A community resource benchmarking predictions of peptide binding to MHC-I molecules

Peters, B; Bui, HH; Pletscher-Frankild, Sune; Nielsen, Morten; Lundegaard, Claus; Kostem, E; Basch, D; Lamberth, K; Harndahl, M; Fleri, W; Wilson, SS; Sidney, J; Lund, Ole; Buus, S; Sette, Alessandro

Published in:
P L o S Computational Biology (Online)

Link to article, DOI:
10.1371/journal.pcbi.0020065

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Peters, B., Bui, H. H., Pletscher-Frankild, S., Nielsen, M., Lundegaard, C., Kostem, E., ... Sette, A. (2006). A community resource benchmarking predictions of peptide binding to MHC-I molecules. P L o S Computational Biology (Online), 2(6), 574-584. DOI: 10.1371/journal.pcbi.0020065

DTU Library
Technical Information Center of Denmark

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules

Bjoern Peters1*, Huynh-Hoa Bui1, Sune Frankild2, Morten Nielsen2, Claus Lundegaard2, Emrah Kostem1, Derek Basch1, Kasper Lamberth3, Mikkel Harndahi3, Ward Fleri1, Stephen S. Wilson1, John Sidney1, Ole Lund2, Soren Buus3, Alessandro Sette1

1 La Jolla Institute for Allergy and Immunology, San Diego, California, United States of America, 2 Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark, 3 Department of Experimental Immunology, Institute of Medical Microbiology and Immunology, University of Copenhagen, Copenhagen, Denmark

Recognition of peptides bound to major histocompatibility complex (MHC) class I molecules by T lymphocytes is an essential part of immune surveillance. Each MHC allele has a characteristic peptide binding preference, which can be captured in prediction algorithms, allowing for the rapid scan of entire pathogen proteomes for peptide likely to bind MHC. Here we make public a large set of 48,828 quantitative peptide-binding affinity measurements relating to 48 different mouse, human, macaque, and chimpanzee MHC class I alleles. We use this data to establish a set of benchmark predictions with one neural network method and two matrix-based prediction methods extensively utilized in our groups. In general, the neural network outperforms the matrix-based predictions mainly due to its ability to generalize even on a small amount of data. We also retrieved predictions from tools publicly available on the internet. While differences in the data used to generate these predictions hamper direct comparisons, we do conclude that tools based on combinatorial peptide libraries perform remarkably well. The transparent prediction evaluation on this dataset provides tool developers with a benchmark for comparison of newly developed prediction methods. In addition, to generate and evaluate our own prediction methods, we have established an easily extensible web-based prediction framework that allows automated side-by-side comparisons of prediction methods implemented by experts. This is an advance over the current practice of tool developers having to generate reference predictions themselves, which can lead to underestimating the performance of prediction methods they are not as familiar with as their own. The overall goal of this effort is to provide a transparent prediction evaluation allowing bioinformaticians to identify promising features of prediction methods and providing guidance to immunologists regarding the reliability of prediction tools.

Introduction

Cytotoxic T lymphocytes of the vertebrate immune system monitor cells for infection by viruses or intracellular bacteria by scanning their surface for peptides bound to major histocompatibility complex (MHC) class I molecules (reviewed in [1]). The presented peptides are generated within the cells during the degradation of intracellular proteins. Cells presenting peptides derived from nonself proteins, such as viruses or bacteria, can trigger a T-cell immune response leading to the destruction of the cell. Likewise, this peptide presentation mechanism is utilized to detect cancerous cells [2] and—when malfunctioning—is implicated in several autoimmune diseases [3].

Peptides bound to MHC molecules that trigger an immune response are referred to as T-cell epitopes. Identifying such epitopes is of high importance to immunologists, because it allows the development of diagnostics, evaluation of the efficacy of subunit vaccines, and even the development of peptide-based vaccines. Many computational algorithms have been created to predict which peptides contained in a pathogen are likely T-cell epitopes [4–25]. Such tools allow for the rapid scan of the proteome of a pathogen, and are being widely used in the immunological community. Many of them are freely available on the internet.

Multiple factors influence whether a peptide contained in the proteome of a pathogen is an epitope (i.e., whether it can trigger an immune response). For T-cell epitopes, the most selective requirement is the ability to bind to an MHC molecule with high affinity. Binding is also the most straightforward factor to characterize experimentally as well as model computationally, since the ability of a peptide to bind an MHC molecule is encoded in its primary amino acid sequence. Predictions for peptide cleavage by the proteasomal and peptide transport by the transporter associated with antigen presentation (TAP) have been developed as well [8,15,26–31], but the influence of these processes on peptide
Synopsis

In higher organisms, major histocompatibility complex (MHC) class I molecules are present on nearly all cell surfaces, where they present peptides to T lymphocytes of the immune system. The peptides are derived from proteins expressed inside the cell, and thereby allow the immune system to “peek inside” cells to detect infections or cancerous cells. Different MHC molecules exist, each with a distinct peptide binding specificity. Many algorithms have been developed that can predict which peptides bind to a given MHC molecule. These algorithms are used by immunologists to, for example, scan the proteome of a given virus for peptides likely to be presented on infected cells. In this paper, the authors provide a large-scale experimental dataset of quantitative MHC–peptide binding data. Using this dataset, they compare how well different approaches are able to identify binding peptides. This comparison identifies an artificial neural network as the most successful approach to peptide binding prediction currently available. This comparison serves as a benchmark for future tool development, allowing bioinformaticians to document advances in tool development as well as guiding immunologists to choose good prediction algorithms.

