Propuesta de la Investigación

1. Título

Desarrollo de una herramienta para la detección *in silico* de neoantígenos a partir de datos genómicos.

2. Lineas de investigación

Tecnologías avanzadas y computación / Inteligencia Artificial.

3. Breve estado de la cuestion

3.1. Estado del arte

El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo y depende de la correcta detección de neoantígenos (ver Figura 1 del material complementario). Estos neoantígenos son péptidos que solo están presentes en las células cancerosas. De esta forma, el objetivo de un tratamiento basado en vacunas personalizadas, es entrenar a los linfocitos del paciente (células T) para reconocer los neoantígenos y activar el sistema inmunológico [4, 7]. El proceso consiste en:

1. ETAPA I: Obtener muestras de tejido canceroso y saludable, Luego se secuencia ambos tejidos para obtener el ADN y/o ARN. Algunas propuestas incluyen información inmunopeptidoma de *Mass Spectrometry* (MS).

2. ETAPA II: Aquí realiza alineamiento de secuencias, se desarrolla un *llamado de variantes* para detectar las variantes y/o mutaciones; y se anotan dichas variantes (detección de posibles neoantígenos). Esta etapa cuenta con varias herramientas con buen desempeño.

3. ETAPA III: En esta etapa *in-silico* se priorizan neoantígenos. Esta etapa es crucial y ha tenido bastante investigación los últimos años debido a su complejidad y la baja efectividad de propuestas actuales. Aquí, se toman los neoantígenos candidatos (péptidos) de la etapa anterior y se predice su afinidad con el *Major Histocompatibility Complex* (MHC), este problema se conoce como *pMHC binding*. Luego, se evalúa la afinidad del pMHC para enlazarse al T-cell Receptor (TCR). Al finalizar esta etapa, se obtienen los neoantígenos.

4. ETAPA IV: En esta etapa *in-vitro*, se induce en laboratorio a las células T del paciente a reconocer los neoantígenos. Aquí, se desarrollan las vacunas. Generalmente, esta etapa es desarrollada por biotecnólogos y biólogos.

5. ETAPA V: Finalmente, el médico oncólogo realiza la evaluación clínica de la vacuna.

La detección *in-silico* de neoantígenos se basa en la ETAPA II y III (ver Figura 1 del material complementario). En este contexto, debido a la complejidad del proceso y la cantidad de métodos existentes, se han desarrollado software y *pipelines* para facilitar el uso de estas herramientas. En la Tabla 1 del material complementario, presentamos los *pipelines* publicados a partir del 2018. Estos *pipelines* utilizan diferentes tipos de información como entrada, así PGV Pipeline [19] y PEPPRMINT [21] utilizan DNA-seq; sin embargo, otras herramientas como PGNneo [22], NAP-CNB [23], NaoANT-HILL [24], ProGeo-neo [25], ScanNeo [26] y Neopepse [12] utilizan RNA-seq porque estas secuencias encapsulan mejor la información de mutaciones y *non-coding regions* de ADN [22].

Con el objetivo de reducir la complejidad de los *pipelines*, otras propuestas han optado por utilizar Variant Calling Format (VCF), como entrada. Estos archivos, contienen información de las mutaciones y son obtenidas a partir de métodos de alineamiento y llamado de mutaciones (ETAPA II.1 y ETAPA II.3 de la Figura 1 del material complementario). De esta forma, herramientas como Valid-Neo [27], HLA3D [28], Neoepiscope [20] , pVACtools [29] y NeoPredPipe [30], reducen la cantidad de herramientas utilizadas en la deteccion de neoantígenos; sin embargo, los resultados obtenidos, pueden ser inferiores comparado con herramientas que usan DNA-seq y RNA-seq.

Adicionalmente, para una correcta detección de neoantígenos, es necesario contar con la secuenciación de proteínas Major Histocompatibility Complex (MHC) o Human Leukocyte Antigens (HLA). Es necesario contar con estas proteínas porque, son utilizadas para predecir la unión entre posibles neoantígenos al MHC (pMHC: ETAPA III.1 de la Figura 1 del material complementario). Estas proteínas son codificadas por genes altamente polimórficos, esto

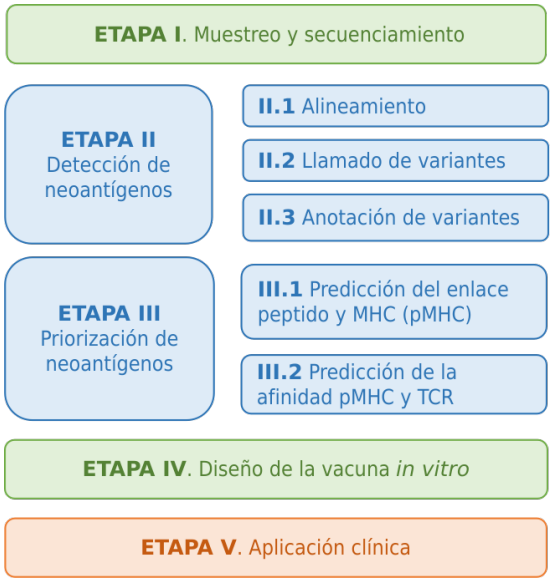

Figura 1: Marco de desarrollo para la elaboración de vacunas personalizadas contra el cáncer basadas en neoantígenos. Se detalla cada fase enfatizando el desarrollo *in-silico*. Fuente: Elaboración propia.

proporciona una variación sustancial en la unión de péptidos (neoantígenos), influyendo de esta manera en el conjunto de péptidos presentados a las células T. [31]. En este contexto, los *pipelines* Valid-NEO [27] y NeoPredPipe [30] y Neopepsee [12] solicitan como entrada estas proteínas (HLA); mientras que las otras predicen esta información a partir de DNA-seq. Desde un punto de vista de usabilidad, obtener los tipos de HLA, implica un esfuerzo innecesario para el usuario.

