Machine Learning-Enhanced T Cell Neoepitope Discovery for Immunotherapy Design

Cancer Informatics
Volume 18: 1–2

DOI: 10.1177/1176935119852081

Joana Martins[1,2], Carlos Magalhães[1,2], Miguel Rocha[3] and Nuno S Osório[1,2]

[1]Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal. [2]ICVS/3B PT Government Associate Laboratory, Braga/Guimarães, Portugal. [3]Centre of Biological Engineering, University of Minho, Braga, Portugal.

ABSTRACT: Immune responses mediated by T cells are aimed at specific peptides, designated T cell epitopes, that are recognized when bound to human leukocyte antigen (HLA) molecules. The HLA genes are remarkably polymorphic in the human population allowing a broad and fine-tuned capacity to bind a wide array of peptide sequences. Polymorphisms might generate neoepitopes by impacting the HLA-peptide interaction and potentially alter the level and type of generated T cell responses. Multiple algorithms and tools based on machine learning (ML) have been implemented and are able to predict HLA-peptide binding affinity with considerable accuracy. Challenges in this field include the availability of adequate epitope datasets for training and benchmarking and the development of fully integrated pipelines going from next-generation sequencing to neoepitope prediction and quality analysis metrics. Effectively predicting neoepitopes from *in silico* data is a demanding task that has been facilitated by ML and will be of great value for the future of personalized immunotherapies against cancer and other diseases.

KEYWORDS: neoepitopes, T cells, immunotherapy, machine learning, epitope prediction

RECEIVED: April 26, 2019. **ACCEPTED:** April 29, 2019.

TYPE: Short Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research received funding from Fundação para a Ciência e a Tecnologia (FCT) contract IF/00474/2014; PhD scholarship SFRH/BD/132797/2017.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Nuno S Osório, Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal.
Email: nosorio@med.uminho.pt

Main text

The immune system is a pervasive network of molecular mediators, cells, tissues, and organs with central relevance in human health. Importantly, immunotherapy designed to modulate the activity of CD8 or CD4 T cell lymphocytes is a promising and increasingly relevant pillar for cancer treatment.[1] T cells recognize epitopes in the context of human leukocyte antigen (HLA) molecules that are highly diverse and have a broad but fine-tuned capacity of binding non-self-peptides to contribute for appropriate immune responses. The search for the motifs responsible for the binding of peptides to HLA molecules started several decades ago but the discovered complexity of HLA-peptide interactions prompted continuous development, namely using in silico approaches and artificial intelligence. Machine learning (ML) is a fundamental branch of artificial intelligence, which can be defined as a set of models and respective induction algorithms, able to learn complex relationships or patterns from empirical data and capable of making accurate decisions. Thus, ML intends to train models from large amounts of data, through specific algorithms that give the machine the ability to learn how to perform a specific task from data without being explicitly programmed. This approach can be used to analyse, interpret and predict the outcomes for unseen data, achieving results that would not be possible in many cases through conventional statistics. ML has several applications in medicine, including the field of immunoinformatics, which focuses on the in silico analysis and modelling of immunological data and problems. Immunoinformatics applications are growing and, consequently, are becoming more important to immunological research.[2] Experimental HLA-binding assays require synthesis and testing of overlapping peptides, including the full-length sequences of interest, which on a large-scale is an expensive and time-consuming laboratorial task. The primary objective of immunoinformatics research is to design efficient algorithms for the mapping of potential B cell and T cell epitopes. These tools can determine the sequence regions with potential binding sites, which in turn accelerates the development of novel immunotherapies.[3]

These methods were developed using models such as artificial neural networks (ANNs).[4] ANNs are one of the main ML models, providing a computational approach mimicking the information processing of the brain. These models can be used to solve problems of classification or regression.[5] Their architecture is composed of a set of 'artificial neurons' distributed in layers: the input layer receiving the initial data, the hidden or intermediate layer(s) that are responsible for extracting the patterns associated with the data, and the output layer, which presents the final result of the process. The learning capability is one of the most important characteristics of neural networks. Thus, from a sample, the neural network learns the relationship between the inputs and outputs and can produce solutions for any new example. The learning process consists of a gradient-descent based algorithm that iteratively changes the synaptic weights associated with the neurons of the network seeking to minimize a given cost function, typically based on an error metric (loss function) computed over the training examples.[5]

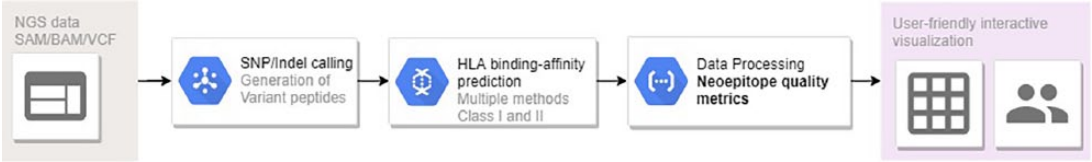

Figure 1. Workflow for automated and integrated bioinformatics frameworks going from next-generation sequencing data inputs to neoepitope prediction, quality analysis and visualization.