Results

Assembling the Dataset

We have collected measured peptide affinities to MHC class I molecules from two sources: the group of Alessandro Sette at the La Jolla Institute for Allergy and Immunology [43], and the group of Søren Buus at the University of Copenhagen [44]. The assays used by the two groups are different in several aspects, such as the indicator used to detect binding (bound radioactive ligand vs. quantitative enzyme-linked immunosorbent assay), what is detected (competitive binding vs. refolding), the way the MHC molecules are prepared (isolated from homozgyous cell lines vs. recombinant MHC), and the purity of peptides used (crude synthesis vs. purified peptide).

The type of data generated, however, is the same: each peptide gets assigned an affinity to a given MHC allele in terms of IC_{50}/EC_{50} nM (for brevity, we will refer to EC_{50} as IC_{50} in the following). Peptides with an affinity worse than the experimental sensitivity threshold are assigned an upper limit of detectable IC_{50} (Sette: >50,000 nM or higher; Buus: >20,000 nM). If affinities for the same peptide to the same MHC molecule were recorded in multiple assays, the geometric mean of the IC_{50} values was taken as the consensus value in the final dataset.

The final dataset is heterogeneous with regard to the peptide sequence tested for binding to each allele. On average, 84% of the peptides in each dataset differed in at least two residues with every other peptide in the set. No additional homology reduction was performed on the peptide sequences, because this should be done by the tool developers, who may prefer to use different homology-reduction approaches that are best optimized for their specific methods. Our purpose is to provide a complete training dataset to the public.

Table 1 gives an overview of the data, comprising 48,828 recorded affinities of peptides for a total of 48 different
The amount of data available per allele varies greatly from 51 recorded affinities for 11-mer peptides binding to the mouse MHC allele H-2 K$k to 3,089 affinities for 9-mer peptides binding to the well-studied human allele HLA-A*0201. The entire dataset is available for download at http://mhcbindingpredictions.immuneepitope.org.

Compared to other public databases, this is a much more homogenous set of data, as all of it was generated in one of only two assay systems. At the same time, the amount of data in our set is much greater than what was previously available. By comparison, the largest set of quantitative peptide affinities to MHC class I molecules currently available is found in the AntiJen database, which contains 12,190 data-points that are compiled from the literature and were derived with a large variety of different assays.

To evaluate how comparable the IC50 values between the two assays are, we have exchanged sets of peptides and experimentally measured their affinity to MHC alleles available in both assay systems. The scatterplot in Figure 1A shows that there is good agreement between the two assays for intermediate- and low-affinity peptides, less so for high-affinity peptides. To quantify the level of agreement between the two assays, we utilized Matthew’s correlation coefficients as a measure of classification agreement, which yield values of 1.0 for perfect agreement and 0.0 for uncorrelated classifications (Figure 1N). For IC50 higher than 150 nM, the correlation coefficient is consistently above 0.65, indicating good agreement between the two assays. Conveniently, at the IC50 = 500 nM cutoff, which is commonly used to classify peptides into binders (IC50 < 500 nM) or nonbinders (IC50 ≥ 500 nM), the correlation coefficient is close to 1.0.

| Organism | MHC Allele | Peptide Length | Peptide Counts |
|----------|------------|----------------|---------------|
| Human    | HLA A*0101 | 9              | 976           |
|          |            | 10             | 56            |
|          | HLA A*0201 | 9              | 970           |
|          |            | 10             | 1316          |
|          | HLA A*0202 | 9              | 1447          |
|          |            | 10             | 1056          |
|          | HLA A*0203 | 9              | 1443          |
|          |            | 10             | 1055          |
|          | HLA A*0206 | 9              | 1437          |
|          |            | 10             | 1054          |
|          | HLA A*0301 | 9              | 986           |
|          |            | 10             | 1082          |
|          | HLA A*1101 | 9              | 865           |
|          |            | 10             | 1093          |
|          | HLA A*2301 | 9              | —             |
|          | HLA A*2402 | 9              | 197           |
|          | HLA A*2403 | 9              | 254           |
|          | HLA A*2601 | 9              | 593           |
|          | HLA A*2902 | 9              | —             |
|          |            | 10             | 160           |
|          | HLA A*3001 | 9              | 539           |
|          |            | 9              | 92            |
|          | HLA A*3101 | 9              | 800           |
|          |            | 10             | 1057          |
|          | HLA A*3301 | 9              | —             |
|          |            | 10             | 1055          |
|          | HLA A*6801 | 9              | —             |
|          |            | 10             | 1055          |
|          | HLA A*6802 | 9              | 1434          |
|          | HLA A*6901 | 9              | 833           |
|          | HLA B*0702 | 9              | 985           |
|          |            | 10             | 205           |
|          | HLA B*0801 | 9              | 698           |
|          | HLA B*1501 | 9              | 975           |
|          | HLA B*1801 | 9              | 118           |
|          | HLA B*2705 | 9              | 969           |
|          | HLA B*3501 | 9              | 484           |
|          |            | 10             | 177           |
|          | HLA B*4001 | 9              | 962           |
|          | HLA B*4002 | 9              | 118           |
|          | HLA B*4402 | 9              | 119           |
|          | HLA B*4403 | 9              | 119           |
|          | HLA B*4501 | 9              | 114           |
|          | HLA B*5101 | 9              | 244           |
|          | HLA B*5101 | 10             | 177           |
|          | HLA B*5301 | 9              | 254           |
|          |            | 10             | 177           |
|          | HLA B*5401 | 9              | 255           |
|          |            | 10             | 177           |
|          | HLA B*5701 | 9              | 59            |
|          | HLA B*5801 | 9              | 930           |
|          |            | 9              | 988           |
| Mouse    | H-2 Db     | 9              | 303           |
|          |            | 10             | 134           |
|          | H-2 Dd     | 9              | 85            |
|          |            | 10             | 75            |
|          | H-2 Kb     | 8              | 480           |
|          |            | 9              | 223           |
|          | H-2 Kd     | 9              | 176           |
|          |            | 10             | 70            |
|          | H-2 Kk     | 8              | 80            |
|          |            | 9              | 164           |
|          |            | 10             | 57            |
|          | H-2 Ld     | 9              | 102           |
determining KD based on IC50 values may no longer be observed differences is that for very-high–affinity peptides, coefficients below 0.37. One explanation consistent with the two assays show much less agreement, with correlation coefficients drawn from the same sample [47]. The corresponding linear regression curves are included in Figure 2.

