# Evaluation of MHC Class I Peptide Binding Prediction Servers: Applications for Vaccine Research

The Harvard community has made this article openly available. **Please share** how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Lin, Hong Huang, Surajit Ray, Songsak Tongchusak, Ellis L. Reinherz, and Vladimir Brusic. 2008. Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research. BMC Immunology 9: 8.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1186/1471-2172-9-8</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:5310900">http://nrs.harvard.edu/urn-3:HUL.InstRepos:5310900</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research

Hong Huang Lin1, Surajit Ray2, Songsak Tongchusak1, Ellis L Reinherz1 and Vladimir Brusic*1,3

Address: 1Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA, 2Department of Mathematics and Statistics, Boston University, Boston, MA, USA and 3School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia

Email: Hong Huang Lin - Honghuang_Lin@dfci.harvard.edu; Surajit Ray - Sray@math.bu.edu; Songsak Tongchusak - Songsak_Tongchusak@dfci.harvard.edu; Ellis L Reinherz - Ellis_Reinherz@dfci.harvard.edu; Vladimir Brusic* - Vladimir_Brusic@dfci.harvard.edu

* Corresponding author

Abstract

Background: Protein antigens and their specific epitopes are formulation targets for epitope-based vaccines. A number of prediction servers are available for identification of peptides that bind major histocompatibility complex class I (MHC-I) molecules. The lack of standardized methodology and large number of human MHC-I molecules make the selection of appropriate prediction servers difficult. This study reports a comparative evaluation of thirty prediction servers for seven human MHC-I molecules.

Results: Of 147 individual predictors 39 have shown excellent, 47 good, 33 marginal, and 28 poor ability to classify binders from non-binders. The classifiers for HLA-A*0201, A*0301, A*1101, B*0702, B*0801, and B*1501 have excellent, and for A*2402 moderate classification accuracy. Sixteen prediction servers predict peptide binding affinity to MHC-I molecules with high accuracy; correlation coefficients ranging from $r = 0.55$ (B*0801) to $r = 0.87$ (A*0201).

Conclusion: Non-linear predictors outperform matrix-based predictors. Most predictors can be improved by non-linear transformations of their raw prediction scores. The best predictors of peptide binding are also best in prediction of T-cell epitopes. We propose a new standard for MHC-I binding prediction – a common scale for normalization of prediction scores, applicable to both experimental and predicted data. The results of this study provide assistance to researchers in selection of most adequate prediction tools and selection criteria that suit the needs of their projects.

Background

Vaccines are the most effective immunologic intervention in controlling infectious disease [1] and offer a great promise for control of emerging infectious disease, cancer, allergies, and autoimmunity [2]. Peptide-based vaccines offer means for safe and precisely-directed immune intervention; more than 30 peptide-based vaccines are currently under development, including several that are in phase III clinical trials [3]. Most of these vaccines contain various forms of pathogen-derived or tumor-associated antigens. Various strategies for formulation (e.g. cells, whole antigens, subunits, or peptides) as well as for deliv-
high magnitude CD8+ T-cell immunity is a major direc-
vaccines that enable efficient priming, long-lasting and
by T-helper cells (Th). The development of multivalent
cells and bound by MHC class II molecules are recognized
external proteins internalized by the antigen presenting
phocytes (CTL). Peptides derived from degradation of
MHC-I molecules are recognized by cytotoxic T lym-
for recognition by T-cells of the immune system. Peptides
derived from degradation of internal proteins that bind
MHC-I molecules are recognized by cytotoxic T lymphocytes (CTL). Peptides derived from degradation of
external proteins internalized by the antigen presenting
cells and bound by MHC class II molecules are recognized
by T-helper cells (Th). The development of multivalent
vaccines that enable efficient priming, long-lasting and
high magnitude CD8+ T-cell immunity is a major direc-
ion in the current vaccine research [5]. CTL epitopes
induce specific responses against infected or malignant
cells, while Th epitopes initiate and regulate immune
responses.

T-cell epitopes are peptides that induce immune
responses when bound by major histocompatibility com-
plex (MHC) molecules and presented on the cell surface
for recognition by T-cells of the immune system. Peptides
derived from degradation of internal proteins that bind
MHC-I molecules are recognized by cytotoxic T lymphocytes (CTL). Peptides derived from degradation of
external proteins internalized by the antigen presenting
cells and bound by MHC class II molecules are recognized
by T-helper cells (Th). The development of multivalent
vaccines that enable efficient priming, long-lasting and
high magnitude CD8+ T-cell immunity is a major direc-

Antigens from pathogens or tumors represent suitable tar-
gets for immunotherapies and vaccines. Synthetic pep-
tides offer advantages for therapeutic use [6]; they are easy
to produce even for a clinical grade, are free from patho-
gen contamination, have minimal oncogenic potential,
and are chemically stable. Peptide epitopes have been
used in various formulations of vaccines [7-9]. While
some successes of epitope-based cancer vaccines have
been reported [10-12], the clinical applications of
epitope-based vaccines lag behind and the correlation
between responses to T-cell epitopes and clinical out-
comes has not been established [13,14]. The formulations
for cancer immunotherapies include tumor-specific tar-
gets, immune response enhancers, and immune evasion
suppressors [15]. Recent clinical studies indicate that high
level of tumor infiltration by activated CD8+ T-cells com-
bined with a low number of regulatory T-cells (Treg) is a
significant positive prognostic factor for patient survival
in cancers [16-19].

