This information is current as of October 4, 2017.

NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data

Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters and Morten Nielsen

J Immunol published online 4 October 2017
http://www.jimmunol.org/content/early/2017/10/04/jimmunol.1700893

Supplementary Material **http://www.jimmunol.org/content/suppl/2017/10/04/jimmunol.1700893.DCSupplemental**

Subscription Information about subscribing to *The Journal of Immunology* is online at: http://jimmunol.org/subscription

Permissions Submit copyright permission requests at: http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts Receive free email-alerts when new articles cite this article. Sign up at: http://jimmunol.org/alerts

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852
Copyright © 2017 by The American Association of Immunologists, Inc. All rights reserved.
Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

Published October 4, 2017, doi:10.4049/jimmunol.1700893

NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data

Vanessa Jurtz,* Sinu Paul,† Massimo Andreatta,‡ Paolo Marcatili,* Bjoern Peters,† and Morten Nielsen*,‡

Cytotoxic T cells are of central importance in the immune system's response to disease. They recognize defective cells by binding to peptides presented on the cell surface by MHC class I molecules. Peptide binding to MHC molecules is the single most selective step in the Ag-presentation pathway. Therefore, in the quest for T cell epitopes, the prediction of peptide binding to MHC molecules has attracted widespread attention. In the past, predictors of peptide–MHC interactions have primarily been trained on binding affinity data. Recently, an increasing number of MHC-presented peptides identified by mass spectrometry have been reported containing information about peptide-processing steps in the presentation pathway and the length distribution of naturally presented peptides. In this article, we present NetMHCpan-4.0, a method trained on binding affinity and eluted ligand data leveraging the information from both data types. Large-scale benchmarking of the method demonstrates an increase in predictive performance compared with state-of-the-art methods when it comes to identification of naturally processed ligands, cancer neoantigens, and T cell epitopes. ***The Journal of Immunology*****, 2017, 199: 000–000.**

Cytotoxic T cells play a central role in the immune regulation of pathogenesis and malignancy. They perform the task of scrutinizing the surface of cells for the non-self peptides presented in complex with MHC molecules. In cases in which such peptides are recognized, an immune response can be initiated, potentially leading to killing of the infected (malfunctioning) cell. Binding to MHC is the most selective step in the pathway leading to this peptide presentation.

Over the last decades, several efforts have been dedicated to the development of computational methods capable of accurately predicting this event. The accuracy of these methods has improved substantially over the last years, and most recent benchmark results demonstrate that >90% of naturally presented MHC ligands are identified at an impressive specificity of 98% (1). This gain in performance was achieved, in part, by the extended experimental binding data sets made available in the Immune Epitope Database (IEDB) (2) and, in part, by the development of novel machine-learning algorithms capable of capturing the information in the experimental binding data in a more effective manner. One such novel method is NNAlign-2.0, allowing the integration of peptides of variable length into the machine-learning framework (3). This novel training approach allows the incorporation of a larger set of training data and, maybe more importantly, enables the method to directly learn the length preference of presented peptides for each MHC molecule from the experimental binding data (4). Although most presented MHC class I ligands are 9 aa in length, the ability to incorporate length preferences directly into the model is critical, because experimental data demonstrate that the length profiles of presented ligands can vary substantially between MHC molecules. Prominent examples are the mouse H-2-Kb, with a preference for 8 aa–long peptides (5), and HLA-A*01:01, for which nearly one third of MHC-presented peptides have a length > 9 aa (6).

Some of the most well-documented and applied methods for predicting peptide binding to MHC class I include NetMHC (4, 7) and NetMHCpan (1, 8). Over the last years, these tools have garnered increasing interest because of the recent focus on neoantigen identification within the field of personalized immunotherapy (9, 10). However, as underlined in several studies, including the recent Nature Biotechnology Editorial (11), "neoantigen discovery and validation remains a daunting problem," primarily as a result of the relatively high false positive rate of predicted epitopes.

One potential cause for this relatively high rate of false positive epitope predictions is the fact that most methods are trained on binding affinity (BA) data and, as a consequence, only model the single event of peptide–MHC binding. As stated above, this binding to MHC is the most selective step in peptide Ag presentation. However, other factors, including Ag processing (12) and the stability of the peptide–MHC complex (13), could influence the likelihood of a given peptide to be presented as an MHC ligand. Similarly, the length distribution of peptides available for binding to MHC molecules is impacted by other steps in the processing and presentation pathway, such as TAP transport and ERAP trimming, which are not reflected in binding data in itself (6). Advances in mass spectrometry (MS) have allowed the field of MS peptidomics

*Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Lyngby, Denmark; †Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037; and ‡Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina

ORCIDs: 0000-0002-8964-0244 (S.P.); 0000-0002-8036-2647 (M.A.); 0000-0003-2615-5695 (P.M.); 0000-0002-8457-6693 (B.P.); 0000-0001-7885-4311 (M.N.).

Received for publication June 19, 2017. Accepted for publication September 6, 2017.

This work was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under Contract HHSN272201200010C and by the Agencia Nacional de Promoción Científica y Tecnológica, Argentina (Grant PICT-2012-0115).

Address correspondence and reprint requests to Dr. Morten Nielsen, Technical University of Denmark, Building 208, DK-2200 Lyngby, Denmark. E-mail address: mniel@cbs.dtu.dk

The online version of this article contains supplemental material.

Abbreviations used in this article: AUC, area under the receiver operating curve; BA, binding affinity; BA+EL, model trained on BA and EL data; BA+EL BA, BA predictions of model trained on combined BA and EL data; BA+EL EL, EL likelihood predictions of model trained on combined BA and EL data; EL, eluted ligand; IEDB, Immune Epitope Database; LOO, leave-one-out; MS, mass spectrometry.

Copyright © 2017 by The American Association of Immunologists, Inc. 0022-1767/17/$35.00

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

to move forward. In this context, recent studies (14–16) have suggested that training prediction methods on such data, rather than BA data, could improve the ability to accurately identify MHC ligands. As such, MS peptidome data would contain the comprehensive signal of Ag processing and presentation rather than just MHC BA. Moreover, MS peptidome data generated by immunopeptidomic studies would contain precise information about the allele-specific peptide-length profile preferences that is not available in the MHC BA data sets.

Thus, identification of MHC-bound peptides by MS holds great promise for the generation of large-scale data sets characterizing the peptidome specific for individual MHC molecules (15, 17) and potentially for the identification of T cell epitopes (18). However, it is clear that, within the foreseeable future, the number of MHC molecules characterized by such MS studies will remain limited. In this context, significant efforts over the last decades have been dedicated to experimentally characterizing the peptide-binding space of MHC molecules using semi–high-throughput MHC–peptide BA assays (19, 20), enabling binding-specificity characterization of a large set of MHC molecules from different species.

The IEDB contains a comprehensive set of MHC-binding and eluted ligand (EL) data available in the public domain. Although this data set contains BA data characterizing >150 MHC class I molecules (from human, nonhuman primates, mouse, and livestock), at the onset of this study only 55 MHC class I molecules were characterized by MS peptidome data. This imbalance caused us to suggest a novel machine-learning approach integrating information from both types of data (BA and MS ligands) into a combined framework benefitting from information from the two worlds. The proposed framework is "pan specific," because it can leverage information across MHC molecules, data types, and peptide lengths into a single model. Hence, we expect this approach to achieve superior predictive performance compared with models trained on the two data types individually, as well as achieve an improved performance when it comes to predicting length profile preferences of different MHC molecules.

