identificación inicial de candidatos, seguida de su posterior priorización. En esta sección, explicaremos el proceso de detección de candidatos a neoantígenos (etapa 2 en la Figura 3.1).

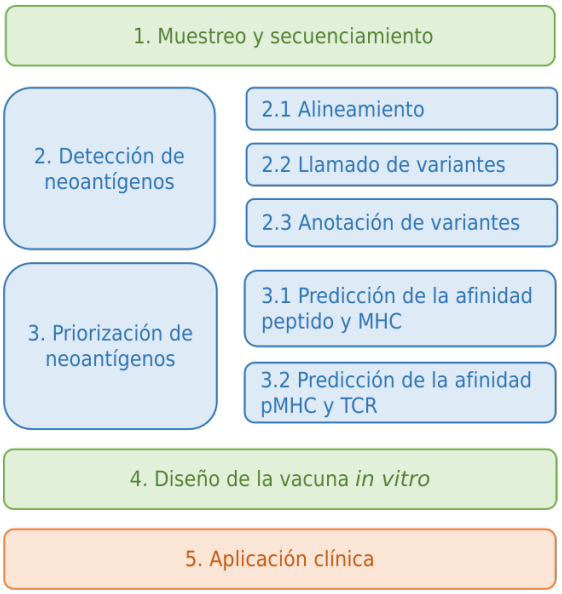


FIGURA 3.1: Una visión general de cada fase del proceso de generación de vacunas personalizadas basadas en neoantígenos.

Durante esta etapa, se utilizan datos de secuenciación de ADN (DNA-seq) y de ARN (RNA-seq) para identificar candidatos a neoantígenos. Sin embargo, en este campo, se han adoptado ampliamente varias herramientas bien establecidas, y los Transformers no se utilizan regularmente. En primer lugar, se encarga de tomar datos de DNA-seq, RNA-seq y *Mass Spectrometry* (MS) como entrada. Luego, procede a alinear estas secuencias utilizando herramientas como BWA-MEM y Bowtie2. Además, STAR podría ser utilizado porque alinea muestras de tumores de manera más efectiva (Rubinsteyn et al., 2018). La salida de esta etapa consiste en archivos de alineación BAM. Para la llamada de variantes, se podrían emplear MuTect y Strelka. Posteriormente, la información de ambos métodos se podría combinar, siguiendo el enfoque utilizado por Zhou et al. (2021) y Rubinsteyn et al. (2018). La salida consiste en archivos VCF. A continuación, está la etapa de anotación de variantes, donde se utilizan archivos con formato VCF para derivar péptidos generados a partir de estas variaciones o mutaciones; Isovar y ANNOVAR podrían ser utilizados en esta tarea. Finalmente, para determinar el tipo de HLA del paciente, la herramienta OptiType es una opción. Al final, tenemos varios candidatos a neoantígenos y los tipos de HLA del paciente.

3.2.2. Priorización de neoantígenos

La priorización de neoantígenos es la tercera etapa en el desarrollo de vacunas contra el cáncer (Figura 3.1). En esta etapa, se toman los candidatos a neoantígenos y se predice su afinidad con el MHC, un problema conocido como predicción del enlace pMHC. Luego, este complejo pMHC se utiliza para predecir la interacción con el TCR. Ambos

TABLA 3.1: Cadenas de búsqueda utilizadas en la RSL para cada fase de detección de neoantígenos.

Categoría	**Cadena de búsqueda**
Priorización de neoantígenos	(mhc OR hla) AND (peptide OR epitope OR antigen) AND (specificity OR immunogenicity OR binding OR affinity OR predict* OR detection OR presentation OR classification) AND (transformer* OR bert* OR attention OR 'transfer learning' OR method* OR predict*)", (tcr OR 't cell' OR t-cell) AND (mhc OR peptide OR epitope OR antigen) AND (specificity OR immunogenicity OR binding OR affinity OR predict* OR detection OR presentation OR classification) AND (transformer* OR bert* OR attention OR 'transfer learning' OR method* OR predict*)
Pipelines	(pipeline OR toolkit) AND (tcr OR 't cell' OR t-cell OR mhc OR hla OR peptide OR epitope OR antigen* OR neoantigen*) (pipeline OR tool* OR workflow OR application OR web*) AND (peptide OR epitope OR antigen* OR neoantigen* OR neoepito*) AND (immunotherapy OR detection OR identify* OR predict* OR presentation*)
Ensayos clínicos	(neoantigen OR neoepitope OR denditric cell) AND (vaccines OR immunology)

problemas toman dos secuencias de proteínas como entrada, y el objetivo es predecir su afinidad (regresión) o unión (clasificación).

3.2.2.1. Bases de datos

Para priorizar neoantígenos, los investigadores a menudo recopilan muestras de diversas fuentes, generalmente extrayendo datos de estudios previos y recursos similares. Sin embargo, existen conjuntos de datos públicos disponibles, como se enumeran en la Tabla 3.2, que se centran específicamente en la interacción entre péptidos y MHC (péptido-MHC) (Wu et al., 2018; Zhou et al., 2019; Tan et al., 2020; Lu et al., 2022), así como en la interacción entre pMHC y TCR (Shugay et al., 2018; Bagaev et al., 2020). Es importante destacar que un estudio reciente proporciona estructuras tridimensionales de péptidos y HLA, lo que introduce una nueva perspectiva de investigación. Finalmente, la *Immune Epitope Database* (IEDB) (Vita et al., 2018) se destaca como un recurso ejemplar en este campo.

Tabla 3.2: Bases de datos públicas de unión pMHC e interacción pMHC-TCR

Nombre	**Año**	**Ref.**	**Descripción**
VDJdb	2018	Shugay et al. (2018); Bagaev et al. (2020)	Base de datos de unión del TCR al pMHC, contiene 5491 muestras.
IEDB	2018	Vita et al. (2018)	Es la base de datos más grande que contiene información de *epitopes* de células T de humanos y otros organismos.
TSNAdb	2018	Wu et al. (2018)	Involucra 7748 muestras de mutaciones y HLA de 16 tipos de cáncer.
NeoPeptide	2019	Zhou et al. (2019)	Incorpora muestras de neoantígenos resultantes de mutaciones somáticas y elementos relacionados. También contiene 1818137 *epitopes* de más de 36000 neoantígenos.
pHLA3D	2019	Oliveira et al. (2019)	Presenta 106 estructuras 3D de las cadenas α, $\beta 2M$ y péptidos de las moléculas HLA-I.
dbPepNeo	2020	Tan et al. (2020)	Contiene muestras validadas de la unión del pMHC a partir de MS. Incluye 407794 muestras de baja calidad, 247 de calidad media y 295 de alta calidad.
dbPepNeo2.0	2022	Lu et al. (2022)	Recopila una lista de neoantígenos y moléculas HLA. Presenta 801 HLAs de alta calidad y 842,289 de baja calidad. Además, 55 neoantígenos de clase II y 630 neoantígenos con unión al TCR.
IntroSpect	2022	Zhang et al. (2022a)	Es una herramienta para construir bases de datos sobre la unión pMHC. Utiliza datos de MS.
IPD-IMGT	2022	Robinson et al. (2020)	Tiene 25000 moléculas MHC y 45 *alleles*.