Identification of MHC-binding peptides and their subset
of T-cell epitopes helps improve our understanding of
specificity of immune responses. It is important for dis-
covery of vaccines and immunotherapies [3,6,20]. Tens of
thousands of protein variants have been characterized in
viruses, such as HIV, influenza, or dengue. The numbers
of bacterial, fungal, and parasite antigens are even larger.
Several hundred of tumor-related antigens and their vari-
ants have been reported [21,22]. More than two thousand
variants of human MHC (HLA) have been characterized
to date [23]. Given the significant number of antigens and
their variants, and a large number of HLA variants, sys-
tematic experimental testing of binding capacity of these
peptides is impractical. A number of computational meth-
ods have been developed to facilitate the identification
of HLA binding peptides [24-26]. More than thirty predic-
sion servers have been developed and are accessible via
the Internet. These methods use a variety of statistical and
machine learning approaches making computational pre-
screening of antigens for CTL epitopes a standard
approach in epitope-mapping studies. However, with so
many choices of prediction servers, new questions have
arisen: how to select the best server for a particular HLA
allele; can they be used to predict binding affinity of pep-
tides rather than classify into binders and non-binders;
and how to use predictions to identify T-cell epitopes
amongst HLA ligands? Lack of standards for the develop-
ment of MHC-I binding predictors resulted in servers that
show differences in predictions values and the wide scale
of prediction values. Comparisons of methods for prediction
of MHC-binding peptides have been reported, indi-
cating high accuracy of binding predictions [27-29].
Predictions of T-cell epitopes, which are a subset of MHC-
binders are less accurate and more difficult to model than
peptide binding predictions. In a recent study using HLA
class I (HLA-I) transgenic mice, 40 candidate T-cell
epitopes were identified from computational screening of
some 2,900 peptides. Of these 21 were identified as T-cell
epitopes and 17 were high-affinity HLA-binders [30]. A
new generation of predictive models that combine predic-
tions of multiple antigen processing and presentation
steps: HLA binding, peptide binding to transported asso-
ciated with antigen processing (TAP) and predictions of
proteasomal cleavage have been developed (reviewed in
[31]). While combination of HLA predictions and TAP
predictions offers improvement of predictions in some
cases [32], it eliminates TAP-independent peptides from
further analysis, such as those produced by vacuolar [33],
lysosomal [34], or endosomal [35], among others, path-
ways. Proteasomal cleavage predictions are of much lower
accuracy [32] than HLA-binding or TAP-binding predic-
tions; proteasomal cleavage methods have not yet been
adequately validated [31]. The utility and the mode of
usage of combined predictors are yet to be determined. In
the meantime, HLA-binding predictions remain the most
useful computational tools for mapping of HLA ligands
and T-cell epitopes.
Peters et al. [28] developed a community resource benchmarking prediction on a dataset comprising of 48,828 quantitative peptide-binding affinity measurements. However, a large fraction of this dataset has already been employed by some groups to develop their prediction servers, which may invalidate the comparison with those that did not employ this dataset. Trost et al. [29], on the other hand, compared the performance of sixteen servers, and combined predictions by a number of tools into a more accurate combined method. Because of the lack of adequate independent test sets, the comparison studies performed to date have been based on assessing predictive performance using pre-defined sets of peptides, rather than full-overlapping studies of complete antigens. In this study we compared the performance of 30 servers by first normalizing the predictions to a common scale and then assessing the performance using the data from a full-overlapping binding study of 9-mer peptides to seven HLA-I molecules. These peptides were derived from a tumor antigen and from a fragment of a viral antigen. We compared all the servers to find whether any of them produce identical predictions. The main part of the study explored the classification (prediction into binders and non-binders) vs. peptide binding affinity prediction capabilities of these servers. We analyzed their prediction performances on two sets of well-defined T-cell epitopes. Finally we explored the non-linear issues of post-processing the prediction values as possible means for improving predictions.

Results
Classification
While not all of these servers were designed for the specific purpose of peptide binding predictions, all of them have peptide binding predictions implemented as specific modules. For example MAPPP and ProPred1 predict multiple steps of antigen processing. MULTIPRED predicts peptide binding to HLA supertypes, and BIMAS predicts peptide binding as half-time dissociation (off-rate). Some servers have advanced options, for example MHCPred enables the specification of anchor positions. For this analysis we used the simplest prediction method available at each server. After performing all predictions using the test set, we first calculated the Pearson correlation coefficient for all the servers and found that MAPPP (BIMAS) and MAPPP (SYFPEITHI) showed identical predictions ($r = 1$) to BIMAS and SYFPEITHI, respectively. The ProPred1 and BIMAS predictions showed $r \geq 0.998$ for six HLA-I molecules, $r = 0.25$ for B*0702, while B*1501 were not available in ProPred1. This was as expected because BIMAS and SYFPEITHI matrices were adopted by MAPPP servers and BIMAS matrices in ProPred1 as HLA-I binding prediction tools. We therefore excluded MAPPP and ProPred1 from further analysis. The numbers of the servers we studied were: A*0201 – 27; A*0301 – 26; A*1101 – 25; A*2402 – 17; B*0702 – 23; B*0801 – 19; and B*1501 – 12. The mutual analysis of predictors by calculating correlation coefficient indicates that these predictors are independent, and predict different subsets of HLA-I binding peptides. These predicted sets are largely overlapping for predictors that employ similar prediction algorithm and show very high accuracy, for example IEDB_ANN and NETM_ANN where $r = 0.912$.

