Universidad La Salle

Desarrollo de una aplicación Web para la detección de neoantígenos en el marco de desarrollo de vacunas personalizadas para tratar el Cáncer

Proyecto interno de la universidad

Dr. Vicente Machaca Arceda
Mg. Richart Escobedo
Jose Grados
Kristhyan Lazarte

2024

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

DNA
Localización

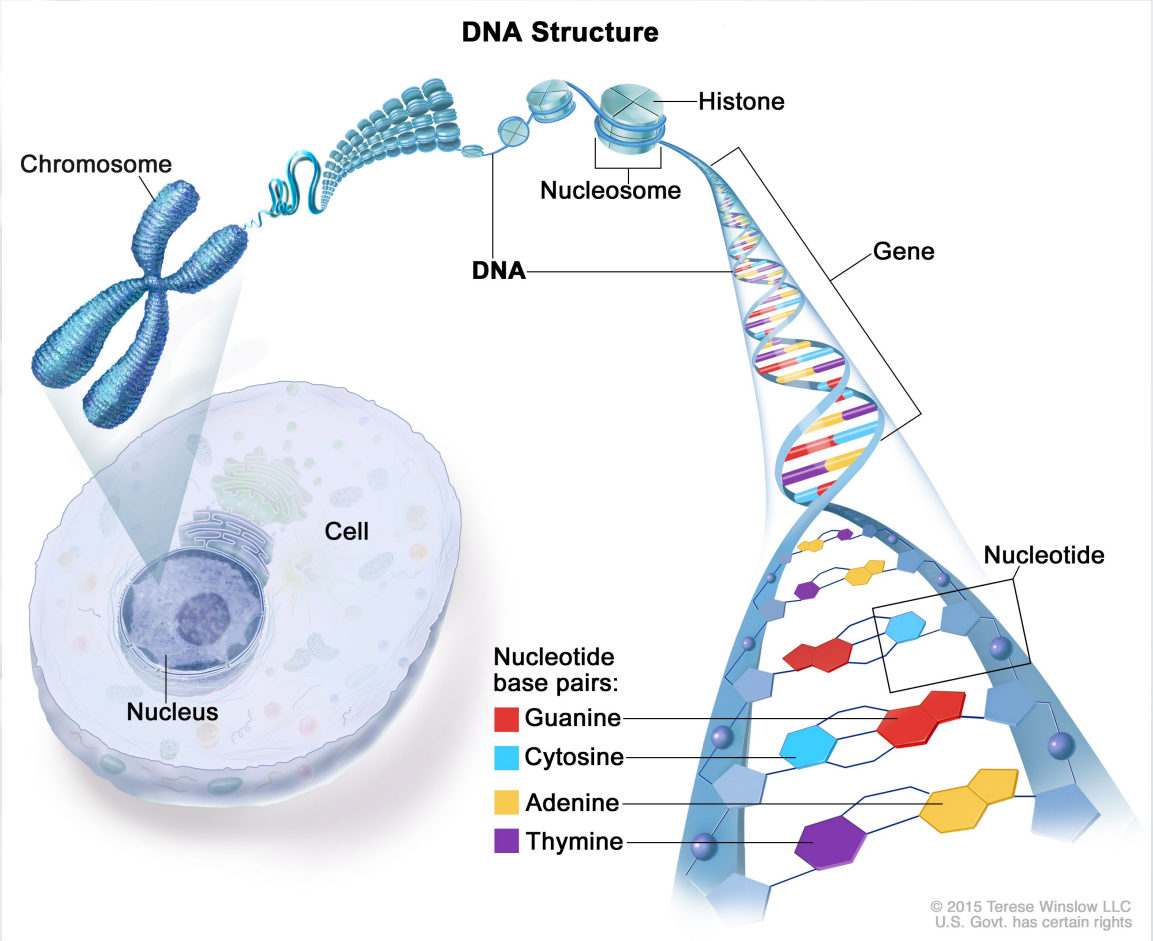

Figure: Where DNA is located [1].

