Quantitative modeling of peptide binding to TAP using support vector machine

Carmen M. Diez-Rivero,1 Bernardo Chenlo,1 Pilar Zuluaga,2 and Pedro A. Reche1*

1 Laboratorio de InmunoMedicina, Departamento de Microbiologia I-Immunologia, Facultad de Medicina, Universidad Complutense, Madrid 28040, Spain
2 Departamento de Estadística e Investigación Operativa, Facultad de Medicina, Universidad Complutense, Madrid 28040, Spain

ABSTRACT

The transport of peptides to the endoplasmic reticulum by the transporter associated with antigen processing (TAP) is a necessary step towards determining CD8 T cell epitopes. In this work, we have studied the predictive performance of support vector machine models trained on single residue positions and residue combinations drawn from a large dataset consisting of 613 nonamer peptides of known affinity to TAP. Predictive performance of these TAP affinity models was evaluated under 10-fold cross-validation experiments and measured using Pearson's correlation coefficients ($R_p$). Our results show that every peptide position (P1–P9) contributes to TAP binding (minimum $R_p$ of 0.26 ± 0.11 was achieved by a model trained on the P6 residue), although the largest contributions to binding correspond to the C-terminal end ($R_p = 0.68 ± 0.06$) and the P1 ($R_p = 0.51 ± 0.09$) and P2 (0.57 ± 0.08) residues of the peptide. Training the models on additional peptide residues generally improved their predictive performance and a maximum correlation ($R_p = 0.89 ± 0.03$) was achieved by a model trained on the full-length sequences or a residue selection consisting of the first 5 N- and last 3 C-terminal residues of the peptides included in the training set. A system for predicting the binding affinity of peptides to TAP using the methods described here is readily available for free public use at http://imed.med.ucm.es/Tools/tapreg/.

Key words: antigen processing; peptide; TAP; prediction; WEKA; SVM.

INTRODUCTION

CD8 T cells play a key role in tumor immunosurveillance and clearing of intracellular infectious agents, and a subset of them known as cytotoxic T lymphocytes (CTLs) are capable of directly killing infected and tumor cells.1 CTLs discriminate between normal and damaged cells using their T cell receptor (TCR) to monitor the peptides presented by major histocompatibility class I (MHCI) molecules on the cell surface. T cells recognizing self-peptides are eliminated during the process of thymic selection, and, thereby, T cell immune responses are triggered by the recognition of MHCI molecules incorporating foreign or antigenic peptides (T cell epitopes).2 T cell epitopes result from the degradation of proteins through pathways that determine the repertoire of peptides that are available for binding to MHC and recognition by T cells. The dominant pathway for class I antigen processing is reviewed next.

MHCI molecules preferably bind peptides nine residues long that generally originate from endogenous proteins that are degraded in the cytosol of the cell by the proteolytic activity of the proteasome.3,4 Peptide fragments cleaved by proteasomes are shuttled to the lumen of the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), where they can bind to newly assembling MHCI molecules.5,6 Before MHCI binding, peptides can also undergo an optional N-terminal trimming by ER-associated amino peptidases (ERAAP).7 Finally, peptide–MHCI complexes are exported to the cell surface for presentation to the CD8 T cells.5,6 There is evidence supporting that these processing steps limit/shape the peptides that can be presented by MHCI molecules in vivo,7-9 thus explaining the numerous observations of high affinity MHCI binding peptides that are unable to elicit CTL responses.10,11 Nonetheless, peptide transport by TAP represents the single most selective step in T cell epitope processing.12 In addition, TAP is also important for presentation of epitopes derived from exogenous antigens.13

Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.
Carmen M. Diez-Rivero and Bernardo Chenlo contributed equally to this work.
Grant sponsor: Ministerio de Ciencia e Innovacion (MICINN) of Spain; Grant number: SAF2006-07879;
Grant sponsor: Universidad Complutense de Madrid (U.C.M); Grant number: CCG08-UCM/BIO-3769.
*Correspondence to: Pedro A. Reche, Laboratorio de InmunoMedicina, Departamento de Microbiologia I-Immunologia, Facultad de Medicina, Universidad Complutense, Madrid, Ave. Complutense s/n, Madrid 28040, Spain. E-mail: pareche@imed.ucm.es
Received 8 April 2009; Revised 2 July 2009; Accepted 7 July 2009
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22535

