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

Detección de neo antígenos utilizando *deep learning* en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del Cáncer

Tesis presentada por el Magister:
Vicente Enrique Machaca Arceda

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

Asesor:
Prof. Dr. Juan Carlos Gutiérrez Cáceres

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 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; gracias a ellos he logrado cumplir mis objetivos. De igual forma, dedico este trabajo a mi esposa Pamela Laguna Laura, quien me ha acompañado durante todo este proceso, me ha motivado a seguir y sobre todo me ha dado su amor, que me ha ayudado a prevalecer y siempre seguir adelante.

Abstract

En desarrollo...

Índice general

Índice de figuras

Índice de tablas

Abreviaciones

ANN	Artificial Neural Network
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
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
tRNA	Transfer Ribonucleic Acid
TCR	T-cell receptor

Capítulo 1

Introducción

1.1. Motivación

El cáncer representa el mayor problema de salud mundial (Siegel et al., 2022) y es el causante líder de muertes, solo en el 2020 se registraron alrededor de 10 millones de muertes y aproximadamente cada año 400000 niños desarrollan cáncer (WHO, 2022). Lamentablemente, a pesar de muchos esfuerzos por mitigar las muertes causadas por esta enfermedad, los métodos tradicionales basados en cirugías, radioterapias y quimioterapias tienen baja efectividad (Peng et al., 2019). En este contexto, surge el desarrollo de la inmunoterapia del cáncer, el cuál tiene el objetivo estimular el sistema inmune de un paciente. La idea es que nuestro propio sistema inmune sea capaz de reconocer las células de cáncer como agentes extraños y por consiguiente elimine dichas células. Existen varios enfoques y metodologías en la inmunoterapia del cáncer, de estos, la de mayor estudio y efectividad es el desarrollo de vacunas personalizadas (Borden et al., 2022).

El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo y depende de una correcta detección de neo antígenos. Estos neo antígenos son péptidos[1] que solo se presentan en células cancerosas; entonces, el objetivo es entrenar a los linfocitos (células T) de un paciente para que estos puedan reconocer los neo antígenos y asi activar el sistema inmune.

Determinar qué estrategia o método de detección de neo antígenos es el adecuado o en qué circunstancias conviene la aplicación de alguno, es muy importante para el desarrollo de vacunas personalizadas (Mattos et al., 2020; Peng et al., 2019). Sin embargo, a pesar de los esfuerzos de los investigadores en desarrollar métodos y herramientas, menos del

[1] Secuencias cortas de aminoacidos.

3 % de los neo antígenos detectados logran activar a las células T (sistema inmune) (Mattos et al., 2020). De esta forma, es relevante que se continue con la investigación y desarrollo de nuevos métodos que permitan detectar neo antígenos.

1.2. Problema

Los neo antígenos son peptídos mutados específicos de tumores y son considerados los principales causantes de una respuesta inmune (Borden et al., 2022; Chen et al., 2021a; Gopanenko et al., 2020). Es así que surgen varios esfuerzos e investigación en la Inmunoterapia del cáncer, concentradas en el estudio y detección de neo antígenos. En la actualidad existen tres clases de tratamientos basados en la representación y expresión de neo antígenos: vacunas personalizadas, terapias adoptivas de células T y *immune checkpoint inhibitors*. De los métodos mencionados anteriormente, el desarrollo de vacunas personalizadas 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.

Según lo mencionado anteriormente, la detección de neo antígenos es un factor clave en el desarrollo de vacunas personalizadas. En este proceso el compuesto *Major Histocompatibility Complex* (MHC), juega un papel muy importante, es el encargado de presentar los péptidos a la células T (Hashemi et al., 2022). Para el caso de células humanas el gen MHC es conocido como Human Leukocyte Antigens (HLA) y es polimórfico, se cree que existen las 10000 diferentes *HLA-I alleles* (Abelin et al., 2017), esto complica mucho más la detección de neo antígenos.

El ciclo de vida de un neo antígeno para células con núcleo podría resumirse como: primero una proteína es degradada en péptidos en el citoplasma de las células, luego los péptidos se enlazan a la molecula MHC (*pMHC binding*), luego este compuesto sigue un trayecto hasta llegar a la membrana de la célula (*pMHC presentation*), finalmente el compuesto pMHC es reconocido por el T-cell Receptor (TCR) de las células T y así si activaría el sistema inmune. Además, el número de posibles péptidos enlazables a MHC son entre 1000 a 10000, esto es el 0.1 % de los posibles péptidos de 9 aminoacidos[2] (Abelin et al., 2017). En este proceso, el objetivo es detectar los péptidos (neo antígenos) que llegan a la membrana de la célula, luego con ayuda de procedimientos de biotecnología, se entrena a las células T de un paciente para que aprenda a reconocer los neo antígenos.

[2]La mayoría de péptidos enlazados a moléculas MHC-I tienen 9 aminoácidos, se suele utilizar el termino *n-mer* para referirse a péptidos de n aminoácidos.

El problema de *pMHC binding* está casi solucionado con una precisión de 0.98 por parte de la herramienta NetMHCPan 4.1 (Reynisson et al., 2020). Sin embargo, no es bueno limitar la detección de neo antígenos solo al problema de *pMHC binding*, porque la mayoría de estos compuestos no llegan a la membrana (Mill et al., 2022), a este problema se le conoce como *pMHC presentation*. Por ejemplo, se sabe que menos del 5 % de péptidos detectados llegan a la membrana (Mattos et al., 2020; Mill et al., 2022; Bulik-Sullivan et al., 2019; Bassani-Sternberg et al., 2015; Yadav et al., 2014). Además, existen herramientas como NeyMHC, NetMHCpan y MHCFlurry que tienen un buen desemepeño en *pMHC binding*, pero con resultados pobres en *pMHC presentation* (Bulik-Sullivan et al., 2019).