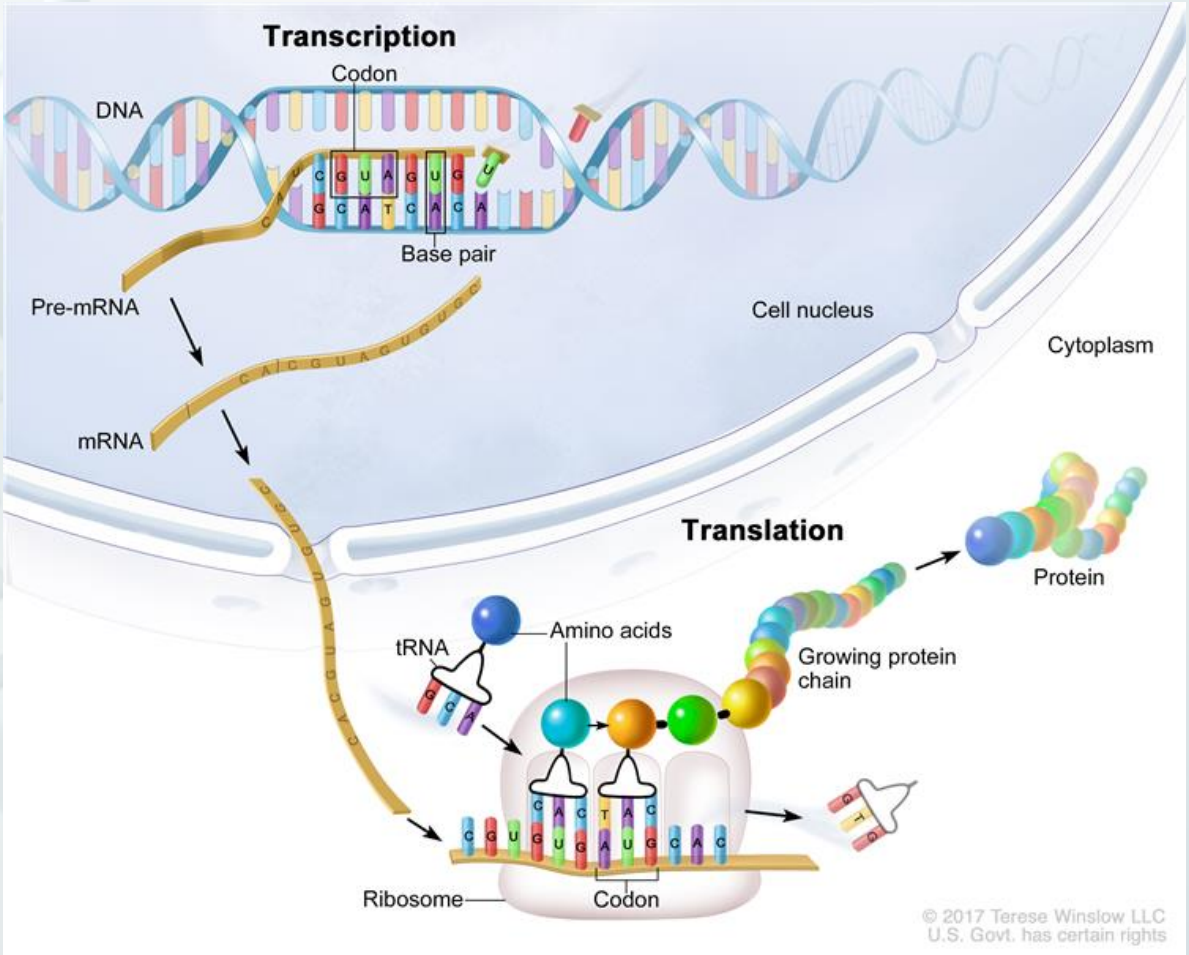

Figure: Transcription and translation [2].

