ESCUELA PROFESIONAL DE INGENIERÍA DE SOFTWARE

APLICATIVO WEB PARA LA PREDICCIÓN DE UNIÓN PÉPTIDO-MHC EMPLEANDO UN MODELO DE REDES DE ATENCIÓN GRÁFICA

Jose Alfredo Grados Chuquitaype

[Asesor]

Bachiller

AREQUIPA – PERÚ

2024

Dedicatoria

Agradecimientos

Índice General

Índice de Abreviaturas y Siglas

Índice de Tablas

Índice de Figuras

Glosario de Términos

Resumen

El resumen posee un conjunto de elementos los cuales dan una visión clara del trabajo. El resumen debe contener lo siguiente:

- Delimitación de la Investigación
- El propósito de la investigación.
- Forma de lograrlo.
- Criterios que justifican al estudio.
- Fundamentación teórica empleada en el estudio.
- Metodología de investigación empleada, se hace mención al tipo de estudio, el diseño de investigación, la modalidad de investigación (si es necesario), técnica empleada para la recolección de datos, tipo de instrumento de recolección usado, validez y confiabilidad.
- Muy breve referencia a los resultados.
- Señalamiento de las conclusiones más significativas

La forma de redacción es en pasado impersonal, por ejemplo: una vez que recabamos la información (esta forma es incorrecta), en vez de ello se debe decir: una vez que se recabó la información (estas la forma correcta).

Abstract

Palabras clave

De 3 a 5 palabras clave del trabajo.

Capítulo I - Problemática del Proyecto

1.1. Contexto del problema

El cáncer representa el mayor problema de salud mundial y es la principal causa de muerte, en el 2022 según GLOBOCAN(Global Cancer Observatory) se reportaron 19 976 499 casos de cáncer y de estos casos para ese mismo año se reportaron 9 743 832 de personas fallecidas[1]. En la figura 1 podemos ver que la mayor cantidad de casos reportados se sitúan en Asia pero eso no quita la enorme cantidad que hay de casos en todo el mundo.

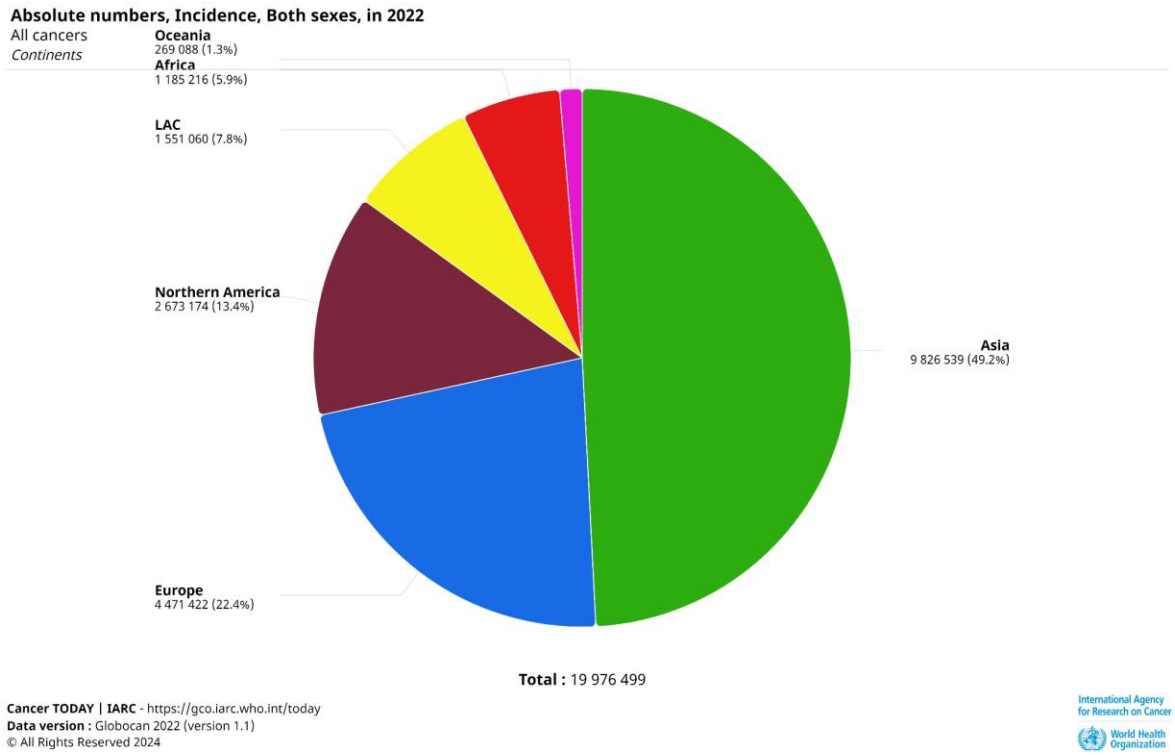

Figura1. Incidencia de casos de cáncer en 2022 en Latinoamérica.

Además, los métodos tradicionales basados en cirugías, radioterapias y quimioterapias tienen baja efectividad[7]. Sin embargo, en las últimas décadas, se ha experimentado un cambio de paradigma notable gracias al surgimiento de la inmunoterapia contra el cáncer. Este enfoque terapéutico innovador se basa en aprovechar el poder del sistema inmunitario del propio paciente para combatir y controlar el crecimiento tumoral[2].

