Peptide Binding at Class I Major Histocompatibility Complex Scored with Linear Functions and Support Vector Machines

Henning Riedesel
riedesel@chemie.fu-berlin.de

Björn Kolbeck
bjko@chemie.fu-berlin.de

Oliver Schmetzer
o.schmetzer@mdc-berlin.de

Ernst-Walter Knapp
knapp@chemie.fu-berlin.de

Institute of Chemistry, Free University of Berlin, Takustrasse 6, Berlin 14195, Germany

Abstract

We explore two different methods to predict the binding ability of nonapeptides at the class I major histocompatibility complex using a general linear scoring function that defines a separating hyperplane in the feature space of sequences. In absence of suitable data on non-binding nonapeptides we generated sequences randomly from a selected set of proteins from the protein data bank. The parameters of the scoring function were determined by a generalized least square optimization (LSM) and alternatively by the support vector machine (SVM). With the generalized LSM impaired data for learning with a small set of binding peptides and a large set of non-binding peptides can be treated in a balanced way rendering LSM more successful than SVM, while for symmetric data sets SVM has a slight advantage compared to LSM.

Keywords: major histocompatibility complex, peptide binding, separating hyperplane, support vector machine, learning and predicting, scoring function

1 Introduction

Every adaptive immune reaction is based on the specific detection of foreign substances by lymphocytes. These lymphocytes than destroy infected cells and/or stimulate an antibody response, which generally leads to the complete removal of an invading microorganism from the body. Absolutely essential for such a successful immune response is the presentation of the foreign substances, which are in most cases peptides derived for instance from a replicating virus. These peptides are generated from the proteasome and transported to the endoplasmatic reticulum, where they are loaded in the major histocompatibility complex (MHC) [11, 13, 19]. This complex together with the peptide is transferred to the cell surface and can be recognized by T-cells via the T-cell-receptor (TCR) [12, 28]. Without presentation of peptides, no immune response against viruses can be initiated which leads to death of the organism and is the strategy of many pathogens [29]. Not all peptides can be presented in the MHC. The binding depends on so-called anchor-amino-acids, which bind often with low specificity to the MHC, leaving the residual peptide exposed to the TCR [2, 10].

The development of vaccines, immunotherapies and the understanding of a pathogen crucially depend on the know-ledge of the immuno-dominat peptides from a target organism. Identification of these peptides can be done by binding assays in vitro after all possible peptides have been synthesized [16, 25]. This is an extremely expensive approach, because even a very small virus encodes a considerable number of medium size proteins. For each of these proteins hundreds of peptides have to be synthesized and their ability to bind at the MHC must be probed in experiment. This often shows that only very few peptides can indeed bind to the MHC and that from thousands of screened peptides only one or two bind with high affinity, which is required for a functional immune response.

To simplify the search for immuno-dominant peptides, several groups collected data of peptides that bind at MHC to generate a database, which can serve as starting point of computer-based methods to predict the ability of peptides to bind at MHC in silico. These approaches can help to reduce the number of peptides, which have to be tested in vitro. The most often used database of MHC binding peptides is the SYFPEITHI-database (SYF) [24]. Another database for MHC binding peptides that offers however no prediction scheme is MHCPEP (PEP) [6]. The SYF database refers to published data only. It contains about 3,500 MHC binding peptides, which are natural ligands to T-cell epitopes. The MHCPEP database is with about 13,000 MHC binding peptides considerably larger than SYF but may be less reliable, since it allows also for direct submission of data.

Generally, there are sequence based and structure based approaches to predict the ability of peptides to bind at the MHC. The latter uses X-ray structures of MHC or even better of the MHC-peptide complex as a starting point to model the binding geometry of different peptides [26]. The structure based approach has the advantage to require only knowledge of one or at most a few crystal structures to study the peptide binding and provides a deeper understanding of the importance of specific interactions between peptides and the MHC. For peptides that bind to the MHC HLA-A*0201 it is evident from crystal structures that the binding peptides are often nonamers, which possess typically two hydrophobic anchor residues Lys at position 2 and Val at position 9 [27]. This knowledge was a starting point to design empirical scoring functions that use also informations from sequence databases [16]. However, a disadvantage of the structure based approach is the difficulty to estimate an error margin.