Variantes y Mutaciones

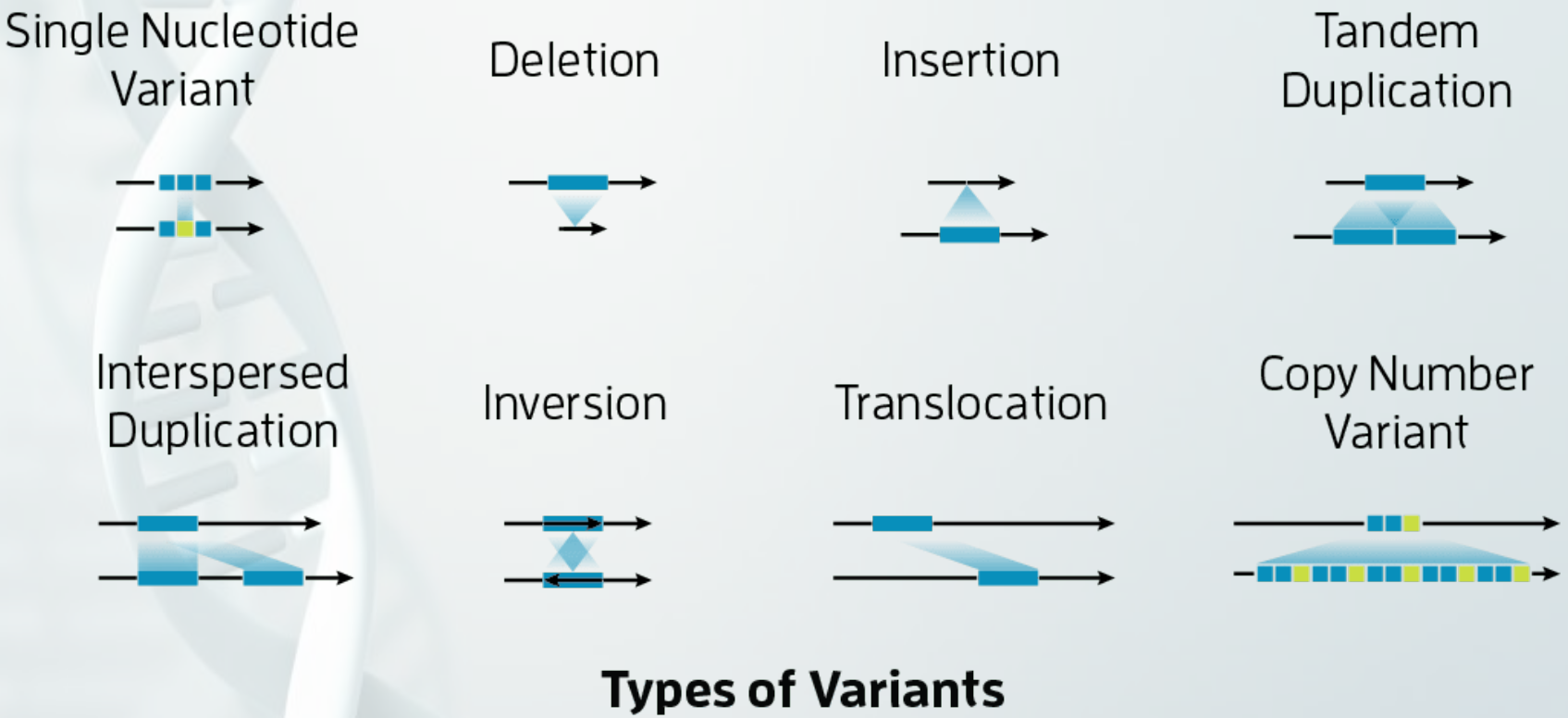

Figure: Example of structural variants. Source: [3]

Inmunoterapia del Cáncer

Es un tipo de tratamiento contra el Cáncer que estimula las defensas naturales del cuerpo para combatir el Cáncer [4].

