Universidad La Salle

Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context

MSc. Vicente Machaca Arceda

2023

Content

Introduction

- Immunotherapy to Treat Cancer
- Problem

Related Works

Proposal

Preliminary Results

- Models and databases
- Comparison of LINEAR, RNN and RNN-ATT
- Comparison of pre-trained BERT models

Conclusions

Content

- Introduction
  - Immunotherapy to Treat Cancer
  - Problem
- Related Works
- Proposal
- Preliminary Results
  - Models and databases
  - Comparison of LINEAR, RNN and RNN-ATT
  - Comparison of pre-trained BERT models
- Conclusions

Immunotherapy to Treat Cancer

Immunotherapy is a type of cancer treatment that helps your immune system fight cancer [1].

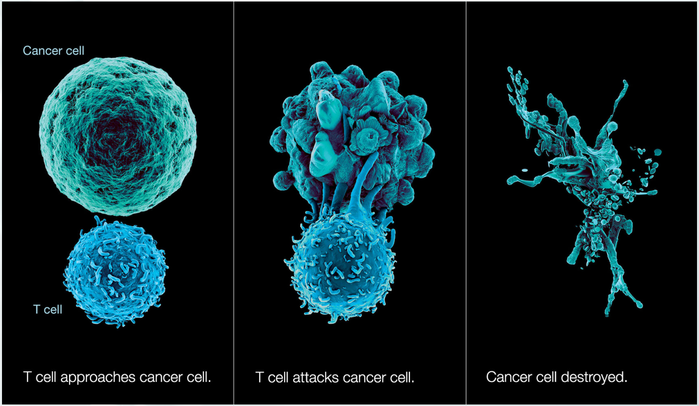


Figure: Example of how a T cell attack a cancer cell [2].

Immunotherapy to Treat Cancer

Neoantigen

Neoantigen

A new protein that forms on cancer cells when certain mutations occur in tumor DNA. Neoantigens used in vaccines and other types of immunotherapy are being studied in the treatment of many types of cancer [3, 4].

Currently, there is a lot of methods to detect neoantigens; however, only a small number of them manage to stimulate the immune system [5, 6].

Immunotherapy for Cancer

Personalized Vaccines

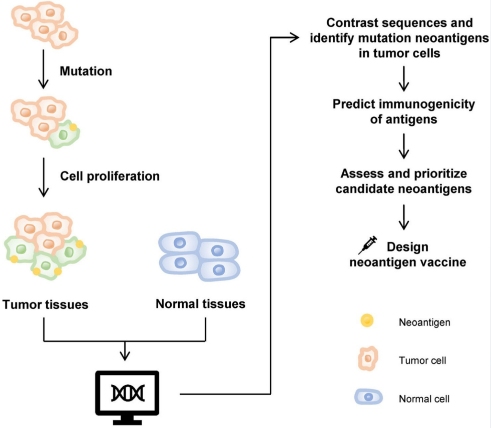

Mutation
Cell proliferation
Tumor tissues
Normal tissues
Contrast sequences and identify mutation neoantigens in tumor cells
Predict immunogenicity of antigens
Assess and prioritize candidate neoantigens
Design neoantigen vaccine
Neoantigen
Tumor cell
Normal cell

Figure: Personalized vaccines process for Cancer [7].

pMHC binding and presentation prediction

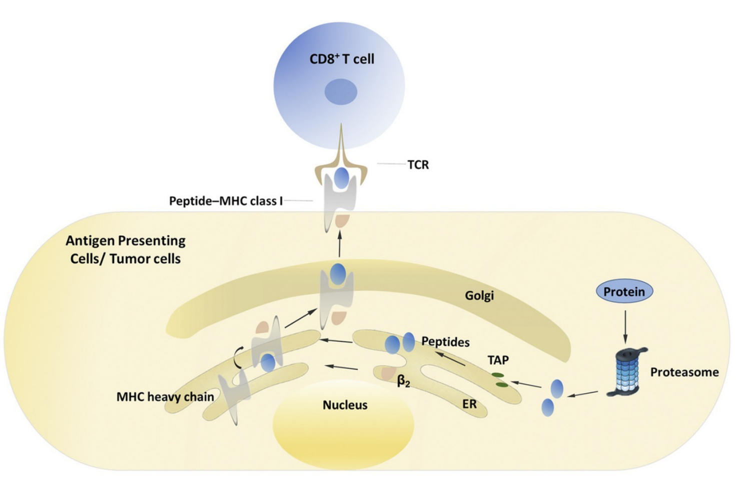


Figure: pMHC presentation process in MHC class I [8].

