Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2: toward universal blueprints for vaccine designs

Brandon Malone
NEC Research Laboratories Europe https://orcid.org/0000-0002-7027-3157
Boris Simovski
NEC OncoImmunity AS https://orcid.org/0000-0002-3629-1570
Clément Moliné
NEC OncoImmunity AS https://orcid.org/0000-0003-2699-3226
Jun Cheng
NEC Research Laboratories Europe https://orcid.org/0000-0001-5573-9791
Marius Gheorghe
NEC OncoImmunity AS https://orcid.org/0000-0002-9666-0809
Hugues Fontenelle
NEC OncoImmunity AS https://orcid.org/0000-0003-0416-3096
Ioannis Vardaxis
NEC OncoImmunity AS https://orcid.org/0000-0002-0296-1687
Simen Tennøe
NEC OncoImmunity AS https://orcid.org/0000-0002-0120-8992
Jenny-Ann Malmberg
NEC OncoImmunity AS
Richard Stratford
NEC OncoImmunity AS
Trevor Clancy (✉ trevor@oncoimmunity.com)
NEC OncoImmunity AS https://orcid.org/0000-0001-9896-0613

Method Article

Keywords: vaccine design, immune informatics, sars-cov-2

Posted Date: November 26th, 2020

DOI: https://doi.org/10.21203/rs.3.pex-1006/v1

Abstract

This protocol predicts blueprints for vaccine design that contain a broad repertoire of T-cell epitopes optimized for the global population. The protocol first requires a screening of the SARS-CoV-2 proteome using immunogenicity predictors to generate comprehensive epitope maps. Then, these epitope maps are used as input to Monte Carlo simulations designed to identify statistically significant “epitope hotspot” regions in the virus that are most likely to be immunogenic. The epitope hotspots that share significant homology with proteins in the human proteome are removed to reduce the chance of inducing off-target autoimmune responses. Finally, a database of the actual HLA genotypes of citizens is used to develop a “digital twin” type simulation to model how effective different combinations of hotspots would work in a diverse human population. The approach identifies an optimal constellation of epitope hotspots that could provide maximum coverage in the human population.

Introduction

The global population is at present suffering from a pandemic of Coronavirus disease 2019 (COVID-19), caused by the novel coronavirus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The goal of this study was to use artificial intelligence (AI) to predict blueprints for designing universal vaccines against SARS-CoV-2, that contain a sufficiently broad repertoire of T-cell epitopes capable of providing coverage and protection across the global population. To help achieve these aims, we profiled the entire SARS-CoV-2 proteome across the most frequent 100 HLA-A, HLA-B and HLA-DR alleles in the human population, using host-infected cell surface antigen presentation and immunogenicity predictors from the *NEC Immune Profiler* suite of tools, and generated comprehensive epitope maps. We then used these epitope maps as input for a Monte Carlo simulation designed to identify statistically significant “epitope hotspot” regions in the virus that are most likely to be immunogenic across a broad spectrum of HLA types. We then removed epitope hotspots that shared significant homology with proteins in the human proteome to reduce the chance of inducing off-target autoimmune responses. We also analyzed the antigen presentation and immunogenic landscape of all the nonsynonymous mutations across 3400 different sequences of the virus, to identify a trend whereby SARS-COV-2 mutations are predicted to have reduced potential to be presented by host-infected cells, and consequently detected by the host immune system. A sequence conservation analysis then removed epitope hotspots that occurred in less-conserved regions of the viral proteome. Finally, we used a database of the HLA genotypes of approximately 22 000 individuals to develop a “digital twin” type simulation to model how effective different combinations of hotspots would work in a diverse human population; the approach identified an optimal constellation of epitope hotspots that could provide maximum coverage in the global population. By combining the antigen presentation to the infected-host cell surface and immunogenicity predictions of the *NEC Immune Profiler* with a robust Monte Carlo and digital twin simulation, we have profiled the entire SARS-CoV-2 proteome and identified a subset of epitope hotspots that could be harnessed in a vaccine formulation to provide a broad coverage across the global population.

Reagents

Equipment

Procedure

Module 1: Generate epitope maps

For a given HLA allele, the score allocated to an amino acid corresponds to the best score obtained by an epitope prediction overlapping with this amino acid. For Class I HLA alleles, the epitope lengths are 8, 9, 10 and 11, and predicted for antigen presentation (AP) or immune presentation (IP) of the viral peptide to host-infected cell surface, generated using the NEC Immune Profiler software. These Class I scores range between 0 and 1, where by 1 is the best score, i.e., higher likelihood of being naturally presented on the cell surface (AP) or being recognized by a T cell (IP). For Class II HLA alleles, we only consider 15mers. The Class II were predictions were percentile rank binding affinity scores (not antigen presentation), so the lower scores are best (the scores range from 0 to 100, with 0 being the best score).