3.2.2.2. Predicción de la unión pMHC

Los enfoques para predecir la unión pMHC se pueden clasificar ampliamente en dos categorías: métodos *allele-specific* y métodos *pan-specific*. Los métodos *allele-specific* implican entrenar un modelo distinto para cada *allele* específico, mientras que los métodos *pan-specific* implican el entrenamiento de un modelo universal aplicable a una variedad de *alleles*. Luego, en la Tabla 3.3, presentamos una comparación de modelos Transformer y métodos de aprendizaje profundo que utilizan mecanismos de atención.

Dado que trabajamos con entradas de proteínas, cada aminoácido se representa utilizando una fila de la matriz BLOSUM. Algunos estudios han utilizado BLOSUM62 (Jin et al., 2021; Ye et al., 2021; Zhao et al., 2019; O'Donnell et al., 2018) y BLOSUM50 (Yang et al., 2021; Hu et al., 2019). Además, ciertos autores han utilizado una combinación de codificación one-hot y codificación BLOSUM (Liu et al., 2021; Jokinen et al., 2021; Zeng and Gifford, 2019b,a). Alternativamente, se han empleado métodos como el codificador universal de Google (Kubick and Mickael, 2021), AAindex (Kawashima and Kanehisa, 2000; Li et al., 2021) (una base de datos de índices numéricos que representan propiedades fisicoquímicas y bioquímicas de los aminoácidos), coordenadas tridimensionales de aminoácidos (Shi et al., 2020), y la consideración de las propiedades fisicoquímicas de aminoácidos individuales (Moris et al., 2021; Montemurro et al., 2021; Luu et al., 2021). Más recientemente, algunos estudios han incorporado ligandos eluídos de la membrana celular, extraídos mediante datos MS (Zhou et al., 2022; Reynisson et al., 2020a,b; O'Donnell et al., 2020; Alvarez et al., 2019).

Actualmente, NetMHCPan4.1 (Reynisson et al., 2020a) es un método de referencia, este es una red neuronal artificial profunda que consiste en 40 redes neuronales artificiales ensambladas; cabe destacar que maneja eficazmente conjuntos de datos de MS, al igual que el MHCflurry2.0 (O'Donnell et al., 2020).

Existen modelos de *Convolutional Neural Networks* (CNN) que incorporan un mecanismo de atención, como ACME (Hu et al., 2019). ACME utiliza una CNN con un módulo de atención que asigna pesos a posiciones de residuos individuales, con el objetivo de asignar mayores pesos a los residuos de mayor importancia en las interacciones pMHC. ACME logró un Coeficiente de Correlación de Rango de Spearman (SRCC) de 0.569, lo cual es superior a NetMHCpan 4.0. A continuación, tenemos MHCAttNet (Venkatesh et al., 2020), que utiliza una CNN seguida de una capa de atención. La capa de atención se utiliza para generar un mapa de calor sobre los aminoácidos, indicando las subsecuencias importantes presentes en la secuencia de aminoácidos. Otro modelo basado en CNN es DeepAttentionPan (Jin et al., 2021), que utiliza una CNN profunda para codificar péptidos y MHC en vectores de dimensiones $40 \times 10 \times 11$ antes de emplear un módulo de atención para calcular pesos posicionales. También contamos con DeepNetBim (Yang et al., 2021), que incorpora un módulo de atención similar a ACME y DeepAttentionPan. Sin embargo, utiliza dos CNN separadas para predecir la unión pMHC y la inmunogenicidad, que luego se combinan en las capas finales. Además, en su estudio sobre SpConvM (Chen et al., 2021c), los autores demostraron que la incorporación de núcleos globales en CNN con atención produjo un mejor rendimiento. Además, sus experimentos incluyeron una comparación de diferentes métodos de codificación de aminoácidos, incluyendo onehot, BLOSUM y Deep. Según sus hallazgos, la combinación de onehot, BLOSUM y Deep juntos dio como resultado mejores resultados. Recientemente, ha surgido el uso de

Capsule Neural Network (CapsNet) para modelar relaciones jerárquicas. CapsNet-MHC (Kalemati et al., 2023) se propone para predecir la unión pMHC-I, y superó a otras herramientas como HLAB, ACME, Anthem y NetMHCpan4.1 para péptidos pequeños de 8 a 11 mers.

Además, se han introducido varias *Recurrent Neural Networks* (RNN), como DeepHLApan (Wu et al., 2019), que es un modelo *allele-specific* que considera datos de unión pMHC e inmunogenicidad. El modelo presenta tres capas de *Bidirectional Gated Recurrent Unit* (BiGRU) y una capa de atención, produciendo finalmente las predicciones de unión e inmunogenicidad. Además, este enfoque incorporó epitopos de células T CD8+ y datos de MS; logró una precisión que supera 0.9 para 43 alelos HLA. Además, el modelo *allele-specific* DeepSeqPanII (Liu et al., 2021) utilizó una combinación de codificación BLOSUM62 y one-hot, con un enfoque específico en MHC-II. El modelo incluyó dos capas de *Long Short-Term Memory* (LSTM) con 100 unidades y un bloque de atención para extraer información ponderada. El bloque de atención consistía en cuatro capas de convolución 1-D, y se emplearon tres capas completamente conectadas para predecir la afinidad. DeepSeqPanII superó a NetMHCIIpan 3.2 para 26 de los 54 *alleles*. Otra RNN es MATHLA (Ye et al., 2021), que utilizó una BiLSTM para aprender las dependencias entre los residuos de aminoácidos y aplicó *multi-head self-attention* para obtener información posicional para la salida de BiLSTM. La salida se procesó aún más a través de capas convolucionales 2-D. MATHLA logró un puntaje de AUC de 0.964, superando el rendimiento de NetMHCpan 4.0, MHCflurry y ACME, que obtuvieron puntajes de 0.945, 0.925 y 0.905, respectivamente. Recientemente, el modelo *allele-specific* DapNet-HLA (Jing et al., 2023) introdujo un conjunto de datos adicional de Swiss-Prot para muestras negativas. El método utilizó un método de *embedding* para cada token y su posición absoluta, que se comparó con varias técnicas de codificación, incluyendo la *Desviación de Dipeptide Deviation from Expected mean* (DDE), *Amino Acid Composition* (AAC), *Dipeptide Composition* (DPC), y *Encoding based on Grouped Weight* (EGBW). Recientemente, DapNet-HLA combinó las ventajas de CNN, SENet (para agrupamiento) y LSTM, logrando buenos resultados, aunque no se comparó directamente con métodos de vanguardia.