0.03)
MATERIAL AND METHODS

Peptide datasets

The main dataset used in this study to analyze the peptide selectivity of TAP consisted of 613 unique nonamer (9-mer) peptides of known binding affinity to human TAP relative to the reference peptide RRYNASTEL (IC50relative). The lower the IC50relative the stronger the peptide binds to TAP. This dataset encompasses 435 peptides, kindly provided by Dr. Peter van Endert (INSERM U580, Paris Descartes University, Paris, France)—IC50relative already referenced to RRYNASTEL—plus 178 peptides parsed from the TAP binding affinity peptide collection of the Antigen Database, kindly provided by Dr. Darren Flower (The Jenner Institute, Compton, UK). To combine the peptides into a single dataset, the TAP binding affinity (IC50) of peptides collected from the Antigen Database was also referenced to the peptide RRYNASTEL. For peptides obtained from the Antigen Database that were identical in sequence but had different TAP binding affinities, median values were considered before referencing. This dataset is provided as Supporting Information in Table 1S. We thank to Dr. Peter van Endert and Dr. Darren Flower for showing no inconvenience in that we provided Table 1S as Supporting Information.

Peptide datasets with reduced sequence similarity were generated from the 613-peptide dataset using the purge utility of the Gibbs Sampler with an exhaustive method and maximum blosum 62 relatedness scores of 25, 30, 35, and 37. The resulting datasets had 293, 332, 465, and 530 peptides and are provided as Supporting Information (Table 2S, Table 3S, Table 4S, and Table 5S, respectively).

To compare TAP affinity scores predicted by available methods, we used a set of 723 unique 9-mer CD8 T cell epitopes obtained from the IMMUNEEPITOPE and EPIMHC databases (provided as Supporting Information in Table 6S).

Model building and evaluation

Predictive models of TAP affinity were trained and evaluated under the EXPERIMETER application of the Waikato Environment for Knowledge Analysis (WEKA) package. WEKA provides a framework for data classification, clustering, and feature selection using a large collection of machine-learning algorithms. In this study, we have selected kernel-based SVMs. Specifically, we used a radial basis function (RBF) as the kernel in combination with Alex Smola and Bernhard Scholkopf’s sequential minimal optimization algorithm for training SVMs (SMOreg algorithm in WEKA). Model refinement was achieved by varying the C (0.2, 0.4, 0.8, 1, 2, 4, 8, 10) and gamma (0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5) values of the RBF kernel. Predictive models were generated from distinct training sets, consisting of different residue selections drawn from the peptide sequences of the training set and encoded using sparse and blosum representations. In the sparse encoding, each amino acid is coded by the relevant amino acid symbol, whereas in the blosum encoding, it is represented by 20 digits corresponding to the relevant amino acid.
acid substitution scores given by the BLOSUM62 substitution matrix.\textsuperscript{38} TAP affinity (IC\textsubscript{50relative}) values of the training sets were provided to WEKA as logIC\textsubscript{50relative} values. Pearson’s correlation coefficient (R\textsubscript{p}) was used to measure the performance of SVMs to fit the experimental data. Since SVM models were built and evaluated using 10-fold cross-validation experiments that were repeated 10 times, R\textsubscript{p} mean values and standard deviations were computed from 100 different values. Predicted peptide affinity scores yielded by the models generated with WEKA were transformed to IC\textsubscript{50} values by considering an IC\textsubscript{50} for the reference peptide RRINASTEL of 400 nM.

**Sequence similarity analyses**

Sequence similarity in peptide datasets was analyzed from pairwise sequence alignments between all peptides in the dataset. Sequence alignments were obtained using the Needleman-Wunsch global alignment algorithm implemented with the needle application that is included in the EMBOSS package.\textsuperscript{39} Alignments with peptide positions shifted were not evaluated (e.g., residues 1–4 of a peptide aligned with residues 3–7 of another peptide). Generally, for any given peptide (query) in the dataset, one could find several peptides that shared sequence similarity with it (hits), but the majority of the peptides in the dataset had no similarity with the query. In this study, we have computed average sequence similarities in the peptide datasets in two ways: globally, considering all possible pairwise comparisons between the peptide sequences but those with themselves (for a dataset with N peptides there will be N x (N-1) comparisons), and using only the hits.

For a given query peptide in the dataset, the relationship between sequence similarity and binding affinity was studied by correlating sequence similarity with hits and differences in binding affinity (logIC\textsubscript{50relative}) using Spearman’s rank correlation (R\textsubscript{s}). For instance, let us consider the peptide PLAKAAAAV (logIC\textsubscript{50relative} = 8.370) had the following hits:

