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. Reche[1]*

[1] Laboratorio de InmunoMedicina, Departamento de Microbiología I-Immunología, 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 MHC 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 Innovación (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 Microbiología I-Immunología, Facultad de Medicina, Universidad Complutense, de Madrid, Ave. Complutense s/n, Madrid 28040, Spain. E-mail: parecheg@med.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

TAP belongs to the ATP-dependent binding cassette (ABC) transporter superfamily, and it is expressed as a heterodimer consisting of the TAP1 and TAP2 proteins subunits.[14,15] Both TAP1 and TAP2 proteins encode one hydrophobic transmembrane domain and one ATP binding domain. Transport of peptides by TAP proceeds in two sequential steps, where peptide binding to TAP occurs first followed by a translocation step consuming ATP.[16-18] Peptide transport rate by TAP is governed by the initial binding step.[19,20] Likewise, TAP preselection of peptides available for MHCI presentation is also controlled by their affinity to TAP. Selectivity of TAP has been studied from data generated using assays that determine peptide binding to TAP or peptide accumulation in the ER.[17,18] TAP preferentially transports peptides with a length of 8–16 residues,[14,21] whereas longer peptides may be transported but with much lower efficiency. Besides peptide length preferences, the first three N-terminal residues and the C-terminal end of the peptides have also been shown to be important for binding to TAP.[12,22] Furthermore, a peptide-binding motif for TAP has been defined by van Endert et al.,[22] which indicates a TAP preference for hydrophobic aromatic residues at the C-terminus, hydrophobic residues at position 3 (P3), and charged and hydrophobic residues at position 2 (P2). On the other end, aromatic or acidic residues at P1 and prolines at P1 and P2 have strong deleterious effects.

A number of methods have also been applied for predicting and analyzing the binding affinity of peptides to TAP, such as artificial neural networks,[23-25] support vector machines (SVMs),[26,27] and matrices generated using the Stabilized Matrix Method[28] and the additive method.[29,30] The majority of these methods were trained on the same training set of ~435 nonamer (9-mer) peptides of known affinity to TAP made available by Dr. van Endert, and until now their performance has not been compared in an independent testing set. In contrast, here we have used a much larger training set, encompassing 178 new peptides, to analyze TAP binding preferences using SVMs. Interestingly, our results indicate that each peptide residue has a significant contribution to TAP binding. Moreover, we have generated TAP binding affinity models that in cross-validation experiments achieved a correlation between experimental and predicted values of 0.89 ± 0.03, which is stronger than that of related methods. Based on these results, we have implemented a system, TAPREG, for predicting affinity of peptides to TAP that is available for free public use at http://imed.med.ucm.es/Tools/tapreg/.

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 ($IC_{50relative}$). The lower the $IC_{50relative}$, the stronger the peptide binds to TAP. This dataset encompasses 435 peptides, kindly provided by Dr. Peter van Endert[23] (INSERM U580, Paris Descartes University, Paris, France)—$IC_{50relative}$ already referenced to RRYNASTEL—plus 178 peptides parsed from the TAP binding affinity peptide collection of the Antijen Database,[31] kindly provided by Dr. Darren Flower (The Jenner Institute, Compton, UK). To combine the peptides into a single dataset, the TAP binding affinity (IC_{50}) 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[32] 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[33] and EPIMHC[34] 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.[35] 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).[36,37] 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 substitution scores given by the BLOSUM62 substitution matrix.[38] TAP affinity ($IC_{50relative}$) values of the training sets were provided to WEKA as $logIC_{50relative}$ values. Pearson's correlation coefficient (R_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_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_{50} values by considering an IC_{50} for the reference peptide RRYNASTEL of 400 n*M*.