Although recent studies have demonstrated the improved ability to identify MHC ligands using methods trained on MS peptidome data (14, 15), limited data are available with regard to their impact for the identification of T cell epitopes. In this article, we focus on demonstrating the improved prediction performance on large sets of MS peptidome data, as well as on T cell epitope data independent from the data used to train the new predictor.

Materials and Methods

Data sets

Data on all class I MHC ligand elution assays available in the IEDB database (http://www.iedb.org) were collected, including the ligand sequence, details of the source protein, position of the ligand in the source protein, and the restricting allele of the ligand. There were 160,527 distinct assays in total, and the length of the ligands ranged from 4 to 37 aa. All lengths that were associated with ≥0.5% of total ligands were selected for further analysis; this included lengths 8–15 aa and included 99% of the assay entries.

The restricting MHC molecule of the ligands was analyzed, and entries with alleles listed unambiguously were selected. For example, some entries for which the HLA alleles are listed as just the gene name, as well as alleles from chicken, horse, cow, and mouse for which we did not have binding prediction algorithms, were excluded. Representative alleles were assigned for entries where only supertypes were listed (e.g., HLA-A*26:01 for HLA-A26). Thus, there were 127 class I molecules from human and mouse in the selected data set. Redundant entries with the same ligand sequence and MHC molecule were removed, and MHC molecules with ≥50 ligand entries were selected. This included 55 class I molecules, and the number of available ligands per molecule varied widely from 50 to 9500.

We hypothesized that some of the ligands could be artifacts of the elution assays and, therefore, their source proteins could be false positive as Ags. A protocol was designed to identify such false positive Ags and exclude them from the final data selected. The protocol identified proteins that had a significantly smaller number of predicted binders among ligands than expected for random peptides using binomial probability distribution. Five sets of random peptides were generated from the ligand sequences by shuffling the amino acid residues within the ligands. BA was then predicted for the original ligands and random peptide sets for their corresponding alleles. The median of the predicted percentile ranks of the five random sets was estimated and assigned as the BA of the random peptides. Based on a predicted BA cut-off of percentile rank 1.0, the number of predicted binders among the original ligands and the random peptide sets was estimated. Thus, five proteins were identified as false positives, and ligand entries from these proteins were excluded from the data set.

The final data set had 85,217 entries in total, with ligand length ranging from 8 to 15 aa. The ligands originated from 14,797 source Ags and were restricted by 55 unique HLA molecules.

Random artificial negatives were generated for each MHC molecule covered by EL data by randomly sampling 10*N peptides of each length (8–15 aa) from the Ag source protein sequences, where N is the number of 9-mer ligands for the given MHC molecule.

Neural network training

The NNAlign training approach with insertions and deletions (3) was extended by adding a second output neuron, as shown in Fig. 1. This was done to allow combined training on BA and MS EL data. BA values are measured as IC_{50} values in nanomolar (aff) and can be rescaled to the interval [0,1] by applying $1 - \log(\text{aff})/\log(50{,}000)$, representing continuous target values (21). For ELs, the strength of the interaction between peptide and MHC molecules is unknown; therefore, a target value of 1 is assigned to binders, and 0 is assigned to artificial negative peptides (see above).

In this network architecture, weights between the input and hidden layer are shared between the two input types (BA/EL), and weights connecting the hidden and output layer are specific for each input type. During neural network training, an example is randomly selected from either data set and submitted to forward and back propagation, according to the NNAlign algorithm (3). In this setting, we define one training epoch as the average number of iterations needed to process each data point in the smaller data set once.

A neural network ensemble was trained, as described by Nielsen and Andreatta (1), using 5-fold nested cross-validation. Networks with 60 and 70 hidden neurons were trained, leading to an ensemble of 40 networks in total.

The inputs to the neural networks consisted of the peptide and the MHC molecule in terms of a pseudo-sequence (8). All peptides were represented as 9-mer binding cores by the use of insertions and deletions, as described by Andreatta and Nielsen (4), and were encoded using BLOSUM encoding (21). As in the earlier work by Andreatta and Nielsen (4), additional features for the encoding of peptides included the length of the deletion/insertion; the length of peptide-flanking regions, which are >0 in the case of a predicted extension of the peptide outside either terminus of the binding groove; and the length (L) of the peptide, encoded with four input neurons corresponding to the four cases $L \leq 8$, $L = 9$, $L = 10$, and $L \geq 11$.

Performance

To benchmark the combined training method described above (referred to as a model trained on BA and EL data [BA+EL]), additional methods with only one output, but an otherwise identical setup, were trained on BA data only (BA method) and EL data only (EL method). Performance was measured as area under the receiver operating curve (AUC); AUC = 0.5 indicates random model performance, whereas AUC = 1 represents a perfect model. AUC values were calculated for each MHC allele separately; subsequently, binomial tests were performed to compare the different models.

Length preference of MHC molecules

For all MHC molecules shared between the BA and EL data sets, we generated predictions for 80,000 random natural peptides of lengths 8–15 aa (10,000 of each length). From the top 2% of predictions, the frequency of each peptide length was estimated. Subsequently, the Pearson correlation coefficient was calculated between the frequencies observed in the EL data set and the frequencies predicted by four models (BA method, EL method, BA predictions of model trained on combined BA and EL data [BA+EL BA], and EL likelihood predictions of model trained on combined BA and EL data [BA+EL EL]).

Leave-one-out validation

Leave-one-out (LOO) experiments were performed for all MHC molecules present in the EL data set. For this, a given MHC molecule was removed from the EL data set and then the BA+EL method was trained in 5-fold

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

cross-validation, as described above, omitting multiple random initializations, resulting in an ensemble of 10 networks. Performance of the LOO models is compared with an ensemble of neural networks of the same size trained on the complete data set. Further predictions are made for 80,000 peptides of lengths 8–15 aa derived from natural proteins to evaluate a model's ability to predict the length preference of an MHC allele that was not part of the EL training data.

The final NetMHCpan-4.0 method implementation

The final neural network ensemble of the NetMHCpan-4.0 method is trained on BA and EL data, as described above using 5-fold cross-validation. Networks with 56 and 66 hidden neurons (in accordance with earlier NetMHCpan implementations) were trained using 10 distinct random initial configurations, leading to an ensemble of 100 networks in total.

Percentile rank scores were estimated from predicted EL and BA binding values from a set of 125,000 8–12-mer random natural peptides (25,000 of each length)

Validation on external data sets

A data set of ELs was obtained from Pearson et al. (17). Also, a set of positive CD8 epitopes was downloaded from the IEDB. The epitope set was identified using the following search criteria "T cell assays: IFN-γ," "positive assays only," "MHC restriction Type: Class I." Only entries with fully typed HLA restriction, peptides length in the range 8–14 aa, and with annotated source protein sequence were included. Positive entries with a predicted rank score $>$ 10% using NetMHCpan-3.0 were excluded to filter out likely noise (6). For the T cell epitope and EL data sets, negative peptides were obtained by extracting all 8–14-mer peptides from the source proteins of the ELs and subsequently excluding peptide–MHC combinations found with an exact match in the training data (BA and EL data sets). The final eluted data set contained 15,965 positive ligands restricted to 27 HLA molecules, and the IEDB T cell epitope data set contained 1,251 positive T cell epitopes restricted to 80 HLA molecules.