- Hit: ALAKAAAAV; Identity: 88.9%; Similarity: 88.9%; logIC\textsubscript{50relative}: 3.984; Dif: 4.386
- Hit: ALAKAAAAAL; Identity: 77.8%; Similarity: 88.9%; logIC\textsubscript{50relative}: 0.688; Dif: 7.682
- Hit: AAAASAAAAAF; Identity: 66.7%; Similarity: 77.8%; logIC\textsubscript{50relative}: -0.734; Dif: 9.104
- Hit: ALAKAAAAAF; Identity: 55.6%; Similarity: 66.7%; logIC\textsubscript{50relative}: 0.332; Dif: 8.038
- Hit: GRQKGAGSV; Identity: 33.3%; Similarity: 44.4%; logIC\textsubscript{50relative}: 6.215; Dif: 2.155

Then, for peptide PLAKAAAAV, an R\textsubscript{s} value was computed by correlating the similarity/identity with its peptide hits (88.9, 77.8, 66.7, 55.6, 33.3) and the differences in logIC\textsubscript{50relative} values (4.386, 7.682, 9.104, 8.038, 2.155). R\textsubscript{s} values were thus computed for each peptide in the dataset. Peptides with less than five hits were discarded from this analysis. These peptide-specific R\textsubscript{s} values were determined considering all peptide hits and only those with an identity ≥ 50%.

**Statistical analyses**

To assess whether the correlation achieved by a given SVM model, i, during training was stronger than that of another SVM model, j, we used one-sided two-sample t-test to examine if the differences of the relevant R\textsubscript{p} mean values were significantly above 0 (H\textsubscript{0}: R\textsubscript{p}i − R\textsubscript{pj} = 0; P ≤ 0.05). To evaluate if R\textsubscript{p} values were statistically significant (H\textsubscript{0}: R\textsubscript{p} = 0), we computed the statistics given by Eq. (1), which follows a t-Student distribution with N – 2 degrees of freedom, and tested subsequently (P < 0.05).

\[
t = \frac{R_p}{\sqrt{1 - R_p^2} \frac{N - 2}{N - 1}}
\]  

To evaluate the correlation coefficients obtaining by comparing the TAP affinity scores predicted by different methods with each other or with experimental data, we applied the test for comparing overlapping correlation coefficients described by Meng et al.,\textsuperscript{40} as implemented in the R package compOverlapCorr by Ka-Lon Li (http://cran.us.r-project.org/web/packages/compOverlapCorr/index.html). Briefly, Fisher’s Z-transform is applied first (Eq. (2)).

\[
Z_i = \frac{1}{2} \ln \left( \frac{1 + R_i}{1 - R_i} \right)
\]  

Next, a statistics Z, which follows a normal distribution is computed using Eq. (3), and tested subsequently (P < 0.05).

\[
Z = (z_i - z_j) \sqrt{\frac{N - 3}{2(1 - R_{ij})h}}
\]

In Eq. (3), R\textsubscript{ij} is the correlation between the predicted values by the methods i and j being compared, and h = (1 − f R\textsuperscript{2})/(1 − R\textsuperscript{2}), with R\textsuperscript{2} = (R\textsuperscript{2}i + R\textsuperscript{2}j)/2 and f = (1 − R\textsuperscript{2}i)/(1 − R\textsuperscript{2}j).

**Web server implementation**

The TAPREG Web server for predicting the binding affinity of peptides to TAP was implemented on an Apache Web server under the Mac OSX operating system. The TAPREG core consists of a PERL CGI (Common Gateway Interface) script that executes the predictions on
user-provided input data and returns the results to the browser. In addition, the TAPREG web interface uses JavaScript for handling and verification of input data before submission.

RESULTS

Quantitative analysis of TAP selectivity using TAP affinity models

We have approached the study of TAP selectivity using a large dataset consisting of 613 9-mer peptides (DS613) of known affinity to TAP (logIC50relative) and SVMs under a regression schema. SVMs are among the most widely used methods for solving common data mining problems in bioinformatics and were chosen because of their solid theoretical foundations and proven generalization ability. A key feature of SVMs is the use of nonlinear functions (kernels) to map the input onto a higher dimensional space in which an optimal separation is achieved—in the regression task—using a linear regression conducted with an \(\epsilon\)-insensitive loss function for error minimization. In this study, we have selected RBF kernels (Material and Methods) because in preliminary training experiments they outperformed the alternative linear and polynomial kernels (data not shown). Moreover, we have chosen two peptide sequence representations, sparse and blosum (Material and Methods), as input for SVMs. The evolutionary relationships between amino acids are taken into consideration with blosum representations of peptide sequences, which may enhance the generalization power of the resulting models. Using WEKA as the framework for model building and parameter optimization (Material and Methods), we first evaluated the ability of SVM models to predict TAP affinity data when trained on individual peptide residues (P1–P9), judging from the relevant Pearson's correlation coefficient (\(R_p\)). No differences were observed for models generated on blosum or sparse encoded sequences. Interestingly, for each peptide residue position, it was possible to generate SVM models that fitted the data with \(R_p\) values [Fig. 1(A)] that are significant for a linear correlation (\(P < 0.05\), Material and Methods). The lowest correlation was obtained with a model trained on the P6 residue (\(R_p = 0.26 \pm 0.11\)), whereas the largest correlation corresponded to a model trained on the C-terminal end of the peptide (\(R_p = 0.68 \pm 0.06\)) followed by the models trained on the P2 (0.56 \(\pm\) 0.08) and the P1 (\(R_p = 0.51 \pm 0.09\)) residues of the peptide. Systematic pairwise comparisons between the predictive performance of the different position-specific TAP affinity models using one-side \(t\)-tests over the relevant \(R_p\) means (Material and Methods) showed the following peptide residue position relevance to TAP binding: (P6 = P5) \(< (P8 = P7) \leq (P3 = P4) \leq P1 \leq P2 \leq P9\) (C-terminal end).