Content

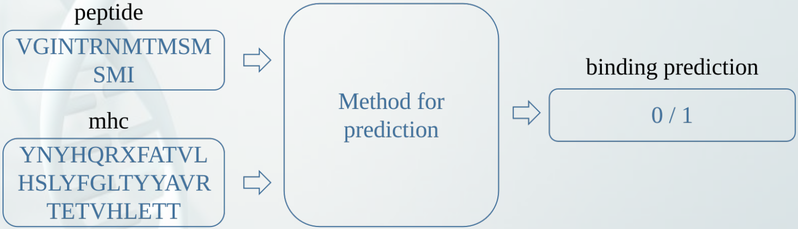
- Introduction
  - Immunotherapy to Treat Cancer
  - Problem
- Related Works
- Proposal
- Preliminary Results
  - Models and databases
  - Comparison of LINEAR, RNN and RNN-ATT
  - Comparison of pre-trained BERT models
- Conclusions

Problem

Less than 5% of detected neoantigens (peptides binded to MHC) succeed in activating the immune system [9].

This is a **binary classification problem**. A peptide could be represented like: $p = \{A, ..., Q\}$ and a MHC like: $q = \{A, N, ..., Q, E\}$. Finally, we need to know the probability of affinity between p and q (pMHC)

Problem



peptide

VGINTRNMTMSMSMI

mhc

YNYHQRXFATVLHSLYFGLTYYAVRTETVHLETT

Method for prediction

binding prediction

0 / 1

Figure: pMHC binding prediction problem.

Related Works

Transformers

Table: Recent works based on transformers and transfer learning.

Year	Ref.	Name	Method
2022	[10]	**HLAB**	Uses protBert model incascade with a RNN with attention
2022	[11]	MHCRoBERTa	Five encoders with 12 multiple-head self-attention pre-trainned with self-supervision
2022	[12]	**TransPHLA**	Based on four modules: an embedding block, an encoder block (multiple self-attention), a feature optimization block (FC layer), and a projection block (FC layer used to predict)
2021	[13]	BERTMHC	Uses TAPE model followed by a linear layer.
2021	[14]	ImmunoBERT	The same as BERTMHC focused on MHC-class I

Proposal

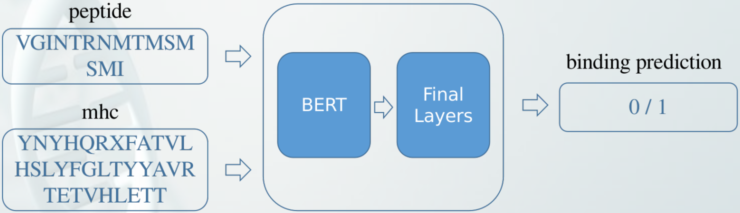


Figure: Proposal for pMHC binding and presentation prediction.

Content

Introduction

- Immunotherapy to Treat Cancer
- Problem

Related Works

Proposal

Preliminary Results

- Models and databases
- Comparison of LINEAR, RNN and RNN-ATT
- Comparison of pre-trained BERT models

Conclusions

Databases

We used the dataset from NetMHCIIpan3.2 [15] and HLAB [10].

Table: Number of samples used in training, evaluation and testing.

	NetMHCIIpan3.2	HLAB
Train	107424	539019
Validation	13428	179673
Testing	13429	172580

Models

We are going to evaluate these BERT models: ESM1-b [16], PortBert [17], ESM2 [18], and TAPE [19]. Moreover, the Bi-LSTM with attention layer is based on HLAB [10].

Table: Final layers in cascade after the BERT architecture.

	Description
LINEAR	BERT architecture followed by a linear layer
RNN	BERT architecture followed by a BiLSTM layer and then a Linear layer
RNN-ATT	BERT architecture followed by a BiLSTM layer with attention and then a Linear layer

Content

Introduction

- Immunotherapy to Treat Cancer
- Problem

Related Works

Proposal

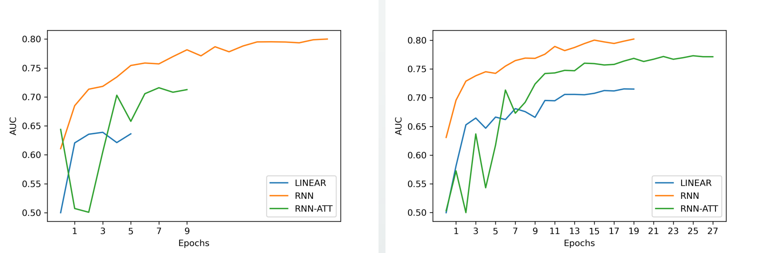
Preliminary Results

- Models and databases
- Comparison of LINEAR, RNN and RNN-ATT
- Comparison of pre-trained BERT models

Conclusions

Training

Comparison of LINEAR, RNN and RNN-ATT



(a) Training in ESM2_t6_8M

(b) Training in ESM2_t30_150M

Figure: Training comparison of ESM2 models in NetMHCIIpan3.2 dataset. We used 30 epochs with early stooping.