Pero a pesar de los avances significativos en el campo de la inmunoterapia contra el cáncer, el diseño y la optimización de vacunas especializadas siguen siendo un desafío importante que todavía no se ha podido superar. Uno de los principales obstáculos en este sentido es la predicción precisa

de la unión péptido-MHC, un proceso fundamental para la activación del sistema inmunitario adaptativo. Los enfoques convencionales basados en técnicas bioinformáticas y modelos de aprendizaje automático han demostrado limitaciones en la captura de la complejidad y la dinámica de estas interacciones moleculares.

1.2. Antecedentes y/o estado del arte

1.2.1. Antígenos y neoantígenos

En [9] los autores recopilaron toda la historia de los neoantígenos desde sus primeros descubrimientos hace aproximadamente 20 años hasta la actualidad donde ya se sabe la gran importancia que tienen en tratamientos contra el cáncer. Menciona que al inicio los primeros indicios vinieron de muestras que se hicieron en ratones y de ahí se descubrieron las funciones del sistema inmunitario a base de las células B y T, además de que estas reconocen a los neoantígenos para poder combatirlos y en el mejor de los casos eliminarlos. Posteriormente se descubrió que en diferentes tipos de cáncer (melanoma, cáncer de pulmón , etc.) se generaban neoantígenos y probaron introduciendo ciertas cantidades de neoantígenos, dando como resultado que en ciertos casos habían mejoras. También estudios recientes de Lennerz(1995) dieron como conclusión que hay un alto grado de individualidad en las respuestas antitumorales, en pocas palabras, los tratamientos debían hacerse individualmente ya que los neoantígenos generados no eran los mismos en todas las personas y también dependían del tipo de cáncer que esta misma tenía.

1.2.2. Las células T

Según [3], [4], [5], [6], [7], [8], [9] y [10] mencionan que las células encargadas de este proceso de respuesta inmune contra neoantígenos son las células T. Según investigaciones [9] mientras se realizaban las pruebas se detectó que las células T CD4 y CD8 tenían respuestas contra antígenos mutados en pacientes con melanoma. Además de estas también se hablaba de la célula T CTLA-4 la cuál demostró buenos resultados en pacientes con melanoma metastásico pero estos datos se quedaron en casos muy reducidos. Posteriormente, también se ha demostrado la reactividad de las

células T específicas de neoantígeno en otras neoplasias malignas humanas, incluido el NSCLC, el cáncer de ovario, el carcinoma de células escamosas de cabeza y cuello, el colangiocarcinoma y el cáncer colorrectal[9], [3].

En [12] nos muestran un sistema microscópico robótico para examinar las células T y que puede servir para terapias de inserción de células T. donde se trata de examinar a la misma usando un microscopio siguiendo una secuencia de pasos establecidos para ayudar con la automatización de experimentos de activación de células T.

1.2.3. Secuenciación de Nueva Generación (NGS)

En [6] se menciona una nueva tecnología llamada NGS que hace referencia a una nueva tecnología capaz de analizar gran cantidad de ADN de forma masiva y trata de darle un enfoque para usarlo en predicciones para la rama. Combinado con la mayoría de enfoques existentes se puede llegar a tener resultados positivos.

1.2.4. Proceso de predicción de neoantígenos reconocidos por células T

Tenemos que saber que sólo una fracción menor de las mutaciones en tumores humanos conducen a la inducción de reactividad de las células T, por lo que se necesita realizar un proceso en el cuál podamos saber qué neoantígenos potenciales pueden llegar a ser neoantígenos validados para ser reconocidos por las células T. El proceso completo lo explican en [9]:

1. **Identificación de mutaciones tumorales:** Primero se tiene como entrada las dos muestras de ADN, normal y tumoral, de los cuáles se secuencian sus exomas para identificar las mutaciones. En [2] se muestran procesos sobre la recolección y secuenciación de muestras incluidos los tejidos necesarios, el método de recolección de tejidos y los tipos de secuenciación(Cada).
2. **Análisis de expresión de mutaciones:** Luego se analiza la expresión de variantes tumorales utilizando datos de secuenciación de ARN.

3. **Predicción de neoantígenos:** En este paso se predicen los futuros neoantígenos dividiendo las entradas a pedazos más pequeños llamados péptidos y comparando las muestras normales con las tumorales. Esta paso es opcional y se toma más como un filtrado para el siguiente paso; para este paso se pueden usar más de un enfoque ya sea sólo eligiendo una o haciendo que se complementen entre todos, entre los más conocidos tenemos:
    a. **Análisis Computacional In Silico:** Este enfoque intenta identificar posibles neoantígenos dividiendo las cuestiones en dos preguntas. ¿Es probable que una mutación particular codifique un péptido que pueda ser presentado por una molécula de MHC de un paciente determinado? Si es así, ¿es probable que este complejo péptido-MHC sea reconocido por el repertorio de TCR disponible? En [5] mencionan que este proceso ha dinamizado la identificación de neoantígenos en una variedad de tumores.
    b. **Espectrometría de masas:** Este enfoque intenta identificar neoantígenos en los espectros de masas obtenidos de miles de secuencias peptídicas de otros pacientes, estos se comparan con los espectros esperados de todos los neoantígenos posibles, según los datos de secuenciación de ARN/exoma del cáncer (es decir, en comparación con la lista de todos los péptidos recién formados resultantes de alteraciones del ADN que también se utiliza como entrada para canales de predicción computacional). Según [6] se centra en el análisis directo de las proteínas y péptidos presentes en las células tumorales eso permite determinar su secuencia y sus características, lo que permite no solo identificar los neoantígenos derivados de mutaciones somáticas, sino también aquellos generados por eventos como el splicing proteasómico.
    c. **Enfoques basados en la genómica y líneas bioinformáticas actuales:** Según [6] se centra en utilizar técnicas de secuenciación de nueva generación (NGS) para analizar el genoma y el transcriptoma de las células tumorales. Este enfoque implica identificar mutaciones somáticas, variantes de splicing, fusiones génicas y otros eventos genómicos que puedan dar lugar a la generación de neoantígenos. Proporciona una visión detallada y directa de los péptidos presentados en la superficie celular, lo que complementa los enfoques genómicos.

d. **Enfoques basados en estructuras:** Según [6] se centra en el análisis de la interacción entre péptidos y moléculas de MHC a nivel molecular. Este enfoque utiliza modelos computacionales y técnicas de bioinformática para predecir la capacidad de un péptido específico de unirse a una molécula de MHC y ser reconocido por receptores de células T.