To evaluate the contribution of several peptide residues to TAP binding and to improve the correlation results, SVMs were trained on peptide fragments consisting of residue combinations drawn from the peptides of the training set. A total of 20 SVM models were generated.
and named after the specific peptide residue selection used for training (model $iNjC$ was generated from a fragment of $i + j$ residues, consisting of the first $i$ N-terminal and last $j$ C-terminal residues of the peptides of the training set). $R_p$ values achieved by these models on the training set together with those achieved by the models trained on just the C-terminus and the full-length peptide sequences (9-mers) are shown in Figure 1(B). Few or no differences were observed between SVMs trained using different sequence representations: sparse [gray bars in Fig. 1(B)] and blosum [black bars in Fig. 1(B)]. However, when differences were found, correlations obtained with the models trained on sparse encoded sequences were always larger than their blosum counterparts and were significantly stronger ($P < 0.05$) for models 3N2C, 4N1C, 4N2C, 5N2C, 4N3C, 4N4C, 3N5C, 5N3C, and ALL (trained on the full-length sequences). Several other general features emerged upon a detailed analysis of these results. Increasing the number of selected residues in the training sets (drawn from the peptides of known affinity to TAP) significantly improved the correlations achieved by the models [Fig. 2(A)], which went from an $R_p$ value of 0.68 ± 0.06 for a model trained on just the C-terminal end of the peptides of the training set to an $R_p$ of 0.89 ± 0.03 for the model trained on the full-length sequences (non-amer). Interestingly, a model trained on just eight residues (5N3C) achieved the same or better correlation (for blosum encoding) than models trained on the full-length peptide sequences [Figs. 1(B) and 2]. Nevertheless, for each fragment size, the best correlations were obtained with models trained on fragments encompassing more N-terminal than C-terminal peptide residue selections (2N1C, 3N1C, 4N2C, 4N3C, and 5N3C) [Fig. 2(A)], and these correlations were significantly stronger ($P < 0.05$) than those obtained with models with reversed N-terminal and C-terminal residue selections (1N2C, 1N3C, 2N4C, 3N4C, and 3N5C) [Fig. 2(B)]. This observation supports a larger contribution of the N-terminal half of the peptide to TAP binding when compared with its C-terminal half.

Sequence similarity in peptide datasets and predictive performance of SVM models

To explore the predictive performance of SVM models in relation to the sequence similarity between testing and training sets, we generated four peptide datasets of 293, 332, 465, and 530 peptides (DS293, DS332, DS465, DS530 respectively) by discarding similar sequences from the original DS613 dataset (Material and Methods). The global sequence identity in percentage in these datasets varied from 1 to 6% in the DS293 dataset to 9 ± 23% in the DS530 dataset, whereas in the DS613 dataset it was 10 ± 25% (Table I). In the 435-peptide dataset provided by Peter van Endert (PVE435) the global identity is 5 ± 16%. The overall low sequence similarity in the datasets reflects that the peptides do not belong to a single class or group related by a given property. On the contrary, each peptide is linked to a different numeric value (logI$_{C50}$relative). The average number of similarity hits per peptide in the datasets varied from nine peptides in the DS293 dataset to 110 hits in the DS613 dataset (Table I). Sequence identity between hits was considerably larger.
and ranged from 23% in the DS293 dataset to 59% in the DS613 dataset (Table I).