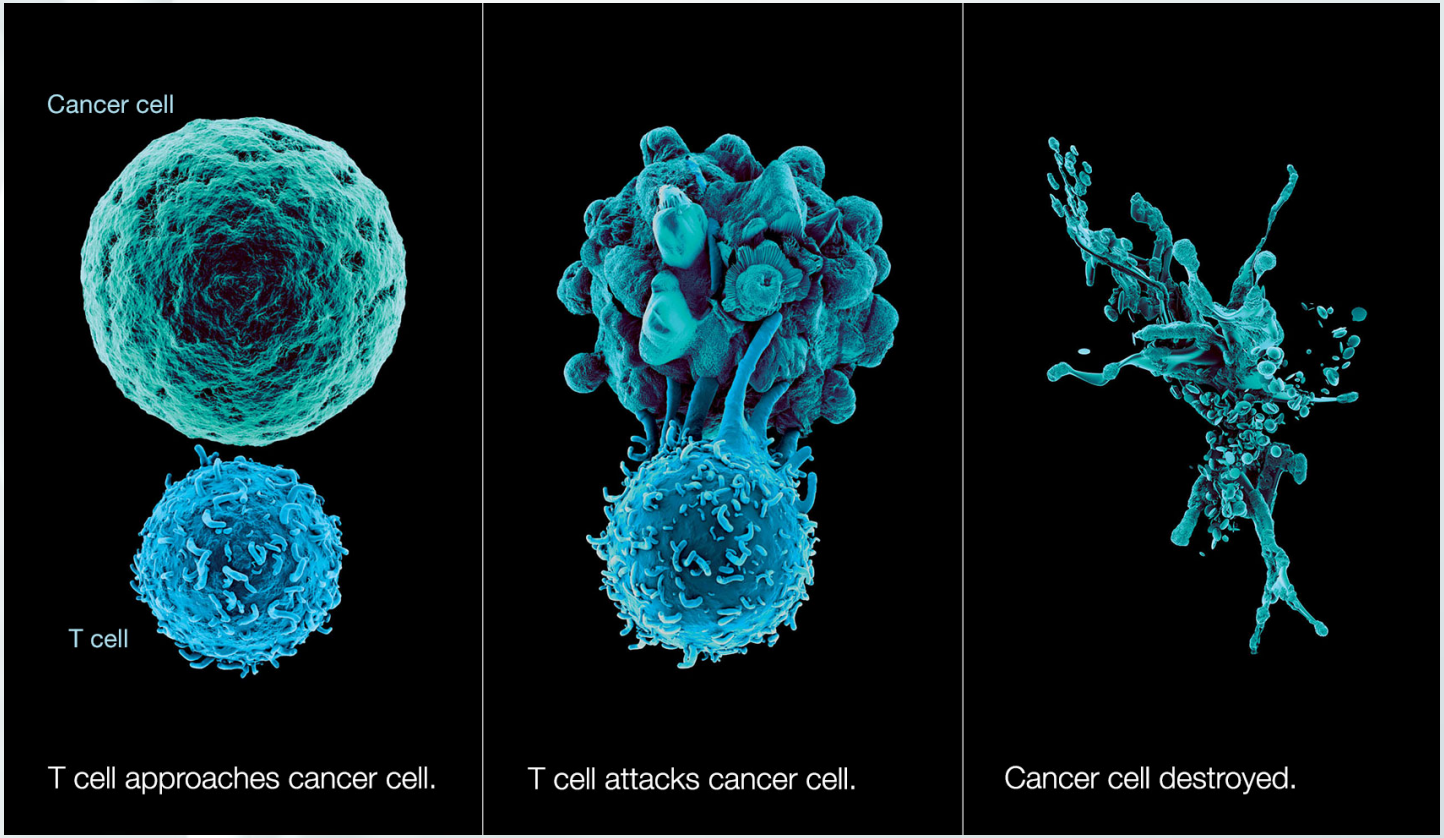

Figure: Ejemplo de como una célula T destruye células del cancer [5].

Inmunoterapia del Cáncer

Neo antígenos

Es una **proteína** que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA, cumplen un rol importante al **estimular una respuesta inmune** [1, 6].

En la actualidad hay varios métodos para detectar a predecir neo antígenos, pero **solo una pequeña cantidad de ellos** logran estimular al sistema inmune [7, 8].

