PredIG: a predictor of T-cell immunogenicity

R. Farriol-Duran[#1], M. Vazquez[*2], E. Porta[2], V. Guallar[#3]

[#] *Electronic and Atomic Protein Modelling, Life Sciences Department, Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona (Spain)*

[1]roc.farriol@bsc.es, [3]victor.guallar@bsc.es

[*] Cancer Immunogenomics, Josep Carreras Leukaemia Research Institute, Josep Carreras Building, Ctra de Can Ruti, Camí de les Escoles, s/n, 08916 Badalona, Barcelona

[2]eduard.porta@bsc.es

***Keywords*— T-cell immunogenicity, HLA & TCR**

EXTENDED ABSTRACT

The identification of immunogenic epitopes (such as fragments of proteins, in particular peptides, that can trigger an immune response) is a fundamental need for immune-based therapies. A computational tool that could predict such immunogenic epitopes would have vast potential applications in biomedicine ranging, from vaccine design against viruses or bacteria to therapeutic vaccination of cancer patients. While there are several methods that predict whether a peptide will be shown to the immune system via the Human Leukocyte Antigen (HLA) proteins of a patient, most of them cannot predict whether such presentation will indeed trigger an immune response. Additionally, T-cell immunogenicity is determined by multiple cellular processes, some of which are often overlooked by the current state-of-the-art immunogenicity predictors.

The aim of this project is to build PredIG, an immunogenicity predictor that discriminates immunogenic from non-immunogenic T-cell epitopes given the peptide sequence and the HLA typing. After a careful study of the drivers of antigen processing and presentation on HLA class I molecules and an assessment of the physicochemical factors influencing epitope recognition by T-cell receptors (TCRs), we have used a selection of publicly available tools and in-house developed algorithms to identify the most relevant features that determine epitope immunogenicity.

We then used these features to build an immunogenicity predictor (PredIG) modelled by XGBoost against immunogenically validated epitopes by the ImmunoEpitope DataBase (IEDB)(1), the PRIME dataset(2) and the TANTIGEN database(3). Pondering the feature importance in the model, the in-house developed softwares, NOAH for HLA Binding Affinity and NetCleave for Proteasomal Processing were identified as the major contributors to the performance of the model.

Once PredIG was developed, we benchmarked the capacity to predict the immunogenicity of validated T-cell epitopes versus a set of state-of-the-art methods (Fig.1). Relevantly, PredIG showed a greater performance than the Immunogenicity predictors from Prime(2) and IEDB(4). Additionally, our results confirm that predicting T-cell immunogenicity based on data from T-cell assays is more accurate than using HLA Binding assays, the method mostly used in the field. An AUC value of 0.67 and an enrichment factor in the TOP10 epitopes of 90% outperforms the predictive performance of the available methods.

In the context of the immune response against cancers, T-cell immunogenicity of tumoral mutations has been described as a response biomarker for immunotherapies such as immune checkpoint inhibitors. Similarly, the presence of immune infiltrate in a tumor has been related to a better prognosis for many cancer types. What is missing is the link between T-cell immunogenicity of tumoral mutations and the capacity of a tumor to attract immune cells. For this reason, we correlated the PredIG immunogenicity score obtained in a dataset of the The Cancer Genome Atlas (TCGA) against the tumor infiltrate in such tumors demonstrating that rather the total number of mutations a tumor accumulates, the tumor mutation burden (TMB), it is the number of immunogenic mutations what should be accounted for as biomarker of response.

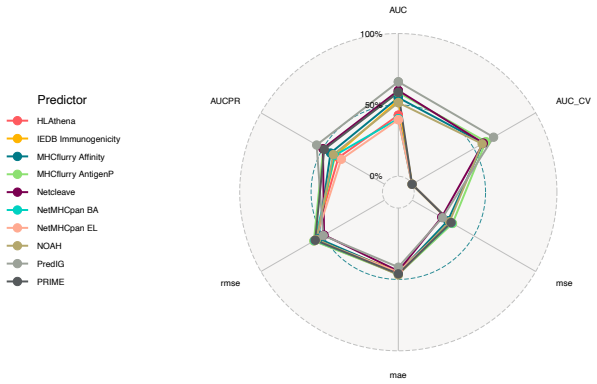

Fig. 1 Radar plot comparing the performance metrics of PredIG versus the benchmarked methods of HLA Binding Affinity (HLAthena(5), MHCflurry(6), NetMHCpan BA & EL(7), NOAH), HLA presentation (MHCflurry Antigen Presentation), Proteasomal Processing (Netcleave)(8) and Immunogenicity (Prime, IEDB immunogenicity). AUC = Area under the ROC Curve. AUC_CV = CrossValidated AUC. RMSE = Root Mean Square Error. MSE = Mean Square Error. MAE = Mean Absoulte Error. AUCPR = Area under the Precision-Recall Curve.

PredIG has shown an increased performance in the prediction of T-cell immunogenicity than the benchmarked state-of-the art methods. Therefore, it can be a useful resource to prioritize T-cell epitopes for their potential capacity to activate an immune response, a functionality that can be applied in vaccine design, cancer immunotherapies or deimmunization of therapeutic proteins.

ACKNOWLEDGEMENTS

I'd like to thank the supervisors of this work, Miguel Vazquez, Eduard Porta and Víctor Guallar for their continuous support and refreshing ideas.

References

1. W. Fleri, *et al.*, The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. *Front. Immunol.* **8**, 278 (2017).
2. J. Schmidt, *et al.*, Prediction of neo-epitope immunogenicity reveals TCR recognition

determinants and provides insight into immunoediting. *Cell Reports Med.* **2**, 100194 (2021).
3. L. R. Olsen, *et al.*, TANTIGEN: a comprehensive database of tumor T cell antigens. *Cancer Immunol. Immunother.* **66**, 731–735 (2017).
4. J. J. A. Calis, *et al.*, Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
5. S. Sarkizova, *et al.*, A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* (2019) https:/doi.org/10.1038/s41587-019-0322-9 (December 23, 2019).
6. T. J. O'Donnell, A. Rubinsteyn, U. Laserson, MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* **11**, 42-48.e7 (2020).
7. V. Jurtz, *et al.*, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).
8. P. Amengual-Rigo, V. Guallar, NetCleave: an open-source algorithm for predicting C-terminal antigen processing for MHC-I and MHC-II. *Sci. Reports 2021 111* **11**, 1–8 (2021).

Author biography

Roc Farriol-Duran was born in Sabadell, Catalunya, in 1995. He received the B.E. degree in Biomedical Sciences from the Autonomous University of Barcelona, Barcelona, Catalunya, in 2018, the Msc. degree in Translational Biomedical Research for the Vall d'Hebron Hospital Campus (VH) Barcelona, Catalunya, in 2019 and the Msc. Degree in Omics Data Analysis for the Vic University, Vic, Catalunya. During 2016-17 he stayed at the Proteomics Group at the Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands. Followingly, he joined the Cellular Immunology Laboratory at the Institute for Biotechnology and Biomedicine, Autonomous University of Barcelona and collaborated closely with the Proteomics Unit at the CSIC-UAB. Then, he joined the lab of Tumor Immunology and Immunotherapy at the Vall d'Hebron Oncology Institute (VHIO). Since May 2020 he has been at the Life Sciences Deparment of the Barcelona Supercomputing Center, first at the Computational Biology group and later at the Electronic and Atomic Protein Modelling group. Currently, he is pursuing a PhD degree in Computational Modelling of T-cell and B-cell immunogenicity.