Figure 2 depicts scatter plots comparing the measured affinities of 3,089 nonamer peptides to HLA-A*0201 with their predicted scores for the three methods. The expected positive correlation between predicted and measured affinities was observed for each method. Note that a large fraction of measured affinities have their value set to the upper detection limit (>20,000 nM), and therefore appear as horizontal lines of dots in the scatter plots. Also note that the three methods handle very high and low predicted values differently. The ANN predictions are limited to values between 1 and 50,000 nM and the ARB predictions are not capped at all, which can lead to predictions outside of the experimentally observable range.

To quantitatively compare prediction quality, we calculated linear correlation coefficients between predicted and measured affinities on a logarithmic scale. For this calculation, all peptides with measured affinities at the upper detection limit were ignored. The resulting correlation coefficients are ARB = 0.55, SMM = 0.62, and ANN = 0.69, making the ANN predictions the best in a statistically significant manner (p < 0.05 using a t test for correlation coefficients drawn from the same sample [47]). The corresponding linear regression curves are included in Figure 2.

An alternative measure of prediction quality is a receiver operating characteristic (ROC) analysis. This evaluates how well the predicted scores classify peptides into binders (experimental IC50 < 500 nM) and nonbinders (experimental IC50 ≥ 500 nM) by plotting the rate of true positive classifications as a function of the false-positive classifications for all possible cutoffs in the prediction output. The overall quality of the prediction is measured by the area under the ROC curve (AUC), which is 1.0 if the prediction is perfect and 0.5 if it is random. This metric has the advantage that (1) it is invariant to different scales of the prediction output and only slightly affected by prediction caps; (2) it is more robust against outliers than a regression analysis; and (3) all differences found when curating from the literature, which is derived by a multitude of different experimental approaches.

Evaluating Prediction Methods

We used this dataset to compare the performance of three prediction methods currently used in-house in our labs: the ARB [5] and SMM [42] methods generate scoring matrices, while the ANN [41] method generates an artificial neural network. All three methods predict the quantitative affinity of a peptide for an MHC molecule. At this time, the ANN method has only been applied to the prediction of peptides of length nine. We are currently working on expanding this algorithm to make prediction for different lengths possible, but in the comparison presented here, we intentionally did not modify any of the three prediction methods from their previously published implementations.

With the dataset described above, we used five-fold cross-validation to generate and evaluate predictions for each of the three methods. For each allele and peptide length combination, the available data were split into five equally sized sets, of which four were used to generate a prediction tool (i.e., a matrix or a neural network). The tool generated was then used to predict the affinities of the peptides in the left-out set. Repeating this five times, each peptide in the original dataset was assigned a predicted score.

For peptides with high affinities of IC50 ≤ 50 nM or better, the two assays show very good agreement with a Matthew’s correlation coefficient of 0.80.

For peptides with high affinities of IC50 = 50 nM or better, the two assays show much less agreement, with correlation coefficients below 0.37. One explanation consistent with the observed differences is that for very-high–affinity peptides, determining K_D based on IC50 values may no longer be reliable as the concentration of MHC molecules is no longer negligible compared to the peptide concentration used for saturation (also known as “ligand depletion”) [46].

The assay comparisons presented herein provide an example of how pooling experimental data from different sources without additional validation can be problematic; the differences encountered between the measurements of the two closely related assays here are small compared to the

Figure 1. Comparability of the Binding Affinities between Assays
(A) Scatter plot comparing measured affinities for peptides to MHC recorded in the Buus (y-axis) and Sette (x-axis) assay systems. (B) The agreement between experimental classifications of peptides as binders/nonbinders at different affinity thresholds (x-axis) is measured by the Matthews correlation coefficient (y-axis). The dashed lines indicates the IC50 = 500 nM cutoff commonly used for classifying peptides into binders and nonbinders, which is used in the ROC analysis. DOI: 10.1371/journal.pcbi.0020065.g001

500 nM) [45], the two assays show very good agreement with a Matthew’s correlation coefficient of 0.80.

For peptides with high affinities of IC50 = 50 nM or better, the two assays show much less agreement, with correlation coefficients below 0.37. One explanation consistent with the observed differences is that for very-high–affinity peptides, determining K_D based on IC50 values may no longer be reliable as the concentration of MHC molecules is no longer negligible compared to the peptide concentration used for saturation (also known as “ligand depletion”) [46].

The assay comparisons presented herein provide an example of how pooling experimental data from different sources without additional validation can be problematic; the differences encountered between the measurements of the two closely related assays here are small compared to the

Evaluating Prediction Methods

We used this dataset to compare the performance of three prediction methods currently used in-house in our labs: the ARB [5] and SMM [42] methods generate scoring matrices, while the ANN [41] method generates an artificial neural network. All three methods predict the quantitative affinity of a peptide for an MHC molecule. At this time, the ANN method has only been applied to the prediction of peptides of length nine. We are currently working on expanding this algorithm to make prediction for different lengths possible, but in the comparison presented here, we intentionally did not modify any of the three prediction methods from their previously published implementations.