4. **Ensayos de validación con células T:** Los ensayos basados en células T tienen la ventaja de probar directamente si el repertorio de células T de un paciente ha captado un neoantígeno presentado por el MHC. También se puede realizar con células de otros pacientes. En [2] a esta parte la llama simplemente validación de neoantígenos.

En [6] concluyen que los enfoques basados en genómica permiten potencialmente descubrir casi todos los tipos de posibles fuentes de neoantígenos, sin embargo, su aplicación está limitada por la falta de evidencia sobre la existencia real de los péptidos predichos y su capacidad para unirse y presentarse mediante moléculas del MHC y ser reconocidos por las células T. En el caso de los enfoques basados en espectrometría de masas nos dicen que son esenciales ya que conducen a una mejor comprensión del proteoma unido al MHC.

1.2.5. Inmunoterapia

Se tiene que saber que la inmunoterapia no se resume sólo a las vacunas sino a los diversos tipos de terapias para las que se usan sustancias a fin de estimular o inhibir el sistema inmunitario y de esta manera ayudar al cuerpo a combatir el cáncer, las infecciones y otras enfermedades[10]. En [7] se menciona que en comparación con los otros tipos de inmunoterapia, la vacuna neoantígena puede inducir una fuerte respuesta inmunitaria y además provocar efectos terapéuticos estables.

Según [9] nos dice que gracias al desarrollo clínico los efectos de la inmunoterapia contra el cáncer, en torno al uso de vacunas, no se limitan al melanoma, sino que también se pueden observar en cánceres como el cáncer de pulmón de células no pequeñas.

En [3] se menciona que en la actualidad la mayoría de neoantígenos se identifican mediante la tecnología de secuenciación de exoma completo. Esta tecnología cuenta con una gran eficacia y a diferencia de hace algunos años la secuenciación de exoma y genoma completo han evolucionado

a gran ritmo. Este artículo nos ofrece también una lista de software para predicción de neoantígenos dónde podemos encontrar su año de despliegue y las funciones que realiza. También se mencionan estrategias para la realización y uso de la inmunoterapia, estas estrategias incluyen la identificación de neoantígenos(donde se puede utilizar el proceso de [2] que es el estándar), la selección de neoantígenos inmunogénicos, el proceso de diseño de vacunas especializadas, el uso de plataformas de vacunación y la evaluación preclínica y clínica de las vacunas de neoantígenos.

Mientras que en [13] se presentan definiciones de los diversos tipos de neoantígenos reportados que se conocían hasta el 2020, además de hablar de la terapia celular adoptiva en la cuál lo que se hace es modificar las células T genéticamente para poder responder ante los neoantígenos que el paciente encuentre. También en [14] se aborda otro tipo de inmunoterapia basada en la terapia celular para tratar la evasión de antígenos en el contexto de la terapia con células T diseñadas con receptores de antígenos quiméricos.

1.2.6. Mutación Somática

Un tema muy importante y que ayuda también a la unión del péptido con el MHC es la mutación somática. La mutación somática es la alteración del ADN que ocurre después de la concepción. Estas se pueden presentar en cualquiera de las células del cuerpo, excepto en las células germinativas (espermatozoides y óvulos), y, por lo tanto, no pasan a los hijos o hijas[11]. En [8] mencionan que gracias a la tecnología de secuenciación de alto rendimiento, ahora se puede obtener un panorama completo de mutaciones somáticas en tumores individuales (el "mutanoma").

1.2.7. Desarrollo de vacunas

En [9] se discute la importancia del proceso que sigue la inmunoterapia. Y así después de poder identificar los neoantígenos necesarios para intentar generar una respuesta inmunitaria se inicia el proceso de desarrollo de la vacuna personalizada. En [7] se menciona que las vacunas tumorales dirigidas a esta rama de los neoantígenos incluyen comúnmente vacunas de ácidos nucleicos, basadas en células dendríticas(DC), células tumorales

y péptidos largos sintéticos(SLP); además se realizaron pruebas que arrojaron que las vacunas basadas en CD, SLP y ARN son seguras y tienen la capacidad de generar respuestas de las células T CD4 y CD8.

1.2.8. Importancia de la unión péptido-MHC

El Complejo Mayor de Histocompatibilidad o mejor conocido como MHC se encarga de realizar la comunicación entre el péptido, unión de dos o más proteínas, con las células T. Según [3,6,7,8,9] estaríamos hablando de que para la inmunoterapia a base de vacunas neoantígenas el ideal es poder saber qué neoantígenos pueden unirse al MHC y así generar respuesta de las células T CD4 y CD8. Según [6] la LC-MS/MS moderna, método analítico moderno utilizado para detectar y cuantificar sustancias en niveles de traza basado en la espectrometría de masas, permite identificar miles de péptidos unidos al MHC en un solo experimento, en comparación con decenas en los primeros estudios.

