DOI 10.18699/SBB-2022-669

Improved MHC–peptide class I interaction prediction with pretrained transformer

Shaburova E.V. [1, 2]*, Antonets D.V.[1, 2]

[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk, Russia*

[2] *Novel Software Systems LLC, Novosibirsk, Russia*

* *shaburova.liz@yandex.ru*

Key words: MHCI-peptide presentation prediction, transformers, vaccines development

Motivation and Aim: Increase in amount of promising neoantigens and viral antigens has paved the way for development of highly specific vaccines. Precise prediction of the binding and presentation of peptides to major histocompatibility complex (MHC) alleles is an important step toward such therapies. The ideal tool for such a task should be panspecific, be able to learn from both binding affinity and presentation data sources, and be interpretable. Here we present a transformer neural network model which leverages self-supervised pretraining from a large corpus of protein sequences with the supervised training for MHCI-peptide interaction prediction, and benchmark it with current SOTA models.

Methods and Algorithms: For training and validation we used publicly available datasets on binding affinity data (BA) and mass-spectrometry data (EL), which reflects both events of binding and presentation respectively [1]. The model takes peptide and pseudosequence of MHC I allele, process it with their own pretrained BERT (we took pretrained protBERT model [2] and additionaly pretrained it with our own data) to get embedding presentation of both amino acid sequences. After embeddings being concatenated, they undergo through 2 dense layers to get a prediction of binding affinity and presentation. As a benchmark models we selected MHCNuggets [3] and MHCFlurry 2.0 [4]. In order to compare performance on the test dataset, we measured accuracy, recall, precision, F-score, MCC (Matthews correlation coefficient) and RocAUC.

Results: Our model show comparable perfomance with both MHCNuggets and MHCFlurry models (MHCFlurry-BA, MHCFlurry-AP and MHCFlurry-PS). Also, being trained on EL dataset only, our model slightly outperformed the other models trained on the data of the same modality (MHCNuggets and MHCFlurry-AP), which suggest the capability of transformer architecture to accumulate more relevant information and generalize better.

Conclusion: The results obtained surmise transformer neural network model is being helpful in peptides binding and presentation prediction, which means they could be presented as a part of vaccine development pipeline. The further research should be focused on the ways of immunogenicity data inclusion to the training dataset and experiments with model architecture. Moreover, implying the fact that presentation of both class I and II epitopes are critical for the induction of a sustained effective immune response, the same model could be trained for both MHCI and MHCII peptide presentation prediction.

References

1. Reynisson B. et al. 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 Res.* 2020;48(1):449-454.
2. Elnaggar A. et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *bioRxiv*. 2020.
3. Shao X.M. et al. High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets. *Cancer Immunol Res.* 2020;8(3):396.
4. O'Donnell T.J., Rubinsteyn A., Laserson U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* 2020;11(1):42-48.