With the dataset described above, we used five-fold cross-validation to generate and evaluate predictions for each of the three methods. For each allele and peptide length combination, the available data were split into five equally sized sets, of which four were used to generate a prediction tool (i.e., a matrix or a neural network). The tool generated was then used to predict the affinities of the peptides in the left-out set. Repeating this five times, each peptide in the original dataset was assigned a predicted score.

For peptides with high affinities of IC50 ≤ 50 nM or better, the two assays show very good agreement with a Matthew’s correlation coefficient of 0.80.

For peptides with high affinities of IC50 = 50 nM or better, the two assays show much less agreement, with correlation coefficients below 0.37. One explanation consistent with the observed differences is that for very-high–affinity peptides, determining K_D based on IC50 values may no longer be reliable as the concentration of MHC molecules is no longer negligible compared to the peptide concentration used for saturation (also known as “ligand depletion”) [46].

The assay comparisons presented herein provide an example of how pooling experimental data from different sources without additional validation can be problematic; the differences encountered between the measurements of the two closely related assays here are small compared to the

Evaluating Prediction Methods

We used this dataset to compare the performance of three prediction methods currently used in-house in our labs: the ARB [5] and SMM [42] methods generate scoring matrices, while the ANN [41] method generates an artificial neural network. All three methods predict the quantitative affinity of a peptide for an MHC molecule. At this time, the ANN method has only been applied to the prediction of peptides of length nine. We are currently working on expanding this algorithm to make prediction for different lengths possible, but in the comparison presented here, we intentionally did not modify any of the three prediction methods from their previously published implementations.

With the dataset described above, we used five-fold cross-validation to generate and evaluate predictions for each of the three methods. For each allele and peptide length combination, the available data were split into five equally sized sets, of which four were used to generate a prediction tool (i.e., a matrix or a neural network). The tool generated was then used to predict the affinities of the peptides in the left-out set. Repeating this five times, each peptide in the original dataset was assigned a predicted score.

Figure 2 depicts scatter plots comparing the measured affinities of 3,089 nonamer peptides to HLA-A*0201 with their predicted scores for the three methods. The expected positive correlation between predicted and measured affinities was observed for each method. Note that a large fraction of measured affinities have their value set to the upper detection limit (>20,000 nM), and therefore appear as horizontal lines of dots in the scatter plots. Also note that the three methods handle very high and low predicted values differently. The ANN predictions are limited to values between 1 and 50,000 nM and the ARB predictions are not capped at all, which can lead to predictions outside of the experimentally observable range.

To quantitatively compare prediction quality, we calculated linear correlation coefficients between predicted and measured affinities on a logarithmic scale. For this calculation, all peptides with measured affinities at the upper detection limit were ignored. The resulting correlation coefficients are ARB = 0.55, SMM = 0.62, and ANN = 0.69, making the ANN predictions the best in a statistically significant manner (p < 0.05 using a t test for correlation coefficients drawn from the same sample [47]). The corresponding linear regression curves are included in Figure 2.

An alternative measure of prediction quality is a receiver operating characteristic (ROC) analysis. This evaluates how well the predicted scores classify peptides into binders (experimental IC50 < 500 nM) and nonbinders (experimental IC50 ≥ 500 nM) by plotting the rate of true positive classifications as a function of the false-positive classifications for all possible cutoffs in the prediction output. The overall quality of the prediction is measured by the area under the ROC curve (AUC), which is 1.0 if the prediction is perfect and 0.5 if it is random. This metric has the advantage that (1) it is invariant to different scales of the prediction output and only slightly affected by prediction caps; (2) it is more robust against outliers than a regression analysis; and (3) all
measurements including peptides without quantitative affinities (e.g., >20,000 nM) can be utilized. Also, our two experimental sources show very good agreement at the IC50 = 500 nM cutoff (Figure 1). This means that an ROC analysis at this cutoff is less prone to artifacts introduced by pooling the two sets of data than the regression analysis.

Figure 2 presents ROC curves for the three methods. Comparing classifications with the same rate of false positives, the ANN predictions always have an equal or higher rate of true positives than the SMM predictions, which in turn outperform the ARB predictions. This is reflected in the AUC values of ARB = 0.934, SMM = 0.952, and ANN = 0.957, which again shows the ANN predictions to be significantly better than the others (p < 0.05 using a paired t test on AUC values generated by bootstrap as described in Materials and Methods).

We repeated the same analysis for all MHC alleles and peptide lengths for which we have binding data available. Table 2 shows the AUC values for each method. Comparing only the predictions for 9-mer peptides, where all three methods were available, shows that the ANN predictions are the best in 30 cases, the SMM predictions in 16, and the ARB predictions in zero cases. The differences between the predictions of the three methods is statistically significant (ARB < SMM < ANN) as evaluated by a paired t test and a Wilcoxon signed-rank test (both with p < 0.05). Comparing the prediction performance utilizing correlation coefficients instead of AUC values gives very similar results, as does repeating the ROC analysis with classification cutoffs of IC50 = 50 and 5,000 nM instead of 500 nM (unpublished data).