More recently, a number of theoretical groups have employed bioinformatic methodology to predict the ability of peptides to bind at MHC based on sequence information. Among these methods are neural networks [7, 14], hidden Markov models [5, 20, 30] and methods based on scoring functions that are optimized by least square fitting [23] or by using the support vector machine [9, 15]. A recent extensive comparison of different methods can be found in Ref. [31]. In this study, we tried to explore a method to predict immuno-dominant epitopes by using a most simple approach employing a linear scoring function in sequence space. The novel aspect of the present approach is that we provide a rigorous scheme in terms of a linear equation system to determine the optimal values of the parameters of the scoring function.

2 Method

Peptide Data Bases. For the set of polypeptide sequences that bind at the MHC, we considered the SYF [24] and PEP [6] data bases in September 2003. There are 268 peptides in SYF that bind to the MHC HLA-A*0201. From these sequences 204 possess the canonical length of 9 residues. The remaining 64 peptides possess sequences longer than 9 residues. The peptides in SYF are sequence aligned i.e. equivalent sequence positions of different peptides were identified such that the corresponding peptide residues are supposed to interact with same residues of the MHC binding groove. We used this information to cut the length of the peptides longer than 9 residues to obtain nonapeptides, which are suitable for our approach. In the PEP database there are 506 peptides that bind to the MHC HLA-A*0201. These peptides were not aligned. Therefore, we considered from this data base nonapeptides only. We merged these two sets of nonapeptides, which after removing the identical peptides yielded a database $\mathbb{S}^+$ of 538 nonapeptides binding at the MHC HLA-A*0201. The sequences of these nonapeptides are listed in Table 6 of the appendix.

There are no explicit non-binding peptides available. We assumed that randomly chosen nonapeptides are unlikely to bind at the MHC. Hence, we generated up to 10,000 different nonapeptides $\mathbb{S}^-$ that were randomly taken from the concatenated sequences of 202 proteins selected from the protein database [1] (Table 1). Care was taken that the selected proteins do not contain nonapeptides that bind at MHC, although this can not be excluded with absolute certainty. The probability of occurrence of the 20 amino acid types in the data base of 10,000 non-binding nonapeptides given in Table 2 is similar to the distribution in other sequence databases as for instance modern vertebrates [4], but, differs in some amino acid types (Ala, Arg, Asp, Glu, Leu, Lys, Val) from the set of binding nonapeptides. It is noticeable that using randomly generated non-binding nonapeptides with the probability of the non-binding peptides provided the same results in recognition and prediction (data not shown).

Data Representation. We assume that two sets of polypeptide sequences are available: one set of binding peptides $\mathbb{S}^+ = \{\overrightarrow{x}_n^+,\ n = 1,\ \ldots,\ N^+\}$ and one set of non-binding peptides $\mathbb{S}^- = \{\overrightarrow{x}_n^-,\ n = 1,\ \ldots,\ N^-\}$, which are obtained as explained above. For the present application all sequences considered are aligned and of equal length say $M = 9$. The polypeptide sequences $\overrightarrow{x}_n$ are represented by M subvectors

$$\overrightarrow{x}_n^t = (\overrightarrow{x}_{1,n}^t,\ \overrightarrow{x}_{2,n}^t,\ \ldots,\ \overrightarrow{x}_{M,n}^t), \tag{1}$$

where each subvector in eq. (1) possesses 20 components

$$\overrightarrow{x}_{m,n}^t = (x_1^{(m,n)},\ x_2^{(m,n)},\ \ldots,\ x_{20}^{(m,n)}), \tag{2}$$

which refer to the 20 different native amino acid types. Note that the superscript t in eq. (1) and (2) refers to a row vector representation. An individual component of a sequence vector $\overrightarrow{x}_n$ denoting the occurrence of amino acid type j at sequence position m will be addressed as $(\overrightarrow{x}_n)_{jm}$. The amino acid type at a particular sequence position is coded by setting the corresponding component of the subvector to unity, while all other 19 components of this subvector contain zero. Thus, from a more general view point the components of each subvector can also be interpreted as a probability distribution to find specific amino acid types at the corresponding sequence position. This interpretation becomes more meaningful, when averages $\langle \overrightarrow{x} \rangle$ of those sequence vectors are considered as is done below.

