Published in final edited form as:

J Immunol Methods. 2013 May 31; 391(0): 163–173. doi:10.1016/j.jim.2013.02.006.

Predictor for the Effect of Amino Acid Composition on CD4+ T Cell Epitopes Preprocessing

Ehud Hoze^a, Lea Tsaban^a, Yaakov Maman^c, and Yoram Louzoun^{a,b,*}

^aDepartment of Mathematics, Bar Ilan University, Ramat Gan 52900, Israel

^bGonda Brain Research Center, Bar Ilan University, Ramat Gan 52900, Israel

^cThe Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat Gan 52900, Israel

Abstract

Predictive tools for all levels of CD8+ T cell epitopes processing have reached a maturation level. Good prediction algorithms have been developed for proteasomal cleavage, TAP and MHC class I peptide binding. The same cannot be said of CD4+ T cell epitopes. While multiple algorithms of varying accuracy have been proposed for MHC class II peptide binding, the preprocessing of CD4+ T cell epitopes is still lacking a good prediction algorithm. CD4+ T cell epitopes generation includes several stages, not all which are well-defined. We here group these stages to produce a generic preprocessing stage predictor for the cleavage processes preceding the presentation of epitopes to CD4+ T cell. The predictor is learnt using a combination of *in vitro* cleavage experiments and observed naturally processed MHC class II binding peptides. The properties of the predictor highlight the effect of different factors on CD4+ T cell epitopes preprocessing. The most important factor emerging from the predictor is the secondary structure of the cleaved region in the protein. The effect of the secondary structure is expected since CD4+ T cell epitopes are not denatured before cleavage. A website developed based on this predictor is available at: http://peptibase.cs.biu.ac.il/PepCleave_cd4/

Keywords

CD4+ T cell; epitope; MHC class II; processing; cleavage

1. Introduction

The cellular immune system recognizes and destroys foreign pathogens. In this process, antigen presenting cells (APCs) internalize, degrade and subsequently express on their surface small antigenic fragments of extracellular proteins on major histocompatibility complex (MHC) class II molecules (Berger and Roche, 2009). Peptide MHC (pMHC) class II complexes bind to appropriate receptors on the surface of CD4+ T cells and lead to CD4+ T cell proliferation and cytokine secretion (O'Garra and Arai, 2000). This activation elicits

© 2012 Elsevier B.V. All rights reserved.

*Corresponding author at: Department of Mathematics, Bar Ilan University, Ramat Gan 52900, Israel, Tel: +972 35317610, louzouy@math.biu.ac.il, (Yoram Louzoun).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

downstream effects that regulates the function of B cells and effector T cells (Jensen, 2007; Berger and Roche, 2009).

The generation of the pMHC class II consists of multiple stages, including internalization, degradation and loading of the peptide onto the MHC class II molecule (Ramachandra et al., 1999). When an APC engulfs extracellular matter that may serve as an Antigen (Ag), the internalized matter is encapsulated to form a phagosome (Ramachandra et al., 1999). Newly synthesized MHC class II molecules are assembled in the Endoplasmic Reticulum (ER) as $\alpha\beta$ heterodimers associated with an invariant chain (Ii), a chaperone protein that stabilizes the MHC class II molecule and prevents interaction of its binding cleft with other molecules (Wolf and Ploegh, 1995). The MHC class II Ii complex (CLIP) is trafficked to the Golgi complex and through the trans Golgi network to the surface of the cell (Lotteau et al., 1990). The complex is then internalized back into endocytic vesicles that are fused with phagosomes. These phagosomes contain both the internalized Ag and the CLIP (Hiltbold and Roche, 2002). The phagosome is fused with the lysosome to establish the phagolysosome (Desjardins et al., 1994). In this compartment, Ii is degraded and removed (Alfonso and Karlsson, 2000). The internalized Ag is also degraded into peptides that are loaded onto the MHC class II molecule to create the pMHC class II (Berger and Roche, 2009).

The peptide cleavage and the removal of the chaperon in the phagolysosome are executed by acid hydrolases and proteins originating from the lysosome (Riese and Chapman, 2000). These hydrolases in the phagolysosome include Cathepsins B, L, D, H and S and dipeptidyl peptidase II (Sturgill-Koszycki et al., 1996; Rathman et al., 1997; Claus et al., 1998). The phagolysosome is also enriched with HLA-DM and HLA-DO that regulate the association of antigen-derived peptides with MHC class II molecules (Busch et al., 2005). Among the known lysosomal proteases, only cathepsin L and cathepsin S have been shown to have a distinct role in the endosomal presentation in vivo (Villadangos and Ploegh, 2000). Besides being engaged in the degradation of the Ii, these enzymes also mediate the processing of antigens in distinct cell types. Specific hydrolases have different expression levels in distinct cell types (Hsing and Rudensky, 2005). These differences induce a variance in the antigenic repertoires displayed by MHC class II molecules among different APCs. (Hsing and Rudensky, 2005). The end-product of these interactions is the presentation of pMHC class II on the surface of an APC.

We here present an algorithm to predict the peptides that can induce proper preprocessing for MHC class II binding. In other words, we are not interested in the MHC class II binding, but in all stages preceding the binding. The peptides resulting from this preprocessing can then be analyzed using existing MHC class II binding prediction algorithms to predict CD4+ T cell epitopes (Meister et al., 1995; Rammensee et al., 1999b; Nielsen et al., 2004). Identifying naturally processed CD4+ T cell epitopes is probably the next step required for the development of future subunit peptide vaccines.

Epitope prediction algorithms have been developed for most stages of epitope presentation and preprocessing for both MHC class I and MHC class II epitopes. MHC class I epitope prediction is accurate for most alleles (Lin et al., 2008) and a large number of prediction tools are exclusively focused on MHC class I binding (e.g., Propred-I (Singh and Raghava, 2003), BIMAS (Parker et al., 1992), NETMHC (Buu et al., 2003; Nielsen et al., 2003; Nielsen et al., 2004) RankPep (Reche et al., 2002; Reche et al., 2004), MLVO (Vider-Shalit and Louzoun, 2011) nHLAPred (Bhasin and Raghava, 2004) and many others). Tools for MHC class I binding peptide preprocessing have also been developed (Kesmír et al., 2002; Ginodi et al., 2008). Finally, integrated servers for MHC class I preprocessing and binding have been developed (EpiJen (2006), IEDB tools (2008) and FRED (2009)). The situation

for MHC class II binding peptides is different. A large number of prediction algorithms have been developed for the MHC class II binding (e.g. SVMHC (Donnes and Elofsson, 2002), NetMHCII (Nielsen et al., 2007; Nielsen and Lund, 2009), NetMHCIIpan (Nielsen et al., 2008), SYFPEITHI (Rammensee et al., 1999a), IEDB_ARB (Bui et al., 2005), IEDB_SMM-align (Nielsen et al., 2007), IEDB_cons (Wang et al., 2008), EpiToolKit (Feldhahn et al., 2008), Tepitope/ProPred (Singh and Raghava, 2001), Rankpep (Reche et al., 2002; Reche et al., 2004), HLA-DR4pred (Bhasin and Raghava, 2004), MultiPred (Brusic et al., 2002; Zhang et al., 2005), etc.), but practically none for the preceding cleavage.