A Frank value was calculated for each positive HLA pair as the ratio of the number of peptides with a prediction score higher than the positive peptide/the number of peptides contained within the source protein. Hence, the Frank value is 0 if the positive peptide has the highest prediction value of all peptides within the source protein and a value of 0.5 in cases in which an equal amount of peptides has a higher and lower prediction value compared with the positive peptide.

An unfiltered EL data set was obtained from Bassani-Sternberg et al. (22). This data set consisted of EL data from six cell lines, each with fully typed HLA-A, -B, and -C alleles. A data set was constructed for each cell line, including all 8–13-mer ligands (N) as positives, and five times N random natural negatives for each length of 8–13 aa (i.e., if a data set contained 5,000 ligands, 5 $\times$ 5,000 = 25,000 random natural peptides of length 8, 9, 10, 11, 12, and 13 aa were added as negatives, arriving at a final data set with 155,000 [5,000 + 6 $\times$ 25,000] peptides).

Results

We trained the NetMHCpan method version 4.0 for the prediction of the interaction of peptides with MHC class I molecules integrating BA and MS EL data. Combined training was achieved by adding a second output neuron (see Fig. 1) to the NNAlign approach described previously (3). In this setup, the first output neuron returns a score of BA, and the second output neuron returns a score of ligand elution. As described in *Materials and Methods*, the model parameters between the input and hidden layer of the artificial neural network are shared between the two input types. Thanks to this network architecture, we expect the model to be able to combine informative patterns found in the two data types, boosting performance for both output neurons. To demonstrate this, we compared the performance of the BA+EL method with the BA method trained only on BA data and the EL method trained only on EL data. Fig. 2 shows the mean performance per MHC allele of the four methods on four data sets given in terms of AUC (for details see Supplemental Table I). From this analysis, it is clear that the BA+EL EL method performs much better on BA data than the EL method. This observation strongly suggests that the EL method, as a result of the small number of MHC molecules (55) included in the EL data set, has limited pan-specific potential compared with the BA+EL EL method trained on data from 169 MHC molecules included in the combined BA and MS EL data set.

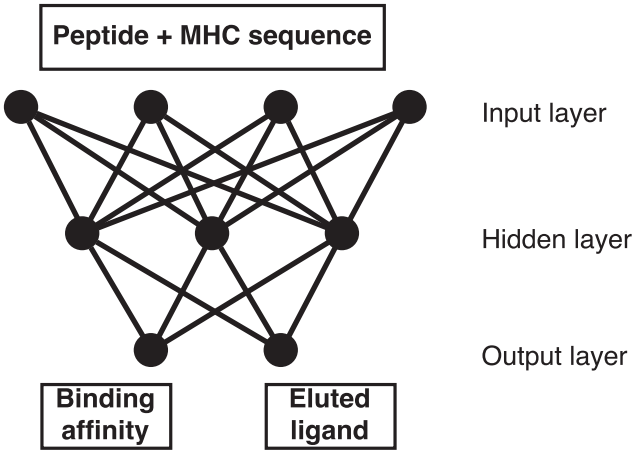

FIGURE 1. Visualization of the neural networks with two output neurons used for combined training on BA and EL data.

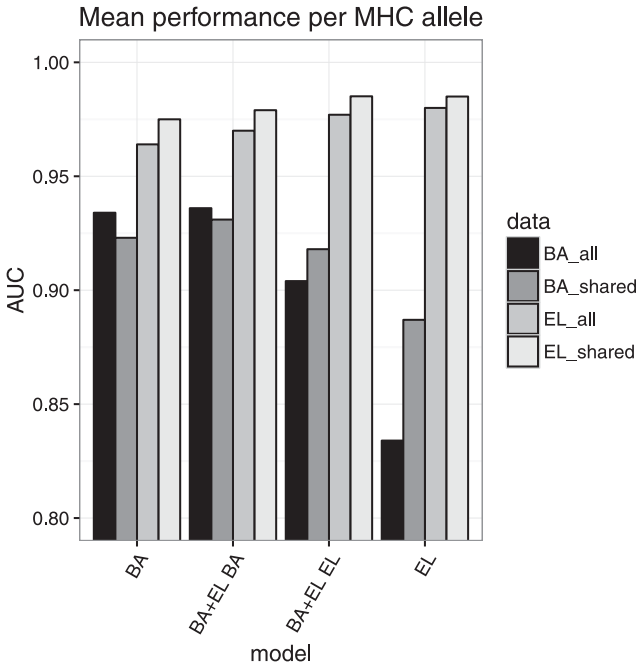

FIGURE 2. Mean performance per MHC molecule measured in terms of AUC for the four methods: BA, EL, BA+EL BA, and BA+EL EL. The methods were evaluated on all BA (BA_all) data and all EL (EL_all) data, including negative peptides derived from source proteins, and on data sets restricted to alleles occurring in BA and EL data sets (BA_shared and EL_shared).

Peptide-length preference of MHC molecules

We next set out to investigate how well the different methods could capture the peptide-length preferences of individual MHC molecules. For this, we predicted binding scores for a set of random natural peptides of lengths 8–15 aa and calculated the frequencies of peptides of different lengths in the top 2% of predictions. In Fig. 3A–C, we show examples of such peptide-length preference profiles predicted by the BA, BA+EL BA, BA+EL EL, and EL methods. The depicted MHC molecules are known to have preferences for different peptide lengths. All HLAs have a preference for 9-mer peptides. However, HLA-A*01:01 has an increased preference for 10-mers compared with average, HLA-A*02:01 has

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

a strong preference for 9-mers only, and HLA-B*51:01 has an increased preference for 8-mers compared with average (6). BA predictors often overestimate the number of binding 10-mer peptides because of their overrepresentation in the BA data set (4), which is also shown in Fig. 3.

Next, we extended the analysis to all MHC molecules included in the EL data set, calculating the correlation between observed and predicted length frequencies for each prediction method. This analysis (Fig. 3D) clearly confirms the results obtained from the three case examples (i.e., that the two methods BA+EL EL and EL show significantly higher power for predicting the peptide-length preference of individual MHC molecules compared with the two methods trained to predict BA [BA and BA+EL BA]).

The predictions for the two EL likelihood models only show low performance for one molecule: HLA-B41:04. However, this molecule is only characterized by 52 ELs, whose length profile forms an unusual bimodal distribution with peaks at 9 and 11 aa (data not shown).

LOO experiments on EL data

In the above experiment, the MHC molecules used for the peptide-length preference evaluation were also included as training data for the EL prediction methods. This naturally leads to a bias in the performance evaluation. To address this, as well as to access the pan-specific potential of the BA+EL EL prediction method, we conducted a LOO experiment. A given MHC molecule was removed from the EL data set, and the BA+EL method was retrained, as described in *Materials and Methods*. Next, the predictive performance (estimated in terms of AUC for separating the known ligands from the artificial negatives) and the ability to predict the peptide-length preference were evaluated. The result of the benchmark is shown in Fig. 4. This figure clearly confirms the pan-specific power of the BA+EL method. In terms of the predictive performance (Fig. 4A), the LOO methods display, as expected, a slight decrease in performance compared with a method trained and evaluated on all data (the all-data method). When looking at the performance for predicting the peptide-length profile (Fig. 4B), the LOO methods display a very high performance. Only in one case does the EL LOO method show a substantial drop in performance for the left out MHC molecule. This case is H2-Kb, the only mouse molecule in the MS ligand data set with a strong preference for 8-mer ligands. The BA+EL EL LOO method is able to predict the length profile of H2-Kb as a result of the H2-Kb affinity data present in the BA training data set.