Table 1:

PDB [a] codes of proteins whose concatenated sequences were used to generate the non-binding nonapeptides

1	7ZNF	1AGQ	1BRX	1C51	1BCC	1EPW	1A75	1E4T	1AQU	1O23	1ISN	2IAD
2	6RLX	1AIR	8TIM	1BQP	1P3H	1E1H	1LF4	1E3E	1ALB	1NMM	1IQ1	2DLF
3	6Q21	1AFO	1A0R	1EIS	1UJL	1DXR	1E0C	1DX1	1AI1	1NCI	1IKN	1SUH
4	1HNE	3MRA	1A12	1C01	1GWY	1GWC	1HRK	1DSV	1AFV	1NAS	1IG3	1HA7
5	1EAD	1AUN	3ZNC	1C4R	1RK4	1G6R	1RIE	1DJ2	1A2Y	1N9P	1IFQ	1GK8
6	1VMO	1AUV	1A38	1GGX	1QKK	1FYT	1UOY	1DF3	1A1H	1MNU	1IFA	1BX7
7	821P	1A04	2SQC	1FHF	1B9C	1L0X	1H1V	1DD7	1914	1MBY	1IAL	1BWK
8	1BOM	1AF6	1BUG	1H4Y	1BFA	1ITZ	1GL5	1CQZ	1R2A	1MBE	1I7W	1BK6
9	1AHL	1A06	1BYO	1BPO	1BD2	1IR1	1GL2	1CL7	1QLX	1M4M	1I7E	1BJT
10	1SRA	1AXM	7PCK	1AB1	1AO7	1LFJ	1G74	1CE6	1PA2	1M3V	1I6Z	1ASZ
11	1DOX	1AZD	1BYY	1H8P	1A2X	1OM0	1FWU	1CDK	1P8J	1LB1	1I07	1AI9
12	1MSP	1AIW	2VSG	1GRW	1A2C	1OED	1FRB	1C2B	1ORS	1KCM	1HQV	1A6R
13	1FAT	1A0D	1B10	1JV1	1D9K	1TCR	1FKW	1BLN	1OMX	1KBQ	1HQ8	1A4H
14	1BGK	1BB9	1C3A	1O7N	1CNE	1QF3	1F93	1BKX	1OKQ	1K2F	1HN3	1A48
15	7UPJ	1A05	1EG5	1H0H	1CJK	1PJU	1F81	1BGX	1OGP	1JJO	1H96	1A2V
16	6UPJ	1BA1	1EHD	1B8M	1FV3	1WGT	1EDH	1BBS	1OCP	1JI9	2ZNC	
17	5UPJ	1BKD	1BQF	1GDJ	1EZF	1BR1	1E4W	1AX8	1OAA	1IWE	2MSS	

[a] Ref. [1]

Scoring Function. The decision that a sequence $\overrightarrow{x}$ is capable to bind or not, is performed with a scoring function, $f(\overrightarrow{x})$, which is linear in sequences space $\mathbb{S}$(feature space) and in parameter space. The most general expression of the linear scoring function $f(\overrightarrow{x})$ is the linear form

$$f(\overrightarrow{x}) = \overrightarrow{w}^t \cdot \overrightarrow{x} + b\ , \tag{3}$$

Table 2: Probability of occurrence of amino acid types.

amino acid type	in modern vertebrates[a]	non-binding peptides[b]	binding peptides[c]
Ala	0.078	0.074	0.110
Arg	0.063	0.047	0.027
Asn	0.034	0.048	0.027
Asp	0.054	0.056	0.027
Cys	0.008	0.018	0.013
Gln	0.032	0.039	0.027
Glu	0.086	0.064	0.039
Gly	0.073	0.076	0.063
His	0.019	0.024	0.019
Ile	0.067	0.055	0.067
Leu	0.089	0.086	0.180
Lys	0.078	0.060	0.038
Met	0.024	0.020	0.024
Phe	0.036	0.042	0.046
Pro	0.044	0.048	0.046
Ser	0.047	0.066	0.054
Thr	0.049	0.058	0.048
Trp	0.010	0.017	0.013
Tyr	0.030	0.037	0.025
Val	0.082	0.066	0.110

