A complete software suite for de novo sequencing and predicting the immunogenicity of class 1 and class 2 neoantigens

Ngoc Hieu Tran, Xin Chen, Lei Xin, Baozhen Shan, Ming Li

Bioinformatics Solutions Inc., Waterloo, Ontario, Canada

Introduction

Tumor-specific neoantigens are only expressed on the surface of cancer cells and hence can help the immune system to distinguish cancer cells from normal cells. Neoantigens may arise from different sources, including missense mutations, non-coding regions, bacteria-derived, which are challenging to identify using a database search approach. Thus, deep learning-based *de novo* sequencing models specialized for human leukocyte antigen (HLA) peptides represent an ideal solution to identify candidate neoantigens from mass spectrometry (MS) immunopeptidomics data. In addition, only a small fraction of HLA peptides can be recognized by T cells to trigger immune responses. The chance of finding such immunogenic neoantigens is remarkably low, usually less than half a dozen out of thousands of somatic mutations detected per patient. Thus, *in silico* methods are essential to accurately predict the immunogenicity and prioritize candidate neoantigens before *in vitro* validations, which often involve time-consuming and costly experiments. *De novo* sequencing of neoantigens and predicting their immunogenicity are essential for cancer immunotherapy and vaccine design.

Methods

DeepNovo-HLA

Here we presented DeepNovo-HLA, a deep learning-based *de novo* sequencing model specialized for HLA peptides (Figure 1).

1. DeepNovo-HLA was trained on a very large, carefully curated dataset from 20 immunopeptidomics studies. The dataset includes nearly 3,000 runs on two major MS instruments Orbitrap and timsTOF, 26M MS/MS spectra, 1.1M HLA-I and HLA-II peptides, and covering >90 alleles for each class.

2. DeepNovo-HLA neural networks were designed to learn both the fragment ion patterns in MS/MS spectra and the amino acid patterns of HLA peptides, especially their allele motifs patterns at the anchor positions.

3. DeepNovo-HLA was integrated in our peptidome workflow for MS-based immunopeptidomics to identify candidate neoantigens from non-canonical sources.

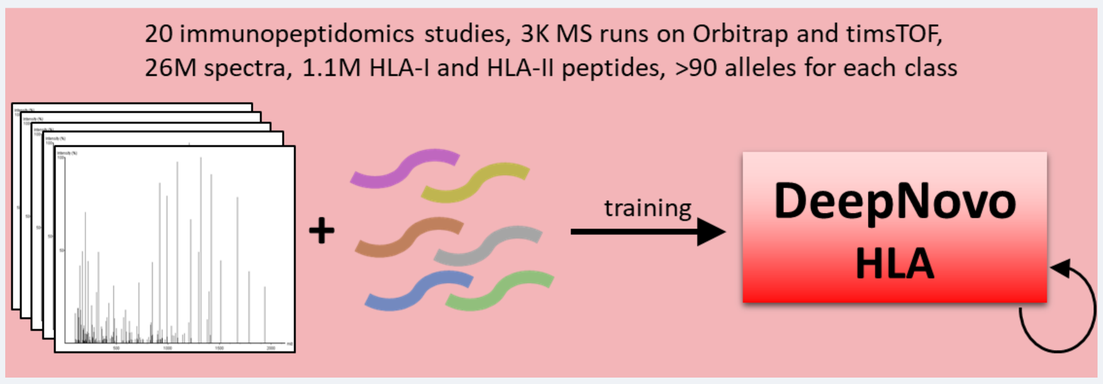

Figure 1. DeepNovo-HLA, a deep learning-based de novo sequencing model specialized for HLA peptides.

DeepImmu

After candidate neoantigens were identified by DeepNovo-HLA, we used DeepImmu to predict their immunogenicity and prioritize top candidates for *in vitro* validations (Figure 2).

1. DeepImmu is a personalized model for immunogenicity prediction based on the central tolerance, i.e. the positive and negative selection of T cells in an individual patient. In the positive selection, T cells are selected by their ability to bind to peptide-HLA complexes. In the negative selection, they are selected ***against*** their ability to bind to self peptides.

2. DeepImmu used HLA self peptides obtained from MS-based immunopeptidomics to resemble the negative selection of T cells in each individual patient. For the positive selection, we collected all epitopes reported in positive T cell assays from the Immune Epitope Database (IEDB) that matched the patient's HLA alleles.