Según [8] La predicción de la afinidad de unión de los péptidos candidatos a las moléculas del MHC es un paso crítico, también menciona que ya hay varios sitios web desarrollados que hacen su mejor intento por predecir. Y nos da dos enfoques:

A. **Método basado en NGS:** Utiliza WES combinado con la secuenciación de ARN(RNA-seq).
B. **Método basado en espectrometría de masas:** Utiliza cromatografía líquida y espectrometría de masas en tándem (LC-MS/MS), que es un método directo para analizar exhaustivamente el repertorio de neoantígenos tumorales como ya se había mencionado antes.

1.2.9. Modelos de Predicción de Unión p-MHC

Según [17] nos dice que existen 2 métodos básicos para estudiar la unión p-MHC: primero los ensayos de afinidad de unión p-MHC (BA), donde de un péptido se miden las preferencias con diferentes moléculas MHC y segundo ligandos eluidos (EL) asociados al MHC generado por espectrometría de masas por cromatografía líquida donde en un solo experimento se identifican un gran número de EL correspondientes a un MHC.

Para toda la teoría como ya se mencionó antes se tienen se han realizado innumerables métodos de predicción, en este caso hablaremos de los métodos relacionados a la predicción de unión péptido-MHC de clase I y de clase II. Primero mostraremos los métodos que hay:

A. **Attention-based Convolutional neural networks for MHC Epitope binding prediction(ACME):** Es un método de predicción de la afinidad de unión entre péptidos y moléculas de clase I del MHC. Este enfoque combina una red neuronal convolucional profunda con un módulo de atención para construir un modelo preciso e interpretable de predicción. ACME usa datos experimentales de un conjunto de datos ampliamente utilizado del IEDB (Immune Epitope Database) para entrenar y validar el modelo. También logró un aumento promedio de 7.0, 4.4 y 2.8 puntos porcentuales en el coeficiente de correlación de Pearson para péptidos de 11, 10 y 9 aminoácidos respectivamente, en comparación con NetMHCpan 3.0. Esto indica una mejora sustancial en la precisión de la predicción. Finalmente según los resultados presentados, ACME logró un promedio de AUROC de 0.90 para los alelos HLA-A y un promedio de 0.88 para los alelos HLA-B [15].
B. **MHCflurry2.0:** Es un predictor mejorado de péptidos presentados por MHC de clase I que incorpora el procesamiento antigénico. Utiliza dos modelos de red neuronal: el predictor MHCflurry BA para la afinidad de unión y el predictor MHCflurry AP para el procesamiento antigénico. El predictor BA se entrena con medidas de afinidad y péptidos MHC identificados por espectrometría de masas, mientras que el predictor AP aprende propiedades residuales de secuencia que distinguen los péptidos identificados por MS de los señuelos entre los péptidos predichos para unirse fuertemente al MHC de clase I. En comparación con otros modelos, MHCflurry 2.0 mostró buen rendimiento en la diferenciación de péptidos MS reales de señuelos, superando a NetMHCpan 4.0 BA, NetMHCpan 4.0 y MixMHCpred 2.0.2 en términos de AUC en la mayoría de los casos. Además, MHCflurry BA demostró una ventaja en precisión positiva predictiva (PPV) en comparación con NetMHCpan BA, con mejoras significativas en PPV en comparación con otros predictores de péptidos MS. Finalmente se menciona que el predictor BA mostró buen rendimiento con un AUC de al menos 0.90 para la mayoría de los alelos evaluados, lo que indica una capacidad sólida para diferenciar entre péptidos reales y señuelo en los datos de entrenamiento y validación [16].
C. **BepiPred3.0:** BepiPred-3.0 es una herramienta de predicción de epítopos de células B que utiliza un modelo de lenguaje de proteínas y datos teóricos para mejorar la precisión de la predicción de epítopos lineales y conformacionales. Utiliza representaciones numéricas del

modelo de lenguaje de proteínas ESM-2 para mejorar la precisión de la predicción. Además, selecciona cuidadosamente la arquitectura del predictor, la estrategia de entrenamiento y variables de entrada adicionales al modelo. Los datos de entrenamiento se extraen de estructuras cristalinas de complejos anticuerpo-antígeno. BepiPred-3.0 se ha comparado con sus predecesores y otros modelos, mostrando una mejora significativa en la capacidad de generalización en conjuntos de datos novedosos. En comparación con BepiPred-2.0, BepiPred-3.0 ha demostrado un rendimiento mejorado en términos de valores AUC, teniendo un aproximado de 0.75, pero lo que argumentan en el artículo es que el valor es una subestimación y que solo al recolectar más datos experimentales será posible evaluar completamente qué tan cerca estamos del límite superior de las herramientas de predicción de epítopos de células B [19].

D. **Redes convolucionales profundas:** Este modelo está basado en un Convolutional Neural Network (CNN) profundo. Utiliza un conjunto de datos de entrenamiento que consiste en 118,174 datos de unión de nonapéptidos-HLA-I(péptidos de longitud 9) para 76 alelos HLA-A y HLA-B. El proceso de predicción implica la conversión de la información de unión de péptidos en un arreglo bidimensional de características de unión de péptidos, que luego es procesado por el CNN para extraer características de bajo nivel y combinarlas en características de alto nivel (motivos) a través de múltiples capas convolucionales y de agrupación. Estas características de alto nivel se utilizan para clasificar el arreglo de características de unión de péptidos de entrada en ligandos o no ligandos a través de capas completamente conectadas. El modelo se entrena utilizando tecnologías de aprendizaje de vanguardia, como las redes generativas adversarias y el aprendizaje por transferencia, para superar la limitación de datos de entrenamiento. Los autores utilizaron la base de datos de epítopos inmunes (IEDB) para entrenar su modelo y las puntuaciones F1 resultantes fueron de 0,86, 0,94 y 0,67 para los alelos HLA-A31:01, HLA-A03:01 y HLA-A*68:01, respectivamente.