Module 2: Monte Carlo simulation detection of epitope epitope hotspots

Input data: The data sets inputted into the statistical framework are epitope maps from module 1 generated for each amino-acid position in all the proteins in the SARS-CoV-2 proteome, for all of the studied 100 HLA alleles. A score for any given amino acid was determined as the maximum AP or IP score that a peptide overlapping that amino acid holds in the epitope map. All peptide lengths of size 8-11 amino acids for Class I, and 15 for Class II were processed, generating one HLA dataset per viral protein. Each row in the dataset represents the amino acid epitope scores predicted for one HLA type.

HLA tracks: The raw input datasets are first transformed into binary tracks. For each Class I HLA dataset, the epitope scores are transformed to binary (0 and 1) values, such that amino-acid positions with predicted epitope scores larger than 0.7 for AP, or larger than 0.5 for IP, are assigned the value 1 (positively predicted epitope), and the rest are assigned the value 0. Similarly, for Class II HLA datasets, amino-acid positions with predicted epitope scores smaller than 10 are assigned the value 1, otherwise 0. Each binary track can effectively be presented as a list of segments, intervals of consecutive ones, and gaps, intervals of consecutive zeros.

Test statistic: For a group of K HLA binary tracks, a test statistic is calculated for each bin of given size dividing the protein in *n* bins. For details of the test statistic please see associated publication.

Null model: A null model is defined as the generative model of the HLA tracks, if they were generated by chance. From the null model, through sampling, arises the null distribution of the test statistic. Amino acids that are positively predicted as epitopes will clump together in segments with minimal length of 8, which is the shortest peptide length for which epitope scores are predicted, and often form longer segments when the source peptides overlap each other. Similarly, non-epitope amino acids will form gaps, with a minimal possible length of 1 amino acid. To preserve these features of the observed HLA tracks in the null model, as a sampling strategy we selected to shuffle the order of the segments and the gaps, respectively, within an HLA track.

Module 3: Filters of non-conserved peptides and human tissue expression

For each protein within the viral genome, the set of unique amino acid sequences was compiled from all the strains available in the GISAID database. These sets were individually processed using the Clustal Omega (v1.2.4) software via the command line interface with default parameter settings.The median conservation score was calculated by sampling 1,000 sub-sequences equal to the epitope hotspot size from the entire consensus sequence of a protein. Each sample was assigned a conservation score, and the median value from all 1,000 conservation scores was calculated. The minimum conservation score was calculated using a sliding window approach, with the window size being equal to the hotspot size. For each increment, a conservation score was calculated and the resulting minimum conservation score was kept.

In addition, to reduce the potential for off-target autoimmune responses against host tissue, we removed hotspots that contained exact sequence matches for all epitope lengths analyzed to proteins in the human proteome**.**

Module 4: The digital twin simulation framework

Figure 1 gives an overview of the problem setting.

(Please see the paper description for the exact equations.)

Step 1. Select a set of candidate vaccine elements

For this work, we select the hotspots filtered through Module 3.

Step 2. Create a set of “digital twin” citizens

- Creating a distribution over genotypes for each region.

We collect full HLA genotypes from actual citizens available from high-quality samples in the Allele Frequency Net Database (AFND). we collect all genotypes observed at least once across all regions; we assign an index *g* to each genotype, and we call the total number of unique genotypes as *G*. Second, we specify a prior distribution over genotypes. We use a symmetric Dirichlet distribution with concentration parameter of 0.5 because this distribution is uninformative in an information theoretic sense and does not reflect strong prior beliefs that any particular genotypes are more likely to appear in any specific region. For each region, we then calculate a posterior distribution over genotypes as a Dirichlet distribution using standard equations.

- Creating a set of “digital twin” citizens.

We create a set of digital twin citizens using a two-step approach. Our method must be given the population size *p*, as well as a distribution over regions. Concretely, the input is a Dirichlet distribution over the regions, as well as *p*. (We note that this Dirichlet is completely independent of those over genotypes discussed in the previous section.) The number of citizens from each region is sampled using the same two-step sampling process described above.

Second, the genotypes for each region are sampled using the posterior distributions over genotypes discussed above.

Step 3. Create a tripartite graph

We next use the vaccine elements and digital twins to construct a tripartite graph that will form the basis of the optimization problem for vaccine design. The graph has three sets of nodes:

1. All candidate vaccine elements identified in Step 1

2. All HLA alleles in all digital twin genotypes

3. All digital twins

The graph also has two sets of weighted edges:

1. An edge from each vaccine element to each HLA allele. The weight of this edge is the likelihood of no response for the allele from that particular vaccine element.

2. An edge from each allele to each citizen which has that allele in its genotype. The weight of these edges is always 1.

As an intuition, we call the edges from a vaccine element to an allele (and, then, from the allele to each patient with that allele) as “active” when the vaccine element is selected. Then, the log likelihood of response for a citizen is the sum of all active incoming edges. That is, the flow from selected vaccine elements to the citizens gives the likelihood of no response for that citizen.

Calculating the likelihood of no response for a given digital twin and vaccine elements.

Most short peptide prediction engines, e.g., netMHC, compute some sort of a score that a peptide will result in some immune response (e.g., binding, presentation, cytokine release, etc.), and this score generally takes into account a specific HLA allele. In some cases, this is already a probability, and in others, it can be converted into a probability using a transformation function, such as a logistic function.