A major limitation in MHC class II epitope prediction is this absence of prediction algorithms for phagosomal and endosomal cleavage. We here develop for the first time a cleavage algorithm for the preprocessing of peptides presented on MHC class II molecules. Such an algorithm is essential for the prediction of CD4+ T cell epitopes, since even if a peptide can bind to an MHC class II molecule, if it is not cleaved it will not be presented.

In the following sections we test multiple score possibilities. We show that using a combination of the amino acid pair (termed diAA) surrounding each of the cleavage sites (at the C and N termini) and the diAA composition within a peptide, a good preprocessed peptide classifier can be obtained. We apply multiple machine learning algorithms on a dataset of naturally processed CD4+ T cell epitopes to produce a family of classifiers, each with different properties. We then study the biochemical properties of peptides emerging from the classifiers.

2. Methods

2.1. statistical dispersion

The comparison between the dispersion of diAA distributions in the peptide body and the cleavage sites was carried out by two steps:

- To get the diAA distributions, the observed/expected ratio of diAA probability was

$$\frac{P(\text{diAA}_{ij})}{P(AA_i) \cdot P(AA_j)}$$

measured for each AA ($P(AA_i) \cdot P(AA_j)$), where i is to the left of the site and j to the right).

- The statistical dispersion of these distributions was then compared by using Average Absolute Deviation (AAD) from the median.

The average absolute deviation (AAD) of a data set is the average of the absolute deviations from the median and is a summary statistic of statistical dispersion or variability. Lower AAD means low dispersion and convergence to specific value.

2.2. Training and validation sets

Epitopes from the IEDB database were used for the positive datasets. We extracted 1027 different epitopes classified as ligand elution (Vita et al.) in the IEDB database (Vita et al.). For each peptide, we found its origin in the appropriate genome and extended it to include the flanking regions at its N and C termini. Peptides that had no explicit origin or peptides where an appropriate origin could not be found were excluded from the data set. As a negative data set, we extracted all the peptides in the source proteins which are not included in the epitope dataset. Out of this pool, we pulled a subset of 20,000 peptides with a length distribution similar to the length distribution of the positive data set (Supp Mat. Figure S1). No peptide was found in more than one dataset. The positive data set was divided into two parts: 856 sequences for the learning set and distinct 171 sequences for the validation set.

We also used a database of 42 sequences of cleavage sites for Cathepsins L and S extracted from the MEROPS database (Rawlings and Barrett, 1999) as an in vitro cleavage positive learning set. We have also used an external test set for further analysis. This additional test set was not used for the learning process or for the choice of the optimal model. This dataset contains a set of 3862 epitopes inducing a T cell response from the IEDB (Vita et al.). The goal of this set is to see whether peptides measured to induce a T cell response are indeed naturally processed.

2.3. Simulated annealing

Optimal weights for each diAA were learnt using a Simulated Annealing (SA) process. The initial configuration of the learnt weights ($20 \text{ AAs} \times 20 \text{ AAs} = 400$ parameters) was initialized at zero. The initial temperature was set to $T_0 = 5$, and was decreased exponentially once during every cycle ($T_n = \lambda^n \cdot T_0$, $\lambda = 0.8$). Each step was a random change of a weight by a random factor uniformly distributed between 2.5 and -2.5; the value of the weights was limited to the domain: 6 to -6. We performed 60 steps of decreasing the temperature; each step composed of 300 cycles per parameter.

2.4. SVM

For the classification of preprocessed peptides, we used a support vector machine (SVM) algorithm with linear and quadratic kernels (Boser et al., 1992; Müller et al., 2001). The SVM finds a maximal margin separator between two data sets with either a linear or quadratic distance kernel. In both cases, the error cost (C) values were set between 10^{-6} – 10^0 and the positive to negative learning set size ratio was varied between 1:1 and 1:3 (through the increase of the negative learning set size). The observations used in the SVM were AA or diAA frequency vectors. The size of each vector was 1×20 and 1×400 respectively for each position. Each domain of the peptide has its unique frequency vector. The AA composition of the peptide was computed using an occupancy vector, where the appropriate AA (or diAA) value was the number of times it appeared, and all others had values of 0. For example, in the diAA formalism, an Ala-Cys combination was represented as a 400 dimensional vector with a value of 1 at position 21 (second row first column) and all other values as 0.

2.5. Performance evaluation

In order to assess the performance of the classifiers, we used the common statistical measures of sensitivity, specificity and accuracy. The sensitivity was computed as the true positive rate, i.e., the preprocessed peptides that were classified correctly (true positive) among all preprocessed peptides either classified correctly (true positive) or classified incorrectly (false negative). The specificity was computed as the true negative rate, i.e., the non preprocessed peptides that were classified correctly (true negative) among all non preprocessed peptides either classified correctly (true negative) or classified incorrectly (false positive). As the real prevalence is not known due to lack of experimental negative data, we used a prevalence of 0.5 in the accuracy calculations.

2.6. Propensity scales analysis

In order to identify properties that correlate with the weights of the cleavage score, we used the Pearson correlation coefficient to selected propensity scales extracted from the AA index database (Kawashima and Kanehisa, 2000). Multiple SVM based classifiers were tested. In all classifiers, we treated the C and N termini as one region by unifying the appropriate weight vectors. Matrices of 20×20 elements were reduced to 20 elements vectors by averaging columns and rows. The resulting averages were compared to the propensity scales to detect the biochemical elements affecting the classification.

2.7. Feature selection

We performed a feature selection using the following score: $\frac{|\mu_{pos} - \mu_{neg}|}{\sqrt{\sigma_{pos}^2 + \sigma_{neg}^2}}$, where μ_{pos} is the average value of the positive set for a given position and σ_{pos} is the standard deviation in the same position, μ_{neg} and σ_{neg} are the parallel for the negative set. This criterion normalizes the absolute value of the distance between the mean frequencies of specific AA (or diAA) in the positive and negative datasets by dividing it with the square root of the sum of standard deviations of specific AA (or diAA) in the positive and negative datasets. Higher values are more meaningful for the classifier, since they represent clear differences between the positive and negative sets. Note that this score was obviously only computed on the learning set and not on the test set, to avoid any information leakage.