1.2.9. Modelos de Lenguaje Proteínico

En [17] se aborda como tema principal la mejora de los modelos de predicción de unión péptido-MHC usando modelos de lenguaje proteínico. Este artículo desarrolla un enfoque de predicción de unión de péptidos MHC Clase I basado en modelos de lenguaje de proteínas pre-entrenados

ESM1b y ESM2(basados en BERT). Estos utilizan una capa Soft-max y Graph Attention Network (GAT) para el ajuste fino. Para el final nos dice que la incorporación de Graph Attention Network (GAT) mejora la capacidad de predicción de péptidos de longitudes diferentes además de tener mejor rendimiento que NetMHCpan 4.1 en predicción de péptidos MHC Clase I.

1.2.10. Graph Neural Networks (GAT)

En 2023 surgió relevancia del uso de las Redes Neuronales Gráficas que han mejorado bastante los modelos que actualmente se usan en esta rama para la predicción p-MHC. En [20] los autores nos presentan un modelo GraphMHC que utiliza una red neuronal gráfica (GNN) compuesta por capas de atención y convolución de gráficos apilados. Se valida con conjuntos de datos intra-validación, inter-validación y aplicación clínica. Destaca principalmente por convertir el enlace MHC-péptido en una estructura gráfica en la cuál nos dicen que las moléculas(péptidos y MHC) de las bases de datos textuales hay que convertirlas en grafos y las uniones en el valor booleano de si se pueden unir o no. En comparación con NetMHCIIpan-4.0, GraphMHC logró un AUC de 92.2%, superando el rendimiento del modelo base [20].

Papers	Historia de Neoantígenos	Respuesta de Células T	Secuenciación de Nueva Generación (NGS)	Modelos de Predicción p-MHC	Desarrollo de Vacunas	Mutación Somática	Espectrometría de Masas	Inmunoterapia	Individualidad en la Respuesta
[9]	x	x			x			x	x

Cancer Neoantigens									
[3] Jointly Learning to Align and Aggregate with Cross Attention Pooling for Peptide-MHC Class I Binding Prediction		x	x	x					
[7] Neoan		x			x			x	

tigen vaccine: An emerging tumor immunotherapy									
[13] Adoptive cell therapy targeting neoantigens: A frontier for cancer research		x						x	
[6] Main Strategies				x		x	x		

for the Identification of Neoantigens									
[12] Arobotic microscope system to examine T cell receptor acuity against tumor neoantigens: A new tool for cancer immunother		x							

apy research									
[8] The Development of Tumor Neoantigen Vaccine Immunotherapy					x	x	x		
[14] Cellular immunotherapy treatment scheduling to address									

antigen escape									
[2] Cancer Neoantigens: Challenges and Future Directions for Prediction, Prioritization, and Validation									x
[15] ACME: Pan-specific peptid				x					

e–MHC class I binding prediction through attention-based deep neural networks									
[17] Improved prediction of MHC-peptide binding using protein langua				x					

ge models									
[16] MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing				x					
[19] BepiP				x					

red-3.0: Improved B-cell epitope prediction using protein language models									
[18] Deep convolutional neural networks for pan-specific peptide-MHC				x					

class I binding prediction									
[20] GraphMHC: Neoantigen prediction model applying the graph neural network to molecular structure				x					

Tabla1. Comparativo de papers según temas que tratan.

1.3. Definición del problema

En este contexto, surge la necesidad de desarrollar un modelo avanzado que pueda aprovechar plenamente la información relacional entre péptidos y moléculas de MHC. Si bien los enfoques existentes han logrado cierto éxito en la predicción de la unión péptido-MHC, la falta de consideración de la estructura subyacente de los datos puede limitar su capacidad para capturar relaciones significativas y generar predicciones precisas.

Por lo tanto, el problema central que aborda este proyecto es la necesidad de diseñar un aplicativo web que use un modelo de aprendizaje profundo que incorpore una capa GAT para mejorar la representación de las interacciones péptido-MHC. Al compararlo con otros modelos que tienen bases de datos textuales, se espera que el modelo propuesto pueda discernir con mayor precisión las asociaciones relevantes desde el punto de vista biológico, facilitando así la identificación de péptidos con alta afinidad y potencial inmunogénico.

Capítulo II – Planteamiento del Proyecto

2.1. Fundamentos teóricos (Dominio - Tecnologías - Metodologías)

2.1.1. Antígenos y neoantígenos

Según el Instituto Nacional del Cáncer [11] antígeno se define como cualquier sustancia que haga que el cuerpo produzca una respuesta inmunitaria contra ella y neoantígeno como una proteína nueva que se produce cuando aparecen ciertas mutaciones en el ADN de un tumor. En pocas palabras los antígenos representan a las células saludables que nuestro sistema inmunitario puede reconocer mientras que los neoantígenos son las células que mutan y que se generan en los tumores.

2.1.2. Secuenciación de Nueva Generación (NGS)

Según el Instituto Nacional del Cáncer [11] es el término que describe los métodos que se usan en un laboratorio para conocer el orden de los componentes básicos (llamados nucleótidos) de millones de fragmentos de ADN y ARN al mismo tiempo. Con la NGS también es posible identificar cambios en ciertas áreas del genoma o en genes específicos.