La presente propuesta ofrece una valiosa contribución a la disciplina de Bioinformática al proponer la integración de datos genómicos RNA-seq y DNA-seq, en combinación con datos de Mass Spectrometry.Además, también se va a incorporar información sobre variaciones estructurales, fusiones de genes y eventos de *alternative splicing*, fenómenos que han sido vinculados a diversos tipos de cáncer [20] y que aún no son considerados por los pipelines examinados en el estado del arte. Adicionalmente, la propuesta presenta una contribución significativa en la sub-area de la Computación y Ciencias de la Información al introducir el desarrollo de un modelo basado en *Transformers* para la predicción del enlace pMHC y pMHC-TCR. Este modelo empleará técnicas de fine-tuning utilizando modelos BERT pre-entrenados en extensas bases de datos de proteínas. Según investigaciones previas, los modelos Transformers han demostrado obtener resultados superiores en comparación con otras herramientas del estado del arte como NetMHCpan4.1 [14] y MHCflurry [13]."

3.2. Resultados o avances previos

La propuesta del proyecto representa la continuación de una serie de iniciativas y publicaciones. Se inició con el PROYECTO 01: "Principales estrategias y métodos basados en deep learning para la detección de neoantígenos en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del cáncer", financiado por las Universidades La Salle y Católica San Pablo. Este proyecto resultó en dos publicaciones: "Deep Learning and Transformers in MHC-Peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review" [32] y "Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context" [33].

Recientemente, hemos completado la ejecución del PROYECTO 02: "Desarrollo de una Aplicación Web para la Detección de Neoantígenos en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer". En este proyecto, hemos creado una aplicación centrada en la detección de neoantígenos, con un enfoque en la predicción del enlace pMHC mediante el uso de modelos Transformers y Transfer Learning. Además, hemos enviado para revisión el artículo "Fine-tuning Transformers for Peptide-MHC Class I Binding Prediction". Además, esta propuesta de PROCIENCIA también representa el trabajo futuro de la tesis de doctorado en Ciencia de la Computación del investigador principal, Vicente Machaca Arceda, titulada: "Detección *in Silico* de Neoantígenos Utilizando Transformers y Transfer Learning en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer". La tesis aborda los mismos objetivos que el PROYECTO 02 y ha logrado la aceptación del artículo "Transformers Meets Neoantigen Detection: A Systematic Literature Review.[en] el Journal of Integrative Bioinformatics (Q2).

Actualmente, estamos llevando a cabo el PROYECTO 03 "NeoArgos-tools: Un Pipeline de Detección In-silico

Tabla 1: Lista de *pipelines* desarrollados desde el 2018 hasta la actualidad para la detección de neoantígenos.GN: Expresión de genes, VA: anotación de variantes.

Nombre	Año	Ref.	Entrada	Salida	Herramientas
PEPPRMINT	2023	[21]	DNA-seq	BWA, Mutect, Strelka, ANNOVAR, OptiType, PEPPRMINT, netMHCpan4.1	Neoantígenos
PGNneo	2023	[22]	VCF, RNA-seq y MS data	Trimmomatic, BWA, SAMtools, GATK, Picard, OptiType, Annovar, Bedtools, MaxQuant, NetMHCpan4.1, Blastp	Neoantígenos
Valid-NEO	2022	[27]	VCF y HLA	Neoantígenos	
HLA3D	2022	[28]	VCF, HLA, SMG y HBV	MHCcluster, SAVES, PROCHECK, CoDockPP, Verify 3D, ERRAT, ClusterW2, 3Dmol, PSRPRED4.0, MHCf lurry, CoDockPP	Neoantígenos
NAP-CNB	2021	[23]	RNA-seq	Star, Picard, GATK, SplitNCigarsReads, MuTect2, Cufinks, EpiSeq, pVAC,seq, Neoantimon, MuPeXI, BLOSUM62	Neoantígenos
NeoANT-HILL	2020	[24]	RNA-seq y VCF	GATK, Mutect2, Optitype, NetMHC, NetMHCpan, NetMHCCcons, NetMHCstapan, PickPoket, SMM, SMMPMBEC, MHCflurry, NetMHCIIpan, NN-align, SMM-align, Sturniolo, Kallisto	Neoantígenos,GE
Neoepiscope	2020	[20]	VCF y BAM	BWA, Bowtie2, Pindel, MuSE, RADIA, SomaticSniper, VarScan2, GATK, HapCUT2	Neoantígenos
ProGeo-neo	2020	[25]	RNA-seq y VCF	SRA Toolkit, BWA, GATK, Bcftools, ANNOVAR, Kallisto, OptiType, NetMHCpan4.0, Picard	Neoantígenos
pVACtools	2020	[29]	VCF	CWL36, Cromwell37, ADNc38, BWA-MEM25, HaplotypeCaller28, MHCflurry14, MHCnuggets15, NetChop17, INTEGRATE-Neo19	Neoantígenos
NeoPredPipe	2019	[30]	VCF y HLA	ANNOVAR, POLYSOLVER, netMHCpan, PeptideMatch	Neoantígenos,VA
ScanNeo	2019	[26]	RNA-seq	HISAT2, BEDTools, BWA-MEM, pVAC-Seq, NetMHC, NetMHCpan	Neoantígenos
Neopepsee	2018	[12]	RNA-seq, VCF, HLA	NetCTLpan, Swiss-Prot	Neoantígenos,GE
PGV Pipeline	2018	[19]	DNA-seq	BWA-MEN, BQSR, MuTect, Strelka, STAR, seq2hla, Vaxrank, Isovar, MHCtools, Varcode, pyEnsembl	Neoantígenos

de Neoantígenos de Cáncer para el Desarrollo de Vacunas Personalizadas", con fecha de finalización en junio de este año y financiamiento de las Universidades La Salle y UCSP. Este proyecto implica el desarrollo de la versión inicial de NeoArgos-tools (la postulación a PROCIENCIA corresponde a la versión 2). En la primera versión, utilizamos archivos Variant Calling File (VCF) como entrada y mejoramos el módulo de predicción del enlace pMHC según investigaciones previas de proyectos anteriores.

Es relevante destacar que en la versión 2 de NeoArgos-tools, las mejoras propuestas incluyen el uso de datos genómicos como entrada al pipeline (RNS-seq y DNA-seq), la incorporación de Mass Spectrometry (MS) para refinar la detección de neoantígenos, la inclusión de información sobre variantes estructurales y fusión de genes, y el desarrollo de una interfaz gráfica.