1.2.1. Formulación del problema

Menos del 5 % de péptidos detectados en *pMHC binding*, llegan a la membrana de la células, para que luego sean reconocidos por las células T. El proceso por el cúal un péptido enlazado a MHC llegue a la membrana es conocido como *pMHC presentation*, pero en este problema las propuestas recientes solo llegan a un 0.61 de precisión y 0.4 de *recall*. En este contexto, la tesis se enfoca en el problema de *pMHC presentation*, considerándolo como un problema de clasificación binaria, y tomando como entrada la secuencia de aminoácidos del péptido y la secuencia de aminoácidos de la proteína MHC.

1.3. Objetivos

1.3.1. Objetivo General

Proponer un método basado en *deep learning* para la detección de neo antígenos, enfocados en el problema de *pMHC presentation*.

1.3.2. Objetivos específicos

(a) Realizar una revisión sistemática de la literatura e implementar los métodos con mejor desempeño en la detección de neo antígenos.

(b) Proponer e implementar un método basado en *deep learning* para la detección de neo antígenos.

(c) Evaluar el método propuesto en bases de datos publicas.

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 *deep learning* para la detección de neo antígenos.

(b) Se ha desarrollado un nuevo método para la detección de neo antígenos, este método utiliza redes neuronales *transformer* y *transfer learning*.

1.5. Organización del Trabajo

En el Capítulo 2 se presentan los conceptos básicos sobre Bioinformaica 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. Este capítulo es el resultado de un *review* utilizando una búsqueda sistemática de la literatura de los métodos basados en *deep learning* para la detección de neo antígenos.

El Capitulo 4, presenta la propuesta de la tesis. Esta se basa en un nuevo método basado en redes neuronales *transformers*. Debido a la falta de muestras, para acelerar el entrenamiento y mejora la generalización, se utilizo *transfer learning* de dos redes neuronales pre entrenadas: TAPE (Rao et al., 2019) y ESM-1b (Rives et al., 2021).

Luego, en el Capítulo 5, se presentan los resultados de la investigación. En este punto se evalúa el método propuesto en una base de datos recolectada de varias investigaciones.

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.

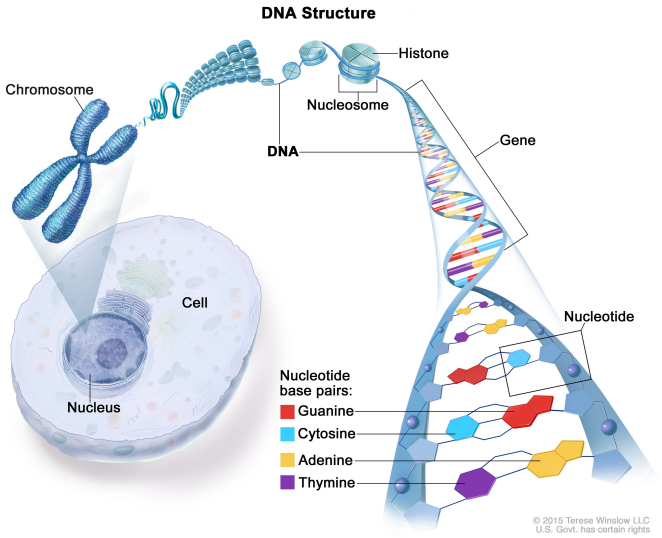

FIGURA 2.1: Localización y estructura del DNA. Fuente: NCI (2022).

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).

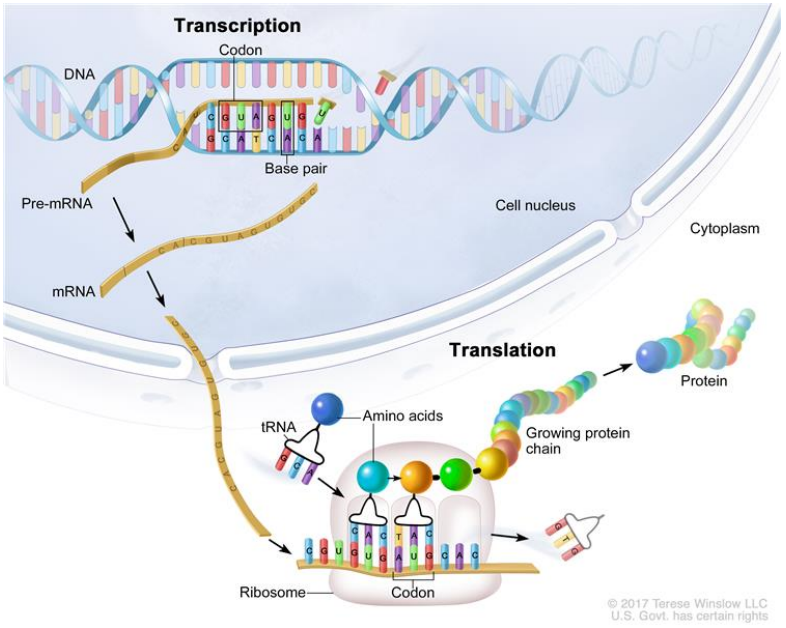

FIGURA 2.2: Transcripción y traducción. Fuente: NCI (2020).

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.