Because we train and evaluate the predictive performance of SVMs using 10-fold cross-validation experiments, and we repeat these experiments 10 times, we can assume that sequence similarity between testing and training sets to be comparable to that in the entire datasets. The correlation between predictions and experimental logIC50relative values achieved by SVMs trained and evaluated on the datasets of reduced sequence similarity (DS293, DS332, DS465, DS530, and PVE435) was significantly lower (P < 0.05; one-sided t-tests) than that obtained in the DS613 dataset (Table I). The smallest Rp was achieved in the DS293 dataset (0.71 ± 0.1), and these values increased significantly (P < 0.05) as the number of peptides in the datasets (Table I). Thus, DS613 Rp > DS530 Rp > DS465 Rp > PVE435 Rp > DS332 Rp > DS293 Rp.

These results may apparently suggest that prediction rates by our SVM models became inflated as sequence similarity in the datasets increased. However, this is an unlikely scenario because Rp values were computed in cross-validation, and the differences in Rp that we observed were statistically significant. For sequence similarity to be responsible for inflating prediction rates, the larger the sequence similarity between peptides in the datasets the closer their binding affinity. As a result, for any given peptide in the dataset one would expect to find a negative correlation between the similarity to its peptide hits and the differences in binding affinity (Material and Methods for details). However, we have not found such a negative correlation for the vast majority of the peptides in any of the datasets, as shown in the boxplot depicted in Figure 3. On the contrary, we have found these correlations to be shifted toward positives values; correlation medians in the DS293, DS332, DS465, DS613, and PVE435 datasets were 0.083, 0.109, 0.102, 0.139, 0.1945, and 0.114, respectively. Notably, the median of the correlation values in the DS613 dataset is significantly larger than those of the remaining datasets (P < 0.05), as judged from Wilcoxon-Mann-Whitney tests. Virtually identical results were obtained when only hits with ≥ 50% identity were considered (data not shown).

These results indicate that sequence similarity between peptides in the datasets does not correlate with proximity in binding affinity—in fact the opposite would appear to be the case. Therefore, the prediction rates obtained with SVMs trained on DS613 dataset are not inflated due to sequence similarity redundancy. Furthermore, similar sequences in the DS613 dataset are not redundant and contribute to the appropriated modeling of TAP binding affinity by SVMs; hence, the enhanced prediction rates achieved by models trained on the DS613 dataset.

### Table I
Predictive Performance of SVMs Trained on Datasets with Different Sequence Similarity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rp</th>
<th>Identity (%)a</th>
<th>Similarity (%)a</th>
<th>Identity (%)b</th>
<th>Similarity (%)b</th>
<th>Hitsb</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS293</td>
<td>0.71 ± 0.1</td>
<td>1 ± 6</td>
<td>2 ± 11</td>
<td>23 ± 11</td>
<td>43 ± 11</td>
<td>9 ± 7</td>
</tr>
<tr>
<td>DS332</td>
<td>0.76 ± 0.09</td>
<td>2 ± 8</td>
<td>3 ± 11</td>
<td>26 ± 18</td>
<td>46 ± 14</td>
<td>11 ± 12</td>
</tr>
<tr>
<td>DS465</td>
<td>0.65 ± 0.05</td>
<td>7 ± 19</td>
<td>8 ± 21</td>
<td>52 ± 25</td>
<td>60 ± 19</td>
<td>59 ± 45</td>
</tr>
<tr>
<td>DS530</td>
<td>0.87 ± 0.03</td>
<td>9 ± 23</td>
<td>10 ± 25</td>
<td>57 ± 24</td>
<td>62 ± 26</td>
<td>86 ± 62</td>
</tr>
<tr>
<td>DS613</td>
<td>0.89 ± 0.03</td>
<td>10 ± 25</td>
<td>11 ± 26</td>
<td>59 ± 23</td>
<td>66 ± 18</td>
<td>110 ± 77</td>
</tr>
<tr>
<td>PVE435</td>
<td>0.83 ± 0.05</td>
<td>5 ± 16</td>
<td>6 ± 18</td>
<td>45 ± 26</td>
<td>56 ± 19</td>
<td>40 ± 33</td>
</tr>
</tbody>
</table>

aIdentity and similarity computed considering all possible pairwise comparisons between the peptides in the datasets.
bIdentity and similarity computed considering only hits (Material and Methods).
cAverage number of similarity hits per peptide in the dataset.

These results indicate that sequence similarity between peptides in the datasets does not correlate with proximity in binding affinity—in fact the opposite would appear to be the case. Therefore, the prediction rates obtained with SVMs trained on DS613 dataset are not inflated due to sequence similarity redundancy. Furthermore, similar sequences in the DS613 dataset are not redundant and contribute to the appropriated modeling of TAP binding affinity by SVMs; hence, the enhanced prediction rates achieved by models trained on the DS613 dataset.