BERTMHC (Cheng et al., 2021) fue uno de los trabajos pioneros en incorporar la arquitectura BERT. Este predictor pan-específico de unión/presentación de pMHC-II utilizó el aprendizaje por transferencia de *Tasks Assessing Protein Embeddings* (TAPE) (Rao et al., 2019), un modelo entrenado con datos de la base de datos Pfam que comprende treinta y un millones de proteínas. Los autores integraron TAPE seguido de una capa *Fully Connected* (FC). En experimentos, BERTMHC superó a NetMHCIIpan3.2 y PUFFIN, logrando un AUC de 0.8822 en comparación con 0.8774. Del mismo modo, ImmunoBERT (Gasser et al., 2021) aprovechó el aprendizaje por transferencia de TAPE,

centrándose en la predicción de pMHC-I. El modelo tambien utilizo capas FC después del modelo TAPE. El análisis de los autores concluyó que los aminoácidos en proximidad a los extremos N/C del péptido son de alta relevancia, según análisis de LIME y SHAP. Además, CapTransformer (Chen et al., 2021a) introdujo un innovador mecanismo de *cross self-attention* que alinea y agrega eficazmente las características de los residuos de pMHC de manera conjunta. Al utilizar tanto la *self-attention* como *cross self-attention*, facilita el aprendizaje de representaciones de características para los residuos individuales y la información global de unión pMHC, lo que resulta en un rendimiento superior en comparación con NetMHCpan4.0.

Otros métodos que utilizaron el aprendizaje por transferencia incluyen MHCRoBERTa (Wang et al., 2022) y HLAB (Zhang et al., 2022b). El primero empleó cinco *encoders* con doce *multi-head self-attention*. Inicialmente, el enfoque utilizó un entrenamiento autosupervisado con datos de las bases de datos UniProtKB y Swiss-Prot. El método también aplicó la tokenización de *subtokens* y superó a NetMHCpan4.0 y MHCflurry2.0, logrando un coeficiente de correlación de Spearman Rank (SRCC) de 0.543. HLAB aprovechó el aprendizaje por transferencia de ProtBert-BFD (Elnaggar et al., 2021), que fue entrenado con datos del conjunto de datos BFD que contiene 2,122 millones de proteínas. HLAB empleó un modelo BiLSTM al final de ProtBert-BFD y logró un rendimiento superior a NetMHCpan4.1. Además, una investigación adicional examinó la aplicación del aprendizaje por transferencia y la aplicación de *padding* (Arceda, 2023). Finalmente, TransPHLA (Chu et al., 2022) aplica la *self-attention* a los péptidos, TransPHLA supero a NetMHCpan4.1, y ofrece la ventaja de ser efectivo para péptidos y *alleles* de MHC de diferentes tamaños.

Una propuesta interesante implica el uso del modelo *Star-Transformer*, SMHCpan (Ye et al., 2023), un modelo liviano en el que la estructura de FC se reemplaza por una topología en forma de estrella. Además, las *Graph Neural Networks* (GNN) se han utilizado en varios problemas de *Protein-Protein Interaction* (PPI) debido a que gestionan las relaciones entre proteínas. En este contexto, surgió una propuesta novedosa, ESM-GAT (Hashemi et al., 2023), que utilizó arquitecturas BERT y aprendizaje por transferencia de los modelos ESM1b y ESM2, luego apiló una *Graph Attention Network* (GAT). Superó a NetMHCpan4.1; sin embargo, los autores no compararon la propuesta con otras herramientas del estado del arte.

3.2.2.3. Predicción de la unión pMHC-TCR

3.2.3. Pipelines

3.2.4. Ensayos clínicos

Tabla 3.3: Transformers y métodos de aprendizaje profundo con mecanismos de atención utilizados para la predicción de la unión pMHC.

Ref.	**Nombre**	**Entrada**	**Modelo**
Hashemi et al. (2023)	ESM-GAT	One-hot	BERT con transferencia de aprendizaje de ESM1b y ESM2 *fine-tuned* con una *Graph Attention Network* (GAT) al final. Superó a NetMHCpan4.1.
Kalemati et al. (2023)	CapsNet-MHC	BLOSUM62	*Capsule Neural Network*, superó a las herramientas de vanguardia para péptidos pequeños de 8 a 11 mer.
Ye et al. (2023)	STMHCpan	One-hot	Un modelo Star-Transformer, útil para péptidos de cualquier longitud y ampliable para predecir respuestas de células T.
Jing et al. (2023)	DapNet-HLA	*Fused word embedding*	Combina las ventajas de CNN, SENet (para agrupación) y LSTM con atención.
Zhang et al. (2022b)	HLAB	One-hot	BERT del modelo pre-entrenado ProtBert seguido de una BiLSTM.
Wang et al. (2022)	MHC RoBERTa	One-hot	RoBERTa pre-entrenado seguido de 12 *multi-head self-attention* y capas totalmente conectadas, superó a NetMHCPan3.0.
Chu et al. (2022)	TransPHLA	One-hot	Utiliza un mecanismo de self-attention basado en cuatro bloques, superó ligeramente a NetMHCpan4.1 y es más rápido en hacer predicciones.
Chen et al. (2021a)	CapTransformer	One-hot	Transformer con *cross self-attention* para capturar información local y global.
Gasser et al. (2021)	ImmunoBERT	One-hot	BERT de TAPE pre-entrenado seguido de una capa lineal. Los autores afirman que los terminales N y C son altamente relevantes después de un análisis con SHAP y LIME.
Cheng et al. (2021)	BERTMHC	One-hot	BERT de TAPE pre-entrenado seguido de una capa lineal. Superó a NetMHCIIpan3.2 y PUFFIN.
Ye et al. (2021)	MATHLA	BLOSUM	Integra BiLSTM con *multi-head self-attention*. Obtuvo una puntuación AUC de 0.964, en comparación con 0.945, 0.925 y 0.905 para NetMHCpan 4.0, MHCflurry y ACME respectivamente
Liu et al. (2021)	DeepSeqPanII	BLOSUM62 y one-hot	Tiene dos capas LSTM, un bloque de atención y tres capas totalmente conectadas. Obtuvo mejores resultados que NetMHCIIpan 3.2 en 26 de 54 alelos.
Yang et al. (2021)	DeepNetBim	BLOSUM50	Utiliza CNN separadas para la predicción de la unión pMHC y *immunogenetic* con un módulo de atención. Obtuvo 0.015 MAE para la unión y 94.7 de precisión para la inmunogenicidad.
Jin et al. (2021)	DeepAttention Pan	BLOSUM62	CNN con un mecanismo de atención. Es *allele-specific* y obtuvo resultados ligeramente mejores que ACME a nivel de *alleles*.
Chen et al. (2021c)	SpConvM	One-hot, BLOSUM y Deep	Capa 1D de CNN, una capa de atención y una capa totalmente conectada. Además, emplearon *global kernels* para mejorar sus resultados, junto con una combinación de onehot, BLOSUM y Deep.
Venkatesh et al. (2020)	MHCAttNet	One-hot	CNN seguido de una capa de atención para generar un mapa de calor sobre los aminoácidos.
Hu et al. (2019)	ACME	BLOSUM50	CNN con atención, extrae patrones interpretables sobre la unión pMHC. Además, obtuvo un SRCC de 0.569, un AUC de 0.9 para HLA-A y 0.88 para HLA-B.
Wu et al. (2019)	DeepHLApan	One-hot	Modelo *allele-specific* con tres capas de GRU Bidireccional (BiGRU) con una capa de atención. Obtuvo una precisión $>$ 0,9 en 43 *alleles* HLA.

Capítulo 4

Propuesta

En este capítulo presetaremos la propuesta y como se relaciona con los métodos tradicionales de detección de neo antígenos.