Step 4. Selecting a set of vaccine elements

Finally, we pose the vaccine design problem as a type of network flow problem through the graph defined in Step 3. In particular, the minimization problem can be posed as an integer linear program (ILP); thus, it can be provably, optimally solved using conventional ILP solvers.

- Handling the minimax problem.

As previously described, our goal is to choose the set of vaccine elements which minimize the log likelihood of no response for each patient. We ignore any interactions among vaccine elements. Standard ILP solvers cannot directly solve this minimax problem; however, we use the standard approach of a set of surrogate variables to address this problem. We then minimize the surrogate variables.

Please see the paper for exact equations and mathematical notation.

Troubleshooting

Time Taken

Epitope maps (Module 1): 6-12 hours

Monte Hotspot detection (Module 2): 2-6 hours

Non conserved filters (Module 3): 2-4 hours

Digital twin simulation (Module 4): 2-4 hours

Anticipated Results

A subset of epitope hotspots that could be harnessed in a vaccine formulation to provide a broad coverage across the global population.

Figures

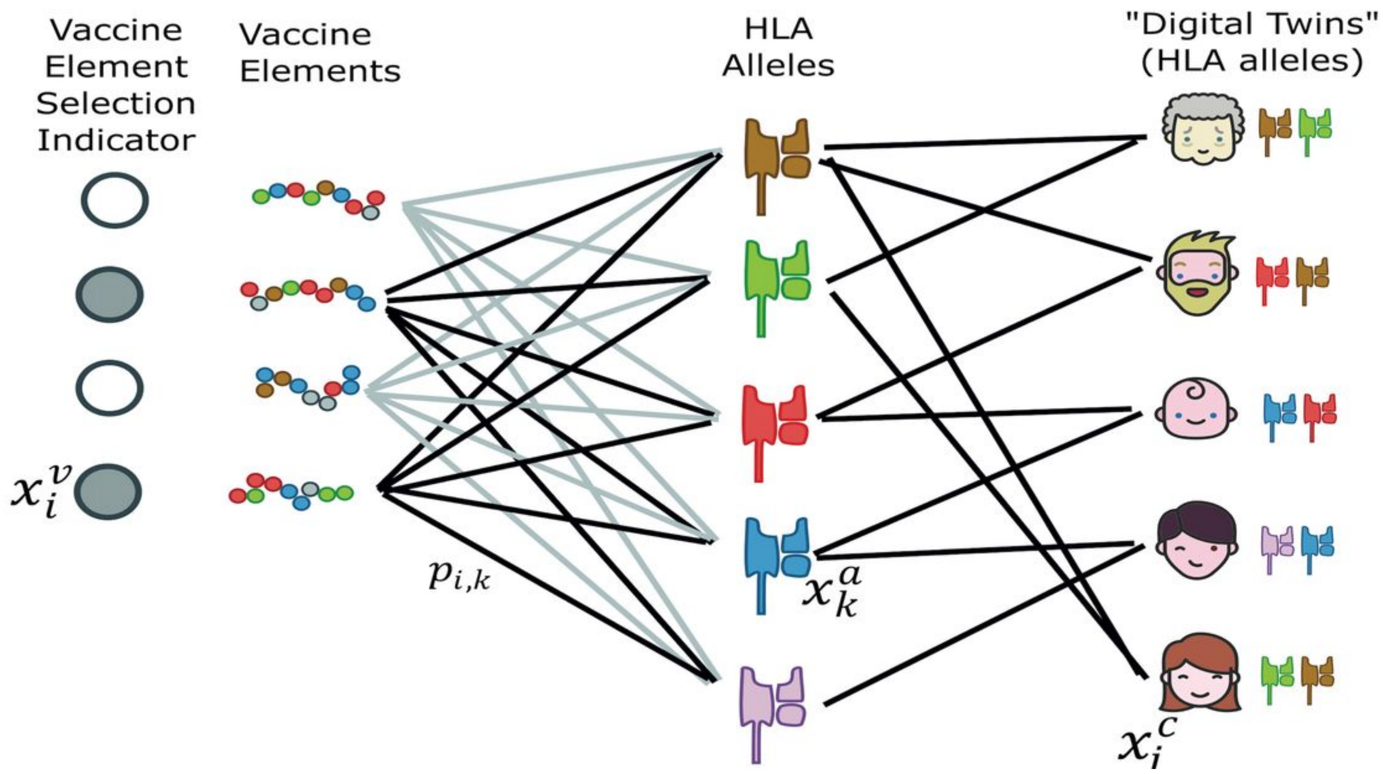
Vaccine Element Selection Indicator
Vaccine Elements
HLA Alleles
"Digital Twins" (HLA alleles)
x_i^v
$p_{i,k}$
x_k^a
x_i^c

Figure 1

Schematic of the problem setting. The vaccines elements were the significant epitope hotspots that emerged from the statistical hotspot detection framework