### Figure 3
Relationship between sequence similarity in peptide datasets and binding affinity proximity. This figure depicts a boxplot of Rp values computed for each peptide in a dataset by correlating its identity with its hits and the difference in logIC50relative values (Material and Methods). Boxplot were generated for peptides in DS293, DS332, DS465, DS530, DS613, and PVE435 datasets. Median Rp values in peptide datasets are indicated with a cross. A negative Rp will indicate that the larger the sequence similarity between peptides the closer their binding affinity. Conversely, a positive correlation will reflect that the larger the sequence similarity between peptides the larger the difference in their binding affinity.
methods of peptide binding affinity to TAP, which are readily available from the relevant publications (those by Peters et al.28 and Doytchinova et al.29) or from dedicated Web services (TAPPRED26 and SVMTAP27). The method developed by Doytchinova et al.29 consists of a matrix generated from 163 poly-Alanine 9-mer peptides of known affinity to TAP using an additive method30; hence, we will refer to this method as ADM. The ADM method achieved a reported $R_p$ between 0.72 and 0.83, depending on the testing set.29 The remaining methods have been trained on the PVE435 dataset.28 Briefly, Peters’ et al.28 method is based on a consensus matrix (CM) that was obtained from three scoring matrices, which included a poly-Alanine derived matrix and a SMM-matrix (generated using the Stabilized Matrix Method) trained on the PVE435 dataset. The CM method achieved a reported $R_p$ of 0.782 on the PVE435 dataset. The TAPPRED26 and SVMTAP27 methods are based on SVMs trained solely on the PVE435 dataset and achieved reported $R_p$ of 0.82 and 0.88, respectively. The TAPPRED method is based on two layers of SVMs, whereas SVMTAP consists of a single SVM model, similar to those trained in this study. We have evaluated all these methods in a testing set consisting of the 178 peptides of known affinity to TAP collected in this study (DS178), using Spearman’s correlation coefficients ($R_p$) (Table II). Interestingly, the lowest $R_p$ values were achieved by TAPPRED and SVMTAP (0.67 and 0.61), the methods with the largest reported correlations. On the other hand, CM achieved an $R_p$ (0.87) comparable to the value achieved by our TAP613 model in cross-validation (0.89), and AMD achieved an intermediate $R_p$ value of 0.74. Statistical comparison of these $R_p$ values (Material and Methods) indicated that the correlations obtained with the CM and TAP613 methods were significantly stronger than those obtained with the remaining methods. However, TAP613 was also trained on the DS178 testing set used for the comparisons, as surely were both the CM and ADM methods (DS178 contains binding affinity data of poly-Alanine peptides).

To further compare these methods, we have used a reference set of 723 MHCI-restricted T cell epitopes and correlated the scores predicted by the different methods (Table III). Interestingly, TAP613 predictions were significantly closer to the predictions by CM ($R_p = 0.86$), a matrix-based method, than to those by TAPPRED (0.29) and SVMTAP (0.76), which are based on SVM. Likewise, ADM predictions also correlated better with TAP613 predictions (0.59) than with those by TAPPRED (0.17) and SVMTAP (0.51). The extreme disparity of TAPPRED predictions with regard to the remaining methods was already noted by Zhang et al. Overall, these results support the view that existing SVM-based methods (TAPPRED and SVM) have suffered to some extent from data over-fitting, particularly TAPPRED, while we do not expect such a problem with our TAP613 model, as it was trained on a much larger dataset.

### Table II

<table>
<thead>
<tr>
<th>Method</th>
<th>$R_p$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP613</td>
<td>0.89 ± 0.03</td>
<td>This study</td>
</tr>
<tr>
<td>SMM</td>
<td>0.87 (0.82)</td>
<td>28</td>
</tr>
<tr>
<td>ADM</td>
<td>0.74 (0.72–0.83)</td>
<td>29</td>
</tr>
<tr>
<td>TAPPRED</td>
<td>0.67 (0.88)</td>
<td>26</td>
</tr>
<tr>
<td>SVMTAP</td>
<td>0.61 (0.82)</td>
<td>27</td>
</tr>
</tbody>
</table>

$R_p$ were computed using a testing set of 178 peptides of known affinity to TAP. For the TAP613 model, $R_p$ shown in the table is that achieved in cross-validation. Correlations reported in the literature for the different methods are shown in parentheses.