To further reduce the dimensionality, we used both Principle Component analysis (PCA) and Partial Least Squares (PLS) algorithms. Both methods project the data on an alternative space composed of components that have either the largest possible variance (PCA) or the maximal covariance with the labels (i.e., cleaved/not cleaved) (PLS).

3. Results

3.1. Amino acids distribution

We have first tested if the amino acid usage of cleaved peptides contains information about the peptide cleavage probability. If the peptide and flanker sequences do not contain such information, sequence based classifiers will probably fail. The information content was estimated by comparing the differences between the AA composition in peptide bodies, at the flanking regions and outside peptides (Figure 1). The distributions of single AAs in each region were calculated as: A) their frequency inside peptides (except for the first and last positions); B) their frequency around the C and N termini of the peptide (the first position surrounding the cleavage site inside the peptide and K positions outside the peptide); C) the single AA frequency in the same proteins outside the cleaved peptides. The same proteins were used in order to avoid biases that may result from the choice of proteins. Similarly, the distribution of diAA was calculated as the frequency of diAA in the above mentioned regions. A clear difference is observed in the single AA distribution around the cleavage site (Figure 1 middle panel) (Chi square test $p < 5 \cdot 10^{-5}$ for all comparisons), within peptides and outside peptides (Figure 1 upper panel) (chi square test $p < 2 \cdot 10^{-131}$). On the other hand, only minor differences are observed between the single AA composition at the second position after the cleavage site and in random regions farther away from the cleavage site (Figure 1 lower panel) (chi square, $p < 6 \cdot 10^{-7}$). Given the limited effect of these regions, we ignored all positions beyond the first flanking residue (K=1) in the learning score.

Beyond the differences in the single AA composition, a correlation between the diAAs may affect the cleavage probability. Such an effect can be determined through the dependence between the diAA composition distributions at neighboring sites. We measured this dependence using the mutual information (I) and Shannon Entropy (H) of each position. The mutual information between the AAs at the two sides of the cleavage site was higher than within the peptide ($I_N = 0.2726$ out of $H_N = 5.3889$ and $I_C = 0.2904$ out of $H_C = 5.3230$ compared with $I_{PB} = 0.0600$ out of $H_{PB} = 5.6438$), suggesting that preferable cleavage sites exist.

In order to get an insight on the source of the effect of diAA on cleavage, we computed the ratio between the observed diAA probability and the expected probability if one assumes that the left and right positions around the cleavage site are independent around cleavage sites, the distribution of hydrophobic amino acids differs from the expected behavior; while

negatively charged and aliphatic amino acids behave as expected. In the peptide body, the behavior of the vast majority of amino acids is much closer to the expected (Supp Mat. Figure S2). In order to measure the statistical dispersion of the above distributions over the cleavage site and the peptide body, we used the Average Absolute Deviation (AAD) from the median (see Methods). Over the cleavage site the AAD was 0.595, while over the peptide body the AAD was 0.398. These values indicate a greater dispersion over the cleavage site than in the peptide body and emphasize the importance of the cleavage site to the characterization of a possible classifier.

To summarize, both the single and amino acid usage and the diAA usage are different around the cleavage site, within the epitope and outside the epitope. Moreover, the AA distribution at the two sides of the cleavage sites is not independent. We further checked whether the different AA distribution can be used to learn a classifier for phagosomal and lysosomal cleavage.

3.2. Single amino acid scores

MHC class II epitope precursor cleavage probably takes place in multiple compartments. We here denote it as “endosomal” cleavage as a simplification. The cleavage process itself may be affected by the full structure of the protein, since proteins are not fully denatured before their cleavage in endosomal compartments (in contrast with proteasomal cleavage in the cytosol) (Collins et al., 1991). However, in the current analysis, we are only interested in relatively short peptides (up to 30 amino acids) that were observed to bind MHC class II molecules (Sercarz and Maverakis, 2003). We assume that within such peptides the secondary structure leaves most AAs exposed to cleavage, and we thus use a sequence-based and not a structure-based score. Note that we do not propose a cleavage site prediction algorithm; rather we proposed a combined score to classify full peptides as either the result of endosomal cleavage or not. The score is based on the peptide sequence as well as the C and N termini flanking residues. We used the end-product of the endosomal process (i.e., naturally processed CD4+ T cell epitopes) as a learning set for the cleavage process.

Cleaved peptide must theoretically fulfill three conditions: cleavage at the C and N termini and no cleavage within the peptide body. We thus propose a score based on the residues in the body of the peptide and in the C and N termini flanking residues. The body of the peptide was defined as all positions, except for the first and the last positions in the peptide. The C and N termini flanking residues are the first positions within and outside the peptide surrounding the cleavage site. We propose a scoring method that contains these three elements: the contribution of the C terminal, the contribution of the N terminal, and the contribution of the body of the peptide (Figure 2). Formally, we assume that:

$$P(\text{cleaved peptide}) = P(C)P(N) \prod_i (1 - P(\text{pep}_i)), \quad (1)$$

where $P(C)$ and $P(N)$ are the cleavage probabilities at the C and N termini, and $P(\text{pep}_i)$ is the cleavage probability at position i of the peptide body. The independence assumption allows for the translation to a sum by taking the log of the two sides of the equation to produce:

$$\begin{aligned} S(\text{peptide}) &= \ln P(\text{cleaved peptide}) = \ln P(C) + \sum_i \ln(1 - P(\text{pep}_i)) + \ln P(N) \\ &= S_C(C) + \sum_i S_P(\text{pep}_i) + S_N(N) \end{aligned} \quad (2)$$

The goal of the machine learning algorithms proposed in the following sections is to produce estimates of $S_C(C) = \ln(P(C))$, $S_P(\text{pep}_i) = \ln(1 - P(\text{pep}_i))$ and $S_N(N) = \ln(P(N))$

(Figure 2). As a first assumption, we test the possibility that the cleavage probability at a given site is the product of the effect of the AA to the left and right of the cleavage site, which leads to:

$$S(\text{peptide}) = S_{CL}(AA_{CL}) + S_{CR}(AA_{CR}) + \sum_i S_P(AA_{Pi}) + S_{NL}(AA_{NL}) + S_{NR}(AA_{NR}), \quad (3)$$

where $S_{CL}(AA_{CL})$ is the contribution of the position to the left of the C terminal to the cleavage, and all other elements follow the same notation. If we further assume that the behavior at the C and N termini are similar, Eq 3 simplifies to:

$$S(\text{peptide}) = S_{C/NL}(AA_{CL}) + S_{C/NR}(AA_{CR}) + \sum_i S_P(AA_{Pi}) + S_{C/NL}(AA_{NL}) + S_{C/NR}(AA_{NR}) \quad (4)$$

where $S_{C/NL}(AA_{CL})$ is the contribution of the position to the left of the C terminal to the cleavage, but the score is similar for the C and N termini.