[a] Probability of occurrence of amino acid types in modern vertebrates according to Ref. [4].
[b] Probability of occurrence of amino acid types in the 10,000 non-binding nonapeptides as explained in text.
[c] Probability of occurrence of amino acid types in the 538 binding nonapeptides as explained in text.

where $\overrightarrow{x} \in \mathbb{S}$ is a 20*M component vector characterizing a particular sequence, $\overrightarrow{w}^t$ is a row vector of the same dimension as $\overrightarrow{x}$ and b is a scalar. The 20*M + 1 free parameters of the scoring function $\overrightarrow{w}^t$ and b are determined for a set of sequences, the so called learning set $\mathbb{S}_{learn}$ such that $f(\overrightarrow{x})$ adopts a value close to +1 for the binding sequences and close to -1 for the non-binding sequences. Hence, setting $f(\overrightarrow{x}) = 0$ defines a hyperplane in the 20*M dimensional sequence space $\mathbb{S}$ with plane normal vector $\overrightarrow{w}$. The hyperplane $f(\overrightarrow{x}) = 0$ separates binding sequences $\overrightarrow{x}^+$ with $f(\overrightarrow{x}^+) > 0$ from non-binding sequences $\overrightarrow{x}^-$ with $f(\overrightarrow{x}^-) < 0$. These criteria can be used to predict the binding ability of peptides.

Least Square Optimization. There are different strategies in the learning phase where the 20*M + 1 free parameters of the scoring function $f(\overrightarrow{x})$, eq. (3), are determined. The most elementary approach is to minimize the scoring function with respect to the sum of least square deviations [least square method (LSM)]

$$L(\overrightarrow{w}, b) = \frac{1}{2N} \sum_{n=1}^{N} (f(\overrightarrow{x}_n) - y_n)^2. \tag{4}$$

The sum in eq. (4) runs over all sequences of the learning set $\mathbb{S}_{learn} = \mathbb{S}^+ \cup \mathbb{S}^-$, where for binding sequences $y_n = +1$ and for non-binding sequences $y_n = -1$. Taking the derivatives of $L(\overrightarrow{w}, b)$ with respect to $\overrightarrow{w}$ and b results in the following set of 20*M linear equations

$$\left\langle (\overrightarrow{x} - \langle \overrightarrow{x} \rangle)(\overrightarrow{x}^t - \left\langle \overrightarrow{x}^t \right\rangle) \right\rangle \cdot \overrightarrow{w} = \langle (y - \langle y \rangle)(\overrightarrow{x} - \langle \overrightarrow{x} \rangle) \rangle \tag{5}$$

and

$$b = \langle y \rangle - \left\langle \overrightarrow{x}^t \right\rangle \cdot \overrightarrow{w} \ . \tag{6}$$

The angular brackets in eq. (5) and (6) denote averages over all sequences of the learning set $\mathbb{S}_{learn}$ as for instance

$$\langle \overrightarrow{x} \rangle = \frac{1}{N} \sum_{n=1}^{N} \overrightarrow{x}_n \ . \tag{7}$$

It is interesting to note that the matrix of the set linear equations (5) is formed from the covariances of the sequence distributions

$$\left\langle (\overrightarrow{x} - \langle \overrightarrow{x} \rangle)(\overrightarrow{x}^t - \left\langle \overrightarrow{x}^t \right\rangle) \right\rangle = \left\langle \overrightarrow{x}\, \overrightarrow{x}^t \right\rangle - \langle \overrightarrow{x} \rangle \left\langle \overrightarrow{x}^t \right\rangle \ , \tag{8}$$

where $\overrightarrow{x}\overrightarrow{x}^t$ denotes the dyadic product of the sequence vector $\overrightarrow{x}$. For instance the matrix element

$$N\left\langle \overrightarrow{x}\overrightarrow{x}^t \right\rangle_{(jm),(j'm')}$$

counts how often in the learning set of sequences $\mathbb{S}_{learn}$ one meets an amino acid of type j at sequence position m while simultaneously at position m' there is an amino acid of type j'. Hence, the matrix of the set of linear equations (5) accounts for such pair correlations. We have developed our own computer program to solve these linear equations.