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.[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 $N \times N\text{-}1$ 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 ($logIC_{50relative}$) using Spearman's rank correlation (R_s). For instance, let us consider the peptide PLAKAAAAV ($logIC_{50relative} = 8.370$) had the following hits:

Hit:ALAKAAAAV; Identity:88.9%; Similarity:88.9%; $logIC_{50relative}$:3.984; Dif:4.386

Hit:ALAKAAAAL; Identity:77.8%; Similarity:88.9%; $logIC_{50relative}$:0.688; Dif:7.682

Hit:AAASAAAAF; Identity:66.7%; Similarity:77.8%; $logIC_{50relative}$:−0.734; Dif:9.104

Hit:ALAKAAAAF; Identity:55.6%; Similarity:66.7%; $logIC_{50relative}$:0.332; Dif:8.038

Hit:GRQKGAGSV; Identity:33.3%; Similarity:44.4%; $logIC_{50relative}$:6.215; Dif:2.155

Then, for peptide PLAKAAAAV, an R_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 $logIC_{50relative}$ values (4.386, 7.682, 9.104, 8.038, 2.155). R_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_s values were determined considering all peptide hits and only those with an identity $\geq$ 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_p mean values were significantly above 0 (Ho: $R_pi - R_pj = 0$; $P \leq 0.05$). To evaluate if R_p values were statistically significant (H_0: $R_p = 0$), we computed the statistics given by Eq. (1), which follows a *t*-Student distribution with $N - 2$ degrees of freedom, and tested subsequently ($P < 0.05$).

$$t = \frac{R_p}{\sqrt{\frac{1-Rp^2}{N-2}}} \tag{1}$$

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.,[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 to the relevant correlation coefficients (R_i) using Eq. (2).

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

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

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

In Eq. (3), R_{ij} is the correlation between the predicted values by the methods *i* and *j* being compared, and $h = (1 - f\overline{R}^2)/(1 - \overline{R}^2)$, with $\overline{R}^2 = (R^2_i + R^2_j)/2$ and $f = (1 - R_{ij})/2(1 - \overline{R}^2)$.

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

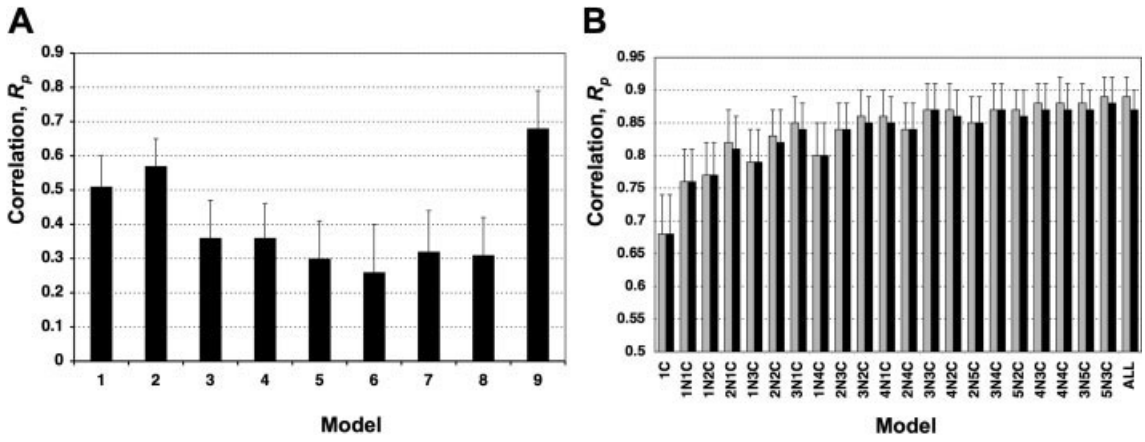

Figure 1

Performance of TAP-affinity prediction models. Models were trained using SVM and their performance was measured using R_p values between predictions and experimental values determined under 10-fold cross-validation experiments that were repeated 10 times. Thus, R_p mean values and standard deviations obtained over 100 measures are represented in the figure. Moreover, plotted R_p values were those achieved by SVMs after parameter optimization. (**A**) Performance of models trained on individual residues of the 9-mer peptides (1–9) included in the training set. (**B**) Performance of models trained on different peptide fragments consisting of the first *i* N-terminal and the last *j* C-terminal residues of the peptides in the training set. Residue selections, *iNjC* are indicated in the abscissa. Grey bars are for SVM models trained on sparse sequence representations and black bars for models trained using blosum sequence representations. There was no difference between sparse and blosum trained models on single peptide residues. Data for making these representations—including the relevant RBF parameters of SVMs—are provided as Supporting Informtion in Table 7S.