The NetMHCpan-4.0 method

Having demonstrated the increased predictive power of the BA+EL method when it comes to prediction of peptide BA (BA+EL BA model), the likelihood of being an EL (BA+EL EL model), and the ability to capture the MHC-specific peptide-length binding preferences (BA+EL EL model), we set out to construct the final NetMHCpan-4.0 method. This method was trained as the BA+EL method, using 5-fold cross-validation, as described in *Materials and Methods*. The method is available at http://www.cbs.dtu.dk/services/NetMHCpan-4.0. The functionality is identical to the earlier NetMHCpan implementations, with the important additional function that two different output options (BA and EL likelihood) are available. By default, the program returns EL likelihood scores. An example of the output of the method is shown in Supplemental Fig. 1.

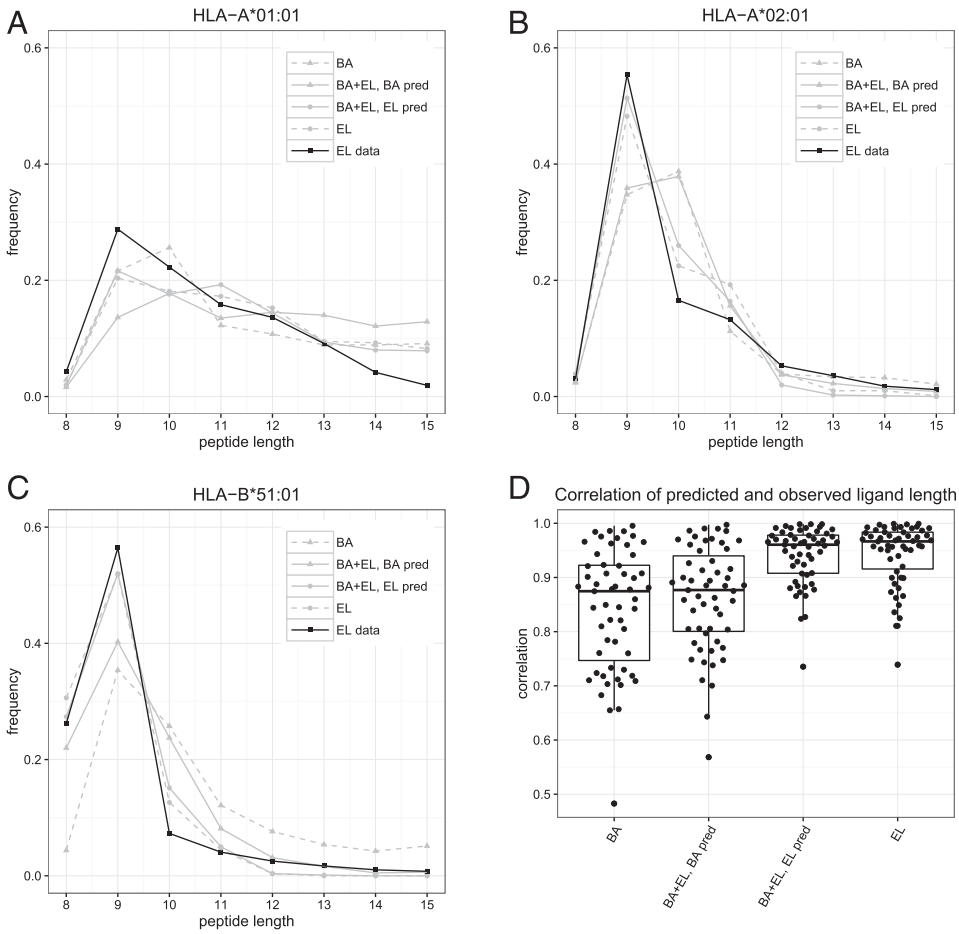

FIGURE 3. (**A**–**C**) Predicted length preference of selected MHC molecules according to different models. Binding to selected HLA molecules was predicted for 80,000 8–15-mer peptides, and the frequency of peptide lengths in the top 2% of predicted peptides was calculated. (**D**) Correlation of predicted and observed ligand length for different models. Binding to all HLA alleles present in both the BA and EL data sets was predicted using the four prediction methods for 80,000 8–15-mer peptides. Subsequently, the occurrence of different peptide lengths in the top 2% of predicted peptides for each molecule was calculated, and the correlation coefficient between these frequencies and the length frequencies in the EL data set was calculated.

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

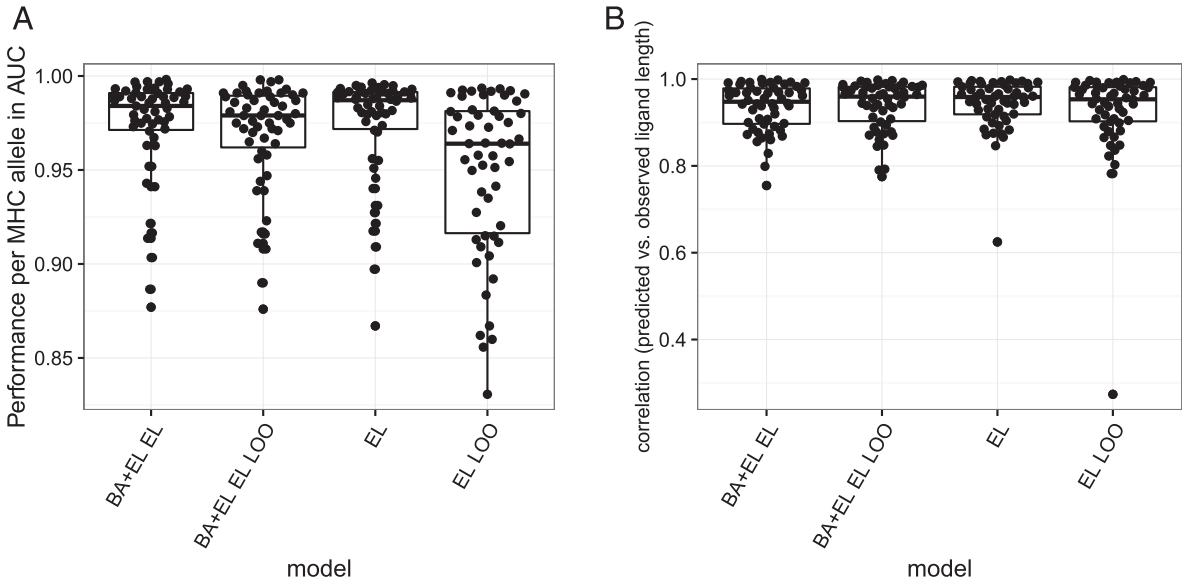

FIGURE 4. EL LOO experiments. (**A**) Performance per MHC allele of a model trained on all data and a model in which the EL data of a given allele was left out of the training process. (**B**) Correlation of predicted and observed ligand length for a model trained on all data and the LOO models.