3. Using this personal dataset of negative and positive peptides, DeepImmu model was trained specifically for that patient to predict his/her T cell responses to candidate neoantigens. A bi-directional long short-term memory (LSTM) network coupled with amino acid embedding was used as the model architecture.

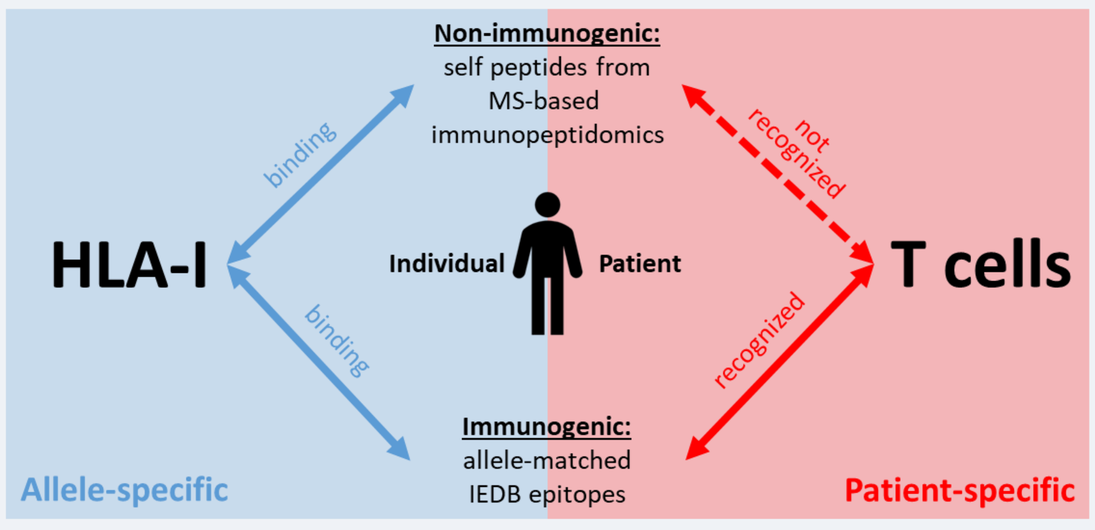

Figure 2. DeepImmu, a personalized immunogenicity prediction model based on the central tolerance of T cells in an individual patient.

Results

Evaluation of DeepNovo-HLA and the peptidome workflow

We evaluated DeepNovo-HLA and the peptidome workflow on the immunopeptidomics datasets of ten cancer patients from previously published studies. HLA-I data was available for all ten patients, whereas HLA-II data was available for patients 1 and 3-6. For comparison, we also performed *de novo* sequencing and non-enzyme-specific database search using our standard proteomics softwares PEAKS De novo and PEAKS DB on these datasets.

Figure 3a shows that the peptide accuracy of DeepNovo-HLA was 53-70% on HLA-I peptides and 21-29% on HLA-II peptides. On average, DeepNovo-HLA achieved 11% higher peptide accuracy than PEAKS De novo on HLA-I peptides and 6% higher on HLA-II peptides. Powered by DeepNovo-HLA, our peptidome workflow identified 21-61% more HLA-I peptides and 23-35% more HLA-II peptides than PEAKS DB (Figure 3b). DeepNovo-HLA and the peptidome workflow consistently outperformed PEAKS across all ten patient datasets and on both HLA classes.

We also double-checked the correctness of the identified HLA peptides by verifying their characteristic length distribution and their binding to the respective patient's alleles using NetMHCpan. Figures 3c,d show an example of the length distribution and the binding %Rank of HLA-I peptides identified by PEAKS DB and the peptidome workflow on patient 1. The improved identifications of the peptidome workflow over PEAKS DB can be observed across different peptide lengths and different levels of binding %Rank. These results again confirm the robust performance of our peptidome workflow.

Figure 3. Evaluation of DeepNovo-HLA and the peptidome workflow on the immunopeptidomics datasets of 10 cancer patients. a, Accuracy of de novo HLA peptides. **b**, Number of identified HLA peptides. **c,d**, Length distribution and NetMHCpan binding prediction for HLA peptides identified from patient 1.

Evaluation of DeepImmu and other immunogenicity prediction tools

We evaluated DeepImmu on the immunopeptidomes and neoantigens of 18 cancer patients from ten previously published studies. The number of neoantigens varied from 2-6 per patient, all of which had been confirmed as immunogenic by T cell assay validation. In the evaluation set of each patient, the ratio between immunogenic neoantigens to random non-immunogenic peptides was set at 1:100, i.e. an immunogenicity prediction tool needs to prioritize 1 neoantigen out of every 100 negative peptides. We also compared DeepImmu to three other tools, including PRIME (ver. 2.0), NetMHCpan (ver. 4.1), and IEDB predictor.