4.1. Detección de neo antígenos (*pipeline*)

Según Gopanenko et al. (2020), la detección de neo antígenos podría clasificarse en tres grupos: (1) basados en genómica, (2) basados en *Mass Spectrometry* (MS) y (3) basados en estructura.

La detección de neo antígenos basada en genómica sigue un proceso muy largo e involucra muchas herramientas, debido a esto se han propuesto bastantes *pipelines*. El proceso general consta de varias etapas presentadas en la Figura 4.1, a continuación detallaremos cada una de ellas y explicaremos en qué fase se ubica la propuesta de esta tesis:

1. **Secuenciamiento**. La primera fase consiste en el secuenciamiento de DNA, en este caso se toman muestras de sangre al tener menos riesgo de no ser contaminadas por un tumor (Borden et al., 2022). Para la secuenciación, se puede obtar por *Whole Genome Sequencing* (WGS) o *Whole Exome Sequencing* (WES), la primera tiene la ventaja de tener mucha más información de mutaciones pero es muy costoso. Esta fase, también puede retroalimentarse con secuenciamiento de RNA (seqRNA). Una tendencia reciente fomenta el uso de *RiboSeq*, este tiene la ventaja de tener más información de las proteínas formadas en los Ribosomas, lamentablemente no se tienen muchas muestras (Borden et al., 2022).

2. **Alineamiento y procesamiento**. En esta fase, se evalúa la calidad del secuenciamiento, se elimina el ruido y se realiza un alineamiento con un genoma base. Como resultado se obtienen archivos BAM (resultado del alineamiento) y FastQC (calidad de cada secuenciación).

3. **Identificación de neo antígenos**. En esta fase se analiza las mutaciones de la secuencia, generalmente se obtienen *Variant Calling Files* (VCF). En esta etapa, es importante secuenciar las proteínas *Human Leukocyte Antigens* (HLA), estas representan las proteínas MHC mencionadas anteriormente. Luego con información del tipo de HLA y mutaciones, se puede identificar los posibles neo antígenos. Esta fase puede ser retroalimentada de *RiboSeq* y datos de MS.

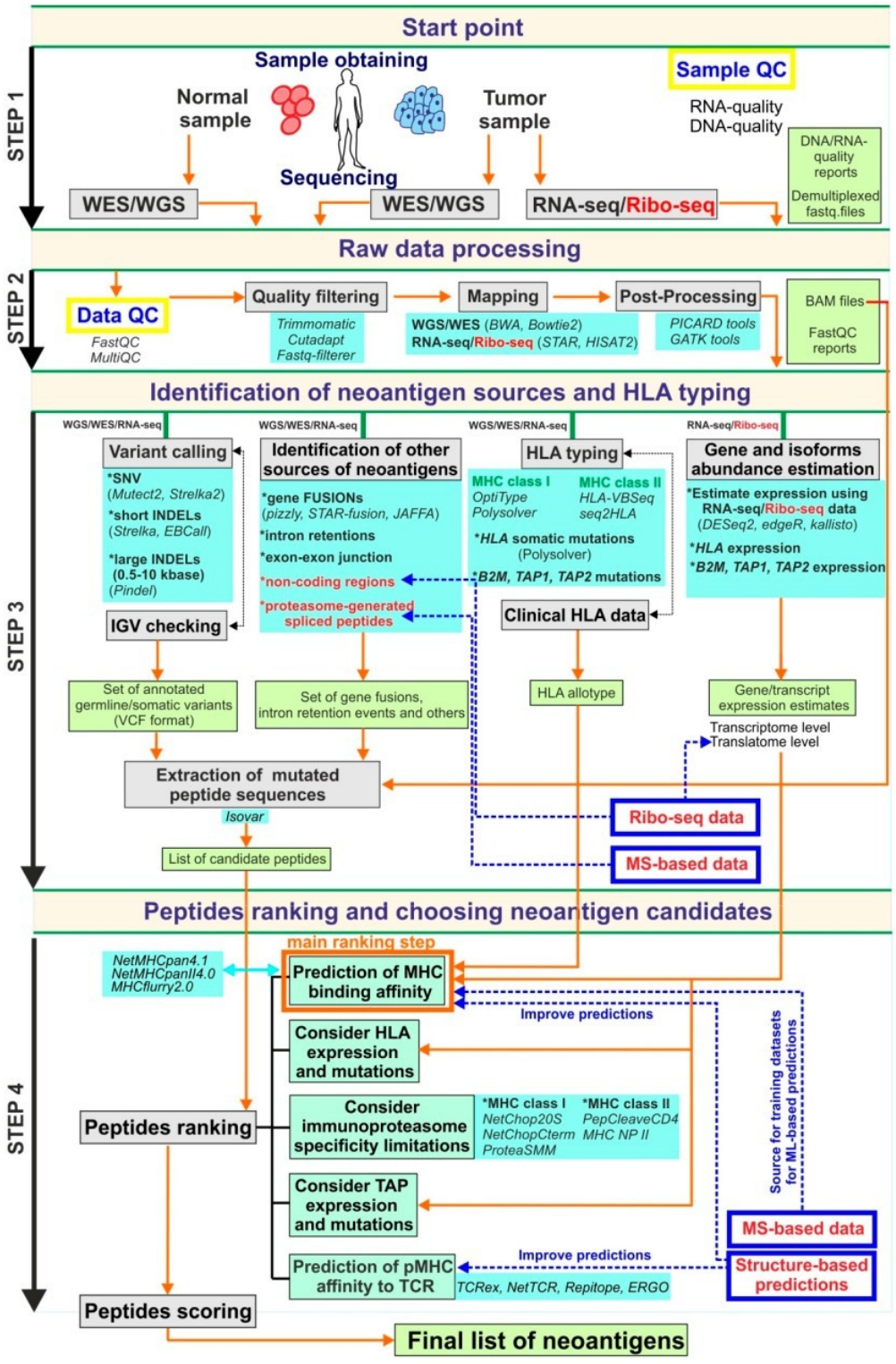


FIGURA 4.1: Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: Gopanenko et al. (2020).

4. **Priorización de neo antígenos**. En esta fase se filtran los neo antígenos identificados anteriormente. Este problema es conocido mayormente como: *MHC-peptide binding*, en este caso se predice el enlace entre el neo antígeno y la proteína MHC

(la propuesta de la tesis se enfoca en esta etapa). Las herramientas con mejor desempeño son *NetMHCpan4.1* y *MHCflurry2.0* según varios *benchmarkings* (Bonsack et al., 2019; Zhao and Sher, 2018; Paul et al., 2020; Trolle et al., 2015). Recientemente una nueva propuesta ha superado a *NetMHCpan4.1*, esta propuesta obtuvo buenos resultados utilizando *protein language models* (Hashemi et al., 2022). Finalmente, se predice la afinidad de T-Cell Receptor (TCR) con pMHC (peptide-MHC binding).

Recientemente, se está utilizando otros enfoques para mejorar la detección de neo antígenos, por ejemplo, se puede utilizar datos MS para mejorar la identificación de neo antígenos. Luego, el enfoque basado en estructura que utiliza información de propiedades químicas y físicas de los péptidos puede ser utilizada para mejorar la predicción de afinidad TCR y pMHC (Borden et al., 2022; Gopanenko et al., 2020).