As is the case for many machine learning problems, a tradeoff must be obtained between the complexity of the score used and the precision obtained during the training stage. We thus tested classifiers with scores of varying complexity. The complexity of the scores is expressed either by the number of features used for the score or in the scoring function itself. The simplest possibility is a score incorporating only the AA distribution, with either different (Eq. 3) or similar (Eq. 4) scores for the C and N termini. The single AA composition exhibits a well-defined distribution for each region (Figure 1), and is a reasonable candidate for a classifier. A Support Vector Machine (SVM) based classifier was built with either linear or quadratic kernel functions. We tested classifiers learnt with different sizes of the negative learning set, and different costs for the classification error of the learning set (C) (see Methods for the precise definition). The best accuracy obtained was 0.75 with sensitivity and specificity of 0.69 and 0.82, respectively. The best sensitivity obtained was 0.81 with specificity and accuracy of 0.64 and 0.73, respectively. The best specificity obtained was 0.99 with sensitivity and accuracy of 0.3 and 0.65, respectively (Table 1).

The main two models of peptide processing in the MHC class II pathway are either that proteins are first cut in their natural form and then the resulting peptides bind MHC class II molecules ('cut first, bind later' model) (Sette et al., 1989; Sercarz and Maverakis, 2003), or that proteins are first partially cut. The resulting long peptides bind MHC-II molecules and then are further trimmed ('bind first, trim later model'/'MHC-guided processing') (Sercarz and Maverakis, 2003). If full cleavage precedes binding, then peptides bound to MHC class II molecules are the product of fully degraded proteins. If this is indeed the case, it would be reasonable to assume similar hydrolyses activity for the N and C termini. This would lead to a similar computed cleavage probability for both termini. However, if the MHC-guided processing model is correct, then a stage of binding between denatured proteins and the MHC molecule exists prior to trimming by exopeptidases. Exopeptidases are roughly divided into two basic families: Aminopeptidases and Carboxypeptidases, and it is only reasonable to assume that the trimming probabilities differ depending on the termini. While the amino acid distribution around the C and N termini were slightly different, using scores based on different cleavage properties for the N and C termini leads to no significant improvement in the precision. Thus, the observed AA distribution around cleavage sites does not explicitly support a MHC-guided processing model.

If the N and C termini do not have the same effect on the cleavage probability, then the score should contain 40 more parameters for the different contribution of the AA flanking the N and C termini, leading to the 100 parameter classifier in Eq. 3, instead of the classifier

of Eq. 4. We have tested such a classifier and obtained a slightly better accuracy than the classifier with N and C termini united (Table 1). We thus do not observe a significant difference between the cleavage properties at the C and N termini.

The three scores produced in the above section highlight important aspects of the cleavage process. When looking at the score separating the C and N cleavage sites, one observes a high correlation between the two sites ($r=0.42$, $p<0.01$). Most of the difference is due to positions with a limited contribution to the cleavage, leading again to the conclusion that the cleavage process is approximately similar in C and N.

If one averages the scores of the C and N termini, one obtains again a 60 parameter score, which is very similar to the 60 parameter scores obtained with and without the *in vitro* cleavage (Figure 3). Since the different scores are quite similar, we averaged them and compared the contribution of each amino acid to the score. The most important contribution is from the cleavage sites, while the body of the peptide has a more limited contribution to the variability of the score, with the main limitation being that a Cysteine in the body of the peptide significantly increases the probability that the peptides should be produced and not cleaved in its center (Figure 4).

At the right of the cleavage site, there is a clear negative contribution of hydrophilic amino acids and a positive contribution of hydrophobic amino acids. This contribution is well demonstrated by correlation of cleavages introduced by individual cathepsins with the more hydrophobic regions within the HEL sequence (Pluger et al., 2002). On the left side of the cleavage, charged amino acids have a negative contribution to the cleavage, while more neutral or aliphatic amino acids increase the cleavage probability.

We further examined the contribution of amino acids at a distance of more than 1 from the cleavage site to the peptide cleavage probability. We repeated the analysis described in the previous section, and extended the score to include flanking regions up to 5 amino acids away from the cleavage site at each side (Supp. Mat., Figure S3). However, no significant improvement in the learning performance was shown (Accuracy of 0.682).

3.3. Amino Acid Pair score

In order to further improve the accuracy of the classifier, we tested scores incorporating not only the single AA distribution, but also the diAA distribution. Note that enlarging the score to include diAA increases the complexity of the problem from 20/40 parameters to 400 variables in each region. The resulting score is:

$$S(\text{peptide}) = S_C(AA_{CL}, AA_{CR}) + \sum_i S_P(AA_{Pi}, AA_{Pi+1}) + S_N(AA_{NL}, AA_{NR}). \quad (5)$$

If one assumes again that the behaviors at the C and N termini are similar, Eq. 5 simplifies to:

$$S(\text{peptide}) = S_{C/N}(AA_{CL}, AA_{CR}) + \sum_i S_P(AA_{Pi}, AA_{Pi+1}) + S_{C/N}(AA_{NL}, AA_{NR}). \quad (6)$$

3.4. Inclusion of *in vitro* cleavage experiments

A second element introduced to further improve the precision of the algorithm is endosomal proteases cleavage sites, as determined by *in vitro* experiments (Rawlings and Barrett, 1999). These *in-vitro* experiments were merged with the current data (naturally processed