It is commonly assumed that scoring matrices can be useful for smaller datasets, while neural networks should outperform them if large training datasets are available [38,48], as they can model higher-order correlations that require a large dataset to be estimated precisely. To analyze how well this is reflected in our results, we plotted the AUC values of each method as a function of peptides in the training set (Figure 3). For datasets with less than 300 peptides available, the ANN method performs best, outperforming the SMM method in 16 of 23 cases. Interestingly, this ratio does not increase for datasets containing more than 300 peptides, for which it outperforms the SMM method in 14 of 23 cases. This indicates that the primary limiting factor for the performance of the SMM method is not its inability to model higher-order correlations, which would have resulted in an increasing performance gap for larger datasets. The same is true for the ARB method, as it gains the most from increasing amounts of data, which again indicates that the matrix representation per se is not the primary reason for its underperformance; rather, it is the accuracy of the determined matrix values that improves as the amount of data increase.

Comparison with Publicly Available Prediction Tools

As far as possible, we also wanted to compare our results with other existing predictions. In October and November 2005, we retrieved predictions from all tools known to us to be freely accessible on the internet for all the peptides in our dataset. Only servers that (1) provided predictions for the alleles in our dataset; (2) were available during that time; and...
did not specifically disallow the use of automated prediction retrieval were taken into account. This included the following 16 tools: arbmatrix [5], bimas [10], hlaigand [16], hla_a2_smm [12], libscore [17], mapp [8], mhcpathway [18], mhcpred [7], multipred [19], netmhc [20,24,41], pepdist [22], predbalbc [23], predep [21], rankpep [14], svmhc [6], and syfpeithi [13]. We always used the default parameter settings for each tool, and we used the immediate tool output as seen by a user. No attempts were made to optimize the results for any tool once meaningful predictions could be retrieved. We are aware that this may lead us to underestimate the performance of some tools (e.g., svmhc provides an alternative output format with quantitative values for nonbinding predictions) that we have initially overlooked.

| Dataset Allele | Length | # Peptides | Method AUCs | External Tool |
|----------------|--------|------------|-------------|---------------|
|                |        |            | ARB         |               |
|                |        |            | SMM         |               |
|                |        |            | ANN         |               |
| H-2 Db         | 9      | 303        | 0.865       | syfpeithi     | 0.892         |
|                | 10     | 134        | 0.715       | syfpeithi     | 0.892         |
| H-2 Dd         | 9      | 85         | 0.696       | predbalbc     | 0.929         |
|                | 10     | 75         | 0.990       | rankpep       | 0.958         |
| H-2 Kb         | 8      | 480        | 0.846       | libscore      | 0.885         |
|                | 9      | 223        | 0.792       | libscore      | 0.866         |
| H-2 Kd         | 9      | 176        | 0.798       | syfpeithi     | 0.903         |
|                | 10     | 70         | 0.486       | syfpeithi     | 0.785         |
| H-2 Kk         | 8      | 80         | 0.782       | rankpep       | 0.825         |
|                | 9      | 164        | 0.758       | bimas         | 0.756         |
|                | 10     | 57         | 0.615       | bimas         | 0.690         |
| H-2 Ld         | 9      | 102        | 0.551       | libscore      | 0.967         |
| HLA A*0101     | 9      | 1157       | 0.964       | hlaigand      | 0.955         |
| HLA A*0201     | 9      | 3089       | 0.934       | hla_a2_smm    | 0.922         |
|                | 10     | 1316       | 0.885       | bimas         | 0.873         |
| HLA A*0202     | 9      | 1447       | 0.875       | multipredann  | 0.793         |
|                | 10     | 1056       | 0.819       | syfpeithi     | 0.500         |
| HLA A*0203     | 9      | 1443       | 0.884       | multipredann  | 0.788         |
|                | 10     | 1055       | 0.775       | syfpeithi     | 0.539         |
| HLA A*0206     | 9      | 1437       | 0.872       | multipredann  | 0.735         |
| HLA A*0301     | 9      | 2094       | 0.908       | rankpep       | 0.724         |
| HLA A*1101     | 9      | 1985       | 0.918       | multipredann  | 0.869         |
|                | 10     | 1093       | 0.859       | bimas         | 0.795         |
| HLA A*2402     | 9      | 197        | 0.718       | syfpeithi     | 0.770         |
|                | 10     | 78         | 0.733       | syfpeithi     | 0.853         |
| HLA A*2601     | 9      | 672        | 0.907       | pepdist       | 0.736         |
| HLA A*2902     | 9      | 160        | 0.755       | rankpep       | 0.597         |
| HLA A*3101     | 9      | 1869       | 0.909       | rankpep       | 0.828         |
|                | 10     | 1057       | 0.832       | bimas         | 0.746         |
| HLA A*3301     | 9      | 1140       | 0.892       | rankpep       | 0.807         |
|                | 10     | 1055       | 0.837       | rankpep       | 0.661         |
| HLA A*6801     | 9      | 1141       | 0.840       | syfpeithi     | 0.772         |
|                | 10     | 1055       | 0.812       | syfpeithi     | 0.779         |
| HLA A*6802     | 9      | 1434       | 0.865       | mhcspred      | 0.643         |
| HLA B*0702     | 9      | 1262       | 0.952       | hlaigand      | 0.942         |
|                | 10     | 205        | 0.758       | bimas         | 0.759         |
| HLA B*0801     | 9      | 708        | 0.936       | pepdist       | 0.766         |
| HLA B*1501     | 9      | 978        | 0.900       | rankpep       | 0.816         |
| HLA B*1801     | 9      | 118        | 0.573       | pepdist       | 0.779         |
| HLA B*2705     | 9      | 969        | 0.915       | rankpep       | 0.926         |
| HLA B*3501     | 9      | 736        | 0.851       | bimas         | 0.792         |
|                | 10     | 177        | 0.758       | rankpep       | 0.793         |
| HLA B*4002     | 9      | 118        | 0.541       | rankpep       | 0.775         |
| HLA B*4402     | 9      | 119        | 0.533       | syfpeithi     | 0.783         |
| HLA B*4403     | 9      | 119        | 0.461       | rankpep       | 0.698         |
| HLA B*5101     | 9      | 244        | 0.822       | rankpep       | 0.820         |
|                | 10     | 177        | 0.727       | bimas         | 0.838         |
| HLA B*5301     | 9      | 254        | 0.871       | rankpep       | 0.861         |
|                | 10     | 177        | 0.595       | rankpep       | 0.701         |
| HLA B*5401     | 9      | 255        | 0.847       | svmhc         | 0.799         |
| HLA B*5701     | 9      | 59         | 0.428       | pepdist       | 0.767         |
| HLA B*5801     | 9      | 988        | 0.889       | bimas         | 0.899         |
|                |        |            | 0.791       |               |
|                |        |            | 0.874       |               |
|                |        |            | 0.900       |               |
| Averages       |        |            | 0.796       |               |