MHC-I

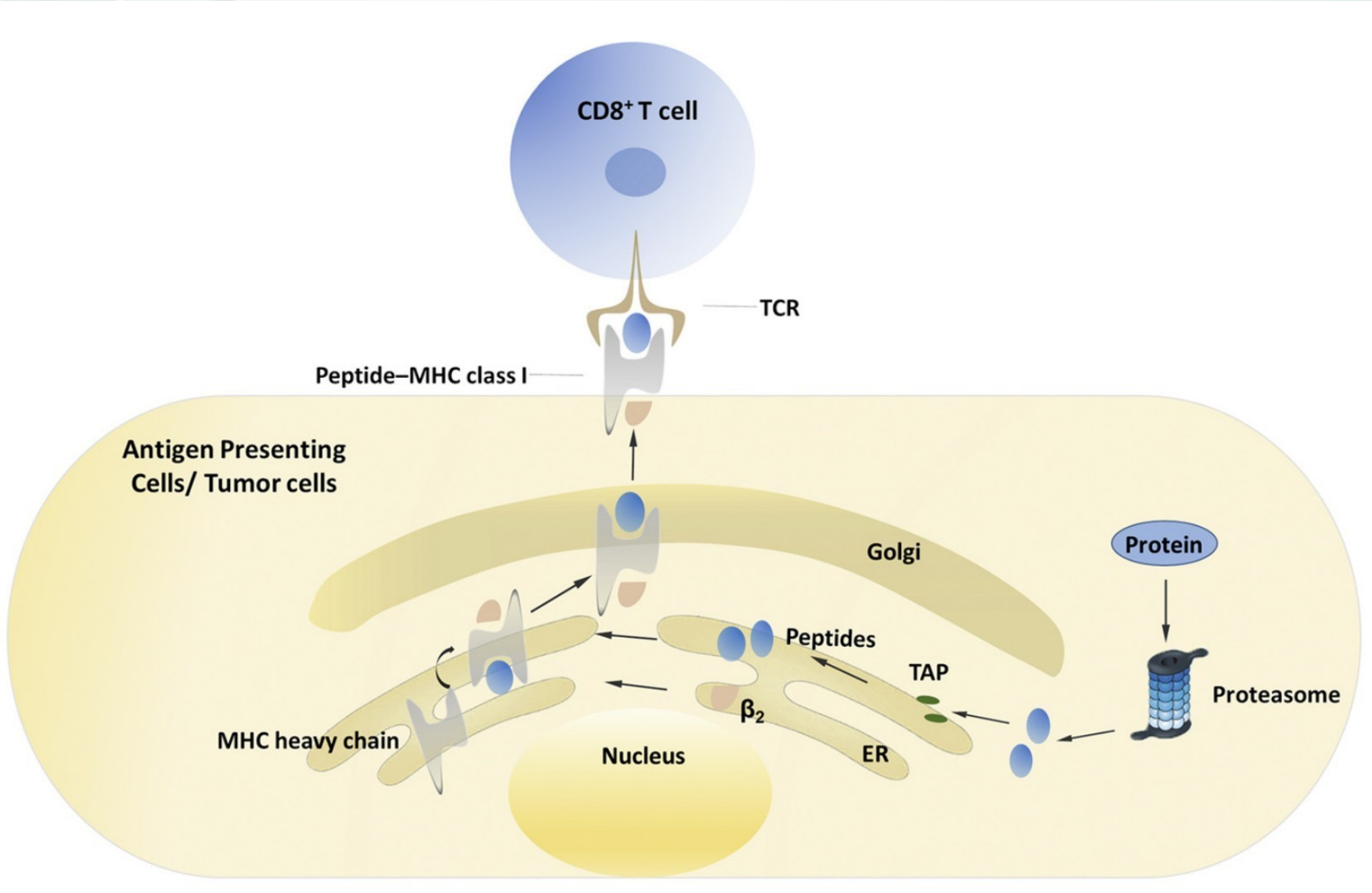

Figure: Presentación de antígenos por MHC-I. Fuente: [9]