Weighting and Regularization. To weight binding and non-binding peptides differently one can generalize the averages, eq. (7), according to

$$\langle \overrightarrow{x} \rangle = \frac{w^+}{N^+}\sum_{n=1}^{N^+} \overrightarrow{x}_n^+ + \frac{w^-}{N^-}\sum_{n=1}^{N^-} \overrightarrow{x}_n^- \ , \tag{9}$$

where $w^+ + w^- = 1$ and $N^+ + N^- = N$ holds. This description allows a weighting of sequences in the learning set $\mathbb{S}_{learn}$, which is independent from the actual number of binding $\mathbb{S}^+$ and non-binding sequences $\mathbb{S}^-$ considered. Good results were obtained using for instance weighting factors of $w^+ = 0.45$ and $w^- = 0.55$ or $w^+ = 0.36$ and $w^- = 0.64$.

In case the number of data is small compared with the set of parameters that are to be optimized a regularization of the optimization procedure has turned out to be useful. This is the so-called ridge regression procedure [17], which is widely used for sequence prediction problems [23]. It can be considered by an additional term in the optimization function, eq. (4), yielding

$$\hat{L}(\overrightarrow{w}, b) = L(\overrightarrow{w}, b) + \lambda \overrightarrow{w}^t \cdot \overrightarrow{w} \ , \tag{10}$$

where λ is an empirical parameter, which needs to be chosen. Since the optimization function $L(\overrightarrow{w}, b)$, eq. (4), is normalized by dividing with N, the number of sequences considered, the value of λ is independent from the size of the learning set. The regularization term eliminates the possible occurrence of singular behavior and contributes to a minimization of the length of the normal vector $\overrightarrow{w}$ of the separating hyperplane that is defined by $f(\overrightarrow{x}) = \overrightarrow{w}^t \cdot \overrightarrow{x} + b = 0$. As a consequence, the sensitivity of this separating hyperplane may increase for moderate values of λ in particular if the set of linear equations (5) is ill-conditioned due to the smallness of the learning set $\mathbb{S}_{learn}$. Interestingly, a support vector machine uses also a strategy to minimize the hyperplane normal vector $\overrightarrow{w}$ to increase the sensitivity [15]. In the present applications of LSM we used $\lambda = 10^{-7}$, which is large enough to prevent singularities in the linear equations (5) but simultaneously small enough to have no influence on results in the absence of singular behavior.

In the set of linear equations the regularization term in the optimization function gives rise to an extra term in the diagonal of the matrix yielding instead of eq. (5)

$$\left\langle (\overrightarrow{x} - \langle \overrightarrow{x} \rangle)(\overrightarrow{x}^t - \left\langle \overrightarrow{x}^t \right\rangle) \right\rangle \cdot \overrightarrow{w} + \lambda \overrightarrow{w} = \langle (y - \langle y \rangle)(\overrightarrow{x} - \langle \overrightarrow{x} \rangle) \rangle \ . \tag{11}$$

In the present applications we have a sufficient number of data, such that an application of the ridge regression method did not offer significant advantages.

Support Vector Machine. An alternative approach to optimize the parameters of the linear scoring function, eq. (3), is to use a support vector machine (SVM). A detailed description of SVM can be found in Refs. [8, 15]. With this method one determines the parameters of the scoring function $f(\overrightarrow{x})$ such that for binding peptides the inequality $f(\overrightarrow{x}^+) \geq +1$ and for non-binding peptides the inequality $f(\overrightarrow{x}^-) \leq -1$ is approximated, while simultaneously the length of the hyperplane normal vector $\overrightarrow{w}$ is minimized. The latter increases the sensitivity to discriminate between binding and non-binding peptides. A crucial point of this method is to consider only a subset of the total learning set $\mathbb{S}_{learn}$ by sorting out data for which the corresponding inequality rigorously holds, i.e. $f(\overrightarrow{x}^+) > +1$ and $f(\overrightarrow{x}^-) < -1$ for binding and non-binding peptides, respectively. Also this selection increases the sensitivity of the method.

A further increase in sensitivity may be achieved by applying a non-linear transformation to the sequence (feature) space of the learning data set and optimizing the discrimination problem in this new feature space. In several test computations a non-linear representation in the feature space did not show any improvement. Thus, we refrain from giving more details. We used the public domain program SVM^{light} [18] to optimize the parameters of the support vector machine.