Validation on external data sets

The performance of the updated NetMHCpan method was assessed on two independent external data sets: one consisting of 15,965 ELs covering 27 HLA molecules and another consisting of 1251 validated CTL epitopes covering 80 HLA molecules reported in the IEDB. The validation data sets were constructed as described in *Materials and Methods*. The source protein sequence was identified for each ligand/epitope, and all overlapping 8–14-mer peptides, with the exception of the ligand/epitope, were annotated as negatives. All data points included in the BA and EL training data sets were excluded from the validation data set. A Frank value was calculated for each positive HLA pair, as described in *Materials and Methods*, as the ratio of the number of peptides with a prediction score higher than the positive peptide/the number of peptides contained within the source protein. In this manner, we can construct the sensitivity curves presented in Fig. 5. Two observations are striking from these results. First and foremost, the results clearly demonstrated the increased predictive power of integrating EL data into the training data of NetMHCpan. In Fig. 5A (the analysis of the EL data), we can observe that the gain in sensitivity at a Frank threshold of 1% for the EL models (NetMHCpan-4.0 EL or EL only) compared with NetMHCpan-3.0 is 10% (95 versus 85%), and it is 15% at a Frank threshold of 0.5% (90 versus 75%). These numbers mean that a ligand will have a prediction score within the top 0.5% of its source protein peptides in 90% of the cases using the EL models, compared with only 75% using NetMHCpan-3.0. However, the results shown in the left panel of Fig. 5 also suggest that the two EL models achieve very similar predictive performance when it comes to identification of ELs. This is in strong contrast to the results obtained from the IEDB epitope data set (Fig. 5B). In this case, only the NetMHCpan-4.0 EL model demonstrates an improved predictive performance compared with NetMHCpan-3.0.

There are several potential explanations for the improved performance of the EL models on the EL evaluation data, including a bias against cysteines specific for the EL training and evaluation data; differences in the MHC-binding motifs contained within the EL and in vitro binding data, as suggested earlier (15); and the improved prediction accuracy of the ligand length preference (see Fig. 3D). To investigate the first possibility, we repeated the experiment displayed in Fig. 5, removing all peptides containing one or more cysteines. If the bias against cysteines in the EL data had any impact on the predictive performance of the proposed method, the bias would be reflected in an altered predictive performance on the reduced data sets. This turned out not to be the case (data not shown), suggesting that cysteine bias is not influencing the relative predictive performance of the different methods. Looking into the differences in the binding motif derived from BA and EL data for specific HLA molecules, we find differences for most MHC molecules. A few examples are shown in Fig. 6.

These results demonstrated that ELs tend to share more conserved anchor motifs compared with affinity-defined binders. This observation is in agreement with earlier findings suggesting that ELs are bound more stably to MHC class I molecules compared with other affinity-matched peptides (13, 23). In summary, these analyses suggest that the gained predictive performance of the EL method on the EL evaluation data is driven by at least two factors: differences in binding preferences between EL and affinity-defined peptide binders and the improved prediction accuracy of ligand length preference of the EL methods.

To be or not to be a ligand

We investigated what prediction threshold to use to best separate ligand from nonligand peptides. Our earlier work suggests that different MHC molecules present peptides at different predicted BA thresholds (1, 24). Given this, it was interesting to investigate to what degree a similar observation could be made for the EL likelihood predictions produced by the NetMHCpan-4.0 method. To address the question, we compared the predicted ligand likelihood scores of all 15,965 ligands in the Pearson data set. The result of this analysis is displayed as box-plots in the left panel of Fig. 7.

This figure reveals that the likelihood prediction scores for known ligands are very different for different HLA molecules. The large difference in prediction values between HLA molecules can be linked, to a high degree, to their absence from the EL training data. The molecules with the lowest median EL likelihood scores in this figure are molecules that are absent from the EL training data set. However, as demonstrated in Figs. 4 and 5, the fact that an HLA molecule has not been characterized with

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

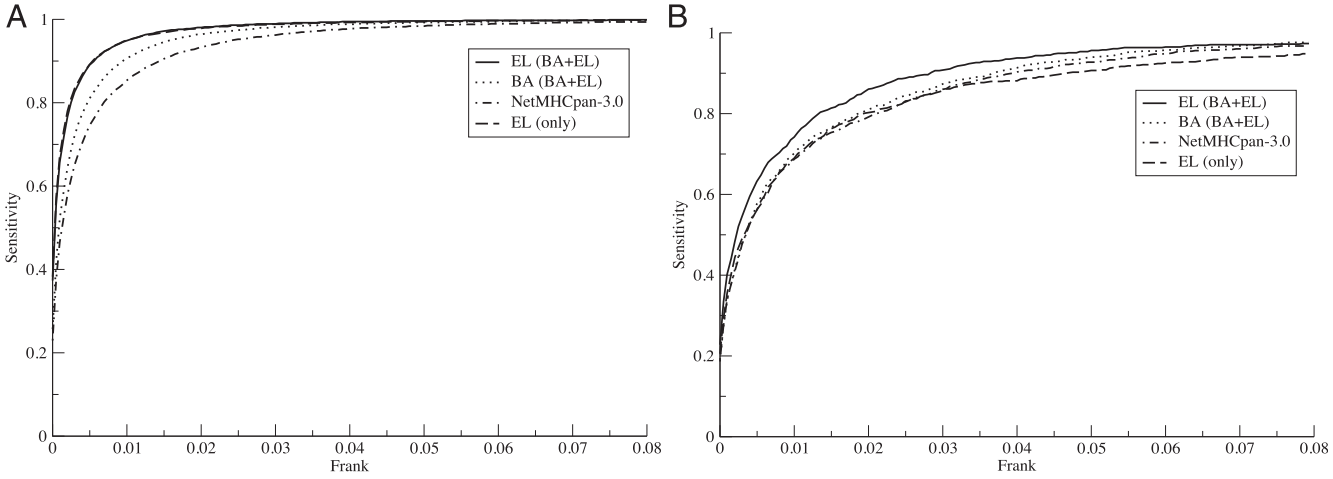

FIGURE 5. Sensitivity of different models as a function of the Frank threshold on (**A**) ELs published by Pearson et al. (17) and (**B**) T cell epitope data downloaded from IEDB.

EL training data does not impair its predictability. Given this, a natural measure to correct for this great imbalance in prediction score is to use percentile rank scores to reconcile and make prediction scores comparable between different MHC molecules. The right panel in Fig. 7 shows the results of such a transformation. In this case, EL likelihood prediction values for each ligand in the Pearson data are transformed to percentile rank scores, and the score distribution is visualized as box plots for each HLA molecule. Given that percentile rank values fall in the range of 0–100%, it is apparent that transforming the prediction values into such rank scores allows for a direct score comparison between HLA molecules.

In light of these results, we next investigated which percentile rank threshold to apply to optimally identify MHC ligands. We assessed this by calculating sensitivity/specificity curves as a function of the percentile rank score threshold for a balanced set (maximum of 100 ligands per HLA) of ELs and source protein negatives from the Pearson evaluation data set. The results are shown in Fig. 8 and confirm earlier findings that the vast majority (96.5%) of natural ligands are identified at a very high specificity (98.5%) using a percentile rank threshold of 2%.