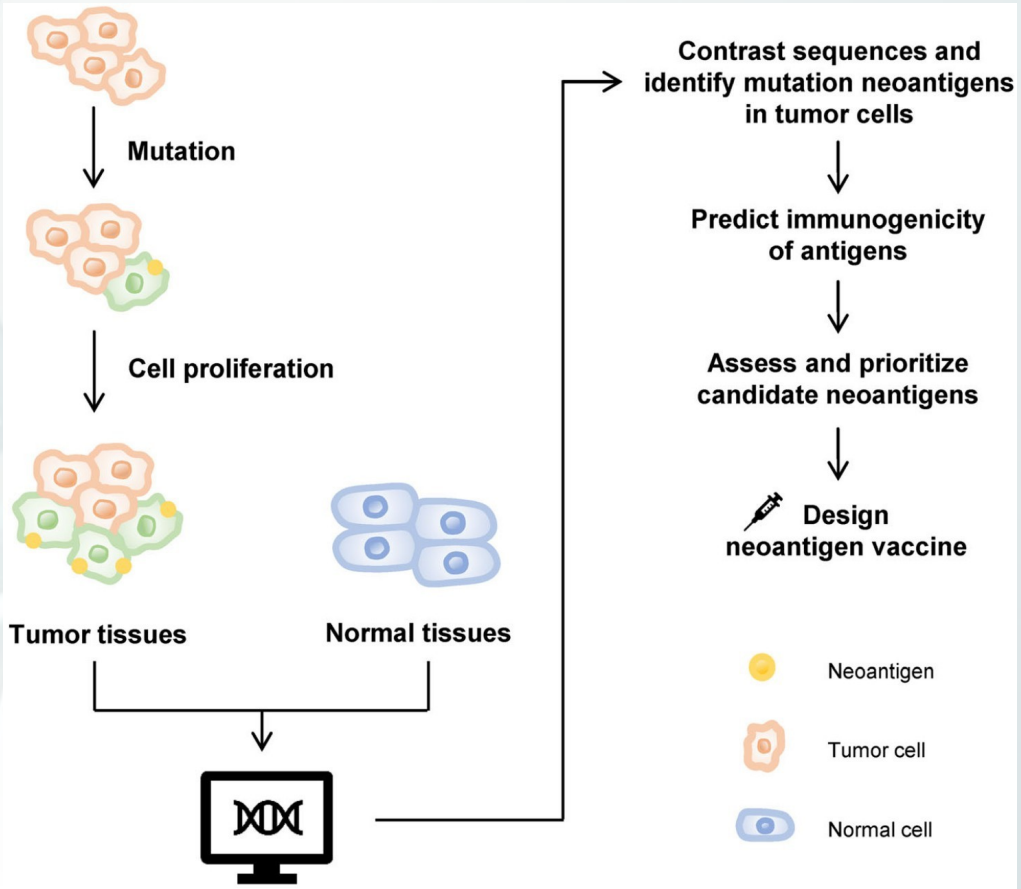

Figure: Proceso para la generación de vacunas personalizadas [10].

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Motivación

El cáncer representa el mayor problema de salud mundial, pero lamentablemente los métodos basados en cirugías, radioterapias, quimioterapias tienen baja efectividad [10].

La inmunoterapia del cáncer es una alternativa para el desarrollo de vacunas personalizadas, pero este proceso depende de una correcta detección de neo antígenos [11, 10].

Problema

Menos del 5% de los neoantígenos detectados logran activar a las células T (sistema inmune) [11, 12, 13, 14, 15].

Objetivos

Desarrollar un método basado en *deep learning* que mejore el acierto de la detección de neoantígenos a partir de la predicción del enlace pMHC.

Desarrollar una aplicación Web que permita realizar la predicción del enlace pMHC.