Quality Control. The performance of learning (predicting) can be characterized by providing simply the fraction of binding and non-binding nonapeptides, which were recognized (predicted) properly or alternatively by the Matthew correlation coefficient (MCC) [21], which is defined as

$$\mathrm{MCC} = \frac{cor^+ cor^- - incor^+ incor^-}{[N^+ N^-(cor^+ + incor^-)(cor^- + incor^+)]^{\frac{1}{2}}} \;, \tag{12}$$

with cor^+ and cor^- as the number of correctly classified binding and non-binding peptides and $incor^+$ and $incor^-$ as the number of incorrectly classified binding and non-binding peptides, respectively. Note that $N^+ = cor^+ + incor^+$ and $N^- = cor^- + incor^-$. The advantage of the MCC measure is to ignore spurious contributions, which are obtained also in the absence of a learning or prediction strategy. In case of a symmetric situation with $cor^+ = cor^- = cor$, $incor^+ = incor^- = incor$ and $N^+ = N^- = N/2$ and a low error margin $cor \gg incor$ the expression(12) simplifies approximately to

$$\mathrm{MCC} \simeq 1 - 2\frac{incor}{cor + incor}$$

valid for binding and non-binding nonapeptides, such that a prediction probability of 0.9 corresponds to an MCC value of 0.8. A prediction probability of 0.5 that can be obtained also in the absence of a learning or prediction strategy yields MCC = 0.

Another widely used method to characterize the quality of learning and predicting are plots of sensitivity (*sens*) versus specificity (*spec*) [3]. The functional dependence $sens(spec)$ can be obtained by varying the threshold t used to classify a peptide of sequence $\overrightarrow{x}$ as binding for $f(\overrightarrow{x}) > t$ and as non-binding for $f(\overrightarrow{x}) < t$ and monitoring $sens(t)$ and $spec(t)$, which are defined as

$$sens(t) = \frac{cor^+(t)}{N^+} \quad and \quad spec(t) = \frac{cor^-(t)}{N^-} \;. \tag{13}$$

The area under the function $sens(spec)$ can be understood as an overall quality measure of recognition/prediction. However, it is preferable to consider the *sens* and *spec* values for a symmetric situation i.e. $sens \cong spec$, which can be achieved by variation of the threshold t value.

3 Results and Discussion

Parameters of the Scoring Function. We have determined the parameters of the scoring function by solving the set of linear equations (5) and by applying the support vector machine for different sets of binding and non-binding nonapeptides. To provide results of a typical application, which can be reproduced, we calculated optimal parameters of the scoring function, eq. (3), for the 269 even numbered binding nonapeptides $\mathbb{S}^+$ listed in Table 6 of the appendix supplemented by the same number of non-binding peptides taken from the set $\mathbb{S}^-$ of 10.000 nonapeptides as described above. The parameters of $\overrightarrow{w}$ and b obtained using the method of minimizing the least square optimization method (LSM) and the support vector machine (SVM) are given in Table 3. Surprisingly, the support vector machine did not provide improved results by using a non-linear transformation from which we can conclude that the problem of peptide binding is likely to be not separable the non-linear feature space of sequences. Consequently, the SVM parameters displayed in Table 3 refer to the linear version.

Table 3: Optimized parameters $\overrightarrow{w}$ of the scoring function $f(\overrightarrow{x}) = \overrightarrow{w}^t \cdot \overrightarrow{x} + b$.