4.2. Predicción de la afinidad peptido-MHC (peptide-MHC binding)

La propuesta se inspira en los trabajos de Cheng et al. (2021) y Hashemi et al. (2022). Ambos proponen el uso de *transfer learning* a partir de los modelos pre-entrenados BERT (Devlin et al., 2018) y ESM-1b (Rives et al., 2021) respectivamente.

El modelo *Bidirectional Encoder Representations from Transformers.* (BERT), fue diseñado para el pre-entrenamiento de representaciones bidireccionales de textos no etiquetados. Este modelo fue diseñado inicialmente para el procesamiento natural del lenguaje, pero en el trabajo de Rao et al. (2019), se planteó su uso para secuencias de aminoácidos. Es así que Rao et al. (2019) entrenan BERT con 31 millones de secuencias de proteínas y llaman a su propuesta *Tasks Assessing Protein Embeddings* (TAPE).

Recientemente, Facebook desarrolla el modelo ESM-1b (Rives et al., 2021). La propuesta se basa en el modelo RoBERTa (Liu et al., 2019a), la cuál es una optimización de BERT. Luego, ESM-1b fue entrenado con la base de datos Uniref50 (Suzek et al., 2015), esta base de datos cuenta con aproximadamente 250 millones de secuencias de proteínas. En este caso, se realizó un entrenamiento no supervisado, se ocultaron las etiquetas referentes a la estructura o función de las proteínas.

Entonces, la propuesta de la tesis se basa en utilizar *transfer learning* del modelo pre-entrenado ESM-1b, luego se va a utilizar otra red neuronal paralela que se alimente de datos físico-químicos de los aminoácidos. Se propone utilizar las propiedades físico-químicas de los aminoácidos, porque en varios ensayos clínicos se ha comprobado que influyen en la predicción *peptide-MHC binding* y *pMHC-TCR presentation* (Gopanenko et al., 2020; Borden et al., 2022). Luego, las dos redes neuronales paralelas se unirán en una red neuronal totalmente conectada (ver Figura 4.2). El objetivo, es aprovechar las propiedades físico-químicas de los aminoácidos para mejorar la afinidad *peptide-MHC*.

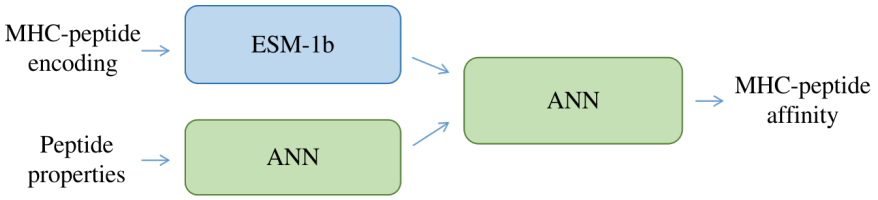


FIGURA 4.2: Propuesta de *transfer learning* de ESM-1b y una red neuronal paralela para la predicción de la afinidad entre un péptido y MHC (peptide MHC binding).

Para los entrenamientos y experimentos se utilizará la base de datos HLA3D (Li et al., 2022), esta contiene información de 1296 aminoácidos. Luego, también utilizaremos las muestras recolectadas de Hashemi et al. (2022).

Capítulo 5

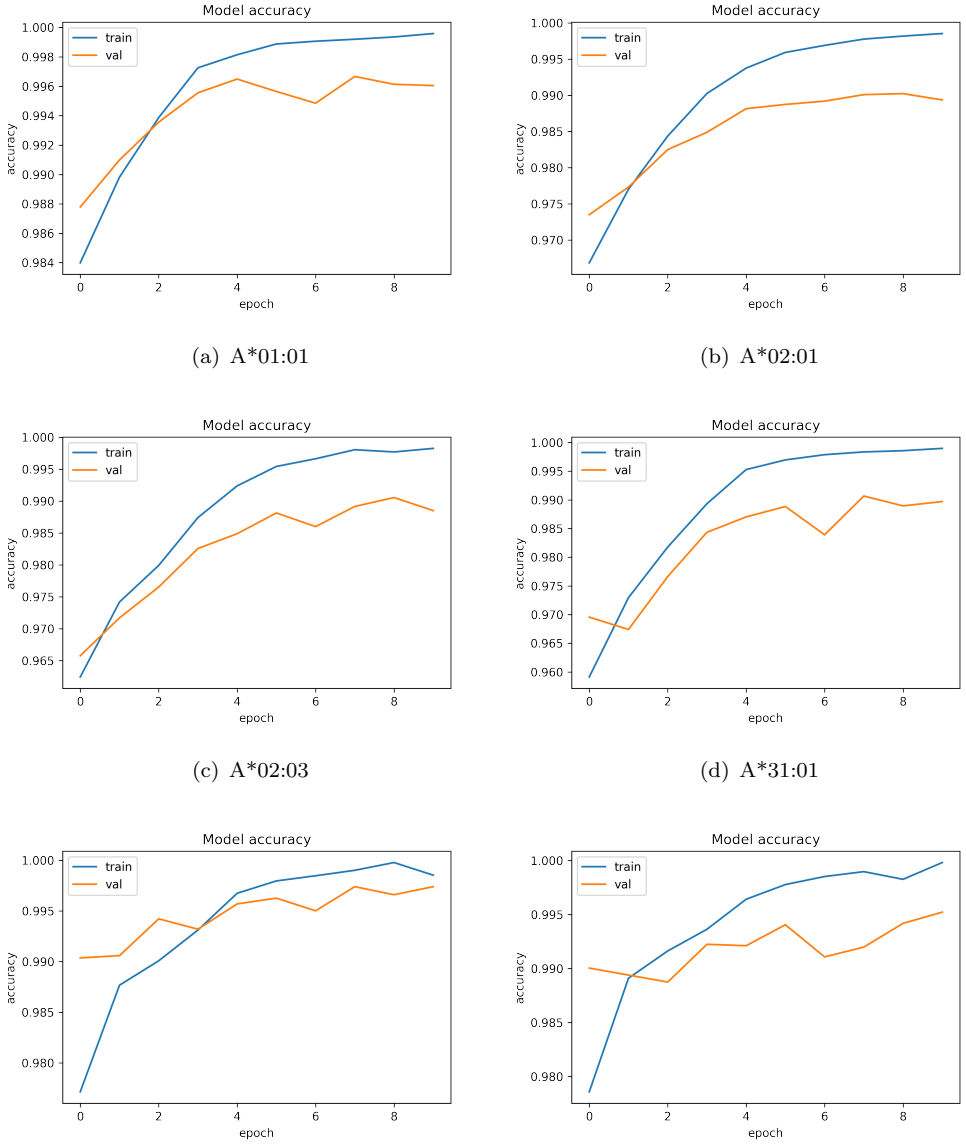
Resultados

En la Tabla 5.1, presentamos el *accuracy, f1 score, precision* y *recall* de cada base de datos (*allele*). Como podemos ver, en todos los casos superamos el 0.9 de *accuracy*, esto valida la propuesta y da origen a seguir trabajando en mejorar la propuesta.