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 (DS_{613}) of known affinity to TAP ($logIC_{50relative}$) and SVMs under a regression schema. SVMs are among the most widely used methods for solving common data mining problems in bioinformatics[41-43] and were chosen because of their solid theoretical foundations and proven generalization ability.[44] 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 ϵ-insensitive loss function for error minimization.[44] 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 ($P \leq 0.05$, Material and Methods). The lowest correlation was obtained with a model trained on the P6 residue (R_p of 0.26 ± 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 ± 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

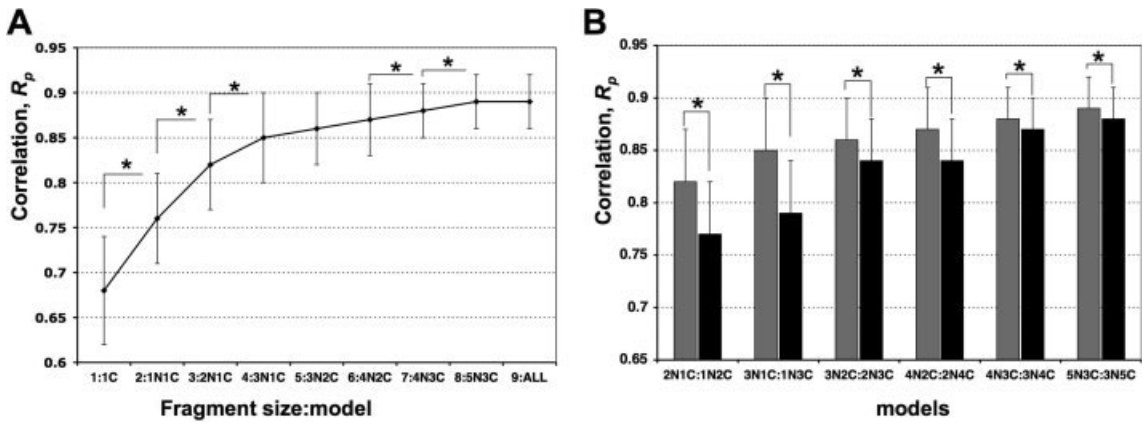

Figure 2

Analysis of TAP selectivity using TAP-affinity prediction models. SVM-Models trained using sparse sequence representation were selected. (**A**) Predictive performance (R_p) of SVM-models with regard to the fragment size used for training (1–9). Only the largest R_p value achieved by a specific model (indicated in the abscissa) at each fragment size is represented. Statistically significant increments between R_p values of neighboring models are indicated with a "*" symbol. (**B**) Predictive performance of the best SVM-models generated upon optimal first *i* N- and last *j* C-terminal residue selections (gray bars) compared with those generated from suboptimal first *j* N- and last *i* C-terminal residue selections (black bars). Statistically significant differences were found between R_p values in all cases (indicated with a "*" symbol). Statistical significance was assessed using *t*-tests (Material and Methods).

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 ($P \leq 0.05$) 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 (nonamers). 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 ($P \leq 0.05$) 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 (DS_{293}, DS_{332}, DS_{465}, DS_{530}, respectively) by discarding similar sequences from the original DS_{613} dataset (Material and Methods). The global sequence identity in percentage in these datasets varied from 1 ± 6% in the DS_{293} dataset to 9 ± 23% in the DS_{530} dataset, whereas in the DS_{613} dataset it was 10 ± 25% (Table I). In the 435-peptide dataset provided by Peter van Endert (PVE_{435}) 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 ($logIC_{50relative}$). The average number of similarity hits per peptide in the datasets varied from nine peptides in the DS_{293} dataset to 110 hits in the DS_{613} dataset (Table I). Sequence identity between hits was considerably larger

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

Dataset	R_p	Identity (%)[a]	Similarity (%)[a]	Identity (%)[b]	Similarity (%)[b]	Hits[c]
DS_{293}	0.71 ± 0.1	1 ± 6	2 ± 10	23 ± 11	43 ± 11	9 ± 7
DS_{332}	0.76 ± 0.09	2 ± 8	3 ± 11	28 ± 18	46 ± 14	14 ± 12
DS_{465}	0.85 ± 0.05	7 ± 19	8 ± 21	52 ± 25	60 ± 19	59 ± 45
DS_{530}	0.87 ± 0.03	9 ± 23	10 ± 25	57 ± 24	62 ± 26	86 ± 62
DS_{613}	0.89 ± 0.03	10 ± 25	11 ± 26	59 ± 23	66 ± 18	110 ± 77
PVE_{435}	0.83 ± 0.05	5 ± 16	6 ± 18	45 ± 26	56 ± 19	40 ± 33

[a]Identity and similarity computed considering all possible pairwise comparisons between the peptides in the datasets.
[b]Identity and similarity computed considering only hits (Material and Methods).
[c]Average number of similarity hits per peptide in the dataset.

and ranged from 23% in the DS_{293} dataset to 59% in the DS_{613} dataset (Table I).