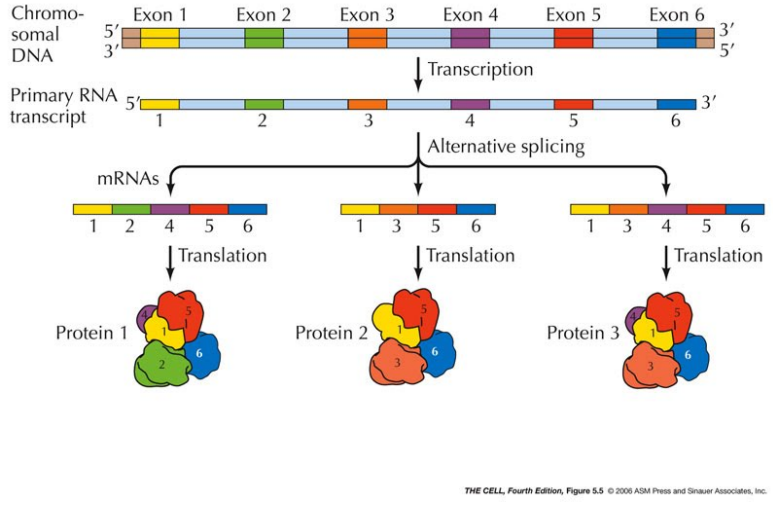

FIGURA 2.3: *Alternative Splicing*. Fuente: NCI (2020).

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., 2021a; 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).

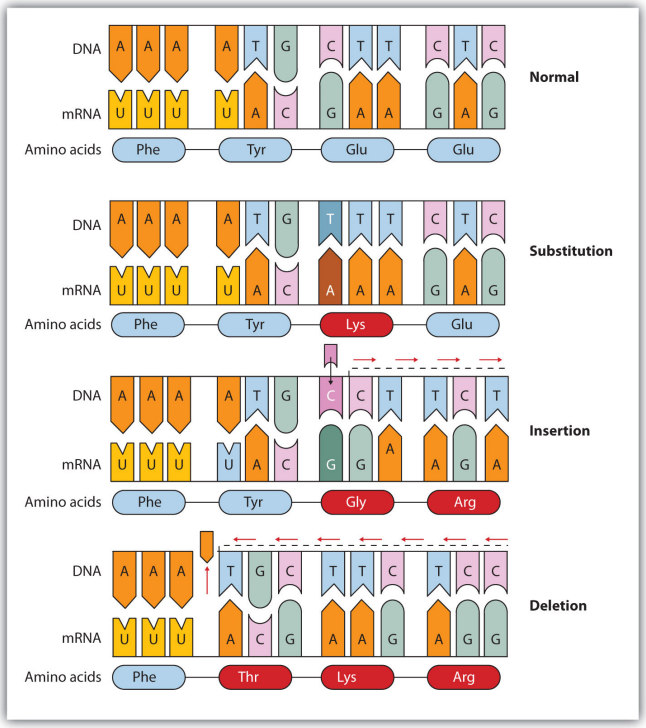

FIGURA 2.4: Ejemplos de SNV en el DNA. Fuente: Socratic.org (2022)

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).

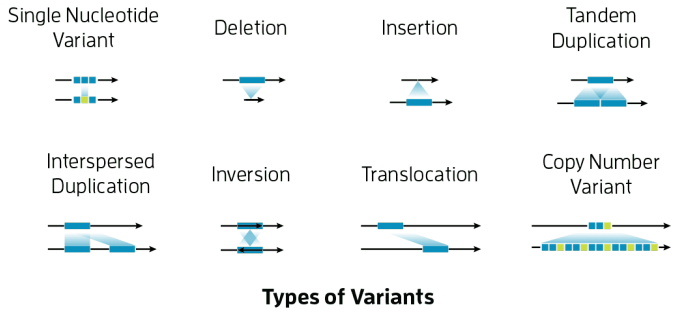

FIGURA 2.5: Ejemplos de variaciones en el DNA. Fuente: PacBio (2021)

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).

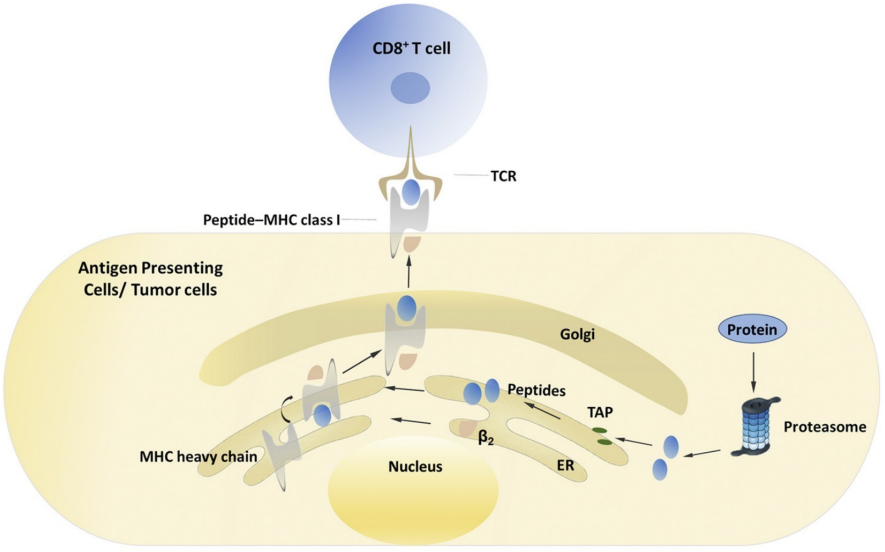

FIGURA 2.6: Presentación de antígenos por MHC-I. Fuente: Zhang et al. (2019)

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).