Testing

Comparison of LINEAR, RNN and RNN-ATT

Table: F1-score comparison of ESM2 (BERT model) followed by a LINEAR, RNN and RNN-ATT layers. It was evaluated in NetMHCIIpan3.2 dataset.

Bert Model	Linear	RNN	RNN-ATT
ESM2_T6_8M	nan	**0.7679**	0.6684
ESM2_T12_35M	0.6638	**0.7734**	0.7367
ESM2_T30_150M	0.6709	**0.7714**	0.7363

Content

Introduction

- Immunotherapy to Treat Cancer
- Problem

Related Works

Proposal

Preliminary Results

- Models and databases
- Comparison of LINEAR, RNN and RNN-ATT
- Comparison of pre-trained BERT models

Conclusions

Comparison of pre-trained BERT models

Table: Comparison of pre-trained BERT mpdels: TAPE, ESM2, and PortBert. We trained these models (followed in cascade by RNN layers) in HLAB dataset for three epochs.

Models	AUC	Precis.	Recall	F1	Acc
tape_freeze	0.9345	0.9283	0.9416	0.9348	0.9345
esm2_t6	**0.9351**	**0.9253**	**0.9464**	**0.9357**	**0.9351**
esm2_t12	0.9344	0.9251	0.9451	0.9350	0.9344
esm2_t30	0.9303	0.9185	0.9440	0.9311	0.9303
esm2_t33	0.6816	0.7139	0.6044	0.6546	0.6818
protbert_bfd	0.9083	0.9176	0.8968	0.9071	0.9083
netMHCpan4.1	0.9006	**0.9586**	0.8372	0.8938	0.9007

Conclusions

We compared the performance of **LINEAR, RNN and RNN-ATT** layers in cascade after ESM2 model trained in NetMHCIIpan3.2 dataset (107424 samples). This experiment shows how the **RNN layer (BiLSTM) outperformed the others**.

Then, we compared **ESM2, TAPE, and ProtBert** models followed by RNN layers. In this case, we trained the models in the HLAB dataset (539019 samples). from these experiment, **ESM2_t6_8M outperformed other models and even NetMHCpan4.1**. It is important to clarify that we froze the BERT architecture to accelerate the training.

References I

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

[2] NortShore,
"Immunotherapy," 2022.

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

[4] 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.

References II

[5] 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.

[6] 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.

[7] 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 III

[8] 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.

[9] 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.

References IV

[10] 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.

[11] Fuxu Wang, Haoyan Wang, Lizhuang Wang, Haoyu Lu, Shizheng Qiu, Tianyi Zang, Xinjun Zhang, and Yang Hu,
"Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences,"
Briefings in Bioinformatics, vol. 23, no. 3, pp. bbab595, 2022.

References V

[12] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Dennis Russell Salahub, Qin Xu, Jianmin Wang, Xue Jiang, et al.,
"A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design,"
Nature Machine Intelligence, vol. 4, no. 3, pp. 300–311, 2022.

[13] 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.

References VI

[14] Hans-Christof Gasser, Georges Bedran, Bo Ren, David Goodlett, Javier Alfaro, and Ajitha Rajan,
“Interpreting bert architecture predictions for peptide presentation by mhc class i proteins,”
arXiv preprint arXiv:2111.07137, 2021.

[15] Kamilla Kjaergaard Jensen, Massimo Andreatta, Paolo Marcatili, Søren Buus, Jason A Greenbaum, Zhen Yan, Alessandro Sette, Bjoern Peters, and Morten Nielsen,
“Improved methods for predicting peptide binding affinity to mhc class ii molecules,”
Immunology, vol. 154, no. 3, pp. 394–406, 2018.

References VII

[16] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al.,
"Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,"
Proceedings of the National Academy of Sciences, vol. 118, no. 15, 2021.

[17] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al.,
"Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arxiv 2020,"
arXiv preprint arXiv:2007.06225, 2007.

References VIII

[18] 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.

[19] 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.