Because we train and evaluate the predictive performance of SMVs 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 $\log IC_{50relative}$ values achieved by SVMs trained and evaluated on the datasets of reduced sequence similarity (DS_{293}, DS_{332}, DS_{465}, DS_{530}, and PVE_{435}) was significantly lower ($P \leq 0.05$; one-sided *t*-tests) than that obtained in the DS_{613} dataset (Table I). The smallest R_p was achieved in the DS_{293} dataset (0.71 ± 0.1), and these values increased significantly ($P \leq 0.05$) as the number of peptides in the datasets (Table I). Thus, $DS_{613}R_p > DS_{530}R_p > DS_{465}R_p > PVE_{435}R_p > DS_{332}R_p > DS_{293}R_p$.

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 R_p values were computed in cross-validation, and the differences in R_p 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 must be. 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 DS_{293}, DS_{332}, DS_{465}, DS_{613}, and PVE_{435} 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 DS_{613} dataset is significantly larger than those of the remaining datasets ($P \leq 0.05$), 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 DS_{613} dataset are not inflated due to sequence similarity redundancy. Furthermore, similar sequences in the DS_{613} 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 DS_{613} dataset.

Comparison of methods for predicting binding affinity of peptides to TAP

We have compared our SVM model trained on 9-mer peptide sequences that achieved an $R_p = 0.89 \pm 0.03$ (hereafter TAP_{613}) with four alternative predictive

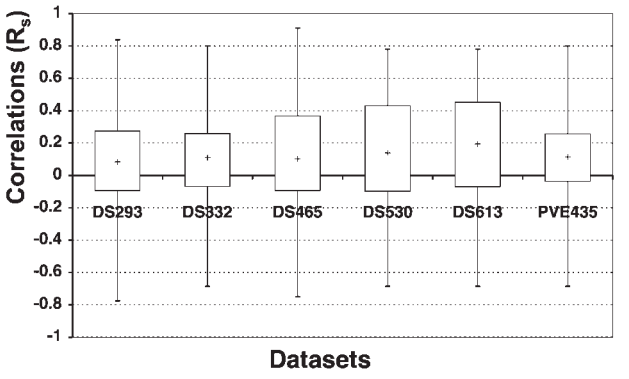

Figure 3
Relationship between sequence similarity in peptide datasets and binding affinity proximity. This figure depicts a boxplot of R_s values computed for each peptide in a dataset by correlating their identity with its hits and the difference in $\log IC_{50relative}$ values (Material and Methods). Boxplot were generated for peptides in DS_{293}, DS_{332}, DS_{465}, DS_{530}, DS_{613}, and PVE_{435} datasets. Median R_s values in peptide datasets are indicated with a cross. A negative R_s 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.

Table II
Correlation Between Experimental TAP Binding Affinities and Predicted Values Using Different Methods

Method	R_s	Reference
TAP_{613}	0.89 ± 0.03	This study
SMM	0.87 (0.82)	28
ADM	0.74 (0.72–0.83)	29
TAPPRED	0.67 (0.88)	26
SVMTAP	0.61 (0.82)	27

R_s were computed using a testing set of 178 peptides of known affinity to TAP. For the TAP_{613} model, R_s shown in the table is that achieved in cross-validation. Correlations reported in the literature for the different methods are shown in parentheses.