Evaluation of unbiased data sets

Most EL data potentially suffer from biases toward current prediction methods. This is because many EL studies, including the Pearson data used in this study, assign MHC restriction based on predicted binding. To address the impact of this bias, we benchmark our method against sets of unfiltered EL data. These data sets were obtained from Bassani-Sternberg et al. (22) and include ELs obtained from six cell lines, each with typed HLA expression. From these data, we constructed six benchmark data sets by enriching each positive EL data set with a set of random natural negative peptides (for details see *Materials and Methods*). After filtering out data included in the training data of NetMHCpan-4.0, we next benchmarked the predictive power of the different prediction methods. The result of the benchmark is shown in Fig. 9.

These results clearly confirm the improved performance of the proposed NetMHCpan-4.0 EL likelihood predictions over the NetMHCpan-4.0 and NetMHCpan-3.0 BA predictions. Also, the results show that, in the majority of cases, the percentile rank predictions achieve improved predictive performance compared with the raw prediction scores.

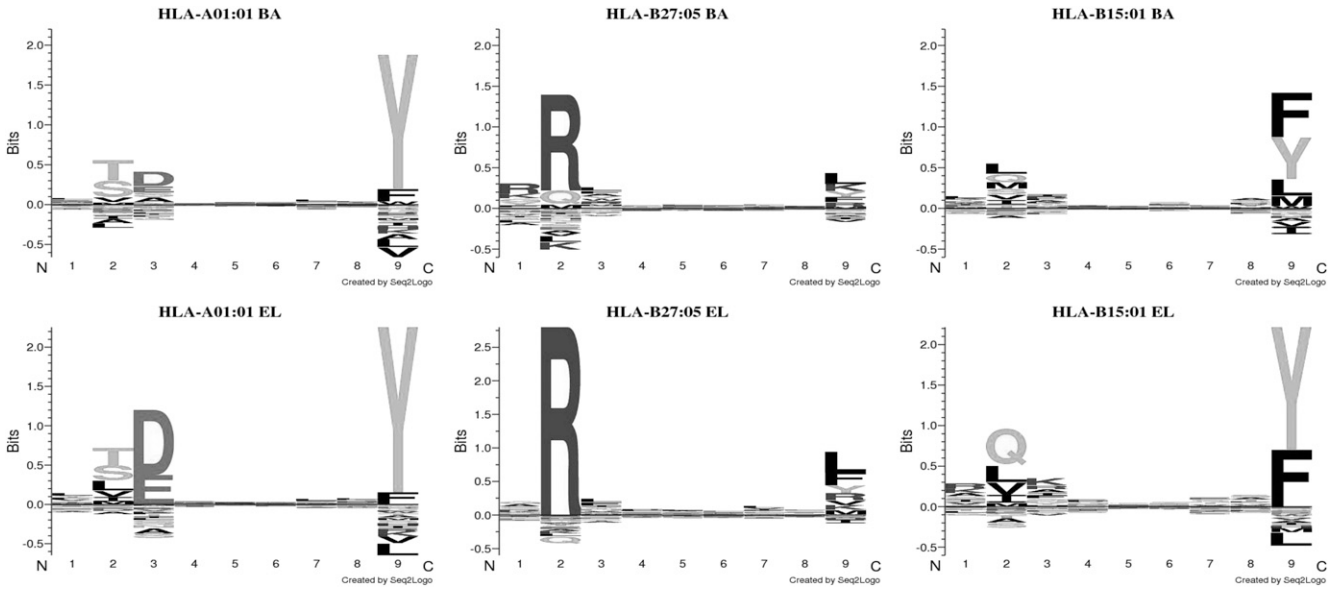

FIGURE 6. Binding motifs for HLA molecules derived from in vitro BA data using a binding threshold of 500 nM (upper panels), and EL data (lower panels). Binding motif plots were made using Seq2Logo with default parameters (30).

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

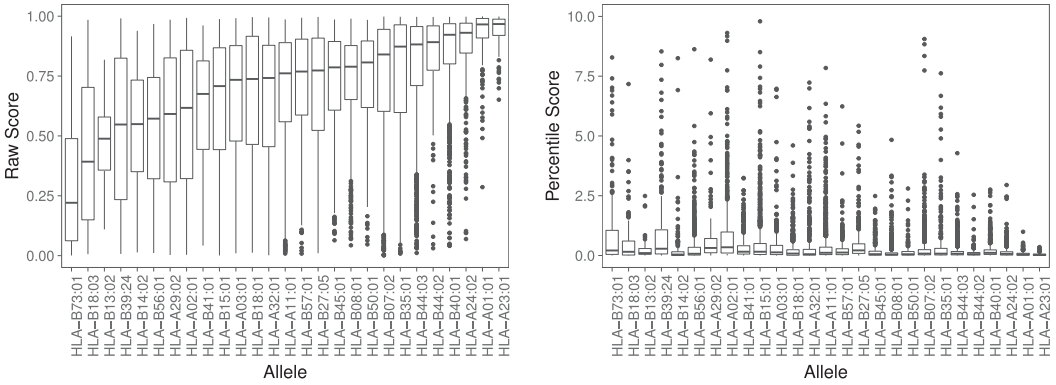

FIGURE 7. Motivation for using percentile rank score predictions. Box-plot representation of prediction values for the ligands in the Pearson data set. EL likelihood prediction scores (left panel) and percentile rank values (right panel).

Identification of cancer neoantigens

A research field in which prediction of naturally processed and presented ELs has attracted significant recent attention is rational identification of cancer neoantigens. In contrast with tumor-associated self-antigens, cancer neoantigens are naturally presented ligands containing tumor-specific mutations. Such neoantigens are attracting great attention because these peptides are new to the immune system and are not found in normal tissues; hence, they are ideal potential cancer vaccine candidates or targets for adoptive T cell therapy. Depending on the mutational load, the number of potential tumor-specific neopeptides (peptides containing one or more missense mutations) can be on the order of many thousands (25). This large number of potential peptide candidates clearly underlines the need for tools to rationally downsize the peptide space in the search for cancer neoepitopes. A recent study by Bassani-Sternberg et al. (14) demonstrated how this downsizing could be effectively achieved by a prediction method trained on a large set of MS ELs. In this study, we repeated this benchmark analysis using NetMHCpan-4.0. The results are shown in Fig. 10 and confirm the findings by Bassani-Sternberg et al. (14) that, in most cases, predictors trained on MS EL data show very high predictive power for the identification of cancer neoantigens. Both NetMHCpan-4.0 and the MixMHCpred method proposed by Bassani-Sternberg et al. (14) identify the known neoantigens within the top 25 peptides in 6 of 10 cases. NetMHCpan-3.0 only achieves this in 2 of 10 cases. The results also confirm the earlier findings presented in this article: that NetMHCpan-4.0 achieves improved performance compared with that of version 3.0 and that the ligands in all cases are predicted with very strong EL likelihood values (all percentile rank values are <1, and the majority are ≤ 0.02).