4. Planteamiento del problema

El cáncer representa el mayor problema de salud mundial [1]. Además, según el instituto de investigación del cáncer del Reino Unido, se ha registrado más de 18 millones de nuevos casos y 10 millones de muertes en el 2020 [2]. Más alarmante aún, se predice que habrá 28 millones de nuevos casos por año alrededor del 2040, si la incidencia se mantiene estable y el crecimiento de la población y el envejecimiento continúan de acuerdo con las tendencias recientes [3]. Esto representa un aumento del 54.9 % con respecto a 2020 y se espera que sea mayor en hombres (aumento del 60.6 %) que en mujeres (aumento del 48.8 %). A todo esto, se sabe que los métodos tradicionales

basados en cirugías, radioterapias y quimioterapias tienen baja efectividad y adversos efectos secundarios [4]. En este contexto, surge el desarrollo de la inmunoterapia de cáncer, que tiene como objetivo estimular el sistema inmunológico de un paciente [5]. Existen varios tratamientos como: vacunas personalizadas; terapias de células T adoptivas; e inhibidores de puntos de control inmunológico. De estos, las vacunas basadas en **neoantígenos** han demostrado un gran potencial, al potenciar las respuestas de las células T y es considerada la de mayor probabilidad de éxito [5]. También, los neoantígenos son utilizados en la terapia de bloqueo de puntos de control inmunológico. En este sentido, los neoantígenos son considerados biomarcadores predictivos y objetivos de tratamiento sinérgico en la inmunoterapia del cáncer [6].

A pesar de varios esfuerzos en el desarrollo de *pipelines* y algoritmos, menos del 5 % de neoantígenos detectados activan el sistema inmune [7–11]. Según los autores de los *pipelines* las razones son:

1. La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de *Mass Spectrometry* (MS) [12]. Por ejemplo, la mayoría de propuestas no utiliza datos de MS; en la actualidad, existe una creciente información de estos datos y se están aplicado a varios campos de la Bioinformática.

2. Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoría de aplicaciones, se basa en el uso de MHCFlurry [13] y NetMHCpan4.1 [14]. Sin embargo, actualmente, se cuenta con herramientas de mejor desempeño como: MixMHCpred [15], Anthem [16], Acme [17] y ESM-GAT [18].

3. Para la etapa 3.2 de la Figura 1 del material complementario, los autores no consideran la predicción del enlace pMHC al TCR (pMHC-TCR) , varios autores consideran incluir esta tarea en trabajos futuros [19].

4. Finalmente y quizás la mas importante es no utilizar información de eventos de *alternative splicing*, variaciones estructurales en el ADN y las mutaciones de fusión de genes, está información esta fuertemente relacionada con varios tipos de cáncer [20].

5. Objetivos del proyecto

5.1. Objetivo general

Desarrollar una herramienta para la detección *in silico* de neoantígenos de cáncer a partir de datos genómicos.

5.2. Objetivos específicos

1. **OBJ 1:** Evaluar alternativas para la integración de datos Mass Spectrometry (MS) con datos genómicos como RNA-seq y DNA-seq.

2. **OBJ 2:** Evaluar las herramientas STAR [36], BWA [37], Bowtie2 [38] y Samtools [39] para determinar su aplicación en el alineamiento de secuencias.

3. **OBJ 3:** Evaluar las herramientas GATK [40] y BFCtools [39] y determinar el uso apropiado de estas en el llamado de variantes.

4. **OBJ 4:** Analizar el uso de información de variaciones estructurales del ADN y mutaciones de fusión de genes. Se evaluará el desempeño de *Arriba* [41] y FusionQ [42].

5. **OBJ 5:** Evaluar la herramienta Isovar [43] y Annovar [44] para la anotación de variantes y la herramienta OptiType [45] para la predicción de tipos de HLA.

6. **OBJ 6:** Implementar un modelo basado en *transformers* para la predicción del enlace pMHC, esto como alternativa a NetMHCpan4.1 [14] o MHCflurry [13]. Ya se cuenta con resultados previos de una propuesta que es superior a otras del estado del arte [33].

7. **OBJ 7:** Evaluar el modelo basado en *transformers* para la predicción del enlace pMHC en otra tarea similar: la predicción del enlace pMHC al TCR (pMHC-TCR).

8. **OBJ 8:** Implementar una interfaz gráfica que sirva como panel de administración para configurar y escoger las herramientas en cada fase del *pipeline*.

9. **OBJ 9:** Comparar el desempeño de la herramienta propuesta con otras herramientas del estado del arte.

6. Importancia del proyecto

6.1. Justificación

El cáncer constituye el principal desafío de salud a nivel global; no obstante, las técnicas convencionales basadas en cirugías, radioterapias y quimioterapias presentan una eficacia limitada [4]. En este escenario, los neoantígenos emergen como elementos cruciales en la concepción de vacunas contra el cáncer [5, 34, 35]. Si se logra desarrollar un enfoque altamente efectivo, la inmunoterapia del cáncer, fundamentada en la creación de vacunas personalizadas, podría posicionarse como una alternativa a procedimientos más tradicionales, como radioterapias y quimioterapias.

En el proyecto se propone realizar dos contribuciones significativas: CONTRIBUCIÓN 01: En el ámbito de la ciencia de la computación, se llevará a cabo el desarrollo de un modelo basado en *Transformers* y Transfer Learning para la predicción del enlace pMHC, con resultados previos que demuestran superar a otras propuestas en el estado del arte. CONTRIBUCIÓN 02: En el ámbito de la Bioinformática, la implementación del *pipeline* plantea un desafío al abordar problemas de integración, alto costo computacional, heterogeneidad y modularidad. Además, este *pipeline* incorporará datos de *Mass Spectrometry* (MS), variaciones estructurales y fusión de genes con el objetivo de obtener resultados más sobresalientes que otros métodos presentes en el estado del arte.

Con la conclusión de este proyecto, se abrirá paso a la segunda fase, que implica una colaboración interdisciplinaria para llevar a cabo el desarrollo *in vitro* de las vacunas de neoantígenos. En una tercera etapa, se contemplarán pruebas clínicas, marcando así un avance integral en la búsqueda de soluciones efectivas en la lucha contra el cáncer.