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 (TAPPRED[26] and SVMTAP[27]). 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 method[30]; hence, we will refer to this method as ADM. The ADM method achieved a reported R_p between 0.72 and 0.83, depending of the testing set.[29] The remaining methods have been trained on the PVE_{435} 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 PVE_{435} dataset. The CM method achieved a reported R_p of 0.782 on the PVE_{435} dataset. The TAPPRED[26] and SVMTAP[27] methods are based on SVMs trained solely on the PVE_{435} 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 (DS_{178}), using Spearman's correlation coefficients (R_s) (Table II). Interestingly, the lowest R_s 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_s (0.87) comparable to the value achieved by our TAP_{613} model in cross-validation (0.89), and AMD achieved an intermediate R_s value of 0.74. Statistical comparison of these R_s values (Material and Methods) indicated that the correlations obtained with the CM and TAP_{613} methods were significantly stronger than those obtained with the remaining methods. However, TAP_{613} was also trained on the DS_{178} testing set used for the comparisons, as surely were both the CM and ADM methods (DS_{178} 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, TAP_{613} predictions were significantly closer to the predictions by CM (R_s = 0.86), a matrix-based method, than to those by TAPRED (0.29) and SVMTAP (0.76), which are based on SVM. Likewise, ADM predictions also correlated better with TAP_{613} 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.[25] 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 TAP_{613} model, as it was trained on a much larger dataset.

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 DS_{613} 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 (IC_{50}). 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
Correlation Between TAP Binding Affinity Predictions by Different Methods

	CM	TAP_{613}	TAPPRED	ADM	SVMTAP
CM	1	0.86	0.26	0.84	0.68
TAP_{613}	0.86	1	0.29	0.59	0.76
ADM	0.84	0.59	0.17	1	0.51
TAPPRED	0.26	0.29	1	0.17	0.34
SVMTAP	0.68	0.76	0.34	0.51	1

Table shows R_s values that were obtained by correlating the TAP binding affinity scores of 723 MHCI-restricted T cell epitopes predicted with the different methods.