The analysis of classification accuracy (binders vs. non-binders) was performed using the cutoff of 30 (measured binding affinity of $\geq 30\%$ of the binding affinity of a positive control) for binders, while other peptides were considered as experimental non-binders. In total 147 individual predictors were tested of which 39 showed excellent, 47 good, 33 marginal, and 28 poor performance. The $A_{ROC}$ values of these predictions are shown in Figure 1.

We also performed the analysis of survivin test set and CMV construct test set and the results were very similar to the combined set predictions (Figures 2 and 3). The intersection values of sensitivity/specificity plots are consistent with the $A_{ROC}$ results. By HLA molecule, the best predictors are for B*0702, where 65% showed excellent classification properties, while approximately 30% of predictors for A*0201, A*0301, and B*1501, and 16% for A*1101 and B*0801 showed excellent classification. The classification accuracy for A*2402 is lower than for other HLA molecules in this study: 18% of predictors showed good classification properties, and the rest showed marginal or poor performance.

The best prediction server across all HLA molecules in this study is NETM_ANN, closely followed by IEDB_ANN and IEDB_SMM. MHCI_MM, MHCI_SM, MULTI_SVM and SVMHC_M also perform well. The best predictors we recommend for classification prediction are shown in Figure 1 as black bars.

Prediction of binding affinity
Prediction scores from various predictors represent a number of measurable entities. Experimental measurements from the iTopia™ are expressed as the concentration of peptide needed to achieve 50% binding (ED$_{50}$ value) and compared as percentage binding affinity relative to the positive control peptide. For example, the binding scores for BIMAS represent off-rates (minutes), IEDB and NETM_ANN servers represent binding affinity on a nanomolar scales, MHC I server predicts “binding energy”, while MULTIPRED server predicts an arbitrary binding score. Large discrepancies are observed even between predictors from the same server. For example the survivin$_{pep}$ peptide MGAPlPPPA is an experimental binder to A*0201 with estimated 94% affinity relative to
the positive control. The respective predictions for IEDB_ANN, IEDB_ARB, and IEDB_SMM are 23441, 365, and 3237 nM, while NETM_ANN predicted value is 8574 nM. Across all predictors a variety of scales and ranges of prediction scores have been observed. Obviously these predictors must be treated as different *in silico* assays and the comparison can be made only by using relative scales of predictions. Using iTopia™ binding assay as the experimental control, we calculated correlation coefficients for all available predictors for three data sets (survivin, CMV construct, and the combined data set). The results show that a high accuracy prediction of peptide binding affinity can be achieved for A*0201 (Figure 4) where IEDB_ANN and NETM_ANN show values of $r > 0.8$ while A*0201 predictors MHCL_MM, MULTI_ANN, MULTI_SVM, NETM_WM, and SYFPEITHI showed a relatively high correlation coefficient of $0.8 < r < 0.7$. The correlation coefficients of predictions for other HLA-I alleles are lower typically $0.6 < r < 0.8$ for the best predictors of binding affinity except for B*0801 where the best predictor had $r = 0.55$. Overall, the best predictors of binding affinity are IEDB_ANN and NETM_ANN. The peptide binding affinities for A*0201 can be predicted *in silico* with high accuracy and, as both the quantity and quality of binding data increases, this will also be achieved for other HLA-I molecules. The best predictors that we recommend for prediction of peptide binding affinity are marked by asterisks (Figure 4).

The peptide binding prediction results across three different datasets show reasonable consistency for indicating that the most predictors generalize well (i.e. predict well across different data sets). For most predictors the prediction accuracy for CMV construct was somewhat higher than for survivin, while the predictions on the combined
set were mostly higher than those for survivin and lower than those for CMV construct values. The BIMAS predictions showed low stability in this test, while recommended predictors show high consistency of predictions across the three test sets.

Non-linear transformations
The predictions of peptide binding classification (Figure 1) show much higher accuracy across different prediction servers than the predictions of binding affinity (Figure 4). For example, the three IEDB predictors and two NetMHC predictors show very similar classification accuracy for A*0201 (Figure 1) while they show significant differences in the prediction of peptide binding affinity where ANN-based predictors are far superior to matrix-based predictors. For each A*0201 predictor, we performed four non-linear transformations and from five sets data selected one that showed the best predictive performance.

The results indicate that the scaling of the output results is a major issue and that it is necessary if linear predictors (matrix-based) are used for prediction of binding affinity. Only four predictors were optimized for output scaling (HLA_LI, IEDB_ANN, PEPC_M, and SVMHC_S), additional fourteen servers showed minor improvements of the correlation coefficient (less than 10% increase relative to the raw predictor output), while the rest of the servers showed sizable improvements (Figure 5). The largest improvements were seen for BIMAS, IEDB_ARB,
IEDB_SMM, MHCP_I and MHCP_AA predictors. These results show that most of predictors can be improved by post-processing the prediction outputs through scaling and non-linear transformations. This correction will not affect classification accuracy (binders vs. non-binders) since classification is threshold-dependent and the relative order of predictions remains the same as in the raw prediction list. While all four transformations are represented in the improved prediction sets, the largest improvements were achieved by the logarithmic transformation of matrix predictions indicating that in these cases an inappropriate formula was used for the definition of matrix coefficients.

Prediction of T-cell epitopes

We performed prediction of peptide binding with tumor antigen T-cell epitopes and viral epitope sets. Both sets showed similar prediction patterns and we proceeded with the analysis of merged data sets. For each server we predicted the binding affinity of all T-cell epitopes in the merged set and determined the threshold at which approximately 90% of the tested T-cell epitopes were predicted as binders and the threshold at which the first false positive appears at the test set of binders. The higher of the thresholds was used for further analysis for the assessment the number of false positive predictions based on the number performance on the survivin/CMV construct set.