CD4+ T cell epitopes). These two data sources are independent and each provide a different type of information. *In vitro* cleavage experiments provide information about the activity of specific proteases. However, such experiments do not provide information about the relative importance of each protease in the combined cleavage process. On the other hand, naturally processed CD4+ T cell epitopes provide a comprehensive end product of the cleavage process, but they are affected by other processes, mainly the MHC class II binding, and as such may be biased. We created a score based on the average diAA frequency in the peptide body (20X20 parameters) and the flanking regions (unified C and N – 20X20 parameters), yielding an overall dimension of 800. The combination of the two data sources produced a better classifier than is obtained by each separately (Table 1). In this case again the SVM results were better than the ones obtained in the SA analysis (data not shown). The SVM algorithm was used with varying error cost factors: 10^{-6} C 10^0 , negative training data set size: 1000 and 3000 peptides. The SVM classifier was applied either with or without the *in vitro* cleavage data and with linear or quadratic kernels. Within the different SVM, the accuracy of linear kernel functions was lower than the quadratic kernel function. These best classification results are summarized in Table 1. The best classifiers are available at http://peptibase.cs.biu.ac.il/PepCleave_cd4/

3.5. Optimal classifier with reduced dimension

The gain obtained in the learning accuracy may be offset by the loss in the generalization resulting from the large number of parameters used. We tested whether an optimal balance between the learning and generalization by testing scores with different numbers of

variables. We ranked the features by the value of $\frac{\mu_{pos} - \mu_{neg}}{\sqrt{\sigma_{pos}^2 + \sigma_{neg}^2}}$ in descending order. An SVM based classifier was built with either linear or quadratic kernel functions with a gradually increasing number of features. We tested classifiers learnt with different sizes of the negative learning set, and different costs for the classification error of the learning set (C) (see Methods for a precise definition). Features were taken from the top of the ranking list created by the above criterion. The thresholds for each classifier were set to maximize accuracy. The best accuracy was achieved with 240 out 800 features and a linear classifier with *in vitro* data included (accuracy of 0.77 with sensitivity and specificity of 0.82 and 0.71 relatively (Table 2); threshold = -0.2023). While this score is less natural from an immunological point of view, since it ignores multiple diAA combinations, given the amount of existing observations, it is one of the best that can be produced. Using data extraction algorithms (see Methods) to further reduce dimensionality didn't improved the accuracy (Figure S4).

3.6. Effect of MHC class II binding

The predicted difference between the cleaved peptides and the non cleaved was based on MHC class II naturally processed eluted peptides. One could claim in principle that the source of the difference is due to the MHC class II binding. This claim is not very probable since, in contrast with MHC class I, the binding core of MHC class II epitopes is limited to a small part of the peptide. Thus, many different cleaved peptides can prime the appropriate T cells. However, in order to ensure that the results of the algorithm are not due to MHC class II binding, we used the learnt classifier (linear SVM learning on 240 out of 800 features description) to evaluate the score of MHC class II binding peptides. The flanking regions of 6169 peptides were extracted from their source protein, and only peptides with a single source protein were used. The fraction of MHC class II binding peptide positives was not significantly different than the fraction of random peptides ((sensitivity of 0.3197 (Table 2)); Figure 5).

3.7. T Cell response epitopes

An interesting question emerging in the context of a classifier is the predicted score of T cell activating peptides (epitopes). In the case of MHC class I molecules, an externally introduced epitope is considered to be naturally processed, since otherwise there would be no T cell clone responding to it. Thus T cell epitopes indeed have high predicted proteasomal cleavage and TAP binding scores as one can see in the case of HIV and Herpes Viruses (Vider-Shalit et al., 2007) (Vider-Shalit et al., 2009). However, in the case of MHC class II, this correlation may break, since binding is based on the properties of the typically nine-amino acid long. The cleavage, on the other hand, depends on positions that may be far from this nine amino acid binding site. Thus, T cell response (of an externally administered peptide) and cleavage may be independent in the case of MHC class II. To test if this is indeed the case, we repeated the analysis performed for MHC class II binding peptides, using T cell epitopes, and indeed there is no difference between the fraction of T cell epitopes defined to be positive in the cleavage score and the fraction of random peptides defined to be positive (Figure 5).

To conclude, the low sensitivity of positive MHC-II binders and T cell responders indicate that the classifier is indeed based on cleavage features. To validate that this observations are not sensitive to the threshold we used (-0.2023), we compared the sensitivity of the same data sets with different thresholds (Figure S5). For any threshold we used, the sets of MHC-II binders and CTL responders had low sensitivity, similar to the random set.

3.8. Biochemical elements affecting cleavage

In order to understand the cleavage mechanism, we have looked for correlations between the score obtained in the previous section and multiple biochemical properties. We have here focused on the cleavage score that yielded the optimal accuracy (240 out of 800 variables). The propensity scales relate to single AA and the weight vector are based on diAA. We have thus averaged the weight vector over all positions on the left of the diAA, yielding a length 20 vector. A parallel analysis was then performed when averaging on the position to the right of the cleavage site. The resulting vectors all showed a high correlation to structural properties of amino acids, such as the linker propensity, which is determined by the rigidity and hydrophobicity of the amino acids (George and Heringa, 2002). The highest correlations coefficient calculated are: for the left side of cleavage site with Linker propensity from non-helical (annotated by DSSP) dataset (George and Heringa, 2002) and with Linker propensity from 1-linker dataset (George and Heringa, 2002) $r=0.803$ $p<0.005$ and $r=0.763$, $p<0.005$, for the right side of cleavage site $r=-0.627$, $p<0.005$, for the left side of diAAs in peptide body with Linker propensity from 3-linker dataset (George and Heringa, 2002) $r=-0.619$, $p<0.005$ and for the left side of diAAs in peptide body with Linker propensity from 3-linker dataset (George and Heringa, 2002) $r=-0.694$, $p<0.005$. Multiple other measures were highly correlated with the score, most of which are related to the secondary structure of the peptide.

These correlations highlight the importance of secondary structural aspects and the available surface area for the cleavage process. The structural effect is probably the most important difference between proteasomal cleavage operating on fully denatured peptides and endosomal cleavage affecting natural proteins. The proposed score was fully sequence based. We expect structure based scores to provide an even better accuracy. However, the limited precision of structure prediction algorithms and the number of features describing the structure make such an algorithm more complex, and would probably require significantly larger learning sets.

3.9. Relation between body of peptide and cleavage site

The AA that have the greatest impact on the cleavage score (highest absolute weight in the cleavage score) are Arg, Gly and Lys on the left side, and Ala, Gly and Ser on the right side. Since the actual cleavage site profile is a consequence of the peptidases participating in the cleavage process, the calculated cleavage profile site should express these peptidases substrates.

We have computed the correlation between the contributions of each amino acid to the experimental profile of the two peptidases shown to have a distinct roles in the endosomal presentation *in vivo*: cathepsins L and S (Villadangos and Ploegh, 2000) and the score vector. We found a significant correlation for the left side of cleavage site, but not for the right side of the cleavage site. We have no clear explanation at this stage for this difference.