### The TAPREG server

We have implemented a Web tool, TAPREG, for predicting the binding affinity of peptides to TAP, which is available for free public use at http://imed.med.ucm.es/Tools/tapreg/ [Fig. 4(A)]. There are two models available at the TAPREG site that were trained both on the DS613 dataset using the entire peptide sequences; one was generated from a sparse representation of peptide sequences and the other from a blosum representation. The model trained on blosum-encoded sequences displayed a somewhat lower predictive performance ($R_p = 0.87 ± 0.03$) than the sparse counterpart ($R_p = 0.89 ± 0.03$), but nonetheless, it is included in the TAPREG server because blosum representation of sequences can often increase the generalization power of predictive models. The input data for TAPREG can consist of either protein sequences or multiple peptide sequences. For the protein sequence, TAPREG returns all 9-mer peptides encompassed by the protein, ranked by their affinity to TAP (IC50). The number of peptides listed in the output can also be limited using a user-defined threshold of binding affinity [Fig. 4(B)]. For the peptide input, the server returns the affinity of each individual peptide [Fig. 4(C)]. As TAP can bind and transport peptides of arbitrary length ranging from eight to 16 residues,14,21 TAPREG will predict the affinity of any peptide within that length range as described below.

### Table III

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>TAP613</th>
<th>TAPPRED</th>
<th>ADM</th>
<th>SVMTAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>1.00</td>
<td>0.86</td>
<td>0.26</td>
<td>0.84</td>
<td>0.68</td>
</tr>
<tr>
<td>TAP613</td>
<td>0.86</td>
<td>1.00</td>
<td>0.29</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td>ADM</td>
<td>0.84</td>
<td>0.59</td>
<td>1.00</td>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td>TAPPRED</td>
<td>0.26</td>
<td>0.29</td>
<td>1.00</td>
<td>0.17</td>
<td>0.34</td>
</tr>
<tr>
<td>SVMTAP</td>
<td>0.68</td>
<td>0.76</td>
<td>0.34</td>
<td>0.51</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table shows $R_p$ values that were obtained by correlating the TAP binding affinity scores of 723 MHCI-restricted T cell epitopes predicted with the different methods.
In general, models generated using machine-learning algorithms require input data of the same format as the data used for training. Therefore, in TAPREG, we have implemented a system to predict the TAP binding affinity of any peptide longer than nine residues, for example, ALRQFDMSERDNAVFL, by applying the model to a peptide fragment encompassing the first five N-terminal and last four C-terminal residues of the longer peptide; in this example, ALRQFAVFL. For peptides of eight residues, for example AVDFSDRS, we simply insert an Alanine at P6, AVDFSADRS, and then predict the binding affinity. Note that the P6 residue had the lower contribution to TAP binding [Fig. 1(A)]. Using the 5N3C model, which achieved the same correlation as the TAP613 model that was trained on the entire 9-mer peptides (DS613)—encompassing 178 new extra peptides—to study TAP selectivity quantitatively, using SVM regression models that were trained on single residue and residue combinations drawn from the peptides in the dataset. Thus, we have been able to recognize that each peptide position has a significant contribution to TAP binding, and that the contribution of the P4 residue is equivalent to that of the P3 residue [Fig. 1(A)]. Previously, only the positions P1, P2, P3, and the C-terminal end of the peptide were thought to be clearly relevant for binding to TAP.12,22,26,28,29 We have confirmed that the C-terminal end of the peptide has the largest quantitative input to TAP binding; a model trained on this residue alone reached an $R_p = 0.68 \pm 0.06$. Nonetheless, we have shown that the N-terminal half of the peptide has a larger contribution to TAP binding than the C-terminal half of the peptide, as judged by the predictive performance of SMVs trained on peptide fragments encompassing a varying number of N-terminal and C-terminal residues of the peptides in the DS613 dataset (Fig. 2).

**DISCUSSION**

The majority of TAP binding models have been derived from the same dataset consisting of ~435 9-mer peptides of known affinity which was made available by Dr. Peter van Endert28 (PVE435). In contrast, in this work, we have used a larger dataset of 613 peptides (DS613)—encompassing 178 new extra peptides—to study TAP selectivity quantitatively, using SVM regression models that were trained on single residue and residue combinations drawn from the peptides in the dataset. Thus, we have been able to recognize that each peptide position has a significant contribution to TAP binding, and that the contribution of the P4 residue is equivalent to that of the P3 residue [Fig. 1(A)]. Previously, only the positions P1, P2, P3, and the C-terminal end of the peptide were thought to be clearly relevant for binding to TAP.12,22,26,28,29 We have confirmed that the C-terminal end of the peptide has the largest quantitative input to TAP binding; a model trained on this residue alone reached an $R_p = 0.68 \pm 0.06$. Nonetheless, we have shown that the N-terminal half of the peptide has a larger contribution to TAP binding than the C-terminal half of the peptide, as judged by the predictive performance of SMVs trained on peptide fragments encompassing a varying number of N-terminal and C-terminal residues of the peptides in the DS613 dataset (Fig. 2).
Optimal modeling of the binding affinity of peptides in the DS613 dataset was achieved by SVM models trained on the full-length peptide sequences (TAP613) or on 8-residue fragments consisting of the first five N-terminal and last three C-terminal residues (5N3C) of the peptides ($R_p = 0.89 \pm 0.03$) [Figs. 1(B) and 2]. These results may reflect the observation that TAP can transport peptides of eight and nine residues with comparable efficiency.\(^{14,21}\) Overall, that optimal fitting of TAP binding affinity data required training on multiple peptide residues also implies that all peptide residues—perhaps with the exception of the P6 residue—have a relevant contribution to TAP binding.