Predictors could be used for different practical purposes. We compared the performance of servers in three scenarios for each predictor (representative results are shown in Tables 1, 2 and 3). These scenarios are represented by the selection of thresholds which correspond to practical application. The first case is the selection of threshold at which ~90% of T-cell epitopes are predicted as binders; the second threshold predicts correctly the majority of binders (31 of 33); and the third threshold does not allow any non-binders to be predicted as binders. The results

Table 1: Prediction performance of selected representative servers in order to correctly predict ~90% of T-cell epitopes

<table>
<thead>
<tr>
<th>Server</th>
<th>Thr</th>
<th>TP (binding)</th>
<th>TN (binding)</th>
<th>FP (binding)</th>
<th>FN (binding)</th>
<th>TP (tumor epitopes)</th>
<th>TP (viral epitopes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIMAS (A)</td>
<td>2</td>
<td>10</td>
<td>143</td>
<td>0</td>
<td>23</td>
<td>76 (89%)</td>
<td>39 (89%)</td>
</tr>
<tr>
<td>MHCP_I (A)</td>
<td>100</td>
<td>31</td>
<td>7</td>
<td>136</td>
<td>2</td>
<td>80 (94%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>IEDB_SMM (B)</td>
<td>1,000</td>
<td>10</td>
<td>143</td>
<td>0</td>
<td>23</td>
<td>77 (91%)</td>
<td>38 (86%)</td>
</tr>
<tr>
<td>NHCP_CP (C)</td>
<td>0</td>
<td>31</td>
<td>126</td>
<td>17</td>
<td>2</td>
<td>79 (93%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>IEDB_ANN (D)</td>
<td>10,000</td>
<td>9</td>
<td>143</td>
<td>0</td>
<td>24</td>
<td>76 (89%)</td>
<td>39 (89%)</td>
</tr>
<tr>
<td>MULTI_SVM (D)</td>
<td>5.5</td>
<td>6</td>
<td>141</td>
<td>2</td>
<td>27</td>
<td>79 (93%)</td>
<td>41 (93%)</td>
</tr>
<tr>
<td>NETM_ANN (D)</td>
<td>10,000</td>
<td>15</td>
<td>143</td>
<td>0</td>
<td>18</td>
<td>80 (94%)</td>
<td>41 (93%)</td>
</tr>
</tbody>
</table>
clearly show that the superior performance, and thus the selection of the best predictor depend on the practical purpose. For example, NETM_ANN has been judged as the best overall A*0201 predictor (Figure 1 and Figure 4). This server also shows the best performance for thresholds that optimize the selection of T-cell epitopes (Table 1) and the threshold which does not allow false positive (Table 3), but it comes as distant second at the threshold that predicts the vast majority of binders (Table 2). The distinct best predictor for high sensitivity threshold (Table 2) is NHP_CP whose overall performance has been assessed as modest. Overall, considering the balance between false positive and false negative and prediction of T-cell epitopes, NETM_ANN is likely to produce the best result in most cases. The selected thresholds represent the extreme scenarios (high sensitivity, or high specificity predictions). In practical applications, the thresholds will be between these extreme values and costs in terms of false positives and false negatives can be assessed. The higher the sensitivity of prediction, the larger the number of false positives. Conversely, the higher the specificity the lower the number of true positives.

Further analysis of results (Figure 6) revealed four main groups of predictors. Group A (BIMAS, MHC_BP, and NHP_ANN) have the majority of predictions clustered at the top of the graph with the nearly horizontal trend line. Although these predictors may provide good prediction of accuracy with carefully selected threshold, however this threshold is difficult to determine. The predictions are of low sensitivity, but relatively high specificity because of a small numbers of TP and FP, and large numbers of TN and FN. Group B (IEDB_SMM, IEDB_ARB, MHCP_I and MHCP_AA) have majority of predictions clustered along the bottom of the graph with the nearly horizontal trend line. Again, these predictions may show good classification accuracy but it is difficult to identify the appropriate threshold. The predictions are typically of high sensitivity and low specificity because of the large number of TP and FP, and small numbers of TN and FN. Group C numbers of (MHC_BPS, MULTI_HMM, NHP_CP, PEPDIST, PREDEP, SVMHC_M, and SVMHC_S) have predictions clustered horizontally or as a cloud with the nearly horizontal trend line. These predictors show moderate accuracy of predictions irrespective of the selected threshold. Finally, the remaining predictors form group D which show the distribution of predictions across the diagonal with a trend line showing slope from non binders to high binders. The accuracy of these predictors is moderate to high with a reasonable balance of TP, TN, FP, and FN. However, these results need to be taken with a note of caution, because some of the T-cell epitopes used for the comparison are likely to be included in the training sets for server development. Nevertheless, it is clear that the servers that are better for prediction of binding affinity are also better in predicting of T-cell epitopes.

In summary, our results have shown that the best predictors of classification also show the best performance in prediction of HLA binding affinity, and prediction of T-cell epitopes, which supports the contention that T-cell epitopes are more likely to be drawn from the highest binding affinity peptides [30,36] and for which quantitative theoretical support has been provided recently [37].