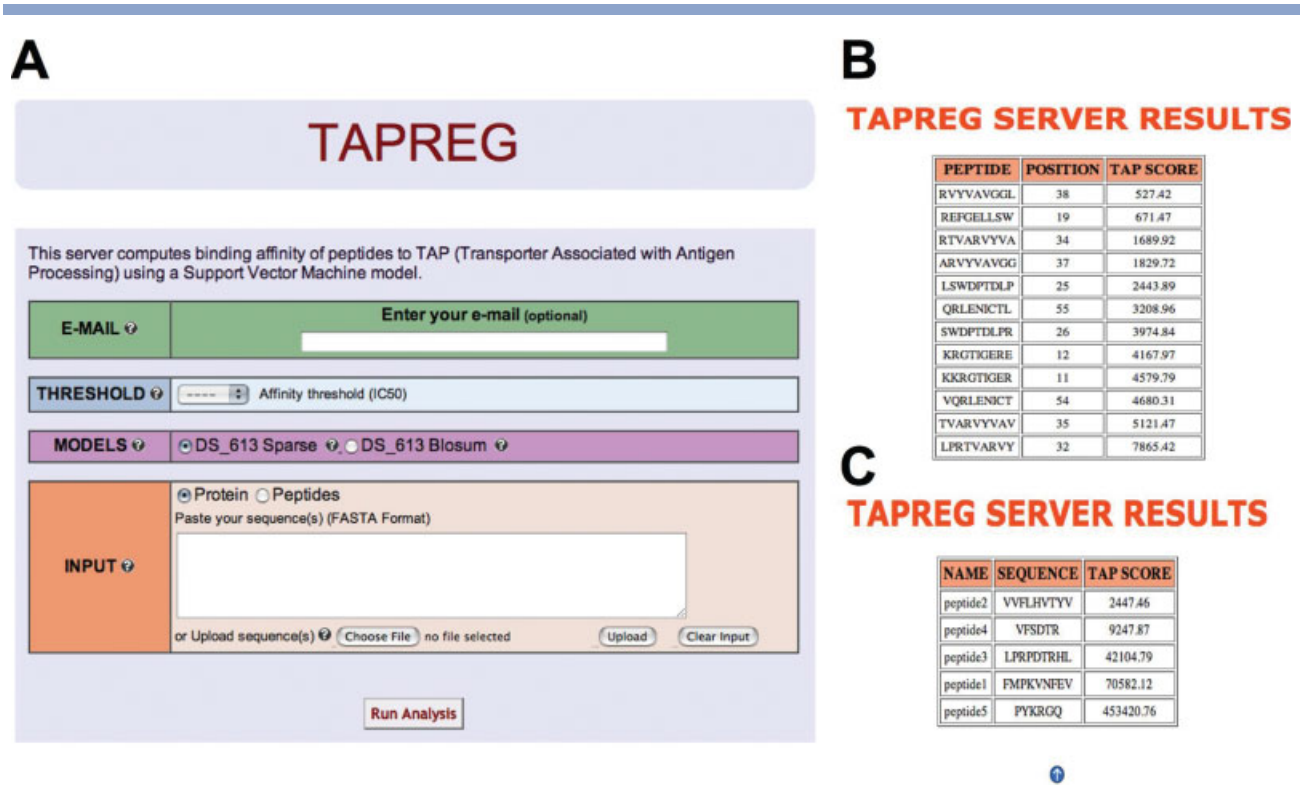

Figure 4

TAPREG server for predicting peptide binding affinity to TAP. (**A**) TAPREG Web interface. TAPREG can take two types of input data consisting of either multiple peptides in FASTA format (size 8 to 16 allowed) or a protein sequence in FASTA format. For protein sequences, TAPREG computes the TAP affinity of all 9-mer peptides in the protein and returns the peptides sorted by their affinity (IC_{50}) (Panel **B**). When multiple peptides are submitted, the program returns the binding affinity to TAP (IC_{50}) of each peptide (Panel **C**).

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, ALRQFDSMERDNAVFL, 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 TAP_{613} model that was trained on the entire 9-mer peptides (Fig. 2), the binding of any peptide longer than eight residues could be predicted by applying the model to a derivative fragment consisting of the first 5 N-terminal and last 3-C terminal residues.

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 Endert[28] (PVE_{435}). In contrast, in this work, we have used a larger dataset of 613 peptides (DS_{613})—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 DS_{613} dataset (Fig. 2).

Optimal modeling of the binding affinity of peptides in the DS_{613} dataset was achieved by SVM models trained on the full-length peptide sequences (TAP_{613}) 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 TAP_{613} and *5N3C*, both trained on the DS_{613} dataset, is larger (0.89 ± 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 DS_{613} 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 DS_{613} dataset (Table I).

Using the new 178 peptides of known affinity to TAP collected in this study as a testing set (DS_{178} 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 PVE_{435} dataset, appear to have suffered to some extent from data overfit; they achieved much lower correlation coefficients in the testing DS_{178} dataset than those reported on the PVE_{435} dataset (Table II). We have also evaluated two matrix-based methods, ADM[29] and CM,[28] on the same DS_{178} 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 DS_{178} dataset, because they were developed using binding affinity data of poly-Alanine peptides, such as those included in the DS_{178} 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/SVMTAP/) 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

1. Paul WE. Fundamental immunology. Philadelphia, PA: Lippincott, Williams & Wilkins; 1998.
2. Von Boehmer H. Positive and negative selection of the ab T cell repertoire in vivo. Curr Opin Immunol 1991;3:210–215.
3. Craiu A, Akopian T, Goldberg A, Rock KL. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. Proc Natl Acad Sci USA 1997;94:10850–10855.
4. Yewdell JW, Haeryfar SM. Understanding presentation of viral antigens to CD8+ T cells in vivo: the key to rational vaccine design. Annu Rev Immunol 2005;26:651–682.