Hay diferentes métodos de NGS, como la secuenciación del genoma completo, la secuenciación del exoma completo, la prueba de genes múltiples y la secuenciación de transcriptomas. La NGS a veces ayuda a los investigadores a entender la causa de ciertas enfermedades, como el cáncer. También se llama secuenciación de última generación, secuenciación masiva en paralelo, SMP y SUG.

2.1.3. Machine Learning

El Machine Learning es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo). Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados.

2.1.4. Modelos de Predicción

Los modelos de predicción son algoritmos o enfoques que se utilizan para prever resultados basados en datos existentes. En el contexto de la predicción de la unión péptido-MHC, se han desarrollado diversos modelos que combinan técnicas avanzadas como redes neuronales convolucionales (CNN), redes de atención (attention-based networks), y redes neuronales gráficas (Graph Neural Networks, GATs). Estos modelos son capaces de analizar secuencias proteicas y predecir la afinidad de unión entre péptidos y moléculas de MHC, lo que es fundamental para el desarrollo de vacunas y tratamientos en inmunoterapia

2.1.5. Redes Neurales Gráficas (Graph Neural Networks)

Las redes neuronales gráficas han surgido como una técnica prometedora para mejorar la predicción de la afinidad de unión entre péptidos y moléculas de MHC. Estas redes convierten las estructuras moleculares en grafos, lo que permite un análisis más detallado y preciso. En [20] mencionan que su uso ha logrado superar a modelos tradicionales en precisión, destacando su potencial en aplicaciones clínicas de inmunoterapia.

2.1.6. Modelos de Lenguaje Proteínico (Protein Language Models)

Los modelos de lenguaje proteínico son herramientas que han demostrado ser efectivas en la predicción de la unión péptido-MHC al utilizar representaciones avanzadas de secuencias proteicas para mejorar la precisión del modelo. Según [17], la incorporación de redes de atención en estos modelos permite una mejor captura de las interacciones a nivel molecular.

2.1.8. Metodología de Minería de Datos para Aplicaciones de Ingeniería(DMME)

La metodología DMME es una extensión holística de CRISP-DM, la cual incorpora la fase de adquisición de datos dentro de los escenarios de producción, con el fin de constituirse como una metodología de minería de datos con enfoque específico de ingeniería y en su fase final incorpora la tarea de implementación técnica del modelo. En la figura 2 podemos ver las fases que comprenden esta metodología.

Figura 2. Fases de la metodología DMME

2.2. Objetivos del Proyecto

a. Objetivo General

Diseñar un modelo de aprendizaje profundo que incorpore una capa GAT para mejorar la representación de las interacciones péptido-MHC y así facilitar la identificación de péptidos con alta afinidad y potencial inmunogénico.

b. Objetivos Específicos

- Elaborar un cuadro comparativo sobre la información del dataset que será considerado en el proyecto de investigación.
- Diseñar y desarrollar el modelo de aprendizaje profundo que incluya una capa GAT para mejorar la representación de las interacciones péptido-MHC.
- Elaborar las métricas que impliquen la efectividad y precisión del modelo propuesto en la predicción de la unión péptido-MHC en comparación con enfoques convencionales.
- Validar las predicciones del modelo en términos de su capacidad para identificar péptidos con alta afinidad y potencial inmunogénico.
- Desarrollar un aplicativo web en el que usando el modelo se pueda realizar la predicción.

2.3. Justificación

En la actualidad los casos de melanoma, enfermedad por la que se forman células malignas (cancerosas) en los melanocitos, son bastante frecuentes y uno de los cánceres más mortales en el mundo. En la Figura 3 se puede ver que los casos no son pocos en el continente. En este contexto una de las terapias que mejores resultados ha tenido contra este cáncer es la inmunoterapia, la cuál tratamos de apoyar en esta investigación.

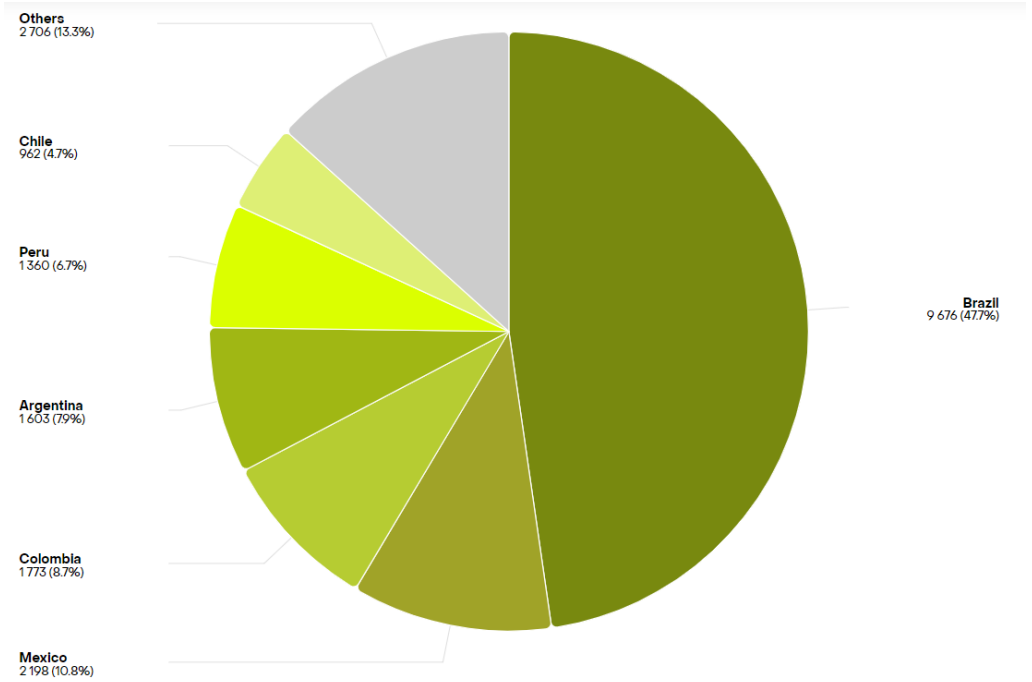

Figura 3. Incidencia de casos de melanoma en 2022 en Latinoamérica.