### Table 2: Prediction performance of selected representative servers in order to correctly predict the majority (95%) of binders

<table>
<thead>
<tr>
<th>Server</th>
<th>Thr2</th>
<th>TP (binding)</th>
<th>TN (binding)</th>
<th>FP (binding)</th>
<th>FN (binding)</th>
<th>TP (tumor epitopes)</th>
<th>TP (viral epitopes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIMAS (A)</td>
<td>0.003</td>
<td>31</td>
<td>105</td>
<td>38</td>
<td>2</td>
<td>84 (99%)</td>
<td>44 (100%)</td>
</tr>
<tr>
<td>MHCP_I (A)</td>
<td>1,000</td>
<td>31</td>
<td>7</td>
<td>136</td>
<td>2</td>
<td>80 (94%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>IEDB_SMM (B)</td>
<td>79,000</td>
<td>31</td>
<td>109</td>
<td>34</td>
<td>2</td>
<td>85 (100%)</td>
<td>44 (100%)</td>
</tr>
<tr>
<td>NHP_CP (C)</td>
<td>0</td>
<td>31</td>
<td>126</td>
<td>17</td>
<td>2</td>
<td>79 (93%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>IEDB_ANN (D)</td>
<td>35,000</td>
<td>31</td>
<td>75</td>
<td>68</td>
<td>2</td>
<td>85(100%)</td>
<td>42 (95%)</td>
</tr>
<tr>
<td>MULTI_SVM (D)</td>
<td>3.9</td>
<td>31</td>
<td>101</td>
<td>42</td>
<td>2</td>
<td>84 (99%)</td>
<td>44 (100%)</td>
</tr>
<tr>
<td>NETM_ANN (D)</td>
<td>40,000</td>
<td>31</td>
<td>113</td>
<td>30</td>
<td>2</td>
<td>85 (100%)</td>
<td>44 (100%)</td>
</tr>
</tbody>
</table>

### Table 3: Prediction performance of selected representative servers in order to exclude all false positives

<table>
<thead>
<tr>
<th>Server</th>
<th>Thr3</th>
<th>TP (binding)</th>
<th>TN (binding)</th>
<th>FP (binding)</th>
<th>FN (binding)</th>
<th>TP (tumor epitopes)</th>
<th>TP (viral epitopes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIMAS (A)</td>
<td>2</td>
<td>10</td>
<td>143</td>
<td>0</td>
<td>23</td>
<td>76 (89%)</td>
<td>39 (89%)</td>
</tr>
<tr>
<td>MHCP_I (A)</td>
<td>10</td>
<td>0</td>
<td>143</td>
<td>0</td>
<td>33</td>
<td>2 (2%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>IEDB_SMM (B)</td>
<td>1,000</td>
<td>10</td>
<td>143</td>
<td>0</td>
<td>23</td>
<td>77 (91%)</td>
<td>38 (86%)</td>
</tr>
<tr>
<td>NHP_CP (C)</td>
<td>0.5</td>
<td>6</td>
<td>143</td>
<td>0</td>
<td>27</td>
<td>76 (89%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>IEDB_ANN (D)</td>
<td>10,000</td>
<td>9</td>
<td>143</td>
<td>0</td>
<td>24</td>
<td>76 (89%)</td>
<td>39 (89%)</td>
</tr>
<tr>
<td>MULTI_SVM (D)</td>
<td>5.8</td>
<td>4</td>
<td>143</td>
<td>0</td>
<td>29</td>
<td>75 (88%)</td>
<td>40 (91%)</td>
</tr>
<tr>
<td>NETM_ANN (D)</td>
<td>10,000</td>
<td>15</td>
<td>143</td>
<td>0</td>
<td>18</td>
<td>80 (94%)</td>
<td>41 (93%)</td>
</tr>
</tbody>
</table>
Representative graphs for A*0201 binding predictions on T-cell epitopes and the test peptide. The thresholds marked by broken lines predict approximately 90% of T-cell epitopes and are used for the assessment of false positives and false negatives in binding predictions. Representative examples of predictor groups are shown. The x-axis in the left figure represents experimental scores of test peptides while y-axis represented their scaled predicted scores. The x-axis in the right figure indicates index of sorted list T-cell epitopes while the y-axis represented their scaled predicted binding scores.
Conclusions and Discussion
This study shows that major advances have recently been achieved in the field of computational immunology and immunoinformatics. These are mainly the results of the collaborative initiatives that focus on the development of computational infrastructure for immunology, such as IEDB or ImmunoGrid. The availability of large high-quality datasets of HLA ligands and T-cell epitopes and advanced algorithms enabled the development of advanced in silico tools that complement experimental research and enable screening collections of pathogen proteomes and large collections of antigens.

We have learnt important lessons about the algorithms that are used to model HLA-peptide interactions. Non-linear algorithms, in particular ANNs appear to offer advantage for prediction of peptide binding affinity. Recently developed algorithms are generally work to be done, since in silico assays that match contemporary experimental accuracy are available only for single HLA*0201 9-mer peptides. We have also identified the problems with some prediction methods (Figure 6): group A predictors suffer from low sensitivity and can be improved by re-training their prediction engines with new data, particularly binders; group B suffers from low specificity and these models can be improved by retraining with larger number of non-binders; group C can be further improved by retraining with larger number of training data; while group D can be improved by further improvement of algorithms, while addition of new data is likely to offer only a small gradual improvement for this group. The combination of predictions from high-accuracy predictors is likely to be a major direction for improvement of predictions other than for A*0201 [29]. A large number of predictors, in particular those from groups A, B and C can be improved by post-processing of raw prediction data, principally non-linear transformation.