amino acid	position 1		position 2		position 3		position 4		position 5		position 6		position 7		position 8		position 9	
type	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM	LSM	SVM
ALA	-0.02	0.16	-0.07	0.12	0.14	0.44	-0.21	-0.21	-0.26	-0.15	-0.19	-0.06	0.07	0.26	-0.16	-0.13	-0.10	-0.05
HIS	-0.01	0.11	-0.26	-0.16	0.18	-0.02	-0.41	-0.11	-0.08	0.08	0.08	-0.06	-0.38	-0.10	-0.13	-0.11	0.10	-0.03
GLU	-0.16	-0.24	-0.20	-0.08	-0.37	-0.34	0.02	0.13	-0.13	-0.10	-0.29	-0.27	-0.38	-0.30	-0.04	0.16	-0.21	-0.13
GLN	-0.31	-0.17	-0.28	-0.11	0.01	0.13	-0.20	-0.14	0.25	0.25	-0.14	-0.05	-0.33	-0.14	-0.07	0.10	0.09	-0.02
ASP	-0.39	-0.22	-0.39	-0.33	-0.06	0.05	0.10	0.12	-0.38	-0.36	-0.36	-0.26	-0.41	-0.39	-0.02	-0.15	-0.09	-0.19
ASN	-0.32	-0.20	-0.80	-0.58	-0.04	0.10	-0.29	-0.26	-0.16	0.01	-0.18	-0.11	-0.23	-0.07	-0.51	-0.38	-0.25	-0.11
LEU	-0.20	-0.15	0.78	1.50	-0.10	0.20	-0.24	-0.11	-0.22	-0.10	0.00	-0.03	0.15	0.40	-0.01	0.24	0.40	0.63
GLY	-0.03	0.04	-0.22	-0.16	-0.17	-0.15	0.06	0.25	-0.03	0.32	-0.09	0.05	-0.30	-0.25	-0.01	0.08	-0.29	-0.37
LYS	-0.19	0.02	-0.54	-0.36	-0.16	-0.12	0.08	0.27	-0.16	-0.10	-0.23	-0.26	-0.44	-0.28	-0.14	0.06	-0.42	-0.34
SER	-0.19	-0.15	-0.39	-0.39	-0.03	0.03	-0.04	0.05	-0.24	-0.32	-0.04	0.12	-0.11	0.04	-0.20	-0.05	-0.09	0.01
VAL	-0.05	0.02	0.05	0.04	-0.03	0.13	-0.09	-0.08	0.05	0.30	0.07	0.38	0.02	0.00	-0.12	-0.06	0.48	0.95
ARG	0.19	0.23	-0.56	-0.14	-0.28	-0.03	-0.09	0.07	0.00	0.13	-0.52	-0.35	-0.32	-0.20	-0.21	-0.04	-0.35	-0.27
THR	-0.16	-0.18	0.08	0.23	0.27	0.19	-0.01	-0.01	-0.09	-0.02	-0.07	-0.09	-0.01	0.16	0.01	0.09	0.01	0.24
PRO	-0.04	0.10	-0.28	-0.16	-0.15	-0.13	0.01	0.22	-0.05	0.18	0.05	0.23	0.13	0.17	0.16	0.28	-0.49	-0.32
ILE	0.04	0.15	0.29	0.35	-0.07	-0.14	0.31	0.29	0.08	0.07	0.14	0.53	-0.03	0.11	-0.18	-0.12	0.29	0.53
MET	-0.13	0.15	0.69	0.88	-0.27	-0.01	-0.25	-0.14	-0.33	-0.18	-0.28	-0.03	-0.12	0.00	0.05	0.04	0.20	0.14
PHE	-0.04	0.10	-0.26	-0.16	-0.07	0.08	-0.29	-0.20	-0.20	0.12	0.19	0.25	0.20	0.39	-0.36	-0.08	-0.28	-0.19
TYR	0.11	0.18	-0.44	-0.44	-0.28	-0.34	-0.46	-0.26	-0.24	-0.13	0.08	-0.01	0.06	-0.06	-0.18	0.03	-0.37	-0.37
CYS	-0.30	-0.07	0.54	0.04	-0.58	-0.22	-0.18	0.08	0.15	0.08	-0.09	-0.02	0.17	0.12	0.14	0.02	-0.17	-0.04
TRP	0.19	0.13	0.28	-0.07	0.05	0.16	0.17	0.02	0.05	-0.07	-0.14	0.03	0.26	0.14	-0.01	0.01	-0.46	-0.07

The 180 parameters are displayed in a two-dimensional array w_{jm} with respect to the 20 amino acid types j and the 9 sequence positions m. The parameters obtained with the least square optimization method (LSM) are given in the left columns the parameters obtained from the support vector machine (SVM) are given in the right columns, respectively. The values of parameter b are $b_{LSD} = 0.22$ and $b_{SVM} = 1.09$. To determine the parameters with LSM we used as learning set every even numbered binding nonapeptide from Table 6 in the appendix, making up a total of 269 peptides. The same number of 269 non-binding peptides was chosen at random from the prepared set of 10,000 non-binding peptides as described in text. The weights used to compute the averages, eq. (9), needed for LSM were $w^+ = 0.45$ and $w^- = 0.55$ for binding and non-binding peptides, respectively.