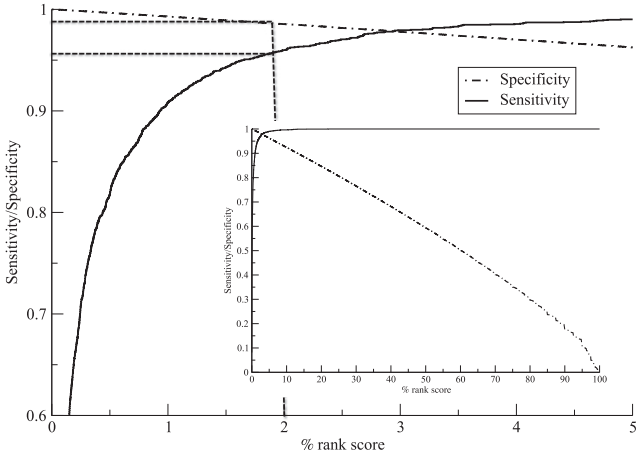

FIGURE 8. Sensitivity and specificity performance curves for the NetMHCpan-4.0 EL likelihood predictions. Curves are estimated from a balanced set of ELs from the data set of Pearson et al. (17). The inset shows the complete sensitivity and specificity curves as a function of the percentile rank score. The main plot shows the curves in the high-scoring range for 0–5 percentile scores. Dashed vertical and horizontal lines indicate sensitivity and specificity and the 2% rank score threshold.

Discussion

In this article, we have demonstrated how a relatively simple pan-specific machine-learning method based on the NNAlign framework can be constructed that integrates information from BA data with MS peptidome data. Benefitting from the larger set of peptide BA data with very broad MHC coverage (>150 molecules) and the additional information contained within MS peptidome data (information about Ag processing and presentation, as well as allele-specific peptide-length profile), we could demonstrate that the proposed method, NetMHCpan-4.0, achieved improved predictive performance with regard to characterizing the binding specificity of a given MHC molecule, as well as for predicting the peptide-length profile. Because of the pan-specific potential of the method, the improved performance was extended beyond the relatively few MHC molecules characterized by MS binding data included in the training of the method. Given this, we conclude that the proposed framework is able to benefit from the best of the two data sets: MHC coverage from the BA data and Ag processing and presentation and allele-specific peptide-length profile from the MS data.

Our benchmarks confirmed earlier findings that prediction values for known ligands vary substantially between MHC molecules (26) and that predictions between different MHC molecules can be readily compared only by the use of percentile rank scores.

The improved peptide–MHC tool is publicly available at http://www.cbs.dtu.dk/services/NetMHCpan-4.0. The tool was benchmarked on two large independent data sets: one consisting of ~16,000 MS-identified MHC-restricted ligands (17) and one consisting of $>1,250$ validated T cell epitopes described in the IEDB. For both data sets, the updated version of NetMHCpan (4.0) significantly outperformed the earlier NetMHCpan-3.0 method. In particular, the benchmark on T cell epitope data demonstrated for the first time, to our knowledge, how integration of MS peptidome data into a prediction method of MHC peptide presentation can lead to improved predictive performance for T cell epitope discovery. The improved performance on this data set was only observed for the method trained on the combined data; it was not observed for the method trained on MS peptidome data alone.

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

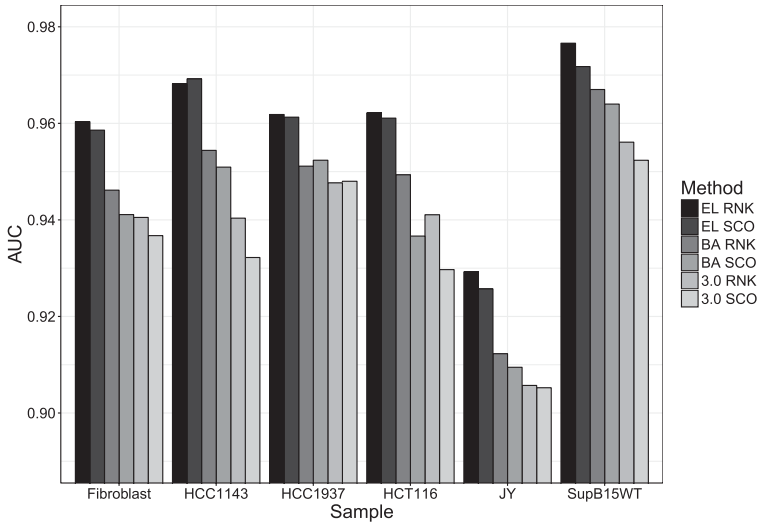

FIGURE 9. Predictive performance measured in terms of AUC on the Bassani-Sternberg unfiltered EL data sets. Prediction values are assigned to each peptide in a given data set as the lowest percentile rank score/highest prediction score to each of the HLA molecules expressed by the given cell line. The six methods included are EL RNK (NetMHCpan-4.0 EL percentile rank), EL SCO (NetMHCpan-4.0 EL likelihood score), BA RNK (NetMHCpan-4.0 BA percentile rank), BA SCO (NetMHCpan-4.0 BA score), 3.0 RNK (NetMHCpan-3.0 percentile rank), and 3.0 SCO (NetMHCpan-3.0 BA score).

This observation underlines the large benefit of merging the two data types.

Investigating potential causes for the observed improved performance of the proposed tool for identification of ELs confirmed earlier findings that ELs share a reduced amino acid diversity at the MHC anchor positions (13). This observation is consistent with the notion that ligands are bound more stably to MHC class I molecules compared with average affinity-defined bound peptides. We postulate that this difference in binding preferences between EL and affinity-defined peptide binders, combined with the improved prediction accuracy of ligand-length preference of the EL methods, are the main factors driving the improved predictive performance.

When benchmarking the predictive performance for identification of T cell epitopes, we observed that only the NetMHCpan-4.0 EL model trained on the combined EL and BA data set demonstrated an improved predictive performance compared with NetMHCpan-3.0. This observation was surprising at first, because we also would expect an improved performance by the method trained on the EL only as a result of the reasons outlined above. One likely explanation for this result is the bias in the T cell epitope data toward predicted BA motifs. Most T cell epitopes have been identified using some kind of HLA-binding predictions as a filter prior to experimental validation, giving a bias toward prediction methods trained based on BA data. Given this, the source of the improved performance of the NetMHCpan-4.0 EL method compared with NetMHCpan-4.0 BA on the T cell epitope benchmark data set is primarily driven by its improved prediction of the ligand-length preference.

It is clear that, even with the improved predictive performance of the NetMHCpan-4.0 tool reported in this article, not all MHC ligands and T cell epitopes will be captured by a prediction workflow. Likewise, it is clear that very few, if any, experimental workflows enable the exhaustive identification of the ligandome or epitope set contained within a given sample. In silico predictions can be used in concert with experimental approaches, effectively boosting the sensitivity of the combined workflow. Such an approach where in silico predictions were used to reduce the search space has been successful in improving the sensitivity of MHC class I ligand discovery (27), and we expect other similar applications to appear in the future.