Our results also suggest that normalization of outputs by scaling onto a common scale (in this study we used the scale of 0–100) would benefit the field by providing a standard in silico scale, which would, in turn, enable mapping of various experimental methods to a common base and fair comparison of the results. In this schema, the negative control peptide maps to 0, while the positive control peptide maps to 100. Binders of higher affinity than the positive control will have binding score greater than 100. The interpretation of the normalized scores is clearer than the raw scores for examples shown in Table 1, 2, 3. Appropriate scaling of outputs also provides practical benefits: a number of predictors that theoretically have good or excellent predictive performance when analyzed in fine detail. However, for those that belong to predictor groups A, B, or C (Figure 6) it is difficult to determine the best threshold for classification predictions because the threshold between "good" and "poor" predictions is narrow, rather than wide as in group D predictors. This makes predictors in groups A, B, and C inferior to those in the group D because chances for making poor predictions due to the sub-optimal, or even poor, selection of prediction thresholds by users are high.

The fields of computational immunology and immunoinformatics [25,38] are growing rapidly. Combining experimental and in silico methods is essential to address combinatorial problems associated with deciphering immune responses and the applications such as design of vaccines and immunotherapies. While identification of HLA ligands and T-cell epitopes is only a step in the whole process of translation of basic immunology research into clinical applications, it is a prime showcase of significant advances that can be achieved by intelligently combining wet-lab experimentation with mathematical modeling and computation.

Methods
We identified 30 servers developed by 19 groups that can predict HLA-I binding peptides and are accessible through the Internet (Table 4). The study included several consecutive steps: a) Independent experimentally measured test data sets were identified; b) predictions of peptide binding were made using up to 30 servers (as available for each of the seven HLA-I molecules); c) the predictions of individual servers were compared whether they are identical and "duplicate servers" were removed from further analysis; d) predictions were normalized to the common scale to facilitate comparison of predictive performances; e) classification accuracy (binders vs. non-binders) was estimated; f) the accuracy of predicted binding affinities was assessed; g) non-linear transformations of prediction scores were performed for the improvement of predictions. Predictive algorithms used in these studies include: binding matrices [39-50], artificial neural networks – ANN [45,51-54], hidden Markov models – HMM [52], support vector machines [55-58], structure-based model [59,60], partial least square function [61], and peptide-peptide distance function [62].

Data sets
In this study we used data sets produced by the iTopia™ Epitope Discovery System. The two data sets included the full overlapping study of 134 9-mer peptides spanning the full length of the tumor antigen survivin (Swiss-Prot: O15392) [63] and the 42 peptides spanning a 50 amino acids long construct containing cytomegalovirus (CMV) internal matrix protein pp65 peptides [64]. These studies produced binding data for eight HLA-I molecules (HLA-A*0101, -A*0201, -A*0301, -A*1101, -A*2402, -B*0702, -B*0801, and -B*1501). Only two
binders within 176 peptides were identified as -A*0101 binders; this molecule was excluded from further study because of insufficient quantity of test data. For binding/non-binding classification we considered as positives those peptides whose binding affinity was \( \geq 30\% \) of the binding affinity of the positive control, as suggested in the iTopia™ technical information. HLA-A*0201 restricted T-cell epitopes have been extracted from the literature and contain 85 well-characterized tumor antigen-related peptides and 44 well-characterized viral T-cell epitopes (see supplemental materials in Additional file 1). Several predictors do not have information on specific genotype alleles but have predictions for serotypes. For instance, the prediction results generated by SMM are actually the binding affinities of peptides to HLA-A2, not exclusively HLA-A*0201. Such approximation may affect their specificity in predicting HLA-A*0201 epitopes to some extent. The data sets used in this study were also deposited in the Dana-Farber Repository for Machine Learning in Immunology [65].

Predictions and comparisons

The two protein sequences were submitted to the prediction servers and the prediction results were recorded. For each HLA molecule two prediction applications were analyzed: classification into binders and non-binders and prediction of peptide binding affinity. For the assessment of classification accuracy we used the analysis of the area under the ROC curve (AROC) [66].

This curve is a plot of the true positive rate TP/(TP+FN) on the vertical axis vs. false positive rate FP/(TN+FP) on the horizontal axis for the full range of the decision thresholds. The values AROC\( \geq 0.9 \) indicate excellent, 0.9>AROC\( \geq 0.8 \) good, 0.8>AROC\( \geq 0.7 \) marginal and 0.7>AROC poor predictions [66]. We also used the sensitivity/specificity plot measure by determining the intersection point of sensitivity and specificity curves for the complete range of thresholds. To assess the accuracy of binding affinity predictions we calculated the Pearson correlation coefficient for experimental measurements \( X \) and a prediction series \( Y \) for the studied set of peptides:

\[
 r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

\( x_i \) and \( \bar{x} \) are experimental individual and average affinities.