6.2. Impacto Científico de la propuesta

Como se menciona en los antecedentes, desarrollar vacunas personalizadas es un proceso que está en franca investigación, y depende mucho de la correcta detección de los antígenos, mostrándose como un gran desafío en estos días. En este sentido la propuesta apunta a mejorar este nivel de detección ya sea por el uso de técnicas de Inteligencia Artificial como Transformer y Transfer Learning, para mejorar la previsión de enlaces pMHC como incorporando más informaciones a considerar en esta detección como son: datos de Espectrometría de Masas (MS), datos genómicos DNA-Seq, RNA-Seq, variaciones estructurales, fusión de genes e incluso eventos de alternative splicing, que aun no se ha visto en el estado del arte como parte de los pipelines. Esta inclusión de nuevas informaciones generarán resultados nuevos, se espera que sean mejores a los actuales en el estado del arte, y que permitan avanzar en el problema de detección de neoantígenos para vacunas personalizadas que al momento tienen una tasa de acierto de menos del 5 % siendo aun poco atractivo para pensar en las siguientes etapas de desarrollo tecnológico.

7. Metodología

Hemos dividido la propuesta en dos módulos: NeoArgosMut y NeoArgosAntigen. NeoArgosMut, se enfoca en el llamado y anotación de variantes, obteniéndose como salida neoantígenos candidatos. Luego, NeoArgosAntigen, prioriza estos antígenos, al predecir su afinidad al MHC (pMHC) y luego la afinidad del pMHC al TCR (pMHC-TCR). En la Figura 2 del material complementario, mostramos estos módulos.

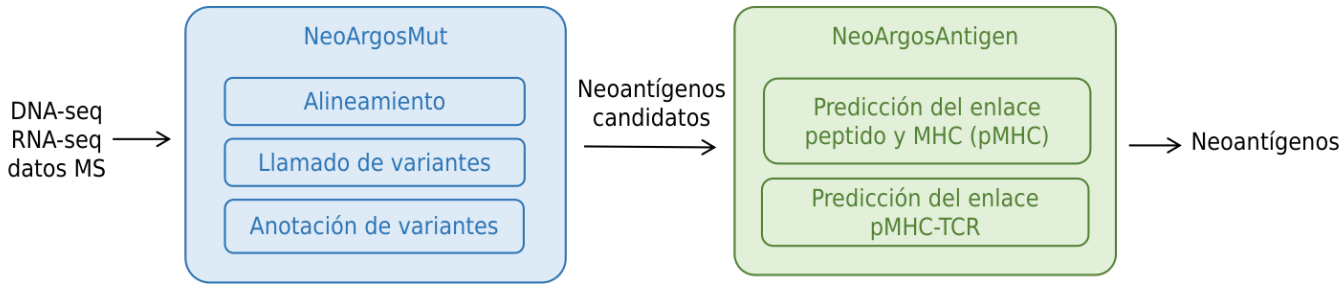

Figura 2: Representación de NeoArgosMut y NeoArgosAntigen para la detección de neoantígenos

7.1. NeoArgosMut

Los objetivos específicos OBJ 1, OBJ 2, OBJ 3, OBJ 4 y OBJ 5 serán desarrollados con este modulo. Así, para el objetivo específico OBJ 1, NeoArgosMut, se encargará de recibir como entrada datos de DNA-seq, RNA-seq y

Mass Spectrometry (MS). Se plantea utilizar la herramienta MaxQuant [46] para identificar las mutaciones a nivel de péptidos con ayuda de información de Mass Spectrometry (MS), esto forma parte de la contribución del trabajo al incluir fuentes adicionales de información como MS.

Referente al objetivo específico OBJ 2, se plantea alinear las secuencias con uso de las herramientas BWA [37], Bowtie2 [38] y Samtools [39]. Adicionalmente, se planteará utilizar STAR, porque alinea mejor muestras tumorales [19]. En esta etapa, se analizará y se determinará la mejor configuración y uso de cada herramienta. Como salida a esta etapa, se obtiene archivos de alineamiento BAM.

Para el objetivo específico OBJ 3 que corresponde al llamado de variantes se utilizará MuTect y Strelka. Se plantea usar la unión de la información de ambos métodos tal como lo hizo [21] y [19]. Como salida, se obtienen archivos VCF. Adicionalmente a otros *pipelines*, en esta etapa también abordaremos el objetivo especifico OBJ 4 y utilizaremos información sobre la fusión de genes que se obtendrán de las herramientas *Arriba* [41] y FusionQ [42]. Esta forma parte de la contribución de este trabajo, porque se sabe que la mayoría de *pipelines* tienen un bajo desempeño debido ausencia de información en su procesos de variantes estructurales y fusión de genes [20].

Finalmente el objetivo específico OBJ 5, se enfoca en la anotación de variantes y predicción del tipo de HLA. En esta etapa se toman los archivos en formato VCF y se obtienen los péptidos generados a partir de estas variaciones o mutaciones. Estos péptidos representan los posibles neoantígenos. Para está tarea se va a utilizar Isovar [43] y ANNOVAR [44], se evaluará su desempeño y se determinará cual de ellas usar bajo varios contextos. Luego, para obtener el tipo de HLA del paciente se va a utilizar la herramienta OptiType [45]. Otros *pipelines* optan por solicitar al usuario la información del tipo de HLA; sin embargo, obtener el HLA a partir de las mismas secuencias de ADN, mejora considerablemente el desempeño general del pipeline y la accesibilidad del usuario. Al finalizar esta etapa, se va a obtener los neoantígenos candidatos y los tipos de HLA.

