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) Una .....;

(b) Una .....;

(c) Una .....;

1.5. Organización del Trabajo

En el Capítulo 2 se presentan los conceptos básicos ....

en el Capítulo 3 se describen los trabajos relacionados a la presente tesis.

.....

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

Capítulo 2

Marco Conceptual

En este capítulo se presentarán conceptos necesarios para el correcto entendimiento de esta tesis.

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).

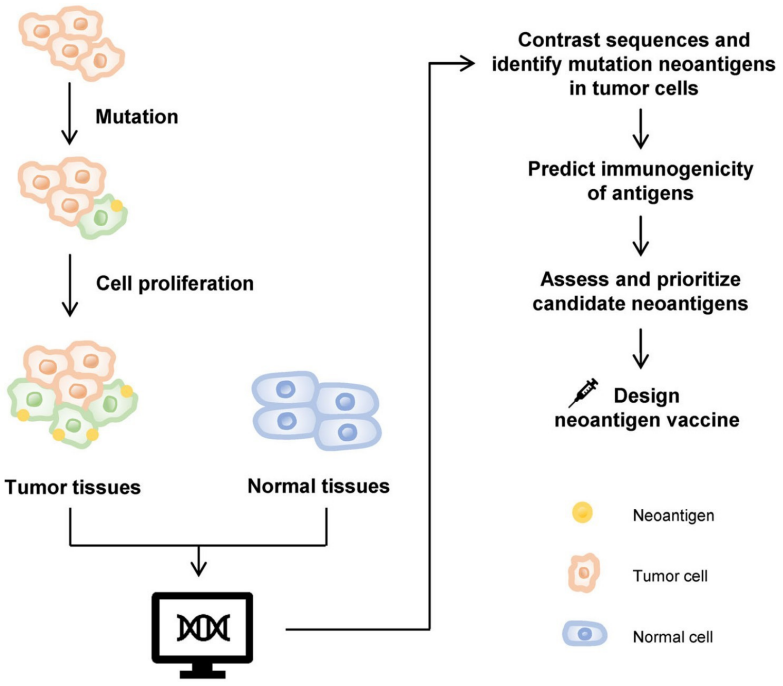

FIGURA 3.1: Proceso para la generación de vacunas personalizadas. Fuente: (Mattos et al., 2020)

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.

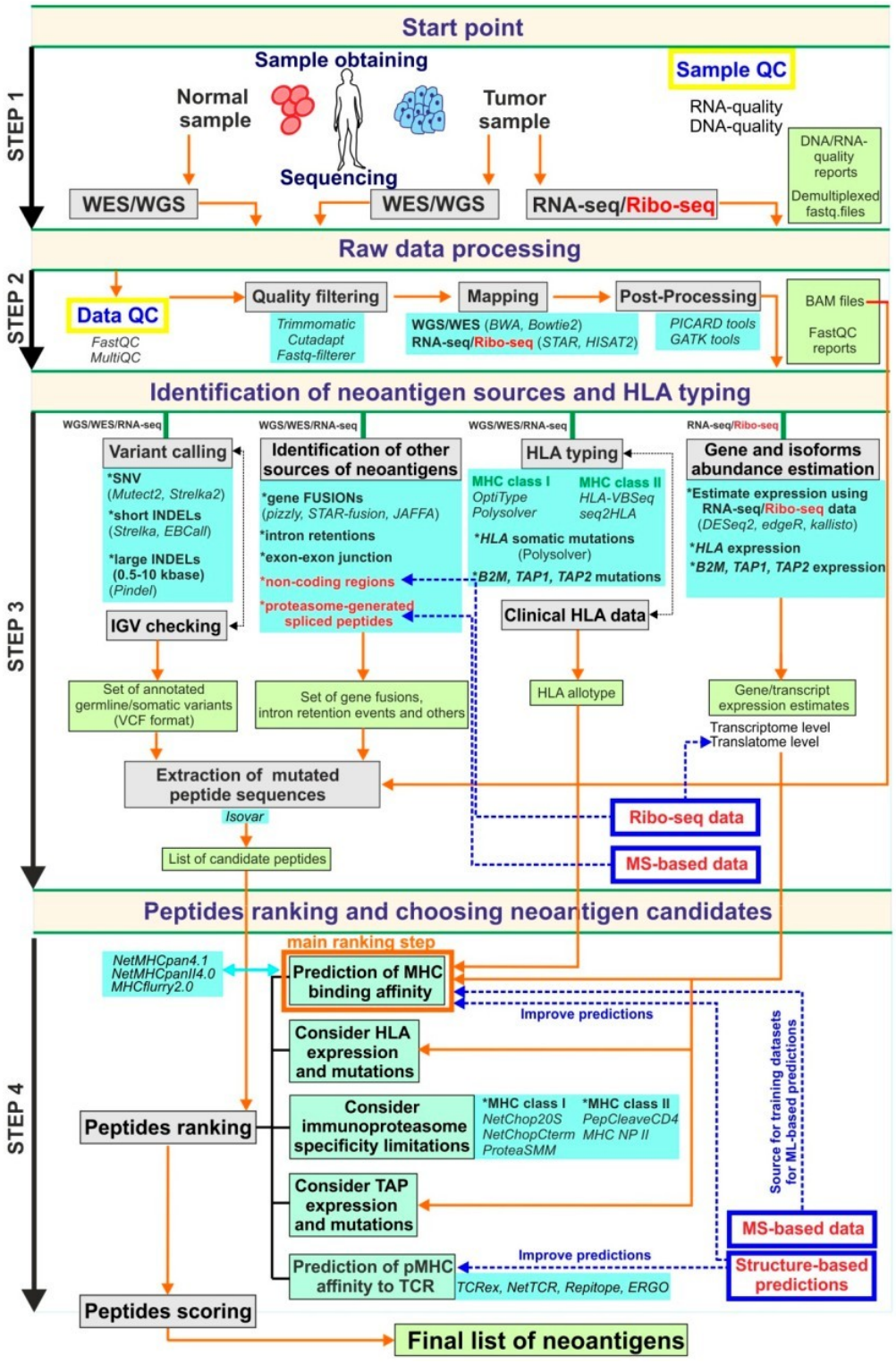

FIGURA 4.1: Proceso general utilizado para la detección de neo antígenos a partir de secuencias de DNA. Fuente: Gopanenko et al. (2020).

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*.

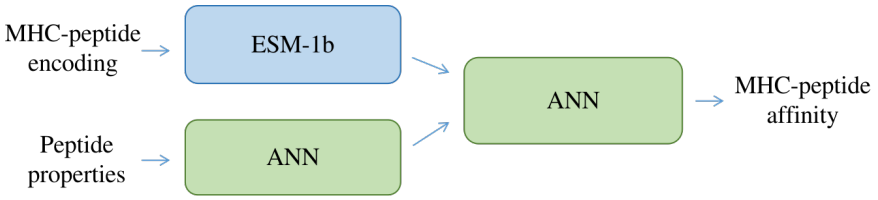

FIGURA 4.2: Propuesta de *transfer learning* de ESM-1b y una red neuronal paralela para la predicción de la afinidad entre un péptido y MHC (peptide MHC binding).

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).

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.

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.

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*.

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.

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*.

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.

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.

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.

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*.

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.

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.

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.

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

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.

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).

WHO (2022). Cancer.

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).

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