### Table 4: List of prediction servers of HLA-I binding peptides, their URLs (as of April 2007), and name abbreviations

<table>
<thead>
<tr>
<th>ID</th>
<th>Servers</th>
<th>Abbreviation</th>
<th>URLs</th>
<th>Prediction algorithm</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BIMAS</td>
<td>BIMAS</td>
<td>[68]</td>
<td>Matrix</td>
<td>[39]</td>
</tr>
<tr>
<td>2</td>
<td>HLA Ligand</td>
<td>HLA_Li</td>
<td>[69]</td>
<td>Matrix</td>
<td>[47]</td>
</tr>
<tr>
<td>3</td>
<td>IEDB (ANN)</td>
<td>IEDB_ANN</td>
<td>[70]</td>
<td>ANN</td>
<td>[54]</td>
</tr>
<tr>
<td>4</td>
<td>IEDB (ARB)</td>
<td>IEDB_ARB</td>
<td>[71]</td>
<td>Matrix</td>
<td>[48]</td>
</tr>
<tr>
<td>5</td>
<td>IEDB (SMM)</td>
<td>IEDB_SMM</td>
<td>[72]</td>
<td>Matrix</td>
<td>[49]</td>
</tr>
<tr>
<td>6</td>
<td>MAPPP (Bimas)</td>
<td>MAPPP_B</td>
<td>[73]</td>
<td>Matrix</td>
<td>[43]</td>
</tr>
<tr>
<td>7</td>
<td>MAPPP (SYFPEITHI)</td>
<td>MAPPP_S</td>
<td>[74]</td>
<td>Matrix</td>
<td>[43]</td>
</tr>
<tr>
<td>8</td>
<td>MHC Binder Prediction</td>
<td>MHC_BP</td>
<td>[75]</td>
<td>Matrix</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>MHC-BPS</td>
<td>MHC_BPS</td>
<td>[76]</td>
<td>SVM</td>
<td>[55]</td>
</tr>
<tr>
<td>10</td>
<td>MHC-I (Multiple matrix)</td>
<td>MHC1_MM</td>
<td>[77]</td>
<td>Structure-based model</td>
<td>[60]</td>
</tr>
<tr>
<td>11</td>
<td>MHC-I (Single matrix)</td>
<td>MHC1_SM</td>
<td>[78]</td>
<td>Structure-based model</td>
<td>[60]</td>
</tr>
<tr>
<td>12</td>
<td>MHCPred (Interactions)</td>
<td>MHCP_I</td>
<td>[79]</td>
<td>Partial least square</td>
<td>[61]</td>
</tr>
<tr>
<td>13</td>
<td>MHCPred (Amino Acids)</td>
<td>MHCP_AA</td>
<td>[80]</td>
<td>Partial least square</td>
<td>[61]</td>
</tr>
<tr>
<td>14</td>
<td>MULTIPRED (ANN)</td>
<td>MULTI_ANN</td>
<td>[81]</td>
<td>ANN</td>
<td>[52]</td>
</tr>
<tr>
<td>15</td>
<td>MULTIPRED (HMM)</td>
<td>MULTI_HMM</td>
<td>[82]</td>
<td>HMM</td>
<td>[52]</td>
</tr>
<tr>
<td>16</td>
<td>MULTIPRED (SVM)</td>
<td>MULTI_SVM</td>
<td>[83]</td>
<td>SVM</td>
<td>[53]</td>
</tr>
<tr>
<td>17</td>
<td>NetMHC (ANN)</td>
<td>NETM_ANN</td>
<td>[84]</td>
<td>ANN</td>
<td>[51]</td>
</tr>
<tr>
<td>18</td>
<td>NetMHC (Weight Matrix)</td>
<td>NETM_WM</td>
<td>[85]</td>
<td>Matrix</td>
<td>[44]</td>
</tr>
<tr>
<td>19</td>
<td>nHLAPred (ANNPred)</td>
<td>NHP_ANN</td>
<td>[86]</td>
<td>ANN</td>
<td>[53]</td>
</tr>
<tr>
<td>20</td>
<td>PepDist</td>
<td>PEPPDIST</td>
<td>[88]</td>
<td>function distance</td>
<td>[57]</td>
</tr>
<tr>
<td>21</td>
<td>PeptideCheck (Matrix)</td>
<td>PEPC_M</td>
<td>[89]</td>
<td>Matrix</td>
<td>[41]</td>
</tr>
<tr>
<td>22</td>
<td>Predp</td>
<td>PREDEP</td>
<td>[90]</td>
<td>Structure-based model</td>
<td>[57]</td>
</tr>
<tr>
<td>23</td>
<td>ProPredI</td>
<td>PROPRED</td>
<td>[91]</td>
<td>Matrix</td>
<td>[42]</td>
</tr>
<tr>
<td>24</td>
<td>Rank peppe</td>
<td>RANKPEP</td>
<td>[92]</td>
<td>Matrix</td>
<td>[41]</td>
</tr>
<tr>
<td>25</td>
<td>SMM</td>
<td>SMM</td>
<td>[93]</td>
<td>Matrix</td>
<td>[50]</td>
</tr>
<tr>
<td>26</td>
<td>SVMHC (MHCPep)</td>
<td>SVMHC_M</td>
<td>[94]</td>
<td>SVM</td>
<td>[56]</td>
</tr>
<tr>
<td>27</td>
<td>SVMHC (SYFPEITHI)</td>
<td>SVMHC_S</td>
<td>[95]</td>
<td>SVM</td>
<td>[56]</td>
</tr>
<tr>
<td>28</td>
<td>SVRMHC</td>
<td>SVRMHC</td>
<td>[96]</td>
<td>SVM</td>
<td>[57]</td>
</tr>
<tr>
<td>29</td>
<td>SYFPEITHI</td>
<td>SYFPEITHI</td>
<td>[97]</td>
<td>Matrix</td>
<td>[40]</td>
</tr>
<tr>
<td>30</td>
<td>SVMHC (SYFPEITHI)</td>
<td>SVMHC_S</td>
<td>[95]</td>
<td>SVM</td>
<td>[56]</td>
</tr>
</tbody>
</table>

This curve is a plot of the true positive rate TP/(TP+FN) on the vertical axis vs. false positive rate FP/(TN+FP) on the horizontal axis for the full range of the decision thresholds. The values AROC\( \geq 0.9 \) indicate excellent, 0.9>AROC\( \geq 0.8 \) good, 0.8>AROC\( \geq 0.7 \) marginal and 0.7>AROC poor predictions [66]. We also used the sensitivity/specificity plot measure by determining the intersection point of sensitivity and specificity curves for the complete range of thresholds. To assess the accuracy of binding affinity predictions we calculated the Pearson correlation coefficient for experimental measurements \( X \) and a prediction series \( Y \) for the studied set of peptides:
and we should interpret these results only as a guide.