La investigación que se está llevando a cabo es de gran relevancia ya que aborda problemas cruciales en el ámbito de la inmunoterapia contra el cáncer, particularmente en la identificación y validación de neoantígenos. Mejorar la precisión y eficacia en la predicción de neoantígenos podría incrementar las tasas de éxito en el tratamiento de diversos tipos de cáncer y disminuir el tiempo de desarrollo de las vacunas, lo cual subraya la importancia de esta investigación.

Como consecuencia esto mejoraría la calidad de vida de los pacientes y optimizaría el uso de recursos en el tratamiento del cáncer, haciéndolo más eficiente y efectivo. En cuanto a la contribución de esta investigación a la sociedad, se espera que mejore significativamente la efectividad de las

terapias contra el cáncer y proporcione una base sólida para futuros avances y aplicaciones en la medicina y la salud pública, generando nuevos conocimientos y tecnologías aplicables en diversos contextos médicos y tecnológicos.

2.4. Viabilidad

Viabilidad económica

Productos	**Item**	**CU**	**CT**
PC(Intel® Core™ i7-11950H7 con Sistema Operativo Ubuntu 20.04)	1	4,999.99	4,999.99
Servicios básicos		1080.00	1080.00
Internet(Claro Hogar)	12	100.00	1200.00
Libros	5		420.00
Artículos	25		350.00
Revistas	25		280.00
TOTAL			8,329.99

Viabilidad técnica

Productos	**Item**		
Laptop(Gamer Hp Intel Core I9 RTX 4070 32GB 1TB SSD 13va Gen 16.1" Omen 16-wf0000la)	SI		

Máquinas virtuales(Paper Space)	Por adquirir		
Internet	SI		
Libros	SI		
Artículos	SI		
Revistas	SI		
TOTAL			

Viabilidad operativa

Productos	**Condición**		
Investigadores	SI		
Médicos Oncológicos	SI		
Semilleros de investigación	SI		

2.5. Limitaciones

El producto no proporciona una solución completa para la predicción de todos los tipos de interacciones moleculares, limitándose a la unión péptido-MHC. No puede garantizar la precisión absoluta en todas las predicciones debido a la variabilidad biológica y la complejidad de las interacciones moleculares. También está limitado a bases de datos gráficas y modelos específicos, lo que podría restringir su aplicabilidad en otros contextos o con diferentes tipos de datos. Por último se limita a la inmunoterapia en casos de cáncer limitados, los cuáles se mencionan en los artículos revisados(melanoma y cáncer de pulmón).

Capítulo III – Metodología de Desarrollo

Se describen las etapas y actividades de acuerdo a la metodología de desarrollo escogida. Se enuncia el tipo de metodología y se explican de forma detallada los procedimientos, procesos o actividades de las que consta.

Se debe de considerar la aplicación de técnicas de análisis de datos para la validación del trabajo.

Capítulo IV – Resultados y Discusión

Mostrar los resultados obtenidos como datos cualitativos o cuantitativos y realizar el análisis correspondiente de tal forma que se atienda el problema de investigación. Las evidencias que confirman las afirmaciones deben ser expuestas.

Capítulo V – Recomendaciones (opcional)

Abordar las implicaciones prácticas de aplicar el estudio en otros ámbitos, o con otra población, y/o sugerir nuevos estudios destinados a investigar otra dimensión del problema.

Conclusiones

Es la parte donde se manifiesta lo más destacado encontrado durante su investigación. Es una parte muy importante de la tesis puesto que en ella se indican los hallazgos y, en consecuencia, la comprobación de los objetivos. Aquí se muestran las aportaciones a la disciplina de estudio.

Es recomendable que el alumno elabore sus conclusiones tratando de cubrir por lo menos algunos de los siguientes puntos, mismos que sólo se presentan como una guía o referencia para una mejor elaboración de esas conclusiones, pero no se describen en el texto. Estas sugerencias se enfocan hacia:

- Resultados encontrados.
- Demostración realizada.
- Conclusión general.
- Conclusiones parciales (útiles al trabajo).
- Aportaciones a su tema (disciplina).

Algunas consideraciones a tener en cuenta son:

- Evitar que las conclusiones sean un resumen de cada capítulo pues en realidad se trata de consecuencias y determinaciones que conllevan una verdad.
- Su redacción debe ser clara, concreta, directa y enfática.
- Es importante hacer conclusiones específicas por cada punto de interés, pero sin abusar de este recurso. Todo estará en función al tema, su importancia y lo relevante de lo encontrado.

- Llegar a una conclusión global, si es posible, en que se concentren los aspectos fundamentales de la investigación, procurando abarcar solamente lo sustantivo y básico del tema.
- Evitar el tono imperativo e impositivo tanto como el tímido y desobligado.