DOI: 10.1371/journal.pcbi.0020065.t002
Figure 4 shows predictions retrieved for peptides binding to HLA-A*0201 from two of the most established prediction methods, bimas and syfpeithi. The prediction output of these two methods are not IC50 values, but estimates of half-life of dissociation in the case of bimas, and an integer score for the syfpeithi predictions. In both cases, higher predictions correspond to better binders, which here means lower IC50 values. An ROC analysis of these predictions is also shown in Figure 4, which gives AUC values of 0.920 for bimas and 0.871 for syfpeithi. A sample collection of AUC values for more alleles and other external prediction tools is shown in Table 3. For each allele and peptide length, the tool with the highest AUC value among the set of external tools is listed in the last two columns of Table 2. The complete listing for all external tools is contained in Table S1, which also contains the URLs from which predictions were obtained.

It has to be stressed that this analysis does not fairly judge the performance of external predictions in all cases. For example, some methods such as syfpeithi do not aim to specifically predict peptide binding to MHC molecules, but rather naturally processed peptide ligands. Also, the amount and quality of training data available to each method are divergent, which disadvantages methods with little access to training data. In contrast, some tools were generated with an appreciable fraction of data that is used here for testing. Such nonblind tool generation leads to an overestimation of performance. These tools are marked with an asterisk (*) in Table 3, and are excluded from Table 2.

In light of the above caveats, we focus on successful external predictions. In total there were 54 allele/peptide length combinations for which we had at least one external prediction tool available (Table 2). For eight of these 54 combinations, one of the external tools performed better than all three of our methods. These eight tools were derived from bimas (one instance), libscore (two instances), predbalbc (one instance), rankpep (one instance), and syfpeithi (three instances). For seven out of eight of these tools, the corresponding datapoints available for training in our affinity set are comparably few (<140), which may explain the low performance of our three methods for these sets. For 9-mers binding to H-2 Kb, however, there are 223 datapoints available, and the libscore predictions, which are based on a combinatorial peptide library, perform the best.

Next, we analyzed if the underperformance of matrix-based tools that we found when comparing in-house
prediction methods could also be seen for external tools. We therefore separated the tools into matrix-based and non-matrix-based (see Materials and Methods for the classification). For 30 allele/peptide length combinations, there were predictions available from tools of both categories. We compared the highest AUC values in each tool category for predictions available from tools of both categories. We compared the highest AUC values in each tool category for all datasets. Using a paired t test and a Wilcoxon signed-rank test, we found that the matrix-based tools significantly outperformed the non–matrix-based tools (both with p < 0.05). This comparison again has to be interpreted with caution, as the access to data of the different tools is probably much more important factor in determining prediction quality than the underlying algorithm used.

A Web-Based Framework for the Generation and Evaluation of Prediction Methods and Tools

When evaluating our three prediction methods, we encountered multiple problems caused by differences in their implementation. All have been implemented in different programming languages: the ANN method is implemented in Fortran, SMM in C++, and ARB in Java. Also, all have different input and output requirements. It became clear that an abstraction layer providing a common interface to prediction tools and methods would be highly beneficial.

As many tools were already implemented as web servers, it was natural to define this abstraction layer as a set of http commands. We defined such a common interface to both query existing prediction tools as well as coordinate the generation of tools by prediction methods. Figure 5 gives an overview of the interactions defined in the abstraction layer.

The framework is designed to be expandable and place minimum requirements on the implementation of outside prediction methods. This will allow tool developers to plug their existing or newly developed prediction methods into the same framework for a transparent, automated comparison with other predictions. This allows controlling for both the training and testing data used, enabling a true side-by-side comparison. Also, all methods implemented this way automatically benefit from increases in the data available to the IEDB.

Discussion

In the present report, we make available what is to date the largest dataset of quantitative peptide-binding affinities for MHC class I molecules. Establishing this dataset is part of the IEDB [39] effort, and was generated specifically with tool developers in mind. While the main part of the IEDB is structured to store a large amount of detailed immunological data, the present dataset is a curated, more homogeneous subset. This allows computer scientists and bioinformaticians to focus on improving prediction algorithms while avoiding common problems in data assembly from the literature such as inconsistent annotation of MHC alleles, handling conflicting data from unrelated assays, errors due to manual entry of the data, and, of course, the effort involved in collecting the data.

Another significant problem in the generation of peptide-MHC binding datasets is that immunologists often consider negative binding data as not interesting enough for publication. This biases the immunological literature to report only positive binding data, and forces tool developers to approximate negative binders with randomly generated peptides. While the use of random peptides is often necessary, previous studies have shown that the use of true nonbinding peptides allows for the generation of better predictions [22,49]. The present set of peptide-binding data removes the need for randomized peptides, as all binding data generated is reported, including plenty of nonbinding peptides.