Figures 4a,b show the areas under the receiver operating characteristic curves (ROC-AUC) of the four tools on each patient and on the combined evaluation set of 18 patients. DeepImmu achieved an overall AUC of 0.69 and outperformed the other tools on 12 of 18 patients. Figure 4c further shows the ranks of neoantigens on top of non-immunogenic peptides. DeepImmu ranked 25% of the neoantigens within its top 10% predictions, and nearly 75% of the neoantigens within its top 30% (Q1 and Q3 quartiles of the boxplot, respectively). That is, for an average evaluation set containing 4 neoantigens and 400 non-immunogenic peptides, DeepImmu's top 40-120 predicted candidates shall contain 1-3 neoantigens of interest.

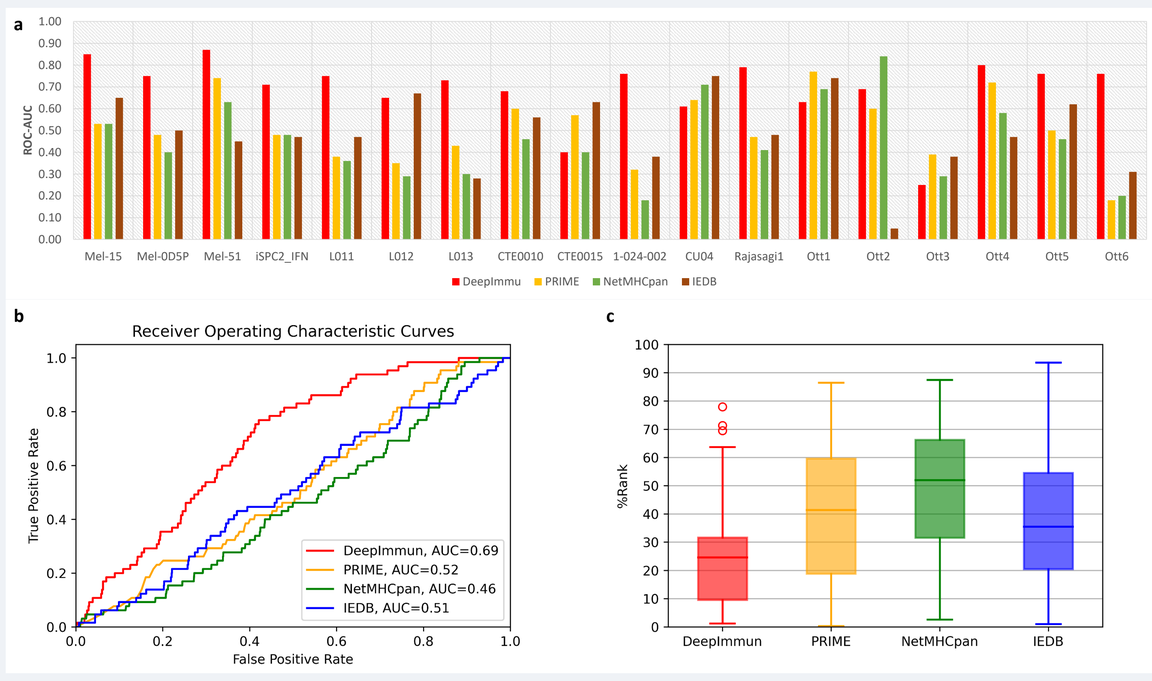

Figure 4. Performance evaluation of four immunogenicity prediction tools on the neoantigens of 18 cancer patients. a, Areas under the receiver operating characteristic curves (ROC-AUC) of the prediction tools on each individual patient. **b**, ROC curves on the combined evaluation set of 18 patients. **c**, Predicted ranks of immunogenic neoantigens versus non-immunogenic peptides (lower % is better).

Scan QR to…

@PEAKSmassspec

@PEAKSProteomics

Bioinformatics Solutions Inc.

Summary

- DeepNovo-HLA and DeepImmu together enabled a complete software suite for *de novo* sequencing and predicting the immunogenicity of class 1 and class 2 neoantigens.
- By taking advantage of a huge amount of MS-based immunopeptidomics data and the idea of T cell central tolerance, our tools substantially improved not only the sensitivity in neoantigen identification but also the accuracy in neoantigen selection for cancer immunotherapy.