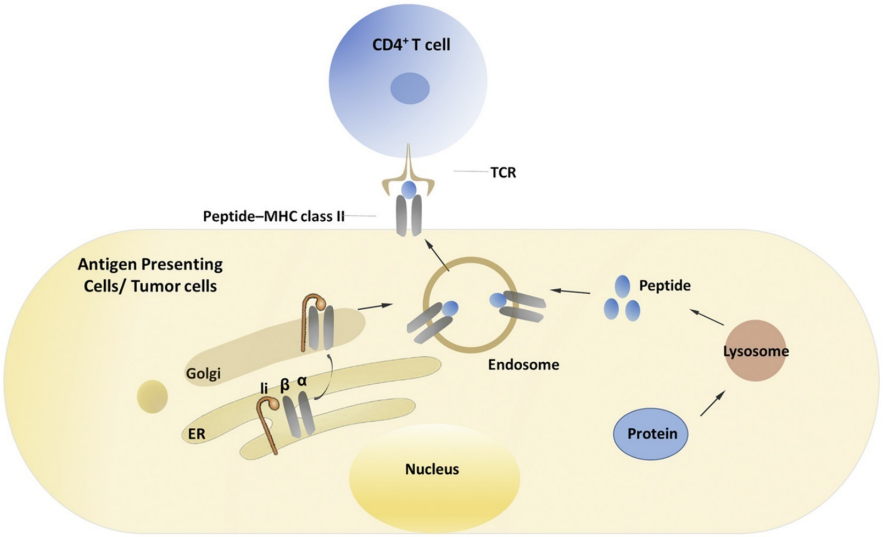

FIGURA 2.7: Presentación de antígenos por MHC-II. Fuente: Zhang et al. (2019)

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. (2021a); Hao et al. (2021).

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

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

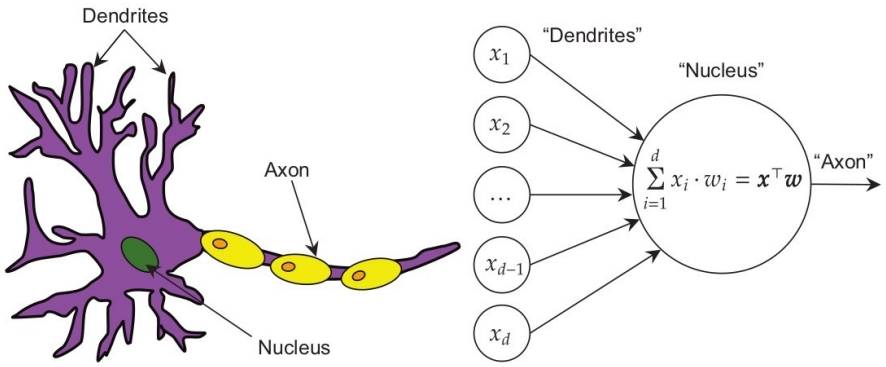

FIGURA 2.8: Representación de una neurona. Fuente: Raff (2022).

El perceptron, es capaz de solucionar varios problemas, pero para casos complejos puede formar una red, como se presenta en la Figura 2.9.

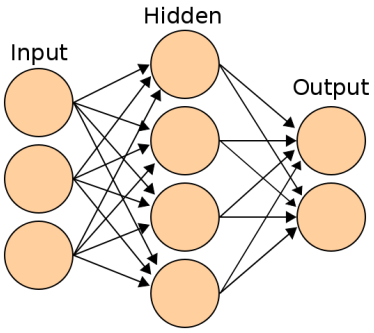

FIGURA 2.9: Representación de una red neuronal.

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 2.10, 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).

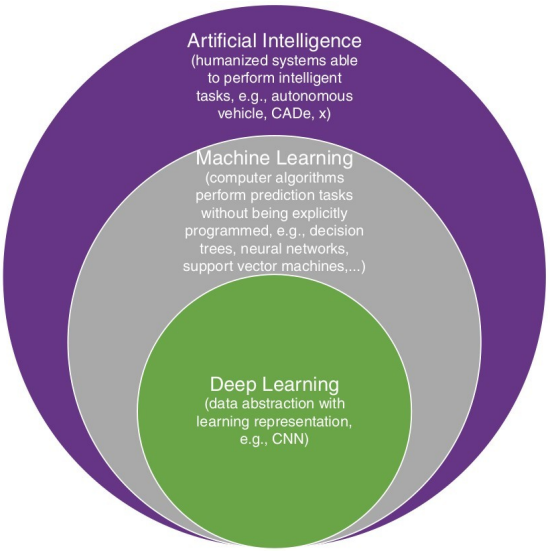

FIGURA 2.10: Relación entre Inteligencia Artificial, *Machine Learning* y *Deep Learning*. Fuente: El Naqa and Murphy (2022).

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 2.11). Existen muchos tipos de *Deep Feedforward networks*, estas serán detalladas en los siguientes apartados.