The data in our set come exclusively from two assay systems established in the Buus and Sette labs. This makes it much more homogeneous than other available datasets, typically

### Table 3. Prediction Quality of Tools Available Online

| Prediction          | MHC Allele/Peptide Length |
|---------------------|---------------------------|
|                     | H-2 Db | H-2 Kd | H-2 Ld | HLA A*0201 | HLA A*0201 | HLA A*1101 | HLA A*2402 | HLA B*0702 |
|                     | 9      | 9      | 9      | 10         | 9          | 9          | 9          | 9          |
| armbmatrixa         | 0.869  | 0.899  | —      | 0.935      | 0.904      | 0.914      | 0.879      | 0.952      |
| bimas              | 0.747  | 0.871  | 0.961  | 0.920      | 0.873      | 0.856      | —          | 0.908      |
| hla_a2_smm         | —      | —      | —      | 0.922      | 0.853      | —          | —          | —          |
| hlagand            | —      | —      | —      | 0.816      | 0.776      | 0.846      | 0.742      | 0.942      |
| libscore           | 0.908  | 0.987  | —      | —          | —          | —          | —          | —          |
| mapppB             | 0.747  | 0.871  | 0.961  | 0.920      | 0.873      | 0.857      | —          | —          |
| mapppS             | 0.803  | 0.904  | 0.957  | 0.871      | 0.676      | —          | 0.941      |
| mhcppathway        | —      | —      | —      | 0.915      | —          | —          | —          | 0.953      |
| mhcppred           | 0.723  | 0.814  | —      | —          | 0.725      | —          | —          | —          |
| multipredann       | —      | —      | —      | 0.883      | —          | 0.869      | —          | —          |
| multipredhmm       | —      | —      | —      | 0.796      | —          | 0.832      | —          | —          |
| netmhcann         | 0.807  | 0.934  | —      | —          | 0.920      | —          | 0.963      |
| netmhcmatrixa      | —      | —      | —      | —          | —          | 0.774      | —          | —          |
| pepdist            | 0.677  | 0.748  | 0.819  | 0.789      | —          | 0.819      | 0.763      | 0.892      |
| predbalbc          | —      | 0.807  | 0.934  | —          | —          | —          | —          | —          |
| predep             | 0.606  | 0.629  | 0.788  | 0.757      | —          | —          | —          | —          |
| rankpep            | 0.748  | 0.883  | 0.939  | 0.836      | 0.749      | 0.861      | 0.723      | 0.935      |
| svmhc              | —      | —      | —      | 0.814      | 0.705      | 0.642      | 0.667      | 0.872      |
| syfpeithi          | 0.803  | 0.903  | 0.957  | 0.871      | 0.801      | —          | 0.770      | 0.937      |

aTraining data for prediction tool are known to substantially overlap with testing data.
DOI: 10.1371/journal.pcbi.0020065.t003
curated from the literature. Moreover, we conducted a set of reference experiments to standardize the quantitative affinities observed in the two assays. This showed that for peptides with IC_{50} values > 400 nM, the measurements of the two assays corresponded very well, less so for high-affinity peptides. We originally hoped to convert IC_{50} values from different sources onto a common scale. However, our analysis suggests that this may not be possible due to differences in sensitivities between the two assay systems. Still, by documenting incompatibilities between assays, these can be taken into account by tool developers. Specifically for the current dataset, we recommend evaluating prediction performance by the ability to classify peptides into binders and nonbinders at a cutoff of 500 nM. We plan to include data from additional sources to this dataset, for which we will carry out a similar process of exchanging peptides and reagents to ensure consistency of the reported affinities.

We have used the dataset to evaluate the prediction performance of three methods that are routinely used by our groups. In this comparison, the ANN method outperformed the two matrix-based predictions ARB and SMM, independent of the size of the training dataset. This surprising result indicates that the primary reason for the superior ANN performance is not its ability to model higher-order sequence correlations, which would result in a larger performance gap for increasing dataset size. This does not imply that higher-order sequence correlations play no role in peptide binding to MHC. Indeed, this is very unlikely, as the peptide must fit into the binding cleft, which is restricted by the available space and contact sites, for which neighboring residues will compete. To directly assess the importance of higher-order correlations, one would need to calculate, for instance, the mutual information by estimating amino acid pair frequencies for the 400 possible pairs at two positions in the peptide [50]. However, the signal-to-noise ratio of such a calculation is still too low for datasets of the size utilized in this study, which are still very small compared to other fields where higher-order correlations definitely do play a role (e.g., secondary structure predictions).

The high performance of the ANN method on small datasets is likely due to the fact that the present ANN method being utilized is a hybrid, where the peptide amino acid sequence is represented according to several different encoding schemes, including conventional sparse encoding, Blosum encoding, and hidden Markov model encoding [41]. This encoding enables the network to generalize the impact on binding of related amino acids.

Multiple comparisons of tool prediction performance have been made before with conflicting outcomes when comparing matrix predictions with neural networks [12,38,48]. The comparison presented here is different in two main aspects. First, the magnitude of data used in this comparison is 10- to 100-fold larger than previous attempts. Second, the three methods in the comparison were all used and optimized as implemented by their developers. This avoids expert bias (i.e., the effect that a tool developer is better able to optimize predictions of methods he/she is familiar with than those he/she is unfamiliar with).

