Supplementary Materials

Predicting Antigen Presentation – what could we learn from a million peptides?

Running title: Predicting Antigen Presentation

David Gfeller\textsuperscript{1,2,*}, Michal Bassani-Sternberg\textsuperscript{3,*}

\textsuperscript{1} Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Switzerland.
\textsuperscript{2} Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.
\textsuperscript{3} Department of Oncology, Ludwig Institute for Cancer Research, University Hospital of Lausanne, Switzerland.

\textsuperscript{*}To whom correspondence should be addressed: david.gfeller@unil.ch, Michal.Bassani@chuv.ch.
Supplementary Materials

HLA peptidomics datasets

Publicly available HLA-I peptidomics datasets from several recent studies in humans were considered in this work (1–12). All HLA-I peptidomics datasets were analyzed with our mixture model algorithm (MixMHCp) to identify and annotate HLA-I binding motifs, as described in our previous studies (1,13). All motifs were manually reviewed to eliminate cases where motifs of different alleles could not be separated. Larger numbers of motifs than the number of alleles were sometimes needed to identify motifs supported by few peptides (e.g., HLA-C motifs). Samples where the motifs could not be annotated were only considered in the list of peptides, but not in the list of interactions reported in Figure 1 (the same applies for samples where the HLA-I typing information was not available). For ref. (10) the raw MS data were reprocessed and were not filtered with existing predictors. For ref. (5), the unfiltered list of peptides was kindly provided to us by the authors and motifs representing endogenous HLA-I alleles (i.e., HLA-C04:01 and HLA-B35:03) and the peptides associated to such motifs were identified with MixMHCp.

Recent HLA-II peptidomics datasets were included in the analysis of the number of class II peptides (3,14–17). Since allelic restriction was in general not known and is more difficult to predict, these data were not included in the list of HLA-II ligand interactions, only in the list of peptides in Figure 1.

IEDB data

IEDB data were downloaded on March 8, 2018 (18). All non-negative data were considered here (i.e., “Positive-High”, “Positive-Intermediate”, “Positive-Low” and “Positive”). MS data were identified based on “ligand presentation” identifier. All the other data (mainly binding affinity assays) were classified as in vitro data. IEDB data were combined with the HLA peptidomics data from the studies mentioned above and the cumulative lists of unique peptides and unique interactions are displayed in Figure 1 as a function of time (years). When computing the number of interactions, only interactions with full information about the HLA allele (e.g., HLA-A01:01) were considered.

HLA-I ligand predictor based on Position Weight Matrices
Here we recall a method to build PWMs describing the binding specificity of HLA-I molecules. This approach has been used in the past by different groups (19–22), including ourselves (2), but the description of the different steps is often scattered across different publications. Therefore we thought it may be useful to review the mathematical aspects of this method, which may also help understand the discussion about potential biases in HLA-I ligand datasets.

Let \( X = (X_1, ..., X_P) \) be the set of \( P \) peptides of length \( L \), i.e. \( X_p = (X_{p1}, ..., X_{pL}) \), interacting with a given HLA-I molecule. Matrices describing the frequency of amino acid \( a \) at position \( l \) are computed as:

\[
M_a^l = \frac{1}{P} \sum_{p=1}^{P} \delta_{a,X_{pl}^l}
\]

where \( \delta_{a,b} \) stands for the standard Kronecker symbol and is equal to 1 if \( a=b \) and 0 otherwise, \( l = 1, ..., L \) standing for the positions along the peptides and \( a = 1, ..., 20 \), standing for the different amino acids.

**Redundant sequences:** correction for redundant sequences can be done in different ways (21). For instance, each sequence \( X_p \) can be given a weight \( w_p \), corresponding to the inverse of the number of sequences in \( X \) with identity with \( X_p \) above a certain threshold. Similarly, sequences can be first clustered and then each sequence within a cluster receives a weight inversely proportional to the size of the cluster. The PWM is then computed as

\[
M_{\text{eff}}^l = \sum_{p=1}^{P} w_p \delta_{a,X_{pl}^l},
\]

with \( P_{\text{eff}} = \sum_{p=1}^{P} w_p \).

**Pseudocounts:** pseudocounts represent a way of smoothing the amino acid frequencies and are especially important for low frequency amino acids. They are equivalent to priors in Maximum Likelihood approaches (technically the exponent of a Dirichlet prior corresponds to a flat pseudocount). In its simplest form, flat pseudocounts can be used and the PWM becomes

\[
\hat{M}_a^l = \frac{p_{\text{eff}} M_a^l + \beta/20}{p_{\text{eff}} + \beta}.
\]

A more powerful approach consists of using the BLOSUM62 matrix. The main idea is that if a given amino acid \( b \) is observed frequently at a given position, the pseudocounts should be larger for amino acids that are similar to \( b \). This idea can be quantified as

\[
\hat{M}_a^l = \frac{p_{\text{eff}} M_a^l + g_a^l}{p_{\text{eff}} + g_a^l}, \quad \text{with} \quad g_a^l = \sum_{b=1}^{20} M_b^l q^{b \rightarrow a},
\]

where \( q^{b \rightarrow a} \) represents the transition probability of amino acid \( b \) into amino acid \( a \), as given by the BLOSUM matrix (23). The score of a new peptide \( Y=(Y_1, ..., Y_L) \) is then given by

\[
S(Y) = \sum_{l=1}^{L} \hat{M}_{Y_l}^l,
\]

with \( f_r(a) \) being the background frequency of amino acid \( a \) (typically including also the pseudocount corrections, for consistency).
\[ \prod_{i=1}^{L} \tilde{\mathbf{M}}_{y_i}, \] or alternatively as \[ \sum_{i=1}^{L} \log(\tilde{\mathbf{M}}_{y_i}) \] (pseudo-counts ensure that all entries of \( \tilde{\mathbf{M}} \) are larger than zero).

**Clustering HLA-I motifs**

Position Weight Matrices for each HLA-I allele were built from all alleles with available ligands (9-mers only). When not enough data were available from unfiltered HLA peptidomics samples (<200 peptides), MS data from IEDB have been included. For alleles with few (<20 peptides) or no MS data available, *in vitro* binding data from IEDB have been used (stars in Figure 2 and 3). Pearson correlation coefficient between 180-dimensional vectors representing each PWM was computed for each pair of HLA-A, respectively HLA-B and HLA-C alleles. Hierarchical clustering (hclust in R, method="ward.D2") was applied using 1 minus the correlation coefficient as the distance between PWMs in the clustering algorithm. Supertypes were used as defined in (24).
### Supplementary Figures

<table>
<thead>
<tr>
<th></th>
<th>Pooled HLA peptidomics</th>
<th>Transfected soluble HLA-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0303</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0602</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0701</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C12:03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1402</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1601</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure S1:** Comparison of HLA-C motifs obtained by deconvolution of pooled HLA peptidomics data (left) and HLA peptidomics profiling of cell lines with transfected soluble HLA-C alleles (right).
Figure S2: Results of the motif deconvolution with MixMHCp on HEK293 cell line HLA peptidomics data from (11). This cell line is homozygous for all HLA-I alleles (HLA-A03:01, HLA-B07:02, HLA-C07:02). Four motifs are needed to see the motif for HLA-C07:02.

References