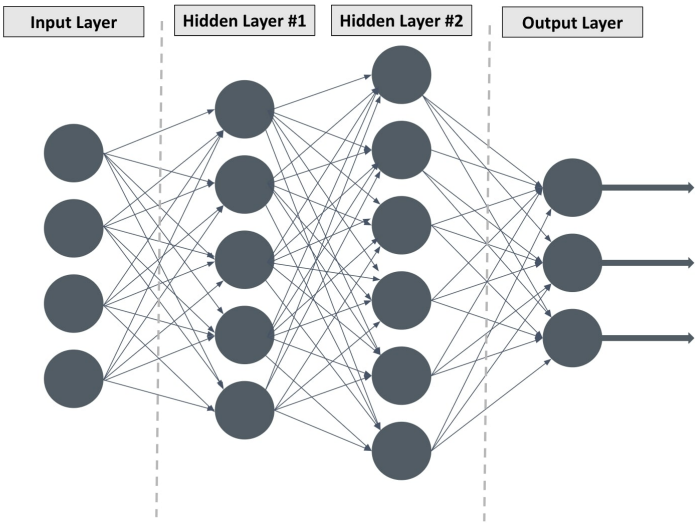

FIGURA 2.11: Representación de un *Deep Feedforward Network*. Fuente: El Naqa and Murphy (2022).

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 2.12. 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.

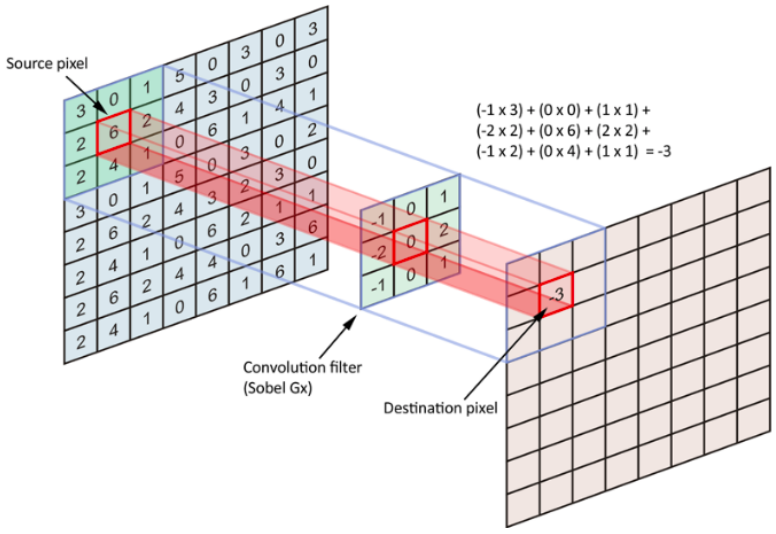

FIGURA 2.12: Ejemplo de una convolución en procesamiento de imágenes. Fuente: Shuchen (2022).

Con inspiración en la operación de convolución, se plantean las CNN por primera vez por LeCun et al. (1998). En la Figura 2.13, 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).

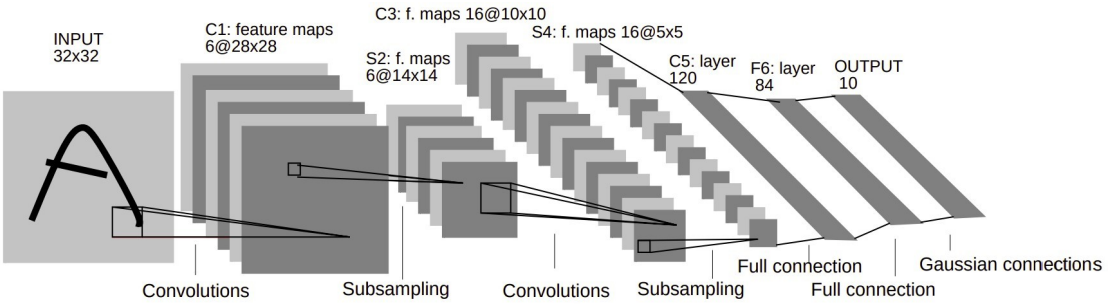

FIGURA 2.13: Arquitectura de LeNet-5, una CNN para el reconocimiento de digitos. Fuente: LeCun et al. (1998).

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 2.14, 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).

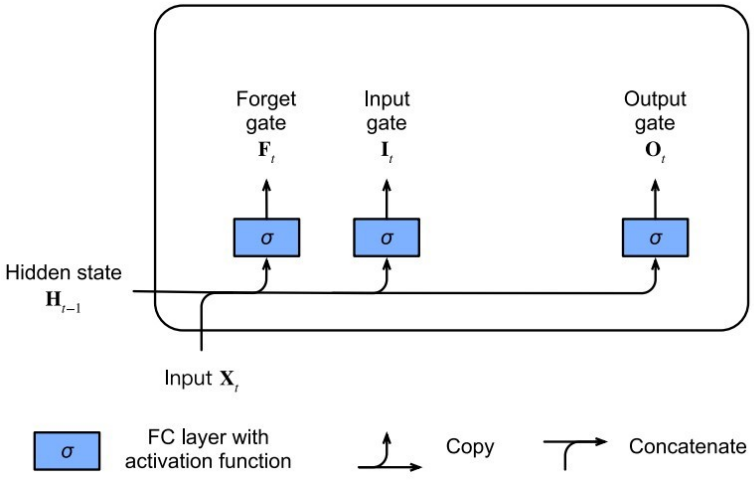

FIGURA 2.14: Ejemplo del procesamiento del *input gate*, *forget gate* y *output gate* de LSTM. Fuente: Zhang et al. (2021).

2.4.4. *Transformers*

Los *Transformers* son propuestas por Vaswani et al. (2017), para dar solución al problema de *long-range dependency*. Por ejemplo el autor comenta: "The Transformer is

the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution". Del enunciado anterior, *transduction* hace referencia a la conversión secuencias de entrada hacia otro formato. Otro termino interesante es *self-attention* (Figura 2.15), este permite al modelo mirar hacia otras palabras en la secuencia de entrada para tener un mejor entendimiento de cierta palabra en la secuencia (Kelvin, 2022)´.