Luego, en la Figura 5.1, presentamos el *accuracy* obtenido durante el entrenamiento de cada base de datos con el conjunto de muestras de entrenamiento y validación. En este caso, utilizamos el 20 % de las muestras de entrenamiento como validación. Como podemos ver, con solo 10 *epochs*, se lograron buenos resultados. Tambien se evaluao con mas *epochs*, pero los resultados no mejoraron.

Tabla 5.1: Resultados obtenidos en cada base de datos.

Allele	***Accuracy***	***F1 score***	***Precision***	***Recall***
A*01:01	0.978	0.917	0.982	0.887
A*0201	0.962	0.956	0.965	0.948
A*02:03	0.992	0.979	0.994	0.969
A*31:01	0.980	0.968	0.989	0.951
B*44:02	0.991	0.981	0.968	0.997
B*44:03	0.992	0.987	0.995	0.980



(a) A*01:01

(b) A*02:01

(c) A*02:03

(d) A*31:01

(e) B*44:02

(f) B*44:03

FIGURA 5.1: *Accuracy* durante cada *epoch*, para cada base de datos. Las bases de datos representan las células HLA A*01:01, A*02:01, A*02:03, A*31:01, B*44:02 y B*44:03.

Capítulo 6

Conclusiones

Primera: Se ha realizado una búsqueda sistemática de la literatura sobre los principales métodos basados en *deep learning*, utilizados para la detección de neo antígenos. Estos métodos involucran las *Shallow Neural Networks*, redes neuronales convolucionales, redes neuronales recurrentes y recientemente las redes *Transformers* y BERT.

Segunda: Se ha presentado un nuevo método basado en redes neuronales BERT y con *trasnfer lerning*, de los modelos pre entrenados TAPE y ESMb-1. El método propuesto ha sido evaluado con colección de varias muestras tomadas de bases de datos públicas y trabajos similares.

Bibliografía

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., et al. (2017). Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*, 46(2):315–326.

Abualrous, E. T., Sticht, J., and Freund, C. (2021). Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation. *Current Opinion in Immunology*, 70:95–104.

Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M., and Nielsen, M. (2019). Nnalign_ma; mhc peptidome deconvolution for accurate mhc binding motif characterization and improved t-cell epitope predictions. *Molecular & Cellular Proteomics*, 18(12):2459–2477.

Arceda, V. E. M. (2023). Neoantigen detection using transformers and transfer learning in the cancer immunology context. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 97–102. Springer.

Bagaev, D. V., Vroomans, R. M., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E. S., Zvyagin, I. V., et al. (2020). Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation*[s]. *Molecular & Cellular Proteomics*, 14(3):658–673.

Bonsack, M., Hoppe, S., Winter, J., Tichy, D., Zeller, C., Küpper, M. D., Schitter, E. C., Blatnik, R., and Riemer, A. B. (2019). Performance evaluation of mhc class-i binding

prediction tools based on an experimentally validated mhc–peptide binding data set. *Cancer immunology research*, 7(5):719–736.

Borden, E. S., Buetow, K. H., Wilson, M. A., and Hastings, K. T. (2022). Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation. *Frontiers in Oncology*, 12.

Bravi, B., Tubiana, J., Cocco, S., Monasson, R., Mora, T., and Walczak, A. M. (2021). Rbm-mhc: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by hla-i alleles. *Cell systems*, 12(2):195–202.

Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2019). Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification. *Nature biotechnology*, 37(1):55–63.

Chen, C., Qiu, Z., Yang, Z., Yu, B., and Cui, X. (2021a). Jointly learning to align and aggregate with cross attention pooling for peptide-mhc class i binding prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 18–23. IEEE.

Chen, I., Chen, M., Goedegebuure, P., and Gillanders, W. (2021b). Challenges targeting cancer neoantigens in 2021: a systematic literature review. *Expert Review of Vaccines*, 20(7):827–837.

Chen, Z., Min, M. R., and Ning, X. (2021c). Ranking-based convolutional neural network models for peptide-mhc class i binding prediction. *Frontiers in Molecular Biosciences*, 8:634836.

Cheng, J., Bendjama, K., Rittner, K., and Malone, B. (2021). Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning. *Bioinformatics*, 37(22):4172–4179.

Chu, Y., Zhang, Y., Wang, Q., Zhang, L., Wang, X., Wang, Y., Salahub, D. R., Xu, Q., Wang, J., Jiang, X., et al. (2022). A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence*, 4(3):300–311.

Clancy, S. (2008). Genetic mutation. *Nature Education*, 1(1):187.

Dalianis, H. (2018). Evaluation metrics and evaluation. In *Clinical text mining*, pages 45–53. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

El Naqa, I. and Murphy, M. J. (2022). Machine and deep learning in oncology, medical physics and radiology.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.

Fang, X., Guo, Z., Liang, J., Wen, J., Liu, Y., Guan, X., and Li, H. (2022). Neoantigens and their potential applications in tumor immunotherapy. *Oncology Letters*, 23(3):1–9.

Gasser, H.-C., Bedran, G., Ren, B., Goodlett, D., Alfaro, J., and Rajan, A. (2021). Interpreting bert architecture predictions for peptide presentation by mhc class i proteins. *arXiv preprint arXiv:2111.07137*.

Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2023). Improved predictions of antigen presentation and tcr recognition with mixmhcpred2. 2 and prime2. 0 reveal potent sars-cov-2 cd8+ t-cell epitopes. *Cell Systems*, 14(1):72–83.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gopanenko, A. V., Kosobokova, E. N., and Kosorukov, V. S. (2020). Main strategies for the identification of neoantigens. *Cancers*, 12(10):2879.

Han, X.-J., Ma, X.-l., Yang, L., Wei, Y.-q., Peng, Y., and Wei, X.-w. (2020). Progress in neoantigen targeted cancer immunotherapies. *Frontiers in Cell and Developmental Biology*, 8:728.

Hao, Q., Wei, P., Shu, Y., Zhang, Y.-G., Xu, H., and Zhao, J.-N. (2021). Improvement of neoantigen identification through convolution neural network. *Frontiers in immunology*, 12.

Hashemi, N., Hao, B., Ignatov, M., Paschalidis, I., Vakili, P., Vajda, S., and Kozakov, D. (2022). Improved predictions of mhc-peptide binding using protein language models. *bioRxiv*.

Hashemi, N., Hao, B., Ignatov, M., Paschalidis, I. C., Vakili, P., Vajda, S., and Kozakov, D. (2023). Improved prediction of mhc-peptide binding using protein language models. *Frontiers in Bioinformatics*, 3.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heyer, E. E. and Blackburn, J. (2020). Sequencing strategies for fusion gene detection. *BioEssays*, 42(7):2000016.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Hu, Y., Wang, Z., Hu, H., Wan, F., Chen, L., Xiong, Y., Wang, X., Zhao, D., Huang, W., and Zeng, J. (2019). Acme: pan-specific peptide–mhc class i binding prediction through attention-based deep neural networks. *Bioinformatics*, 35(23):4946–4954.

Janeway Jr, C. A. (1997). Immunobiology the immune system in health and disease. *Artes Medicas*.