We have also evaluated the performance of external prediction tools on this dataset. As could be expected simply because of differences in the type and amount of data available to the external tools for training, their prediction performance is usually below that recorded by the methods in cross-validation. Specifically, as the set of peptide sequences was not homology-reduced, the performance of the three internal prediction methods is overestimated compared to the external tools. Therefore, we expect that the performance of all external tools will improve significantly when retraining them with the data made available here. Still, for a number of datasets, the best external predictions outperform all three methods tested in cross-validation here. In most cases, these datasets are comparably small (<140 peptides), which could explain why the three methods underperformed. One exception is the H-2 K^b set with 223 peptides, for which the libscore predictions, which are based on characterizing MHC binding combinatorial peptide libraries, perform best. As this requires a comparatively small number of affinity measurements (20X peptide length), this underlines the value of this approach for characterizing new MHC alleles.

All of the data generated in the evaluation process, including the dataset splits and predictions generated in cross-validation, is made publicly available. These data make the evaluation process itself transparent and allow for using them as benchmarks during tool development and testing.

While everyone can work with these benchmark sets in the privacy of their own lab, we hope that promising prediction methods will be integrated into our automated tool generation and evaluation framework. This web-based framework was designed to minimize requirements on hardware and software, and it enables a transparent side-by-side comparison of prediction methods.

Results from such a side-by-side comparison will help bioinformaticians identify which features make a prediction method successful, and they can be used as a basis for further dedicated prediction contests. Importantly, such comparisons will also help immunologists find the most appropriate prediction tools for their intended use.
The present evaluation is solely concerned with the prediction of peptide binding to MHC class I molecules. Binding of a peptide is a prerequisite for recognition during an immune response. However, there are many other factors that make some binding peptides more relevant than others for a given purpose. Examples of such factors include preferring peptides that are able to bind multiple MHC alleles, preferring peptides derived from viral proteins expressed early during infection, or preferring peptides that are efficiently generated from their source protein during antigen processing. For these and other factors, we plan to provide datasets and carry out evaluations similar to the one presented here in future studies. Our overall goal is to communicate problems of immunological relevance to bioinformaticians, and to demonstrate to immunologists how bioinformatics can aid in their work.

Materials and Methods

Peptide-binding assay—Sette. The MHC peptide-binding assay utilized in the Sette lab measures the ability of peptide ligands to inhibit the binding of a radiolabeled peptide to purified MHC molecules, and has been described in detail elsewhere [43,51,52]. Briefly, however, purified MHC molecules, test peptides, and a radiolabeled probe peptide are incubated for 2 d at room temperature in the presence of human B2-microglobulin and a cocktail of protease inhibitors. After the 2-d incubation, binding of the radiolabeled peptide to the corresponding MHC class I molecule is determined by capturing MHC–peptide complexes on W6/32 antibody (anti-HLA A, B, and C)–coated plates, and measuring bound cpm using a microscintillation counter. Alternatively, following the 2-d incubation, the percent of MHC-bound radioactivity can be determined by size exclusion gel filtration chromatography. For competition assays, the concentration of peptide yielding 50% inhibition of the binding of the radiolabeled peptide is calculated. Peptides are typically tested at six different concentrations covering a 100,000-fold dose range, and in three or more independent assays. Under the conditions utilized, where [label] < [MHC] and IC50 > [MHC], the measured IC50 values are reasonable approximations of the true Kp values [48,53].

Peptide-binding assay—Buus. The denatured and purified recombinant HLA heavy chains were diluted into a renaturation buffer containing HLA light chain, B2-microglobulin, and graded concentrations of the peptide to be tested, and incubated at 18 °C for 48 h allowing equilibrium to be reached. We have previously demonstrated that denatured HLA molecules can fold efficiently de novo, allowing equilibrium to be reached. We have previously demonstrated that denatured HLA molecules can fold efficiently de novo, allowing equilibrium to be reached. Denatured and purified HLA molecules, and radiolabeled peptides were incubated for 48 h at 37 °C with peptide concentrations ranging from 500 nM to 500 nM. Given a cutoff for the predicted value, predictions for peptides were separated into positive and negative subsets, allowing for calculation of the number of true-positive and false-positive predictions. Plotting the rates of true-positive predictions as a function of the rate of false-positive predictions gives an ROC curve.

Calculating the AUC provides a highly useful measure of prediction quality, which is 0.5 for random predictions and 1.0 for perfect predictions. The AUC value is equivalent to the probability that the predicted score for a randomly chosen binding peptide is (better) than that of a randomly chosen peptide that is not a binder. To assess if the AUC value of one prediction is significantly better than that of another prediction, we resampled the set of peptides for which predictions were made. Using bootstrapping with replacement, 50 new datasets were generated with a constant ratio of binder to nonbinder peptides. To assess if the AUC value of one prediction is significantly better than that of another prediction, we resampled the set of peptides for which predictions were made. Using bootstrapping with replacement, 50 new datasets were generated with a constant ratio of binder to nonbinder peptides. We then calculated the difference in AUC for the two predictions on each new dataset. One prediction was considered significantly better than another if the distribution of the AUC values was significantly different, which we measured using a paired t test.

Acknowledgments

Author contributions. BP, JS, SB, and AS conceived and designed the experiments. KL, MH, and JS performed the experiments. BP, HHH, SF, MN, CL, EK, JS, OL, SB, and AS analyzed the data. BP, HHH, SF, MN, CL, EK, DB, and OL contributed reagents/materials/analysis tools. BP, HHH, SF, MN, CL, WF, SSW, JS, OL, SB, and AS wrote the paper.

Funding. This work was supported by National Institutes of Health Contracts HHSN2662004006C and HHSN26620040025C.

Competing interests. The authors have declared that no competing interests exist.

Supporting Information

Table S1. Complete List of AUC Values

Found at: DOI: 10.1371/journal.pcbi.0020065.s001 (76 KB XLS).