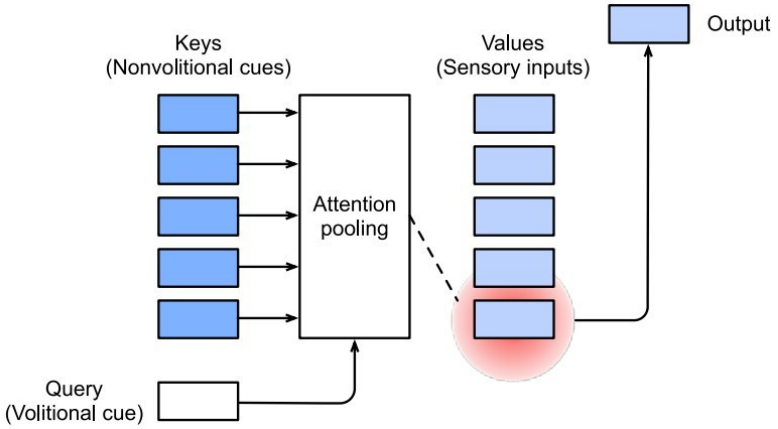

FIGURA 2.15: ejemplo del mecanismo de atención de una red *Transformer*. Fuente: Zhang et al. (2021).

2.4.5. *BERT*

Bidirectional Encoder Representations from Transformers (BERT), propuesta por Devlin et al. (2018), esta 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).

Capítulo 3

Estado del Arte

Con el objetivo de contextualizar las contribuciones de la presente tesis, fueron analisados .....

3.1. Neo antígenos

El cáncer es el mayor problema de salud del mundo y la segunda enfermedad que causa más muertes. Por ejemplo, en el año 2021 se reportaron 1.8 millones de nuevos casos y 608.570 muertes (Siegel et al., 2022). Los tratamientos tradicionales basados en cirugías, radioterapias, quimioterapias tienen baja efectividad (Peng et al., 2019) y se buscan nuevas alternativas para tratar esta enfermedad.

En recientes años, se ha planteado el uso de nuestro propio sistema inmune para eliminar las células cancerosas (inmunoterapia del cáncer). En esta área de estudio, surge la posibilidad de crear vacunas personalizadas que activen el sistema inmune de un paciente y así se elimine las células enfermas. Este proceso 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 3.1).

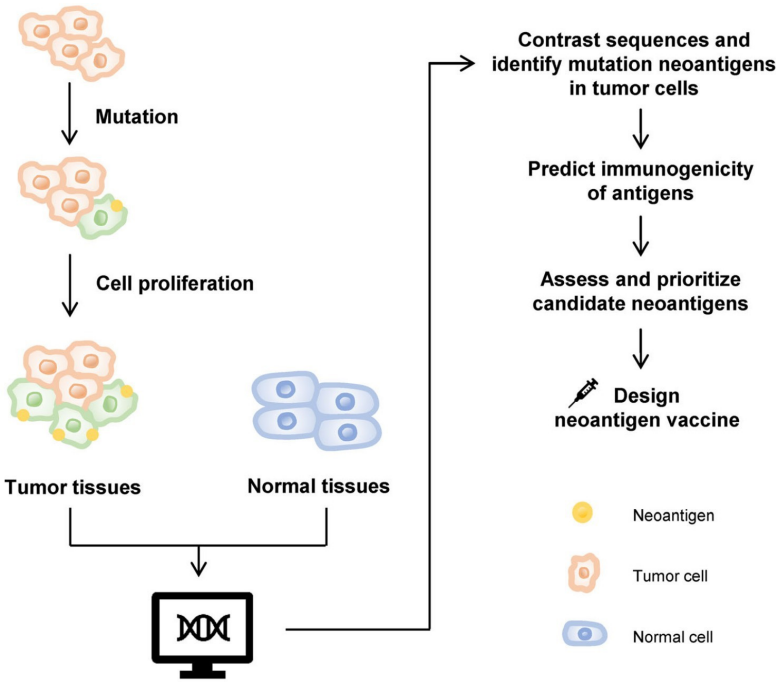

FIGURA 3.1: Proceso para la generación de vacunas personalizadas. Fuente: (Mattos et al., 2020)

Existen herramientas de Software que se basan en la predicción del enlace entre las moléculas Major Histocompatibility Complex (MHC) y péptidos (posibles neo antígenos). La predicción de estos enlaces es importante para determinar qué péptidos pueden representar neo antígenos. Entre las principales propuestas que utilizan Regresión lineal y Redes Neuronales, tenemos: NetMHC4 (Stevanović et al., 2017), NetMHCpan4 (Robbins et al., 2013), PickPocket (Tran et al., 2014), NetMHCcons (Castle et al., 2012), NetMHCIIpan (Yadav et al., 2014). También, existen alternativas como NeonMHC (van Rooij et al., 2013) que utilizan Redes Neuronales Convolucionales. Luego, otras propuestas se basan en la mejorar la predicción de un posible neo antígeno (Lu et al., 2021; Hao et al., 2021; Lang et al., 2021; Chen et al., 2021b; Yang et al., 2021; Li et al., 2021). Una desventaja de estos métodos, es referente a la necesidad de contar de antemano con posibles peptidos, esto complica una propuesta *end-to-end* que tome como entrada una secuencia de ADN.