The machine-learning framework proposed in this article is not limited to the integration of MHC class I peptide BA and MS peptidome data. The approach can readily be extended to integrate other types of relevant data, including MHC-binding stability (28) and epitope data. Also, in its current form, the approach can be directly applied to the MHC class II system. The only critical limitation for such data integration is the criterion that each data point must be associated with a specific MHC element. This information is not always readily available, but in most cases it can be inferred by unsupervised clustering of the available data [using GibbsCluster (29), position weight matrix mixture models (16), or

Sample	Ligand	NetMHCpan-4.0	NetMHCpan-3.0	Bassani-Sternberg et al.	#peptides
Mel8	SPGPVKLEL	**1 (0.0124)**	12	2	1340
Mel5	YIDERFERY	15 (0.0077)	33	**3**	25807
Mel5	ETSKQVTRW	189 (0.1156)	464	**13**	25807
Mel15	GRIAFFLKY	21 (0.0098)	224	**3**	24766
Mel15	LPIQYEPVL	10 (0.0071)	24	7	24766
Mel15	KLKLPIIMK	**6 (0.0061)**	34	21	24766
Mel15	GRTGAGKSFL	1226 (0.6909)	2186	**243**	24766
Mel15	KLILWRGLK	457 (0.2364)	**112**	527	24766
Mel15	ASWVVPIDIK	1629 (0.9723)	**1278**	3978	24766
12T	DANSFLQSV	**19 (0.0205)**	944	38	15750

FIGURE 10. Predictive performance evaluated in terms of rank of neoantigens identified in four melanoma samples. A rank value of 1 corresponds to the ligand obtaining the highest score (lowest percentile rank) of all peptides from the given sample. Data and performance values for MixMHCpred are from Bassani-Sternberg et al. (14). NetMHCpan-4.0 and NetMHCpan-3.0 are performance values obtained by assigning to each peptide in the given data set the lowest percentile rank score to each of the HLA-A and -B molecules expressed by the given cell line. The values in parentheses for NetMHCpan-4.0 are the predicted percentile rank values. Lowest rank value for each ligand is highlighted in bold.

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017

similar approaches], and association of each cluster to an MHC molecule of the given host.

In conclusion, we have described a new framework for training of prediction methods for MHC–peptide presentation prediction integrating information from two data sources (MS EL and peptide BA). The framework was used to develop an updated version of NetMHCpan (version 4.0, available at http://www.cbs.dtu.dk/services/NetMHCpan-4.0) with improved predictive performance for identification of validated ELs, cancer neoantigens, and T cell epitopes.

Disclosures

The authors have no financial conflicts of interest.

References

1. Nielsen, M., and M. Andreatta. 2016. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 8: 33.
2. Vita, R., J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. 2015. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 43: D405–D412.
3. Nielsen, M., and M. Andreatta. 2017. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* DOI 10.1093/nar/gkx276.
4. Andreatta, M., and M. Nielsen. 2015. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics.* 32: 511–517.
5. Deres, K., T. N. Schumacher, K. H. Wiesmüller, S. Stevanović, G. Greiner, G. Jung, and H. L. Ploegh. 1992. Preferred size of peptides that bind to H-2 Kb is sequence dependent. *Eur. J. Immunol.* 22: 1603–1608.
6. Trolle, T., C. P. McMurtrey, J. Sidney, W. Bardet, S. C. Osborn, T. Kaever, A. Sette, W. H. Hildebrand, M. Nielsen, and B. Peters. 2016. The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J. Immunol.* 196: 1480–1487.
7. Lundegaard, C., K. Lamberth, M. Harndahl, S. Buus, O. Lund, and M. Nielsen. 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* 36 (Web Server Issue): W509–W512.
8. Nielsen, M., C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, and S. Buus. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2: e796.
9. Kreiter, S., M. Vormehr, N. van de Roemer, M. Diken, M. Löwer, J. Diekmann, S. Boegel, B. Schrörs, F. Vascotto, J. C. Castle, et al. 2015. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520: 692–696.
10. Gubin, M. M., M. N. Artyomov, E. R. Mardis, and R. D. Schreiber. 2015. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* 125: 3413–3421.
11. 2017. The problem with neoantigen prediction. *Nat. Biotechn.* 35: 97.
12. Tenzer, S., B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P. M. Kloetzel, H. G. Rammensee, H. Schild, and H. G. Holzhütter. 2005. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell. Mol. Life Sci.* 62: 1025–1037.
13. Harndahl, M., M. Rasmussen, G. Roder, I. Dalgaard Pedersen, M. Sørensen, M. Nielsen, and S. Buus. 2012. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* 42: 1405–1416.
14. Bassani-Sternberg, M., C. Chong, P. Guillaume, M. Solleder, H. Pak, P. O. Gannon, L. E. Kandalaft, G. Coukos, and D. Gfeller. 2017. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* 13: e1005725.
15. Abelin, J. G., D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, et al. 2017. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46: 315–326.
16. Bassani-Sternberg, M., and D. Gfeller. 2016. Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-HLA interactions. *J. Immunol.* 197: 2492–2499.
17. Pearson, H., T. Daouda, D. P. Granados, C. Durette, E. Bonneil, M. Courcelles, A. Rodenbrock, J. P. Laverdure, C. Côté, S. Mader, et al. 2016. MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* 126: 4690–4701.
18. Bassani-Sternberg, M., E. Bräunlein, R. Klar, T. Engleitner, P. Sinitcyn, S. Audehm, M. Straub, J. Weber, J. Slotta-Huspenina, K. Specht, et al. 2016. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 7: 13404.
19. Sidney, J., S. Southwood, C. Oseroff, M. F. del Guercio, A. Sette, and H. M. Grey. 2001. Measurement of MHC/peptide interactions by gel filtration. *Curr. Protoc. Immunol.* Chapter 18: Unit 18.3.
20. Harndahl, M., S. Justesen, K. Lamberth, G. Røder, M. Nielsen, and S. Buus. 2009. Peptide binding to HLA class I molecules: homogenous, high-throughput screening, and affinity assays. *J. Biomol. Screen.* 14: 173–180.
21. Nielsen, M., C. Lundegaard, P. Worning, S. L. Lauemøller, K. Lamberth, S. Buus, S. Brunak, and O. Lund. 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12: 1007–1017.
22. Bassani-Sternberg, M., S. Pletscher-Frankild, L. J. Jensen, and M. Mann. 2015. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* 14: 658–673.
23. Jorgensen, K. W., M. Rasmussen, S. Buus, and M. Nielsen. 2014. NetMHCstab - predicting stability of peptide–MHC-I complexes; impacts for CTL epitope discovery. *Immunology* 141: 18–26.
24. Paul, S., D. Weiskopf, M. A. Angelo, J. Sidney, B. Peters, and A. Sette. 2013. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* 191: 5831–5839.
25. Bjerregaard, A.-M., M. Nielsen, S. R. Hadrup, Z. Szallasi, and A. C. Eklund. 2017. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* DOI: 10.1007/s00262-017-2001-3.
26. Karosiene, E., M. Rasmussen, T. Blicher, O. Lund, S. Buus, and M. Nielsen. 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 65: 711–724.
27. Murphy, J. P., P. Konda, D. J. Kowalewski, H. Schuster, D. Clements, Y. Kim, A. M. Cohen, T. Sharif, M. Nielsen, S. Stevanovic, et al. 2017. MHC-I ligand discovery using targeted database searches of mass spectrometry data: implications for T-cell immunotherapies. *J. Proteome Res.* 16: 1806–1816.
28. Rasmussen, M., E. Fenoy, M. Harndahl, A. B. Kristensen, I. K. Nielsen, M. Nielsen, and S. Buus. 2016. *Pan*-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* 197: 1517–1524.
29. Andreatta, M., B. Alvarez, and M. Nielsen. 2017. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* DOI: 10.1093/nar/gkx248.
30. Thomsen, M. C., and M. Nielsen. 2012. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* 40(Web Server Issue): W281–W287.

Downloaded from http://www.jimmunol.org/ by guest on October 4, 2017