5. Pamer E, Cresswell P. Mechanisms of MHC class I--restricted antigen processing. Annu Rev Immunol 1998;16:323–358.
6. York IA, Goldberg AL, Mo XY, Rock KL. Proteolysis and class I major histocompatibility complex antigen presentation. Immunol Rev 1999;172:49–66.
7. Serwold T, Gonzalez F, Kim J, Jacob N. ERAAP customizes peptides for MHC Class I molecules in the endoplasmic reticulum. Nature 2002;419:480–483.
8. Beekman NJ, Van Veelen PA, Van Hall T, Neisig A, Sijts A, Camps M, Kloetzel PM, Neefjes JJ, Melief CJ, Ossendorp F. Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site. J Immunol 2000;164:1898–1905.
9. Smith KD, Lutz CT. Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. J Immunol 1996;156:3755–3764.
10. Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL. Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. J Biol Chem 2003;278:45135–45144.
11. Wang M, Lamberth K, Harndahl M, Røder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. Vaccine 2007;25:2823–2831.
12. Uebel S, Krass P, Kienle S, Wiesmuller KH, Jung G, Tampe R. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. Proc Natl Acad Sci USA 1997;94:8976–8981.
13. Ackerman AL, Cresswell P. Cellular mechanisms governing cross-presentation of exogenous antigens. Nat Immunol 2004;5:678–684.
14. Androlewicz MJ, Cresswell P. Human transporters associated with antigen processing possess a promiscuous peptide-binding site. Immunity 1994;1:7–14.
15. Abele R, Tampe R. The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing. Physiology (Bethesda) 2004;19:216–224.
16. Shepherd JC, Schumacher TN, Ashton-Rickardt PG, Imaeda S, Ploegh HL, Janeway CA, Tonegawa S. TAP1-dependent peptide translocation in vitro is ATP dependent and peptide selective. Cell 1993;74:577–584.
17. Neefjes JJ, Momburg F, Hammerling GJ. Selective and ATP-dependent translocation of peptides by the MHC-encoded transporter. Science 1993;261:769–771.
18. Van Endert PM, Tampe R, Meyer TH, Tisch R, Bach JF, Mcdevitt HO. A sequential model for peptide binding and transport by the transporters associated with antigen processing. Immunity 1994;1:491–500.
19. Gubler B, Daniel S, Armandola EA, Hammer J, Caillat-Zucman S, Van Endert PM. Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. Mol Immunol 1998;35:427–433.
20. Armandola EA, Momburg F, Nijenhuis M, Bulbuc N, Fruh K, Hammerling GJ. A point mutation in the human transporter associated with antigen processing (TAP2) alters the peptide transport specificity. Eur J Immunol 1996;26:1748–1755.
21. Momburg F, Roelse J, Howard JC, Butcher GW, Hammerling GJ, Neefjes JJ. Selectivity of MHC-encoded peptide transporters from human, mouse and rat. Nature 1994;367:648–651.
22. Van Endert PM, Riganelli D, Greco G, Fleischhauer K, Sidney J, Sette A, Bach JF. The peptide-binding motif for the human transporter associated with antigen processing. J Exp Med 1995;182:1883–1895.
23. Daniel S, Brusic V, Caillat-Zucman S, Petrovsky N, Harrison L, Riganelli D, Sinigaglia F, Gallazzi F, Hammer J, Van Endert PM. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J Immunol 1998;161:617–624.
24. Brusic V, Van Endert P, Zeleznikow J, Daniel S, Hammer J, Petrovsky N. A neural network model approach to the study of human TAP transporter. In Silico Biol 1999;1:109–121.
25. Zhang GL, Petrovsky N, Kwoh CK, August JT, Brusic V. PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. Immunome Res 2006;2:3.
26. Bhasin M, Raghava G. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 2004;13:596–607.
27. Donnes P, Kohlbacher O. Integrated modeling of the major events in the MHC class I antigen processing pathway. Protein Sci 2005;14:2132–2140.
28. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 2003;171:1741–1749.
29. Doytchinova I, Hemsley S, Flower DR. Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. J Immunol 2004;173:6813–6819.
30. Doytchinova IA, Blythe MJ, Flower DR. Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. J Proteome Res 2002;1:263–272.
31. Toseland CP, Clayton DJ, Mcsparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res 2005;1:4.
32. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling detection of bacterial outer membrane protein repeats. Protein Sci 1995;4:1618–1632.
33. Sathiamurthy M, Peters B, Bui HH, Sidney J, Mokili J, Wilson SS, Fleri W, Mcguinness DL, Bourne PE, Sette A. An ontology for immune epitopes: Application to the design of a broad scope database of immune reactivities. Immunome Res 2005;1:2.
34. Reche PA, Zhang H, Glutting JP, Reinherz EL. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 2005;21:2140–2141.
35. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–2481.
36. Smola AJ, Scholkopf B. A Tutorial on support vector regression. NC2-TR-1998–030. NTRS. Berlin, Germany: Springer; 1998.
37. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to SMO Algorithm for SVM Regression, Technical Report CD-99–16, 2000.
38. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
39. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;16:276–277.
40. Meng X-L, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull 1992;111:172–175.
41. Yang ZR. Biological applications of support vector machines. Brief Bioinform 2004;5:328–338.
42. Bhasin M, Reinherz EL, Reche PA. Recognition and classification of histones using support vector machine. J Comput Biol 2006;13:102–112.
43. Bhasin M, Zhang H, Reinherz EL, Reche PA. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 2005;579:4302–4308.
44. Vapnik V. The nature of statistical learning theory. New York: Springger-Verlag; 1995.