Debido a la complejidad del proceso y la gran cantidad de métodos desarrollados, se ha desarrollado software y *pipelines* que pretenden facilitar el uso de estas herramientas. Entre las más recientes tenemos: Somaticseq (Fang et al., 2015), NeoPredPipe (Schenck et al., 2019), CloudNeo (Bais et al., 2017), MuPeXI (Bjerregaard et al., 2017), NeoepitopePred (Tran et al., 2015), Neoepiscope (Yossef et al., 2018), pVACtools (Hundal et al., 2020) y NeoFuse (Gros et al., 2016). Estas herramientas en su mayoría toman

como entrada archivos Variant Calling Files (VCF) y archivos de alineamiento Bam, para la detección de mutaciones (inserciones, eliminaciones y fusión de genes) y posibles neo antígenos. Si bien es cierto, los *pipelines* mencionados anteriormente son propuestas *end-to-end*, el acierto es bajo y son dificiles de desplegar.

A pesar de la gran cantidad de métodos y herramientas no existe un método que pueda ser definido como el de mejor desempeño (Mattos et al., 2020), incluso a pesar de ya haberse desarrollado algunos *benchmarkings*. Por ejemplo, en el 2015 se desarrolló una comparativa de los métodos SMM, ANN, ARB y NetMHCpan (Trolle et al., 2015), sin ninguna conclusión sobresaliente. Luego en el 2018 y 2019 se vuelve a intentar realizar otra comparativa (Bonsack et al., 2019; Zhao and Sher, 2018), sin lograr determinar a un método con mayor desempeño. Tambien se han desarrollado *surveys* sobre como los métodos computacionales pueden tener beneficios clínicos (Mattos et al., 2020) y sus principales desafios (Chen et al., 2021a).

Finalmente, en la Tabla 3.1, se presenta un resumen de los métodos basados en *MHC-binding* y *pipelines*. Tambien, indicamos cuales son *open source*.

Tabla 3.1: Resumen de los métodos de detección de neo antígenos.

Nombre	**MHC-binding**	**Método**	**Open source**
NetMHC4	✓	ANN	
NetMHCpan4	✓	ANN	
PickPocket	✓	ANN	
NetMHCcons	✓	ANN	
NetMHCIIpan	✓	ANN	
NeonMHC	✓	CNN	
DeepNetBim	✓	Deep learning	✓
DeepImmuno	✓	CNN	
NeoPredPipe		pipeline	✓
CloudNeo		pipeline	
MuPeXI		pipeline	
NeoepitopePred		pipeline	
Neoepiscope		pipeline	
pVACtools		pipeline	✓
NeoFuse		pipeline	✓

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.

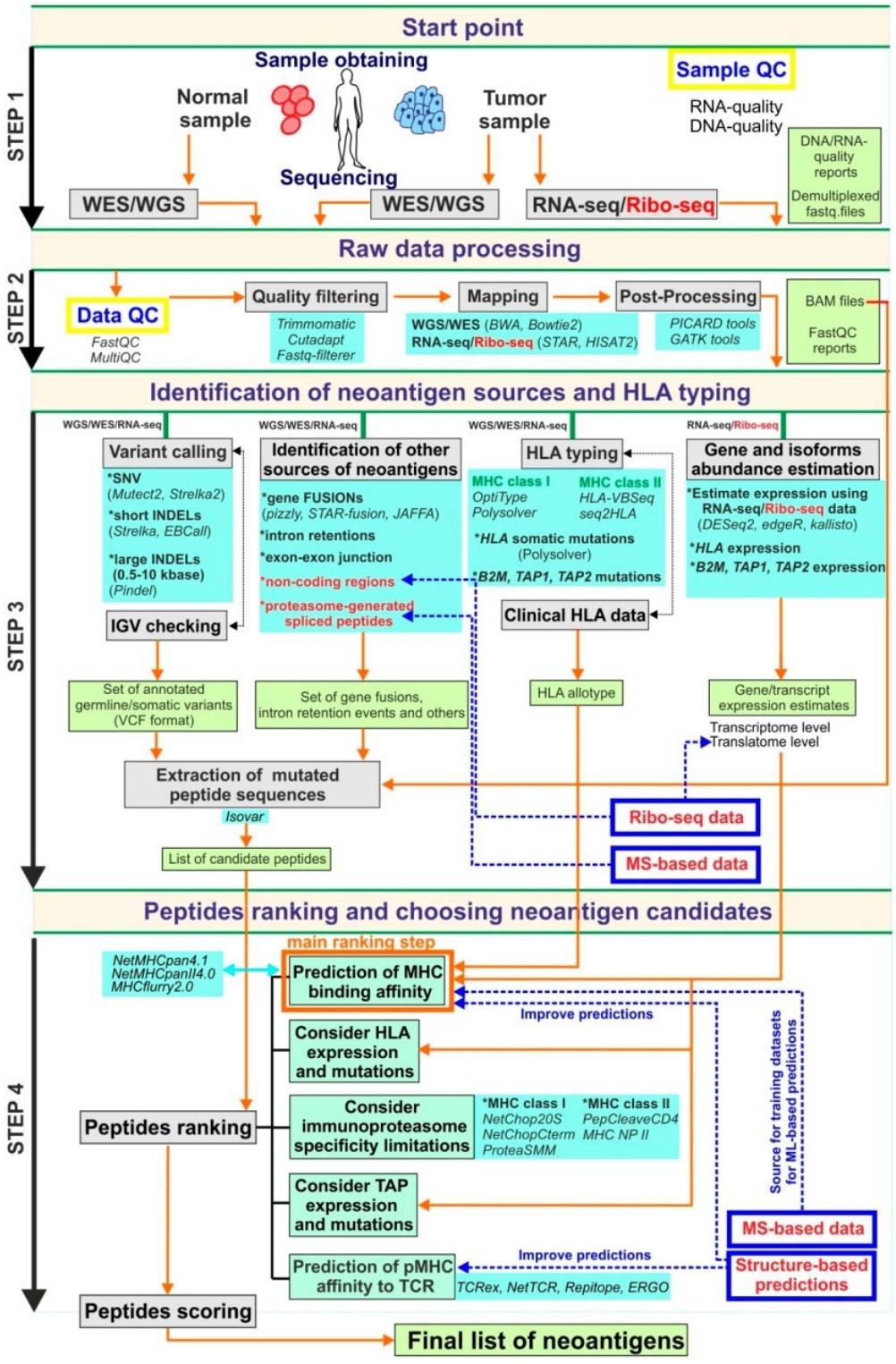

FIGURA 4.1: Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: Gopanenko et al. (2020).