4. Discussion

During the past years several tools have been developed for predicting various stages in the processing of T cell epitopes. An important exception in this field is the preprocessing of MHC class II binding peptides, which has not been extensively studied. This major step in CD4+ T cell epitope presentation is essential for the proper prediction of CD4+ T cell epitopes. MHC class II binding peptides will not become epitopes if they are not cleaved beforehand. We have here shown that naturally processed epitopes and *in vitro* cleavage measurements databases provide enough information to produce for the first time a precise sequence based MHC class II peptide cleavage prediction. A possible reason for the lack of previous prediction algorithms is that the cleavage process producing candidate peptides for MHC class II binding is not well-defined. Proteins can be engulfed by multiple types of APC in different ways (endocytosis, phagocytosis, etc.). These proteins are then passed through multiple compartments, each with a different composition of hydrolases. Thus, the preprocessing of peptides is not a single biological process, but rather an ensemble of processes that eventually leads to the presence of peptides in the phagolysosome. We here avoid the confusion resulting from this combination of processes by lumping the preprocessing stage into a generic event of peptide production, and apply machine learning techniques to predict which peptides can indeed result from this process. Our results show that the current amount of observations on the end product of this process is large enough to compensate for the lack of knowledge on the precise process affecting each peptide. Since the creation of an epitope requires cleavage at certain points and maintaining continuity in others, we tried to identify a different composition at the edges of epitopes and in the body of the epitopes and use this different composition to define preprocessed peptides. Two approaches have been used: either an over-simplified score with a limited number of parameters, or a more complex score with up to 1,200 parameters. Our results show that an optimal balance is obtained with a score with 240 parameters.

Three main caveats limit the accuracy obtained by the machine learning algorithms, the main one being the usage of a sequence based score when the cleavage process may be affected by the full structure of the protein. A second limitation is the linearity assumption of the score. This is equivalent to the assumption that the cleavage process at each amino acid is an independent process. Finally, the learning set is based on epitopes that are also affected by the properties of the MHC class II molecules. Even given these limitations, an accuracy of almost 0.8 has been obtained using this score.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Tal Vider and Prof. Ron Unger for their helpful suggestions, Lea Tsaban for the help in the data accumulation and Miriam Beller for the English editing.

Bibliography

- Alfonso C, Karlsson L. Nonclassical MHC class II molecules. *Annual Review of Immunology*. 2000; 18:113–142.
- Berger AC, Roche PA. MHC class II transport at a glance. *Journal of Cell Science*. 2009; 122:1. [PubMed: 19092054]
- Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*. 2004; 20:421–3. [PubMed: 14960470]
- Boser, BE.; Guyon, IM.; Vapnik, VN. A training algorithm for optimal margin classifiers. *ACM New York; NY, USA*: 1992. p. 144–152.
- Brusic V, Petrovsky N, Zhang G, Bajic VB. Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol*. 2002; 80:280–5. [PubMed: 12067415]
- Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K-A, Mothé BR, Chisari FV, Watkins DI, Sette A. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*. 2005; 57:304–314. [PubMed: 15868141]
- Busch R, Rinderknecht CH, Roh S, Lee AW, Harding JJ, Burster T, Hornell TM, Mellins ED. Achieving stability through editing and chaperoning: regulation of MHC class II peptide binding and expression. *Immunological reviews*. 2005; 207:242–260. [PubMed: 16181341]
- Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S. Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*. 2003; 62:378–84. [PubMed: 14617044]
- Claus V, Jahraus A, Tjelle T, Berg T, Kirschke H, Faulstich H, Griffiths G. Lysosomal enzyme trafficking between phagosomes, endosomes, and lysosomes in J774 macrophages. *Journal of Biological Chemistry*. 1998; 273:9842. [PubMed: 9545324]
- Collins DS, Unanue ER, Harding CV. Reduction of disulfide bonds within lysosomes is a key step in antigen processing. *The Journal of Immunology*. 1991; 147:4054. [PubMed: 1721638]
- Desjardins M, Huber LA, Parton RG, Griffiths G. Biogenesis of phagolysosomes proceeds through a sequential series of interactions with the endocytic apparatus. *Journal of Cell Biology*. 1994; 124:677. [PubMed: 8120091]
- Donnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*. 2002; 3:25. [PubMed: 12225620]
- Feldhahn M, Thiel P, Schuler MM, Hillen N, Stevanovic S, Rammensee H-G, Kohlbacher O. EpiToolKit—a web server for computational immunomics. *Nucleic Acids Research*. 2008; 36:W519–W522. [PubMed: 18440979]
- George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng*. 2002; 15:871–9. [PubMed: 12538906]
- Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*. 2008; 24:477. [PubMed: 18216070]
- Hiltbold EM, Roche PA. Trafficking of MHC class II molecules in the late secretory pathway. *Current Opinion in Immunology*. 2002; 14:30–35. [PubMed: 11790530]
- Hsing LC, Rudensky AY. The lysosomal cysteine proteases in MHC class II antigen presentation. *Immunological reviews*. 2005; 207:229–241. [PubMed: 16181340]
- Jensen PE. Recent advances in antigen processing and presentation. *Nature immunology*. 2007; 8:1041–1048. [PubMed: 17878914]
- Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Research*. 2000; 28:374. [PubMed: 10592278]
- Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering Design and Selection*. 2002; 15:287.

- Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC immunology*. 2008; 9:8. [PubMed: 18366636]
- Lotteau, V.; Teyton, L.; Peleraux, A.; Nilsson, T.; Karlsson, L.; Schmid, SL.; Quaranta, V.; Peterson, PA. Intracellular transport of class II MHC molecules directed by invariant chain. 1990.
- Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B. An Introduction to Kernel-Based Learning Algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS*. 2001; 12:181. [PubMed: 18244377]
- Meister GE, Roberts CGP, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. *Vaccine*. 1995; 13:581–591. [PubMed: 7483779]
- Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*. 2009; 10:296. [PubMed: 19765293]
- Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O. Quantitative Predictions of Peptide Binding to Any HLA-DR Molecule of Known Sequence: NetMHCIIpan. *PLoS Comput Biol*. 2008; 4:e1000107. [PubMed: 18604266]
- Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*. 2007; 8:238. [PubMed: 17608956]
- Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*. 2004; 20:1388–97. [PubMed: 14962912]
- Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. 2003; 12:1007–17. [PubMed: 12717023]
- O'Garra A, Arai N. The molecular basis of T helper 1 and T helper 2 cell differentiation. *Trends in Cell Biology*. 2000; 10:542–550. [PubMed: 11121747]
- Parker KC, Bednarek MA, Hull LK, Utz U, Cunningham B, Zweerink HJ, Biddison WE, Coligan JE. Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. *J Immunol*. 1992; 149:3580–7. [PubMed: 1331239]
- Plüger EBE, Boes M, Alfonso C, Schröter CJ, Kalbacher H, Ploegh HL, Driessen C. Specific role for cathepsin S in the generation of antigenic peptides in vivo. *European journal of immunology*. 2002; 32:467–476. [PubMed: 11813165]
- Ramachandra L, Noss E, Harding CV. Phagocytic processing of antigens for presentation by class II major histocompatibility complex molecules. *Cellular Microbiology*. 1999; 1:205–214. [PubMed: 11207553]
- Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanovi S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999a; 50:213–219. [PubMed: 10602881]
- Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanovi S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999b; 50:213–219. [PubMed: 10602881]
- Rathman M, Barker LP, Falkow S. The unique trafficking pattern of *Salmonella typhimurium*-containing phagosomes in murine macrophages is independent of the mechanism of bacterial entry. *Infection and immunity*. 1997; 65:1475. [PubMed: 9119490]
- Rawlings ND, Barrett AJ. MEROPS: the peptidase database. *Nucleic Acids Research*. 1999; 27:325. [PubMed: 9847218]
- Reche PA, Glutting JP, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol*. 2002; 63:701–9. [PubMed: 12175724]
- Reche PA, Glutting JP, Zhang H, Reinherz EL. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*. 2004; 56:405–19. [PubMed: 15349703]
- Riese RJ, Chapman HA. Cathepsins and compartmentalization in antigen presentation. *Current Opinion in Immunology*. 2000; 12:107–113. [PubMed: 10679409]

- Sercarz EE, Maverakis E. MHC-guided processing: binding of large antigen fragments. *Nature Reviews Immunology*. 2003; 3:621–629.
- Sette A, Adorini L, Colon SM, Buus S, Grey HM. Capacity of intact proteins to bind to MHC class II molecules. *The Journal of Immunology*. 1989; 143:1265. [PubMed: 2787361]
- Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. *Bioinformatics*. 2001; 17:1236–7. [PubMed: 11751237]
- Singh H, Raghava GP. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*. 2003; 19:1009–14. [PubMed: 12761064]
- Sturgill-Koszycki S, Schaible UE, Russell DG. Mycobacterium-containing phagosomes are accessible to early endosomes and reflect a transitional state in normal phagosome biogenesis. *The EMBO journal*. 1996; 15:6960. [PubMed: 9003772]
- Vider-Shalit T, Almani M, Sarid R, Louzoun Y. The HIV hide and seek game: an immunogenomic analysis of the HIV epitope repertoire. *Aids*. 2009; 23:1311. [PubMed: 19550286]
- Vider-Shalit T, Fishbain V, Raffaelli S, Louzoun Y. Phase-dependent immune evasion of herpesviruses. *Journal of virology*. 2007; 81:9536–9545. [PubMed: 17609281]
- Vider-Shalit T, Louzoun Y. MHC-I prediction using a combination of T cell epitopes and MHC-I binding peptides. *J Immunol Methods*. 2011; 374:43–6. [PubMed: 20920507]
- Villadangos JA, Ploegh HL. Proteolysis in MHC Class II Antigen Presentation Who's in Charge? *Immunity*. 2000; 12:233–239. [PubMed: 10755610]
- Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The immune epitope database 2.0. *Nucleic Acids Research*. 38:D854. [PubMed: 19906713]
- Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. *PLoS Comput Biol*. 2008; 4:e1000048. [PubMed: 18389056]
- Wolf PR, Ploegh HL. How MHC class II molecules acquire peptide cargo: biosynthesis and trafficking through the endocytic pathway. *Annual review of cell and developmental biology*. 1995; 11:267–306.
- Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res*. 2005; 33:W172–9. [PubMed: 15980449]

Figure 1. Distribution of single AA frequencies in different regions

The upper panel represents the peptide body AA composition of naturally processed sequences vs. not naturally processed sequences. The peptide backbone is defined as the amino acid sequence of naturally processed sequences (not including in the cleavage site). These results are compared to regions on the same protein at a distance of 3 to 50 AA flanking the naturally processed sequence. The middle panel represents the AA distribution on the left and right of N and C termini cleavage sites. The lower panel represents the AA distribution at the second position away from the cleavage site and the average over positions up to 50 positions away from the cleavage site. No significant difference was found between the distributions in the last panel (Kullback-Leibler divergence of 0.0087). Thus, only the peptide itself and the first position flanking the cleavage site were used in the analysis.

Figure 2. Schematic view of a sample epitope and the applied score

P1 designates the N terminus cleavage site constructed from the last amino acid of the N terminus flanking region and the first amino acid of the peptide (deep gray in left lens). P2 designates the peptide body without the first and last amino acids (light gray letters). P3 designates the C terminus cleavage site constructed from the last amino acid of the peptide and the first amino acid the C terminus flanking region (deep gray letters in right lens).

Figure 3. SVM predictor based on either the 100 parameter score (distinct C and N cleavage sites), the 60 position score (unified C and N cleavage sites) with the in vitro cleavage and the 60 position score without the in vitro cleavage
 In the 100 parameter score, we have averaged the C and N cleavage score (after the full 100 parameter score was produced). One can clearly see that the scores are highly similar.

Figure 4. Contribution of each amino acid at each position (left of cleavage site, right of cleavage site, and peptide body) to the classifier score

The score presented here is the average over the three scores presented in Figure 3.

Figure 5. Sensitivity rate of different data sets calculated using the best yielded classifier of 240 out of 800 variables with linear kernel, learning data included cleavage data
The classifier yielded in the cleavage data used to assess the sensitivity (at threshold -0.2023) of the test set, negative data, T cell response data and binding data.

Table 1

Performance of classifiers

The first column is the number of variables used. The second column is the SVM kernel. The third column represents the presence or absence of in vitro cleavage data in the learning process. The fourth column is the number of samples in the negative learning set. The fifth column is the error coefficient cost in the SVM. The sixth column is the area under curve (AUC). The seventh, eighth and ninth columns are the accuracy, specificity and, sensitivity of each score relatively. The bolded value at each row designates the statistical measure optimized by the specific parameter set. Cost (C) values were set between 10^{-6} and 10^0 , and the positive to negative learning set size was varied between 1000 and 3000 peptides.