Jin, J., Liu, Z., Nasiri, A., Cui, Y., Louis, S.-Y., Zhang, A., Zhao, Y., and Hu, J. (2021). Deep learning pan-specific model for interpretable mhc-i peptide binding prediction with improved attention mechanism. *Proteins: Structure, Function, and Bioinformatics*, 89(7):866–883.

Jing, Y., Zhang, S., and Wang, H. (2023). Dapnet-hla: Adaptive dual-attention mechanism network based on deep learning to predict non-classical hla binding sites. *Analytical Biochemistry*, 666:115075.

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., and Lähdesmäki, H. (2021). Predicting recognition between t cell receptors and epitopes with tcrgp. *PLoS computational biology*, 17(3):e1008814.

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier.

Kalemati, M., Darvishi, S., and Koohi, S. (2023). Capsnet-mhc predicts peptide-mhc class i binding based on capsule neural networks. *Communications Biology*, 6(1):492.

Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374.

Kelvin, J. (2022). Rnns, lstms, cnns, transformers and bert.

Kerbs, P., Vosberg, S., Krebs, S., Graf, A., Blum, H., Swoboda, A., Batcha, A. M., Mansmann, U., Metzler, D., Heckman, C. A., et al. (2022). Fusion gene detection by rna-sequencing complements diagnostics of acute myeloid leukemia and identifies recurring nrip1-mir99ahg rearrangements. *haematologica*, 107(1):100.

Kim, P. and Zhou, X. (2019). Fusiongdb: fusion gene annotation database. *Nucleic acids research*, 47(D1):D994–D1004.

Kim, S., Kim, H. S., Kim, E., Lee, M., Shin, E.-C., and Paik, S. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology*, 29(4):1030–1036.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. (2009). Derivation of an amino acid similarity matrix for peptide: Mhc binding and its application as a bayesian prior. *BMC bioinformatics*, 10:1–11.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kubick, N. and Mickael, M. E. (2021). Predicting epitopes based on tcr sequence using an embedding deep neural network artificial intelligence approach. *bioRxiv*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, G., Iyer, B., Prasath, V. S., Ni, Y., and Salomonis, N. (2021). Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity. *Briefings in bioinformatics*, 22(6):bbab160.

Li, X., Lin, X., Mei, X., Chen, P., Liu, A., Liang, W., Chang, S., and Li, J. (2022). Hla3d: an integrated structure-based computational toolkit for immunotherapy. *Briefings in bioinformatics*, 23(3):bbac076.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Cui, Y., Xiong, Z., Nasiri, A., Zhang, A., and Hu, J. (2019b). Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction. *Scientific reports*, 9(1):1–10.

Liu, Z., Jin, J., Cui, Y., Xiong, Z., Nasiri, A., Zhao, Y., and Hu, J. (2021). Deepseqpanii: an interpretable recurrent neural network model with attention mechanism for peptide-hla class ii binding prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Lu, M., Xu, L., Jian, X., Tan, X., Zhao, J., Liu, Z., Zhang, Y., Liu, C., Chen, L., Lin, Y., et al. (2022). dbpepneo2. 0: A database for human tumor neoantigen peptides from mass spectrometry and tcr recognition. *Frontiers in immunology*, page 1583.

Lucito, R., Suresh, S., Walter, K., Pandey, A., Lakshmi, B., Krasnitz, A., Sebat, J., Wigler, M., Klein, A. P., Brune, K., et al. (2007). Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer biology & therapy*, 6(10):1592–1599.

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(04):346–358.

Luu, A. M., Leistico, J. R., Miller, T., Kim, S., and Song, J. S. (2021). Predicting tcr-epitope binding specificity using deep metric learning and multimodal learning. *Genes*, 12(4):572.

Machaca, V. E., Goyzueta, V., Cruz, M., and Tupac, Y. (2023). Deep learning and transformers in mhc-peptide binding and presentation towards personalized vaccines in cancer immunology: A brief review. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 14–23. Springer.

Marshall, J. S., Warrington, R., Watson, W., and Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy, Asthma & Clinical Immunology*, 14(2):1–10.

Mattos, L., Vazquez, M., Finotello, F., Lepore, R., Porta, E., Hundal, J., Amengual-Rigo, P., Ng, C., Valencia, A., Carrillo, J., et al. (2020). Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group. *Annals of oncology*, 31(8):978–990.

Mei, S., Li, F., Xiang, D., Ayala, R., Faridi, P., Webb, G. I., Illing, P. T., Rossjohn, J., Akutsu, T., Croft, N. P., et al. (2021). Anthem: a user customised tool for fast and accurate prediction of binding between peptides and hla class i molecules. *Briefings in Bioinformatics*, 22(5):bbaa415.

Mill, N. A., Bogaert, C., van Criekinge, W., and Fant, B. (2022). neoms: Attention-based prediction of mhc-i epitope presentation. *bioRxiv*.

Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., Crinklaw, A., Hadrup, S. R., Winther, O., Peters, B., et al. (2021). Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcrα and β sequence data. *Communications biology*, 4(1):1–13.

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., and Meysman, P. (2021). Current challenges for unseen-epitope tcr interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):bbaa318.

NCI (2020). Nci dictionary of cancer terms. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription. Accessed: 2020-03-20.

NCI (2022). National cancer institute dictionary.

Nielsen, M. and Andreatta, M. (2016). Netmhcpan-3.0; improved prediction of binding to mhc class i molecules integrating information from multiple receptor and peptide length datasets. *Genome medicine*, 8(1):1–9.

Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA.

O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018). Mhcflurry: open-source class i mhc binding affinity prediction. *Cell systems*, 7(1):129–132.

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell systems*, 11(1):42–48.

Oliveira, D. M. T., de Serpa Brandão, R. M. S., da Mata Sousa, L. C. D., Lima, F. d. C. A., do Monte, S. J. H., Marroquim, M. S. C., de Sousa Lima, A. V., Coelho, A. G. B., Costa, J. M. S., Ramos, R. M., et al. (2019). phla3d: An online database of predicted three-dimensional structures of hla molecules. *Human Immunology*, 80(10):834–841.

PacBio (2021). Two review articles assess structural variation in human genomes. https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/. Accessed: 2021-05-07.

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., Huang, T., and Cai, Y.-D. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Molecular Genetics and Genomics*, 294(1):95–110.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Patwardhan, N., Marrone, S., and Sansone, C. (2023). Transformers in the real world: A survey on nlp applications. *Information*, 14(4):242.

Paul, S., Croft, N. P., Purcell, A. W., Tscharke, D. C., Sette, A., Nielsen, M., and Peters, B. (2020). Benchmarking predictions of mhc class i restricted t cell epitopes in a comprehensively studied model system. *PLoS computational biology*, 16(5):e1007757.

Peng, M., Mo, Y., Wang, Y., Wu, P., Zhang, Y., Xiong, F., Guo, C., Wu, X., Li, Y., Li, X., et al. (2019). Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular cancer*, 18(1):1–14.

Phloyphisut, P., Pornputtapong, N., Sriswasdi, S., and Chuangsuwanich, E. (2019). Mhcseqnet: a deep neural network model for universal mhc binding prediction. *BMC bioinformatics*, 20(1):1–10.