Referencias

[1] "Global Cancer Observatory". Global Cancer Observatory. Accedido el 17 de abril de 2024. [En línea]. Disponible: https://gco.iarc.fr/en

[2] E. S. Borden, K. H. Buetow, M. A. Wilson y K. T. Hastings, "Cancer Neoantigens: Challenges and Future Directions for Prediction, Prioritization, and Validation", *Frontiers Oncol.*, vol. 12, marzo de 2022. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.3389/fonc.2022.836821

[3] C. Chen, Z. Qiu, Z. Yang, B. Yu y X. Cui, "Jointly Learning to Align and Aggregate with Cross Attention Pooling for Peptide-MHC Class I Binding Prediction", en *2021 IEEE Int. Conf. Bioinf. Biomedicine (BIBM)*, Houston, TX, USA, 9–12 de diciembre de 2021. IEEE, 2021. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1109/bibm52615.2021.9669444

[4] R. H. Murillo Moreno, "Vacunas basadas en neoantígenos y control del cáncer: perspectivas", *Rev. Colomb. Cancerol.*, vol. 24, n.º 4, pp. 178–88, noviembre de 2020. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.35509/01239015.195

[5] C. A. Parra López, "La implementación de las vacunas basadas en neoantígenos tumorales: un desafío para la medicina de precisión en oncología", *Rev. Colomb. Cancerol.*, vol. 24, n.º 4, pp. 154–56, noviembre de 2020. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.35509/01239015.713

[6] A. V. Gopanenko, E. N. Kosobokova y V. S. Kosorukov, "Main Strategies for the Identification of Neoantigens", *Cancers*, vol. 12, n.º 10, p. 2879, octubre de 2020. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.3390/cancers12102879

[7] M. Peng *et al.*, "Neoantigen vaccine: an emerging tumor immunotherapy", *Mol. Cancer*, vol. 18, n.º 1, agosto de 2019. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1186/s12943-019-1055-6

[8] W. Wang, “The Development of Tumor Neoantigen Vaccine Immunotherapy”, *E3S Web Conf.*, vol. 78, p. 01005, 2019. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1051/e3sconf/20197801005

[9] T. N. Schumacher, W. Scheper y P. Kvistborg, “Cancer Neoantigens”, *Annu. Rev. Immunol.*, vol. 37, n.º 1, pp. 173–200, abril de 2019. Accedido el 19 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1146/annurev-immunol-042617-053402

[10] “Neoantígenos: definición y aplicaciones en Oncología - Genotipia”. Genotipia. Accedido el 21 de abril de 2024. [En línea]. Disponible: https://genotipia.com/neoantigenos/

[11] “Diccionario de cáncer del NCI”. Instituto Nacional del Cáncer. Accedido el 21 de abril de 2024. [En línea]. Disponible: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/

[12] L.-L. S. Ong *et al.*, “A robotic microscope system to examine T cell receptor acuity against tumor neoantigens: A new tool for cancer immunotherapy research”, *IEEE Robot. Automat. Lett.*, vol. 4, n.º 2, pp. 1760–1767, abril de 2019. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1109/lra.2019.2894466

[13] Z. Wang y Y. J. Cao, “Adoptive cell therapy targeting neoantigens: A frontier for cancer research”, *Frontiers Immunol.*, vol. 11, marzo de 2020. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.3389/fimmu.2020.00176

[14] N. Dullerud y V. D. Jonsson, “Cellular immunotherapy treatment scheduling to address antigen escape”, en *2020 59th IEEE Conf. Decis. Control (CDC)*, Jeju, Korea (South), 14–18 de diciembre de 2020. IEEE, 2020. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1109/cdc42340.2020.9303848

[15] Y. Hu *et al.*, "ACME: Pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks", *Bioinformatics*, vol. 35, n.º 23, pp. 4946–4954, mayo de 2019. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1093/bioinformatics/btz427

[16] T. J. O'Donnell, A. Rubinsteyn y U. Laserson, "MHCflurry 2.0: Improved pan-allele prediction of MHC class i-presented peptides by incorporating antigen processing", *Cell Syst.*, vol. 11, n.º 1, pp. 42–48.e7, julio de 2020. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1016/j.cels.2020.06.010

[17] N. Hashemi *et al.*, "Improved prediction of MHC-peptide binding using protein language models", *Frontiers Bioinf.*, vol. 3, agosto de 2023. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.3389/fbinf.2023.1207380

[18] Y. Han y D. Kim, "Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction", *BMC Bioinf.*, vol. 18, n.º 1, diciembre de 2017. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1186/s12859-017-1997-x

[19] J. Clifford, M. H. Høie, S. Deleuran, B. Peters, M. Nielsen y P. Marcatili, "BepiPred-3.0: Improved B-cell epitope prediction using protein language models", *Protein Sci.*, noviembre de 2022. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1002/pro.4497

[20] H. Jeong, Y.-R. Cho, J. Gim, S.-K. Cha, M. Kim y D. R. Kang, "GraphMHC: Neoantigen prediction model applying the graph neural network to molecular structure", *Plos One*, vol. 19, n.º 3, marzo de 2024, art. n.º e0291223. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1371/journal.pone.0291223

[21] C. Qu, B. I. Schneider, A. J. Kearsley, W. Keyrouz y T. C. Allison, "Applying graph neural network models to molecular property prediction using high-quality experimental data", *Artif. Intell. Chemistry*, p. 100050, enero de 2024. Accedido el 22 de abril de 2024. [En línea]. Disponible: https://doi.org/10.1016/j.aichem.2024.100050

Anexos