7.2. NeoArgosAntigen

Los objetivos específicos OBJ 6, OBJ 7, OBJ 8, y OBJ 9 serán desarrollados con este modulo. NeoArgosAntigen, toma como entrada los neoantígenos candidatos y los tipos de HLA generados por NeoArgosMut. Esta prioriza estos neoantígenos. Esta priorización la realiza en base a la predicción del enlace de los neoantígenos al MHC, también conocido como predicción del enlace pMHC. Posteriormente se plantea predecir la afinidad del pMHC al TCR. El módulo se divide en dos partes: la predicción del enlace pMHC y la afinidad del pMHC al TCR. Ambas toman como entrada dos secuencias de proteínas, luego se necesita predecir su afinidad (regresión) o el enlace (clasificación). En resumen, las proteínas se pueden representar como $p = \{A, ..., Q\}$ y $q = \{A, N, ..., Q, E, G\}$. Luego, tenemos que predecir la probabilidad del enlace o afinidad entre p y q.

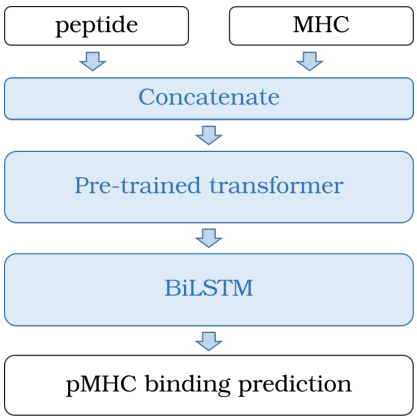

Figura 3: Modelo *transformer* seguido de BiLSTM para predecir el enlace pMHC.

Referente al objetivo específico OBJ 6, sobre el problema de predicción del enlace pMHC, se va a utilizar modelos BERT pre-entrenados y se realizará *fine-tuning* agregando un bloque de capas BiLSTM, y otra basada en grafos. Luego se volverá a entrenar estos modelos con una base de datos compuesta por muestras de [47] y [15]. Se propone la arquitectura de la Figura 3 del material complementario. Como se puede ver, la entrada son dos secuencias de proteínas: el péptido y el MHC. Luego, el modelo basado en Transformers está compuesto por un modelo pre-entrenado y un bloque de capas BiLSTM, esta propuesta se basó en el trabajo de [47]. En esta etapa también. se va a evaluar el desempeño de varios modelos BERT pre-entrenados como: TAPE [48], ProtBERT-BFD [49] y ESM2 [50] cada una con 92 millones, 420 millones, 650 millones parámetros respectivamente. Adicionalmente, TAPE fue entrenado con

30 millones de proteínas, ProtBERT-BFD con 2122 millones de proteínas y 60 millones de proteínas para ESM-2. En base a trabajos anteriores propios, sabemos que el uso de TAPE y el modelo más pequeño de ESM2 tienen buenos resultados [33]. Además, por investigaciones previas propias sabemos que podemos superar el desempeño de las mejores herramientas del estado del arte como NetMHCpan4.1 [14] y MHCflurry [13].

Para el objetivo específico OBJ 7 para la predicción del enlace pMHC y TCR (pMHC-TCR), se utilizará la misma metodología del objetivo 6, según recomendaciones de otros autores [25,51]. Sin embargo, se va reentrenar el modelo para adaptarse a este nuevo problema, se utilizarán muestras de [25] y la base de datos de VDJdb [52]. Al finalizar esta etapa, se obtendrán los neoantígenos priorizados.

Luego, para lograr el objetivo especifico OBJ 8 referente al desarrollo de una interfaz gráfica para la configuración y selección de herramientas se va a desarrollar una herramienta similar a Orange Machine Learning. Por ejemplo, en la Figura 4 del material complementario se muestra el prototipo de la interfaz gráfica. En el panel izquierdo se muestran las posibles herramientas, de las cuales el usuario podrá escoger y configurar. En color azul, se resaltan las herramientas que representan una contribución en este trabajo. Luego, en el panel derecho, se muestra como el usuario puede realizar el flujo de actividades con las herramientas que ha seleccionado.

Finalmente, para lograr desarrollar el objetivo especifico OBJ 9, se va a comparar el resultado de la propuesta con otros pipelines del estado del arte. En la Tabla 1 del material complementario se detallan estas herramientas. La comparación se realizará con las herramientas PEPPRMINT, HLA3d, ProGeoNeo y pVACtools.