The correlation between predictions and experimental binding affinity values achieved by models TAP613 and 5N3C, both trained on the DS613 dataset, is larger ($0.89 \pm 0.03$) than that reported for any predictive model of TAP binding affinity.\(^{26-29}\) It is worth noting that, unlike any of the related studies, we have not only evaluated the predictive performance of our models in cross-validation experiments but have also repeated the experiments 10 times and provided confidence values (standard deviations). Moreover, we have also shown that the enhanced predictive performance obtained with the model trained on the DS613 dataset is not related to sequence similarity redundancy (Fig. 3). In fact, we have found that peptides with high sequence similarity generally differ in their binding affinity (Fig. 3). Therefore, similar sequences are not redundant, and instead of inflating prediction rates, have a genuine contribution to model TAP binding affinity appropriately; hence, the enhanced prediction rates that we have obtained with the model trained in the DS613 dataset (Table I).

Using the new 178 peptides of known affinity to TAP collected in this study as a testing set (DS178 dataset), we have proved that two previous SVM-based methods (TAPPRED\(^{26}\) and SMVTAP\(^{27}\)) for predicting binding affinity of peptides to TAP, which were trained on the PVE435 dataset, appear to have suffered to some extent from data overfit; they achieved much lower correlation coefficients in the testing DS178 dataset than those reported on the PVE435 dataset (Table II). We have also evaluated two matrix-based methods, ADM\(^{29}\) and CM,\(^{28}\) on the same DS178 dataset, and they achieved correlations (0.87 and 0.74, respectively) that were similar to those originally reported by the authors (Table II). However, it is likely that these two matrix-based methods were trained on some of the peptides included in the DS178 dataset, because they were developed using binding affinity data of poly-Alanine peptides, such as those included in the DS178 dataset. In any case, TAP binding affinity predicted by our SVM models correlated more closely with those predicted by CM than with those predicted by related SVM-based methods (Table III). Overall, these results highlight the relevance of identifying and including new data points for training predictive models.

In this study, we have also developed a Web-based tool, TAPREG, to predict the binding affinity of peptides to TAP, which is available for free public use at http://imed.med.ucm.es/Tools/tapreg/. Currently, there are two dedicated web-based tools to predict the binding affinity of peptides to TAP: SMVTAP\(^{27}\) (http://www-bs.informatik.uni-tuebingen.de/Services/SMVTAP/) and TAPPRED\(^{26}\) (http://www.imtech.res.in/raghava/tappred/), both of them based on SVMs. These two resources use a protein sequence as input and report the 9-mer peptides encompassed by the protein, ranked by their predicted binding affinity to TAP. In addition to this task, TAPREG can be used to predict the binding affinity to TAP of multiple peptides with a length ranging from eight to 16 residues,\(^{14,21}\) which is consistent with the transport activity displayed by TAP.

Until now TAP binding affinity of peptides longer than nine residues could only be achieved using quantitative matrices, and only the 3 N-terminal residues and the C-terminus of the peptide were considered to matter for TAP binding.\(^{28}\) In contrast, in TAPREG, we compute the TAP affinity using nine residues selected from the larger peptides—those equivalent to the 9-mer peptides used for training—as we have shown that all residues in a 9-mer peptide contribute to binding. To our knowledge, this is the first machine-learning based approach that can predict the binding affinity to TAP of peptides longer than nine residues.

**CONCLUSIONS**

We have used a large dataset of 9-mer peptides of known affinity to TAP to dissect the TAP binding preferences, concluding that each peptide position has a quantitative contribution to TAP binding. Moreover, we have been able to generate SVM models with enhanced predictive performance as a result of including new peptide binding data. Because accurate modeling of TAP activity is relevant for T cell epitope selection,\(^{12,13}\) we have implemented the Web-based tool TAPREG (http://imed.med.ucm.es/Tools/tapreg/). Unlike any related resource, TAPREG can be used to predict the binding affinity of peptides ranging from eight to 16 residues, in a manner that is consistent with the activity exhibited by TAP.

**REFERENCES**