Currently, several ML-based epitope prediction tools are available. The Immune Epitope Database and Analysis Resource (IEDB) is a widely used resource hosting a database of experimentally validated epitopes and tools for *de novo* prediction. NetMHCpan[6] and NetMHCIIpan[7] are two tools that generate quantitative predictions of peptide binding affinity to class I and class II HLAs, respectively. NetMHCpan and NetMHCIIpan have high accuracy in available datasets but are closed source software with limited licence user agreements. MHCflurry[8] is an ensemble of HLA class I allele-specific predictors, whose accuracy, implementation, and open-source licencing makes it very attractive for large-scale epitope prediction studies. Another ML-based class I HLA prediction method is MHCnuggets[9], which uses gated-recurrent ANNs to process sequences directly and handle peptides of any length without artificial lengthening or shortening. Similarly, needed improvements in class II epitope prediction tools are in development with tools such as MixMHC2pred.[10] Despite of its immense potential, successful ML-based approaches are highly dependent on the datasets available for training and testing.[11] The datasets of validated T cell epitopes found in databases are almost entirely formed of epitopes from bacteria or viruses and were not obtained by standardized experimental methodologies. Improvements in the available epitope datasets will likely boost the performance of ML-based predictors and facilitate tool benchmarking. Recent developments in HLA peptidomics[12] for class I and II HLA molecules have been relevant for this goal, opening doors to evaluate the potential of deep learning tools for T cell epitope predictions. However, neoepitopes are rare and challenging to discover. Maximizing the probability of identifying clinically relevant neoepitopes requires the development of integrated frameworks to generate multiple method epitope prediction and advanced neoepitope quality metrics. Recent webservers such as HABIT (http://habit.evobiomed.com/) address this issue by automating HLA-binding prediction and the interpretation of the impact of amino acid variants in peptide-HLA binding. Future developments should include user-friendly webtools to allow automated treatment of next-generation sequencing (NGS) data, epitope prediction and neoepitope propensity quality analysis with interactive visualization (Figure 1). Overall, the continued development in the field of ML-based epitope prediction is of great value for the rational design of T cell-based immunotherapies to cancer and other relevant diseases.

ORCID iD

Nuno S Osório https://orcid.org/0000-0003-0949-5399

REFERENCES

1. O'Donnell JS, Teng MW, Smyth MJ. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nat Rev Clin Oncol*. 2018;16:151–167. doi:10.1038/s41571-019.
2. Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*. 2015;7:119. doi:10.1186/s13073-015-0245-0.
3. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. *Open Biol*. 2013;3:120139. doi:10.1098/rsob.120139.
4. Maclin PS, Dempsey J, Brooks J, Rand J. Using neural networks to diagnose cancer. *J Med Syst*. 1991;15:11–19. doi:10.1007/BF00993877.
5. Mitchell TM. Does machine learning really work? *AI Magazine*. 1997;18:11. doi:10.1609/aimag.v18i3.1303.
6. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199:3360–3368. doi:10.4049/jimmunol.1700893.
7. Jensen KK, Andreatta M, Marcatili P, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. 2018;154:394–406. doi:10.1111/imm.12889.
8. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*. 2018;7:129-132e4. doi:10.1016/j.cels.2018.05.014.
9. Bhattacharya R, Sivakumar A, Tokheim C, et al. Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. *bioRxiv*. 2017;2017:154757. doi:10.1101/154757.
10. Racle J, Michaux J, Rockinger GA, et al. Deep motif deconvolution of HLA-II peptidomes for robust class II epitope predictions. *bioRxiv*. 2019;2019:539338. doi:10.1101/539338.
11. Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York, NY: Wiley; 2001. doi:10.1142/9789814335461_0005.
12. Kalaora S, Barnea E, Merhavi-Shoham E, et al. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget*. 2016;7:5110. doi:10.18632/oncotarget.6960.