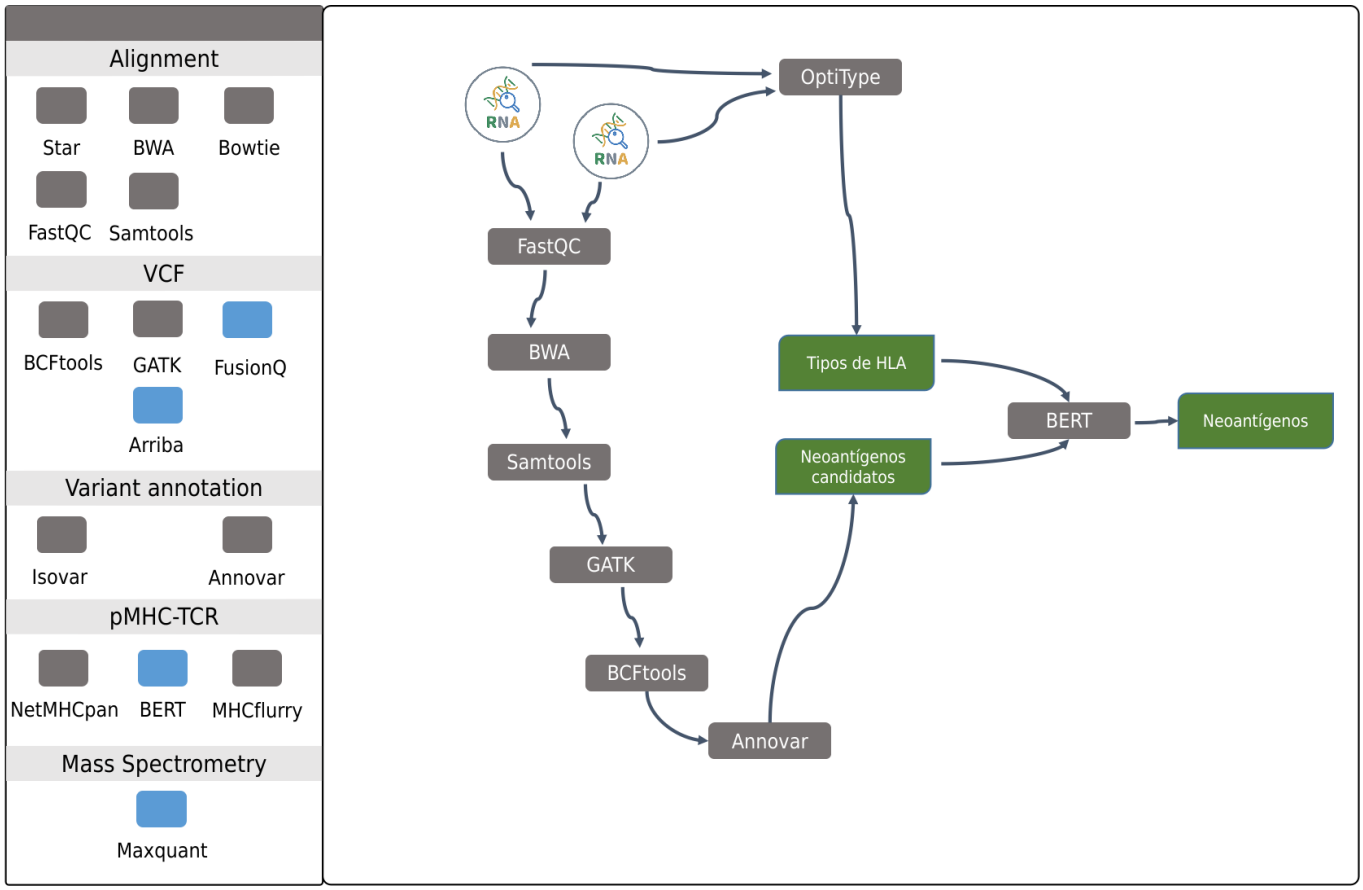

Figura 4: Prototipo de la interfaz gráfica de la herramienta para la detección de neoantígenos. En el panel izquierdo se muestran las posibles herramientas, de las cuales el usuario podrá escoger y configurar. En color azul, se resaltan las herramientas que representan una contribución en este trabajo. Luego, en el panel derecho, se muestra como el usuario puede realizar el flujo de actividades con las herramientas que ha seleccionado.

8. Resultados esperados

- Una aplicación con la funcionalidad de detectar neoantígenos a partir de datos genómicos.
- Una publicación en un Journal.
- Una tesis de pregrado.

- Una tesis de postgrado.

9. Materiales e insumos

- Un Workstation para procesamiento de datos genómicos, valorizada en S/40000.
- Dos laptops para ser utilizadas por el equipó de investigación, valorizadas en S/6000 cada una.
- Creditos en servicios de Cloud: DigitalOceam, Paperspace, AWS, Azure, valorizado en S/5000.
- Suscripción a revistas cientificas y eventos académicos, valorizado en S/1000.
- Pago de licencias de software, valorizado en S/ 4000

10. Cronograma de Actividades

Hito	Fases	Objetivos específicos	Actividades	Encargado	I	II	III	IV	V	VI
I	I.1	OBJ 1: Evaluar alternativas para la integración de datos Mass Spectrometry (MS) con datos genómicos como RNA-seq y DNA-seq.	Investigar y evaluar el uso de RNA-se y DNA.seq	Vicente	x	x				
			Evaluar herramientas para tratar datos de MS			x	x	x	x	
			Integración de datos MS y RNA-seq	Yvan			x	x	x	
	I.2	OBJ 2: Evaluar las herramientas STAR [36], BWA [37], Bowtie2 [38] y Samtools [39] para determinar su aplicación en el alineamiento de secuencias.	Evaluar y comparar cada herramienta de alineamiento	Tesista	x	x	x	x		
			Investigar sobre buenas prácticas y estandares		x	x	x	x		
		OBJ 3: Evaluar las herramientas GATK [40] y BFCtools [39] y determinar el uso apropiado de estas en el llamado de variantes.	Investigar sobre las buenas prácticas de GATK	Julio		x	x			
			Evaluar y comparar la herramienta GATK	Tesista		x	x	x		
			Evaluar la herramienta BCFtools	Tesista			x	x		
	I.3	OBJ 4: Analizar el uso de información de variaciones estructurales del ADN y mutaciones de fusión de genes. Se evaluará el desempeño de Arriba [41] y FusionQ [42].	Investigar sobre variaciones estructurales y fusión de genes	Yvan	x	x	x	x	x	
			Evaluar la herramienta Arriba	Vicente			x			
			Evaluar la herramienta FusionQ	Vicente			x			
			Investigar sobre otras herramientas para la detección de fusion de genes y variaciones estructurales	Yvan		x	x	x	x	
	I.4	OBJ 5: Evaluar la herramienta Isovar [43] y Annovar [44] para la anotación de variantes y la herramienta OptiType [45] para la predicción de tipos de HLA.	Evaluar la herramienta OptiType y semejantes	Yvan			x	x		
			Evaluar la herramienta Isovar				x			
			Evaluar la herramienta Annovar				x			
			Investigar sobre otras herramientas y realizar comparaciones	Yvan			x	x		
			Integrar las herramientas en NeoArgosMut				x	x	x	
II	II.1	OBJ 6: Implementar un modelo basado en transformers para la predicción del enlace pMHC.	Implementar el modelo basado en Transformers	Vicente	x	x	x			
			Entrenamiento del modelo	Vicente		x	x	x		
			Evaluar el desempeño y comparar con otras herramientas	Vicente		x	x	x		
		OBJ 7: Evaluar el modelo basado en transformers para la predicción del enlace pMHC en otra tarea similar: la predicción del enlace pMHC al TCR.	Entrenar el modelo en bases de datos pMHC-TCR	Vicente			x	x		
			Comparar el modelo con otras herramientas	Vicente				x		
	II.2	OBJ 8: Implementar una interfaz gráfica que sirva como panel de administración para configurar y escoger las herramientas en cada fase del pipeline.	Integrar las herramientas en NeoArgosAntigen	Tesista			x	x	x	
			Implementar la interfaz gráfica	Tesista		x	x	x	x	x
			Integración de NeoArgosMut y NeoArgosAnigen					x	x	x
	II.3	OBJ 9: Comparar el desempeño de la herramienta propuesta con otras herramientas del estado del arte.	Evaluar el desempeño del pipeline	Julio			x	x	x	x
			Redacción del artículo cientifico para el Journal						x	x
			Redacción del artículo cientifico para la Conferencia				x	x		
			Evento de difución de resultados							x

Referencias

[1] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal, "Cancer statistics, 2023," *Ca Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.