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Propuesta

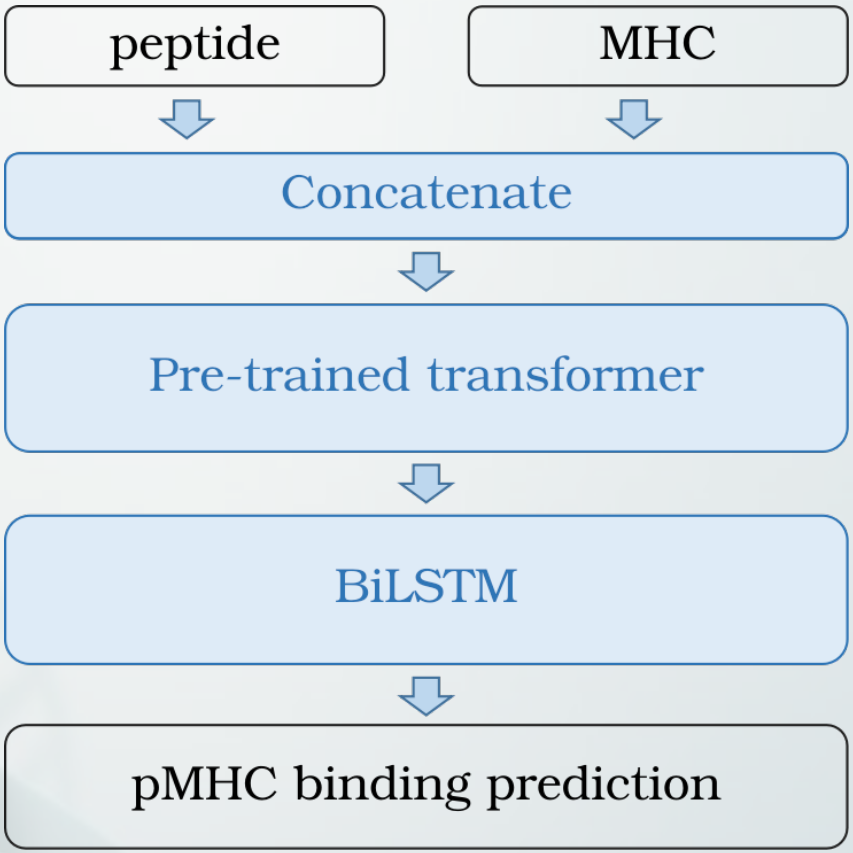

Figure: Propuesta

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Protein Language Models

Table: Pre-trainned BERT models for several protein tasks: TAPE, ProtBert, ESM1, and ESM-2.

Model	Dataset	Samples	Layers	Hidden size	Att. heads	Params.
TAPE	Pfam	30M	12	768	12	92M
ProtBert-BFD	BFD	2122M	30	1024	16	420M
ProtT5-XL	Uniref50, BFD	2122M	24	1024	32	3B
ProtT5-XXL	Uniref50, BFD	2122M	24	1024	128	11B
ESM-1 (6 layers)	Uniref50	60M	6	768	12	43M
ESM-1 (12 layers)	Uniref50	60M	12	768	12	85M
ESM-1 (34 layers)	Uniref50	60M	34	1280	20	670M
ESM-1b	Uniref50	60M	34	1280	20	650M
ESM-2 (6 layers)	Uniref50	60M	6	320	20	8M
ESM-2 (12 layers)	Uniref50	60M	12	480	20	35M
ESM-2 (30 layers)	Uniref50	60M	30	640	20	150M
ESM-2 (33 layers)	Uniref50	60M	33	1280	20	650M
ESM-2 (36 layers)	Uniref50	60M	36	2560	20	3B
ESM-2 (48 layers)	Uniref50	60M	48	5120	20	15B

Hiperparámetros

Table: Hyper-parameters configuration used to train ESM2 models in order to evaluate if they got into vanishing gradient problem.

Configuration	lr	Epochs	Warmup steps	Batch size
c1	4e-4	6	2000	16
c2	4e-4	6	2000	8
c3	2e-5	6	2000	16
c4	1e-5	30	101066	16
c5	2e-6	60	202132	16

Experimentos

1. Evaluar las configuraciones de hiperparámetros al hacer fine-tuning a los modelos ESM2.
2. Comparar el desempeño de LoRA, distillation y un método de congelamiento de capas.
3. Comparar el desempeño de los pLMs: TAPE, ProdBert-BFD y ESM2 para la tarea de pMHC.
4. Comparar el mejor modelo de los experimentos anteriores con los métodos del estado del arte.

Resultados

Table: Resultados obtenidos en cada base de datos.

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

Resultados

Table: AUC entre la propuesta, BERTMHC, NetMHCpan3.2, PUFFIN y MHCnuggets.

Modelo	AUC
Propuesta	0.73
BERTMHC	0.72
NetMHCpan3	0.68
PUFFIN	0.69
MHCnuggets	0.58

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Conclusiones

En esta investigación se propuso el uso de un modelo *transformer* ya entrenado con una base de datos de 30 millones de proteínas. Luego, esta red fue conectada de forma paralela con una red CNN.

El uso de *transfer learning* es una buena opción para suplir la falta de muestras en ciertos problemas y reducir el tiempo de entrenamiento.

La propuesta llego a mejorar los mejores métodos de detección de afinidad entre un péptido y una proteína MHC-II. Como trabajo futuro, se planteará la misma propuesta para proteínas MHC-I.

Predecir la afinidad entre un péptido y una proteína MHC, es uno de los paso mas importantes par calificar al péptido como un neo antígeno, capaz de generar una respuesta inmunitaría.

Contenido

Marco teórico
- Bioinformática y DNA
- Mutaciones
- Neo antígenos

Problema y Objetivos
- Motivación y Problema
- Objetivo

Propuesta

Experimentos y Resultados

Conclusiones

Trabajos futuros

Trabajos futuros

Recientemente un trabajo [23] tambien propone el uso de *transfer learning* pero de un modelo pre-entrenado con 250 millones de proteínas. Entonces, se plantea utilizar la misma red, aumentar la cantidad de muestras y evaluar los resultados.