Prince, S. J. (2023). *UNDERSTANDING DEEP LEARNING.* MIT PRESS.

Raff, E. (2022). *Inside Deep Learning.* Manning Publications Co.

Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., and Stevanović, S. (1999). Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50:213–219.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32.

Reche, P. A., Glutting, J.-P., and Reinherz, E. L. (2002). Prediction of mhc class i binding peptides using profile motifs. *Human immunology*, 63(9):701–709.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020a). Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454.

Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W. H., Peters, B., and Nielsen, M. (2020b). Improved prediction of mhc ii antigen presentation through integration and motif deconvolution of mass spectrometry mhc eluted ligand data. *Journal of proteome research*, 19(6):2304–2315.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling

unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., and Marsh, S. G. (2020). Ipd-imgt/hla database. *Nucleic acids research*, 48(D1):D948–D955.

Rubinsteyn, A., Kodysh, J., Hodes, I., Mondet, S., Aksoy, B. A., Finnigan, J. P., Bhardwaj, N., and Hammerbacher, J. (2018). Computational pipeline for the pgv-001 neoantigen vaccine trial. *Frontiers in immunology*, 8:1807.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development*, 11(6):601–617.

Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., et al. (2020). High-throughput prediction of mhc class i and ii neoantigens with mhcnuggetshigh-throughput prediction of neoantigens with mhcnuggets. *Cancer immunology research*, 8(3):396–408.

Shi, Y., Guo, Z., Su, X., Meng, L., Zhang, M., Sun, J., Wu, C., Zheng, M., Shang, X., Zou, X., et al. (2020). Deepantigen: a novel method for neoantigen prioritization via 3d genome and deep sparse learning. *Bioinformatics*, 36(19):4894–4901.

Shuchen, D. (2022). Understanding deep self-attention mechanism in convolution neural networks.

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., et al. (2018). Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic acids research*, 46(D1):D419–D427.

Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socratic.org (2022). How does a deletion mutation differ from a substitution mutation?

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan, X., Li, D., Huang, P., Jian, X., Wan, H., Wang, G., Li, Y., Ouyang, J., Lin, Y., and Xie, L. (2020). dbpepneo: a manually curated database for human tumor neoantigen peptides. *Database*, 2020.

Trolle, T., Metushi, I. G., Greenbaum, J. A., Kim, Y., Sidney, J., Lund, O., Sette, A., Peters, B., and Nielsen, M. (2015). Automated benchmarking of peptide-mhc class i binding predictions. *Bioinformatics*, 31(13):2174–2181.

UK, C. R. (2023a). Worldwide cancer incidence statistics.

UK, C. R. (2023b). Worldwide cancer statistics.

Vang, Y. S. and Xie, X. (2017). Hla class i binding prediction via convolutional neural networks. *Bioinformatics*, 33(17):2658–2665.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Venkatesh, G., Grover, A., Srinivasaraghavan, G., and Rao, S. (2020). Mhcattnnet: predicting mhc-peptide bindings for mhc alleles classes i and ii using an attention-based deep neural model. *Bioinformatics*, 36(Supplement_1):i399–i406.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2018). The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1):D339–D343.

Wang, F., Wang, H., Wang, L., Lu, H., Qiu, S., Zang, T., Zhang, X., and Hu, Y. (2022). Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences. *Briefings in Bioinformatics*, 23(3):bbab595.

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., and Freund, C. (2017). Major histocompatibility complex (mhc) class i and mhc class ii proteins: conformational plasticity in antigen presentation. *Frontiers in immunology*, 8:292.

Wood, M. A., Nguyen, A., Struck, A. J., Ellrott, K., Nellore, A., and Thompson, R. F. (2020). Neoepiscope improves neoepitope prediction with multivariant phasing. *Bioinformatics*, 36(3):713–720.

Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., Gu, X., Wu, J., Zhou, Z., and Chen, S. (2019). Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity. *Frontiers in Immunology*, page 2559.

Wu, J., Zhao, W., Zhou, B., Su, Z., Gu, X., Zhou, Z., and Chen, S. (2018). Tsnadb: a database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics, proteomics & bioinformatics*, 16(4):276–282.

Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24.

Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, 515(7528):572–576.

Yang, X., Zhao, L., Wei, F., and Li, J. (2021). Deepnetbim: deep learning model for predicting hla-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC bioinformatics*, 22(1):1–16.

Ye, Y., Wang, J., Xu, Y., Wang, Y., Pan, Y., Song, Q., Liu, X., and Wan, J. (2021). Mathla: a robust framework for hla-peptide binding prediction integrating bidirectional lstm and multiple head attention mechanism. *BMC bioinformatics*, 22(1):1–12.

Ye, Z., Li, S., Mi, X., Shao, B., Dai, Z., Ding, B., Feng, S., Sun, B., Shen, Y., and Xiao, Z. (2023). Stmhcpan, an accurate star-transformer-based extensible framework for predicting mhc i allele binding peptides. *Briefings in Bioinformatics*, 24(3):bbad164.

Zeng, H. and Gifford, D. K. (2019a). Deepligand: accurate prediction of mhc class i ligands using peptide embedding. *Bioinformatics*, 35(14):i278–i283.

Zeng, H. and Gifford, D. K. (2019b). Quantification of uncertainty in peptide-mhc binding prediction improves high-affinity peptide selection for therapeutic design. *Cell systems*, 9(2):159–166.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, L., Liu, G., Hou, G., Xiang, H., Zhang, X., Huang, Y., Zhang, X., Li, B., and Lee, L. J. (2022a). Introspect: Motif-guided immunopeptidome database building tool

to improve the sensitivity of hla i binding peptide identification by mass spectrometry. *Biomolecules*, 12(4):579.

Zhang, X., Qi, Y., Zhang, Q., and Liu, W. (2019). Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomedicine & Pharmacotherapy*, 120:109542.

Zhang, Y., Zhu, G., Li, K., Li, F., Huang, L., Duan, M., and Zhou, F. (2022b). Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction. *Briefings in Bioinformatics*.

Zhao, T., Cheng, L., Zang, T., and Hu, Y. (2019). Peptide-major histocompatibility complex class i binding prediction based on deep learning with novel feature. *Frontiers in Genetics*, 10:1191.

Zhao, W. and Sher, X. (2018). Systematically benchmarking peptide-mhc binding predictors: From synthetic to naturally processed epitopes. *PLoS computational biology*, 14(11):e1006457.

Zhou, L. Y., Zou, F., and Sun, W. (2021). Prioritizing candidate peptides for cancer vaccines by pepprmint: a statistical model to predict peptide presentation by hla-i proteins. *bioRxiv*.

Zhou, L. Y., Zou, F., and Sun, W. (2022). Prioritizing candidate peptides for cancer vaccines through predicting peptide presentation by hla-i proteins. *Biometrics*.

Zhou, W.-J., Qu, Z., Song, C.-Y., Sun, Y., Lai, A.-L., Luo, M.-Y., Ying, Y.-Z., Meng, H., Liang, Z., He, Y.-J., et al. (2019). Neopeptide: an immunoinformatic database of t-cell-defined neoantigens. *Database*, 2019.