[2] Cancer Research UK, "Worldwide cancer statistics," 2023.

[3] Cancer Research UK, "Worldwide cancer incidence statistics," 2023.

[4] Miao Peng, Yongzhen Mo, Yian Wang, Pan Wu, Yijie Zhang, Fang Xiong, Can Guo, Xu Wu, Yong Li, Xiaoling Li, et al., "Neoantigen vaccine: an emerging tumor immunotherapy," *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.

[5] Elizabeth S Borden, Kenneth H Buetow, Melissa A Wilson, and Karen Taraszka Hastings, "Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation," *Frontiers in Oncology*, vol. 12, 2022.

[6] Xianzhu Fang, Zhiliang Guo, Jinqing Liang, Jiao Wen, Yuanyuan Liu, Xiumei Guan, and Hong Li, "Neoantigens and their potential applications in tumor immunotherapy," *Oncology Letters*, vol. 23, no. 3, pp. 1–9, 2022.

[7] L Mattos, M Vazquez, F Finotello, R Lepore, E Porta, J Hundal, P Amengual-Rigo, CKY Ng, A Valencia, J Carrillo, et al., "Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group," *Annals of oncology*, vol. 31, no. 8, pp. 978–990, 2020.

[8] Nil Adell Mill, Cedric Bogaert, Wim van Criekinge, and Bruno Fant, "neoms: Attention-based prediction of mhc-i epitope presentation," *bioRxiv*, 2022.

[9] Brendan Bulik-Sullivan, Jennifer Busby, Christine D Palmer, Matthew J Davis, Tyler Murphy, Andrew Clark, Michele Busby, Fujiko Duke, Aaron Yang, Lauren Young, et al., "Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification," *Nature biotechnology*, vol. 37, no. 1, pp. 55–63, 2019.

[10] Michal Bassani-Sternberg, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann, "Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation*[s]," *Molecular & Cellular Proteomics*, vol. 14, no. 3, pp. 658–673, 2015.

[11] Mahesh Yadav, Suchit Jhunjhunwala, Qui T Phung, Patrick Lupardus, Joshua Tanguay, Stephanie Bumbaca, Christian Franci, Tommy K Cheung, Jens Fritsche, Toni Weinschenk, et al., "Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing," *Nature*, vol. 515, no. 7528, pp. 572–576, 2014.

[12] Sora Kim, Han Sang Kim, Eunyoung Kim, MG Lee, E-C Shin, and S Paik, "Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information," *Annals of Oncology*, vol. 29, no. 4, pp. 1030–1036, 2018.

[13] Timothy J O'Donnell, Alex Rubinsteyn, and Uri Laserson, "Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing," *Cell systems*, vol. 11, no. 1, pp. 42–48, 2020.

[14] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen, "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*, vol. 48, no. W1, pp. W449–W454, 2020.

[15] David Gfeller, Julien Schmidt, Giancarlo Croce, Philippe Guillaume, Sara Bobisse, Raphael Genolet, Lise Queiroz, Julien Cesbron, Julien Racle, and Alexandre Harari, "Improved predictions of antigen presentation and tcr recognition with mixmhcpred2. 2 and prime2. 0 reveal potent sars-cov-2 cd8+ t-cell epitopes," *Cell Systems*, vol. 14, no. 1, pp. 72–83, 2023.

[16] Shutao Mei, Fuyi Li, Dongxu Xiang, Rochelle Ayala, Pouya Faridi, Geoffrey I Webb, Patricia T Illing, Jamie Rossjohn, Tatsuya Akutsu, Nathan P Croft, et al., "Anthem: a user customised tool for fast and accurate prediction of binding between peptides and hla class i molecules," *Briefings in Bioinformatics*, vol. 22, no. 5, pp. bbaa415, 2021.

[17] Yan Hu, Ziqiang Wang, Hailin Hu, Fangping Wan, Lin Chen, Yuanpeng Xiong, Xiaoxia Wang, Dan Zhao, Weiren Huang, and Jianyang Zeng, "Acme: pan-specific peptide–mhc class i binding prediction through attention-based deep neural networks," *Bioinformatics*, vol. 35, no. 23, pp. 4946–4954, 2019.

[18] Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Ch Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov, "Improved prediction of mhc-peptide binding using protein language models," *Frontiers in Bioinformatics*, vol. 3, 2023.

[19] Alex Rubinsteyn, Julia Kodysh, Isaac Hodes, Sebastien Mondet, Bulent Arman Aksoy, John P Finnigan, Nina Bhardwaj, and Jeffrey Hammerbacher, "Computational pipeline for the pgv-001 neoantigen vaccine trial," *Frontiers in immunology*, vol. 8, pp. 1807, 2018.

[20] Mary A Wood, Austin Nguyen, Adam J Struck, Kyle Ellrott, Abhinav Nellore, and Reid F Thompson, "Neoepiscope improves neoepitope prediction with multivariant phasing," *Bioinformatics*, vol. 36, no. 3, pp. 713–720, 2020.

[21] Laura Y Zhou, Fei Zou, and Wei Sun, "Prioritizing candidate peptides for cancer vaccines through predicting peptide presentation by hla-i proteins," *Biometrics*, vol. 79, no. 3, pp. 2664–2676, 2023.

[22] Xiaoxiu Tan, Linfeng Xu, Xingxing Jian, Jian Ouyang, Bo Hu, Xinrong Yang, Tao Wang, and Lu Xie, "Pgnneo: A proteogenomics-based neoantigen prediction pipeline in noncoding regions," *Cells*, vol. 12, no. 5, pp. 782, 2023.