Likely included in the training sets for individual servers since some of these peptides are well-known, they are predictions based on the analysis of 176 iTopia™ peptides. We estimated thresholds that identify ~90% of T-cell epitopes as positive predictions (TP) and estimated a number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) at that threshold using predictions based on the analysis of 176 iTopia™ peptides. Since some of these peptides are well-known, they are likely included in the training sets for individual servers and we should interpret these results only as a guide.

Scaling and transformations

To enable visual inspection of prediction comparisons, both experimental measurements and predictions were scaled to a common scale from 0 to 100 using linear transformation of the value ranges using the formula for each value for individual peptide:

\[ y_i^S = \frac{y_i - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \times 100 \]

where \( y_i^S \) is the scaled value, \( y_{\text{min}} \) is the minimum and \( y_{\text{max}} \) is the maximum value.

Furthermore we performed non-linear transformations of the raw predicted values from individual servers to assess whether the scaling and normalization issues affect the accuracy of predictions. In statistics, the “power transform”, also known as "Box-Cox transform" is used to map data to from one space to another for data stabilization procedures such as reduction of data variation, improvement of the correlation between variables, and improving data distribution [67]. We selected four common non-linear transformations and performed them for each predictor (natural logarithm – L, exponential – E, square – S, and square root – R functions):

\[ y_i^{Sn} = \ln(y_i - y_{\text{min}} + \delta) \]
\[ y_i^{Sn} = e^{y_i^{Sn}} \]
\[ y_i^{Sn} = y_i^2 \]
\[ y_i^{Sn} = \sqrt{y_i - y_{\text{min}}} \]

where \( y_i^{Sn} \) is the prediction score for scaled and non-linearly transformed value of raw prediction.

The scaled and transformed predictions were assessed to reveal the predictors have been optimized and those that can be improved by post-processing of prediction values.

**Authors’ contributions**

HHL carried out the study and drafted the manuscript. SR and ELR participated in the design of the study and critically reviewed the manuscript. ST collected and annotated T-cell epitopes, and prepared the manuscript. VB conceived, designed and coordinated the project, and revised the manuscript. All authors read and approved the final version of the manuscript.

**Additional material**

Additional file 1

Data sets used in this study. List of peptide binding affinity and well-characterized T-cell epitopes used in this study.

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2172-9-8-S1.pdf]

**Acknowledgements**

The work was supported by the ImmunoGrid project, under EC contract FP6-2004-IST-4, No. 028069, and NIH grant U19 AI57330.

**References**

12. Wobser M, Keikavoussi P, Kunzmann V, Weininger M, Andersen MH, Becker JC: Complete remission of liver metastasis of pancreatic...
atic cancer under vaccination with a HLA-A2 restricted peptide derived from the universal tumor antigen survivin. 


58. Zhang GL, Bozic I, Kwoh CK, August JT, Brusic V: DFRMLI site


68. BIMAS [http://www-bimas.cit.nih.gov/molbio/hla_bind/]

69. HLA Ligand [http://hlaligand.ohsu.edu/prediction.htm]

70. IEDB (ANN) [http://tools.immuneepitope.org/analyze/html/mhc_binding.htm]

71. MAPPP (Bimas) [http://www.mpib-berlin.mpg.de/MAPPP/binding.html]

72. IEDB (SM) [http://tools.immuneepitope.org/analyze/html/mhc_binding.htm]

73. MAPPP (SYFPEITHI) [http://www.mpib-berlin.mpg.de/MAPPP/binding.html]

74. MHCPred (Interactions) [http://www.jenner.ac.uk/MHCPred/]

75. MHCPred (Amino Acids) [http://www.jenner.ac.uk/MHCPred/]

76. MULTIPRED (ANN) [http://antigen.i2r.a-star.edu.sg/multipred/]

77. MULTIPRED (HMM) [http://antigen.i2r.a-star.edu.sg/multipred/]

78. MULTIPRED (SYF) [http://antigen.i2r.a-star.edu.sg/multipred/]

79. NetMHC (ANN) [http://www.cbs.dtu.dk/services/NetMHC/]

80. NetMHC (Weight Matrix) [http://www.cbs.dtu.dk/services/NetMHC/

81. nHLAPred (ANNPred) [http://www.imtech.res.in/raghava/nhlapred/neural.html]

82. nHLAPred (ComPred) [http://www.imtech.res.in/raghava/nhlapred/como.html]

83. PepDist [http://www.peptidecheck.cs.huiji.ac.il/]

84. PeptideCheck [http://www.peptidecheck.org/]

85. Predep [http://margalit.huji.ac.il/Teppred/mhc-bind/index.html]

86. ProPred [http://www.imtech.res.in/raghava/propred/]

87. Rankpep [http://bio.dfci.harvard.edu/Tools/rankpep.html]

88. SVMHC (SYFPEITHI) [http://www.sbc.su.se/~pierre/svmhc/new.cgi]

89. SYFPEITHI [http://www.sfpeithi.de/Scripts/MHCIserver.dll/EpitopePrediction.htm]