Features	Kernel	In vitro	Neg. data size	C	AUC	Accuracy	Specificity	Sensitivity
60	linear	-	1000	10^{-02}	0.81	0.75	0.68	0.82
60	linear	-	3000	1	0.78	0.7	0.71	0.68
60	linear	-	1000	10^{-02}	0.81	0.75	0.68	0.82
60	linear	+	3000	10^{-02}	0.81	0.75	0.7	0.8
60	linear	+	3000	1	0.81	0.73	0.73	0.73
60	linear	+	1000	10^{-02}	0.82	0.75	0.68	0.81
60	quadratic	-	3000	0	0.77	0.73	0.81	0.65
60	quadratic	-	3000	10^{-02}	0.7	0.64	0.84	0.45
60	quadratic	-	1000	10^{-06}	0.76	0.67	0.69	0.65
60	quadratic	+	2000	0	0.77	0.71	0.8	0.61
60	quadratic	+	2000	0	0.77	0.71	0.8	0.61
60	quadratic	+	1000	10^{-06}	0.77	0.68	0.66	0.69
100	linear	-	1000	10^{-02}	0.81	0.74	0.68	0.81
100	linear	-	2000	0	0.78	0.71	0.72	0.7
100	linear	-	1000	10^{-02}	0.81	0.74	0.68	0.81
100	quadratic	-	1000	10^{-06}	0.77	0.68	0.65	0.72
100	quadratic	-	3000	10^{-02}	0.75	0.65	0.86	0.43
100	quadratic	-	1000	10^{-06}	0.77	0.68	0.65	0.72
420	linear	-	2000	1	0.78	0.72	0.72	0.71
420	linear	-	3000	10^{-06}	0.75	0.68	0.75	0.61
420	linear	-	1000	10^{-02}	0.76	0.7	0.67	0.73
420	linear	+	3000	10^{-06}	0.74	0.67	0.75	0.59

Features	Kernel	In vitro	Neg. data size	C	AUC	Accuracy	Specificity	Sensitivity
420	linear	+	1000	10^{-02}	0.76	0.7	0.67	0.73
420	linear	+	2000	1	0.78	0.72	0.72	0.71
420	quadratic	-	2000	10^{-02}	0.74	0.67	0.85	0.48
420	quadratic	-	3000	0	0.73	0.61	0.92	0.3
420	quadratic	-	1000	10^{-06}	0.77	0.6	0.6	0.61
420	quadratic	+	3000	0	0.75	0.65	0.92	0.38
420	quadratic	+	3000	0	0.75	0.65	0.92	0.38
420	quadratic	+	1000	10^{-06}	0.77	0.63	0.59	0.67
800	linear	-	2000	0	0.82	0.74	0.8	0.68
800	linear	-	2000	10^{-06}	0.77	0.68	0.81	0.56
800	linear	-	1000	1	0.79	0.73	0.71	0.75
800	linear	+	2000	10^{-02}	0.8	0.74	0.77	0.71
800	linear	+	3000	1	0.79	0.71	0.81	0.61
800	linear	+	1000	1	0.79	0.73	0.7	0.75
800	quadratic	-	1000	10^{-02}	0.79	0.75	0.82	0.69
800	quadratic	-	3000	0	0.8	0.65	0.99	0.3
800	quadratic	-	1000	10^{-02}	0.79	0.75	0.82	0.69
800	quadratic	+	1000	10^{-06}	0.83	0.74	0.73	0.75
800	quadratic	+	3000	10^{-02}	0.72	0.67	0.98	0.37
800	quadratic	+	1000	1	0.78	0.73	0.64	0.81
1200	linear	-	3000	10^{-06}	0.82	0.74	0.82	0.66
1200	linear	-	3000	10^{-02}	0.76	0.67	0.82	0.53
1200	linear	-	1000	10^{-06}	0.8	0.73	0.76	0.7
1200	quadratic	-	1000	1	0.82	0.72	0.75	0.68
1200	quadratic	-	3000	10^{-02}	0.78	0.66	0.98	0.33
1200	quadratic	-	1000	1	0.82	0.72	0.75	0.68

Table 2

Performance of reduced classifiers

The first column is the maximal possible number of features. The second column is the SVM kernel. The third column represents the presence or absence of in vitro cleavage data in the learning process. The fourth column is the actual number of features used in the learning process. The fifth column is the number of samples in the negative learning set. The sixth column is the error coefficient cost in the SVM. The seventh column is the area under curve (AUC). The eighth, ninth and tenth columns are the accuracy, specificity and sensitivity, of each score relatively. The two last columns are the sensitivity of T cell response data and negative data. Cost (C) values were set between 10^{-6} and 10^0 , and the positive to negative learning set size was varied between 1000 and 3000 peptides.

Max. Features	Kernel	In vitro	Actual Features	Neg. data size	C	AUC	Acc.	Spe.	Sen.	TCR Sen.	Neg. Sen.
100	linear	-	20	1000	1^{-04}	0.78	0.72	0.65	0.79	0.41	0.35
100	quadratic	-	20	1000	1^{-04}	0.78	0.71	0.53	0.88	0.49	0.48
1200	linear	-	250	3000	1	0.85	0.77	0.71	0.84	0.32	0.30
1200	quadratic	-	80	3000	1^{-04}	0.81	0.72	0.67	0.77	0.34	0.31
420	linear	-	170	3000	1	0.82	0.74	0.66	0.83	0.36	0.34
420	linear	+	170	3000	1	0.82	0.75	0.65	0.84	0.36	0.35
420	quadratic	-	80	1000	1^{-04}	0.80	0.74	0.69	0.80	0.34	0.31
420	quadratic	+	80	1000	1^{-04}	0.80	0.73	0.66	0.80	0.36	0.35
60	linear	-	40	2000	0.01	0.79	0.72	0.65	0.78	0.37	0.36
60	linear	+	40	2000	0.01	0.79	0.72	0.65	0.79	0.36	0.38
60	quadratic	-	50	1000	1^{-04}	0.77	0.71	0.69	0.72	0.32	0.28
60	quadratic	+	50	1000	1^{-04}	0.78	0.71	0.67	0.75	0.33	0.31
800	linear	-	130	2000	0.01	0.82	0.75	0.68	0.83	0.35	0.33
800	linear	+	240	3000	0.01	0.85	0.77	0.71	0.82	0.31	0.29
800	quadratic	-	80	3000	1^{-04}	0.81	0.73	0.67	0.80	0.34	0.30
800	quadratic	+	90	2000	1^{-04}	0.81	0.73	0.68	0.78	0.35	0.31