[23] Carlos Wert-Carvajal, Rubén Sánchez-García, José R Macías, Rebeca Sanz-Pamplona, Almudena Méndez Pérez, Ramon Alemany, Esteban Veiga, Carlos Óscar S Sorzano, and Arrate Muñoz-Barrutia, "Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[24] Ana Carolina MF Coelho, André L Fonseca, Danilo L Martins, Paulo BR Lins, Lucas M da Cunha, and Sandro J de Souza, "neoant-hill: an integrated tool for identification of potential neoantigens," *BMC Medical Genomics*, vol. 13, no. 1, pp. 1–8, 2020.

[25] Yuyu Li, Guangzhi Wang, Xiaoxiu Tan, Jian Ouyang, Menghuan Zhang, Xiaofeng Song, Qi Liu, Qibin Leng, Lanming Chen, and Lu Xie, "Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection," *BMC medical genomics*, vol. 13, no. 5, pp. 1–11, 2020.

[26] Ting-You Wang, Li Wang, Sk Kayum Alam, Luke H Hoeppner, and Rendong Yang, "Scanneo: identifying indel-derived neoantigens using rna-seq data," *Bioinformatics*, vol. 35, no. 20, pp. 4159–4161, 2019.

[27] Yuri Laguna Terai, Chun Huang, Baoli Wang, Xiaonan Kang, Jing Han, Jacqueline Douglass, Emily Han-Chung Hsiue, Ming Zhang, Raj Purohit, Taylor deSilva, et al., "Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples," *Cancers*, vol. 14, no. 5, pp. 1243, 2022.

[28] Xingyu Li, Xue Lin, Xueyin Mei, Pin Chen, Anna Liu, Weicheng Liang, Shan Chang, and Jian Li, "Hla3d: an integrated structure-based computational toolkit for immunotherapy," *Briefings in bioinformatics*, vol. 23, no. 3, pp. bbac076, 2022.

[29] Jasreet Hundal, Susanna Kiwala, Joshua McMichael, Christopher A Miller, Huiming Xia, Alexander T Wollam, Connor J Liu, Sidi Zhao, Yang-Yang Feng, Aaron P Graubert, et al., "pvactools: a computational toolkit to identify and visualize cancer neoantigens," *Cancer immunology research*, vol. 8, no. 3, pp. 409–420, 2020.

[30] Ryan O Schenck, Eszter Lakatos, Chandler Gatenbee, Trevor A Graham, and Alexander RA @miscNCIdictionary2022, author = NCI, title = National Cancer Institute Dictionary, year = 2022, url = https://www.cancer.gov/publications/dictionaries/genetics-dictionary, urldate = 2022-03-20 Anderson, "Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–6, 2019.

[31] Esam T Abualrous, Jana Sticht, and Christian Freund, "Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation," *Current Opinion in Immunology*, vol. 70, pp. 95–104, 2021.

[32] Vicente Enrique Machaca, Valeria Goyzueta, Maria Cruz, and Yvan Tupac, "Deep learning and transformers in mhc-peptide binding and presentation towards personalized vaccines in cancer immunology: A brief review," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 14–23.

[33] Vicente Machaca Arceda Machaca, "Neoantigen detection using transformers and transfer learning in the cancer immunology context," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 97–102.

[34] Ina Chen, Michael Chen, Peter Goedegebuure, and William Gillanders, "Challenges targeting cancer neoantigens in 2021: a systematic literature review," *Expert Review of Vaccines*, vol. 20, no. 7, pp. 827–837, 2021.

[35] Alexander V Gopanenko, Ekaterina N Kosobokova, and Vyacheslav S Kosorukov, "Main strategies for the identification of neoantigens," *Cancers*, vol. 12, no. 10, pp. 2879, 2020.

[36] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[37] Heng Li and Richard Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[38] Ben Langmead, Christopher Wilks, Valentin Antonescu, and Rone Charles, "Scaling read aligners to hundreds of threads on general-purpose processors," *Bioinformatics*, vol. 35, no. 3, pp. 421–432, 2019.

[39] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al., "Twelve years of samtools and bcftools," *Gigascience*, vol. 10, no. 2, pp. giab008, 2021.

[40] Geraldine A Van der Auwera, "Somatic variation discovery with gatk4," *Cancer Research*, vol. 77, no. 13_Supplement, pp. 3590–3590, 2017.

[41] Sebastian Uhrig, Julia Ellermann, Tatjana Walther, Pauline Burkhardt, Martina Fröhlich, Barbara Hutter, Umut H Toprak, Olaf Neumann, Albrecht Stenzinger, Claudia Scholl, et al., "Accurate and efficient detection of gene fusions from rna sequencing data," *Genome research*, vol. 31, no. 3, pp. 448–460, 2021.

[42] Chenglin Liu, Jinwen Ma, ChungChe Jeff Chang, and Xiaobo Zhou, "Fusionq: a novel approach for gene fusion detection and quantification from paired-end rna-seq," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013.

[43] OpenVAX, "Isovar," 2023.

[44] Kai Wang, Mingyao Li, and Hakon Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic acids research*, vol. 38, no. 16, pp. e164–e164, 2010.

[45] András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher, "Optitype: precision hla typing from next-generation sequencing data," *Bioinformatics*, vol. 30, no. 23, pp. 3310–3316, 2014.

[46] Nikita Prianichnikov, Heiner Koch, Scarlet Koch, Markus Lubeck, Raphael Heilig, Sven Brehmer, Roman Fischer, and Jürgen Cox, "Maxquant software for ion mobility enhanced shotgun proteomics," *Molecular & Cellular Proteomics*, vol. 19, no. 6, pp. 1058–1069, 2020.

[47] Yaqi Zhang, Gancheng Zhu, Kewei Li, Fei Li, Lan Huang, Meiyu Duan, and Fengfeng Zhou, "Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction," *Briefings in Bioinformatics*, 2022.

[48] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.

[49] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al., "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[50] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[51] Alexander Myronov, Giovanni Mazzocco, Paulina Krol, and Dariusz Plewczynski, "Bertrand-peptide: Tcr binding prediction using bidirectional encoder representations from transformers augmented with random tcr pairing," *bioRxiv*, pp. 2023–06, 2023.

[52] Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, et al., "Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity," *Nucleic acids research*, vol. 46, no. D1, pp. D419–D427, 2018.