Actualmente se cuenta con una base de datos de proteínas MHC [29], entonces utilizando AlphaFold de Google, se plantea predecir la estructura de varios péptidos y analizar el enlace péptido-MHC desde un punto de vista de la computación gráfica.

References I

[1] NCI,
"National cancer institute dictionary," 2022.

[2] NCI,
"Nci dictionary of cancer terms,"
https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription, 2020,
Accessed: 2020-03-20.

[3] PacBio,
"Two review articles assess structural variation in human genomes," https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-ge
2021,
Accessed: 2021-05-07.

[4] Cancer.net,
"Qué es la inmunoterapia," 2022.

References II

[5] NortShore,
"Immunotherapy," 2022.

[6] 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.

[7] 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.

References III

[8] Qing Hao, Ping Wei, Yang Shu, Yi-Guan Zhang, Heng Xu, and Jun-Ning Zhao,
"Improvement of neoantigen identification through convolution neural network,"
Frontiers in immunology, vol. 12, 2021.

[9] Xiaomei Zhang, Yue Qi, Qi Zhang, and Wei Liu,
"Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy,"
Biomedicine & Pharmacotherapy, vol. 120, pp. 109542, 2019.

[10] 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.

References IV

[11] 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.

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

References V

[13] 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.

[14] 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.

References VI

[15] 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.

[16] 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.

References VII

[17] Guangzhi Wang, Huihui Wan, Xingxing Jian, Yuyu Li, Jian Ouyang, Xiaoxiu Tan, Yong Zhao, Yong Lin, and Lu Xie, "Ineo-epp: a novel t-cell hla class-i immunogenicity or neoantigenic epitope prediction method based on sequence-related amino acid features," *BioMed research international*, vol. 2020, 2020.

[18] 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.

References VIII

[19] 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.

[20] Jingcheng Wu, Wenzhe Wang, Jiucheng Zhang, Binbin Zhou, Wenyi Zhao, Zhixi Su, Xun Gu, Jian Wu, Zhan Zhou, and Shuqing Chen,
"Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity,"
Frontiers in Immunology, p. 2559, 2019.

References IX

[21] 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.

[22] Preeti Bais, Sandeep Namburi, Daniel M Gatti, Xinyu Zhang, and Jeffrey H Chuang,
"Cloudneo: a cloud pipeline for identifying patient-specific tumor neoantigens,"
Bioinformatics, vol. 33, no. 19, pp. 3110–3112, 2017.

[23] Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov,
"Improved predictions of mhc-peptide binding using protein language models,"
bioRxiv, 2022.

References X

[24] Jun Cheng, Kaïdre Bendjama, Karola Rittner, and Brandon Malone,
"Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning,"
Bioinformatics, vol. 37, no. 22, pp. 4172–4179, 2021.

[25] 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.

[26] 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.

References XI

[27] Xiaoshan M Shao, Rohit Bhattacharya, Justin Huang, IK Sivakumar, Collin Tokheim, Lily Zheng, Dylan Hirsch, Benjamin Kaminow, Ashton Omdahl, Maria Bonsack, et al., "High-throughput prediction of mhc class i and ii neoantigens with mhcnuggetshigh-throughput prediction of neoantigens with mhcnuggets," *Cancer immunology research*, vol. 8, no. 3, pp. 396–408, 2020.

[28] Haoyang Zeng and David K Gifford, "Quantification of uncertainty in peptide-mhc binding prediction improves high-affinity peptide selection for therapeutic design," *Cell systems*, vol. 9, no. 2, pp. 159–166, 2019.

References XII

[29] Deylane Menezes Teles Oliveira, Rafael Melo Santos de Serpa Brandão, Luiz Claudio Demes da Mata Sousa, Francisco das Chagas Alves Lima, Semiramis Jamil Hadad do Monte, Mário Sérgio Coelho Marroquim, Antonio Vanildo de Sousa Lima, Antonio Gilberto Borges Coelho, Jhonatan Matheus Sousa Costa, Ricardo Martins Ramos, et al., "phla3d: An online database of predicted three-dimensional structures of hla molecules," *Human Immunology*, vol. 80, no. 10, pp. 834–841, 2019.