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., 2019), 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*.

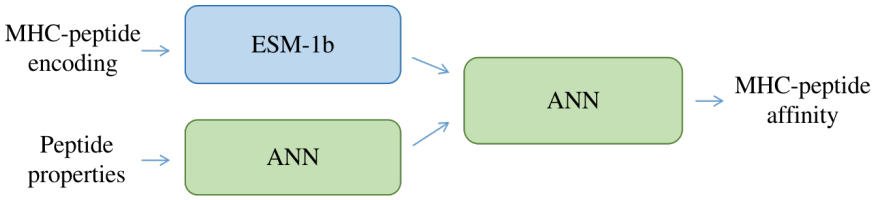

FIGURA 4.2: Propuesta de *transfer learning* de ESM-1b y una red neuronal paralela para la predicción de la afinidad entre un péptido y MHC (peptide MHC binding).

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

Resultaods

Capítulo 6

Conclusiones

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.

Bais, P., Namburi, S., Gatti, D. M., Zhang, X., and Chuang, J. H. (2017). Cloudneo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*, 33(19):3110–3112.

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.

Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z., and Eklund, A. C. (2017). Mupexi: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy*, 66(9):1123–1130.

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.

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.

Castle, J. C., Kreiter, S., Diekmann, J., Löwer, M., Van de Roemer, N., de Graaf, J., Selmi, A., Diken, M., Boegel, S., Paret, C., et al. (2012). Exploiting the mutanome for tumor vaccination. *Cancer research*, 72(5):1081–1091.

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

Chen, R., Fulton, K. M., Twine, S. M., and Li, J. (2021b). Identification of mhc peptides using mass spectrometry for neoantigen discovery and cancer vaccine development. *Mass spectrometry reviews*, 40(2):110–125.

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.

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.

Fang, L. T., Afshar, P. T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J. C., Gibeling, G., Barr, S., Asadi, N. B., Gerstein, M. B., et al. (2015). An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome biology*, 16(1):1–13.

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.

Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., Prickett, T. D., Gartner, J. J., Crystal, J. S., Roberts, I. M., et al. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nature medicine*, 22(4):433–438.

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*.

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.

Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Xia, H., Wollam, A. T., Liu, C. J., Zhao, S., Feng, Y.-Y., Graubert, A. P., et al. (2020). pvactools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer immunology research*, 8(3):409–420.

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

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.

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

Lang, F., Riesgo-Ferreiro, P., Löwer, M., Sahin, U., and Schrörs, B. (2021). Neofox: annotating neoantigen candidates with neoantigen features. *Bioinformatics*, 37(22):4246–4247.

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. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J. V., Gibbons, D. L., Wang, J., et al. (2021). Deep learning-based prediction of the t cell receptor–antigen binding specificity. *Nature Machine Intelligence*, 3(10):864–875.

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.

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.

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.

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

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.

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.

Raff, E. (2022). *Inside Deep Learning*.

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.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). 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.

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).

Robbins, P. F., Lu, Y.-C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., Lin, J. C., Teer, J. K., Cliften, P., Tycksen, E., et al. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive t cells. *Nature medicine*, 19(6):747.

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.

Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A., and Anderson, A. R. (2019). Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC bioinformatics*, 20(1):1–6.

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

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer statistics, 2022. *CA: a cancer journal for clinicians*.

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?

Stevanović, S., Pasetto, A., Helman, S. R., Gartner, J. J., Prickett, T. D., Howie, B., Robins, H. S., Robbins, P. F., Klebanoff, C. A., Rosenberg, S. A., et al. (2017). Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer. *Science*, 356(6334):200–205.

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.

Tran, E., Ahmadzadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P. F., Gartner, J. J., Zheng, Z., Li, Y. F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*, 350(6266):1387–1390.

Tran, E., Turcotte, S., Gros, A., Robbins, P. F., Lu, Y.-C., Dudley, M. E., Wunderlich, J. R., Somerville, R. P., Hogan, K., Hinrichs, C. S., et al. (2014). Cancer immunotherapy based on mutation-specific cd4+ t cells in a patient with epithelial cancer. *Science*, 344(6184):641–645.

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.

van Rooij, N., van Buuren, M. M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., van Dijk, L. J., Behjati, S., Hilkmann, H., El Atmioui, D., et al. (2013). Tumor exome analysis reveals neoantigen-specific t-cell reactivity in an ipilimumab-responsive melanoma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 31(32).

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.

WHO (2022). Cancer.

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.

Yossef, R., Tran, E., Deniger, D. C., Gros, A., Pasetto, A., Parkhurst, M. R., Gartner, J. J., Prickett, T. D., Cafri, G., Robbins, P. F., et al. (2018). Enhanced detection of neoantigen-reactive t cells targeting unique and shared oncogenes for personalized cancer immunotherapy. *JCI insight*, 3(19).

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

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.

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