Learning and Recognizing. In using the scoring function, we discriminate between learning, recognizing and predicting the ability of peptides to bind or not to bind. For the first two procedures we use a learning set $\mathbb{S}_{learn}$ and for the latter we use a predicting set $\mathbb{S}_{predict}$ of binding and non-binding peptides that is disjoint from the learning set. To demonstrate the LSM and SVM methods we determined in a first application the parameters of the scoring function given in Table 3 using a learning set $\mathbb{S}_{learn}$ containing 269 peptides from the total number of 538 available binding nonapeptides (the even numbered in Table 6 of the appendix) and an equal number of non-binding peptides chosen randomly from the set of 10.000 non-binding peptides. For the prediction mode, we considered the remaining 269 non-binding peptides and another 269 peptides randomly chosen from the set of non-binding peptides. With LSM (SVM) the peptides of the learning set $\mathbb{S}_{learn}$ of 538 peptides were recognized to 93.3% (95.9%) and 93.3% (94.4%) for binding and non-binding peptides, respectively. In the prediction mode the LSM (SVM) predicted binding peptides to 92.9% (92.2%) and non-binding peptides to 86.2% (89.2%). The support vector machine was using 115 binding and 123 non-binding nonapeptides yielding a total number of 238 support vectors taken from the 538 data sets used. The absolute numbers of incorrectly as non-binding recognized truly binding peptides are 18 for both SVM and LSM. From these, 14 peptides were incorrectly recognized with both methods. The absolute numbers of incorrectly recognized peptides from the truly non-binding peptides are 15 for SVM and 11 for

LSM. In this case 10 peptides were recognized wrongly as binding peptides with both methods.

Table 4: Number of incorrectly recognized sequences and the corresponding range of values of the scoring function for different weights. See also Figure 1.

weights w^+	incorrectly recognized as binding	values of scoring function	incorrectly recognized as non-binding	values of scoring function
0.8	0	-	17	0.01 to 0.98
0.6	1	at -0.02	7	0.02 to 0.72
0.4	7	-0.36 to -0.07	2	at 0.23, at 0.45
0.2	21	-0.77 to -0.02	1	at 0.11
0.1	37	-1.02 to -0.02	0	-

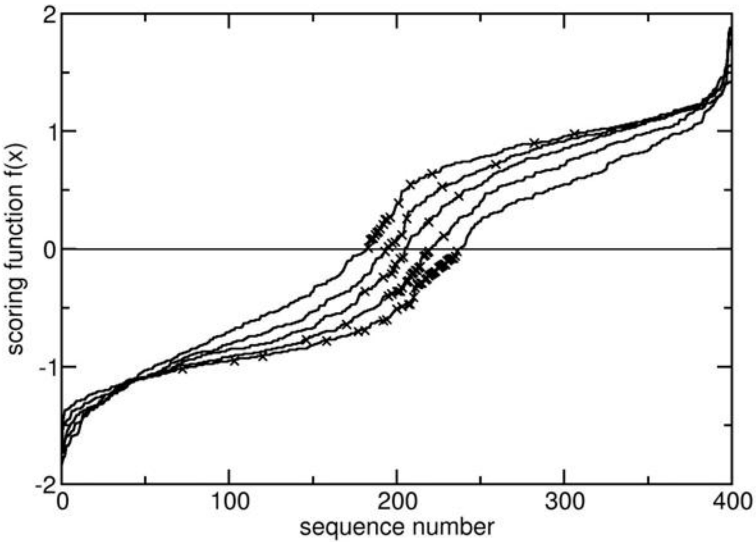

Figure 1: Course of the scoring function $f(\overrightarrow{x})$, eq. (3), for different weights w^+ of the binding peptides. Parameters of the scoring function were determined based on 200 binding and 200 non-binding peptides in the learning mode as explain in text. The scoring function is displayed in recognition mode considering the peptides of the learning set $\mathbb{S}_{learn}$. From top to bottom the scoring functions refer to weights w^+ of the binding peptides of 0.8, 0.6, 0.4, 0.2, 0.1. Crosses mark incorrectly predicted peptides whose statistics are given in Table 4.

Weighting Binding and Non-Binding Peptides. The least square optimization method allows to apply different weights for the set of binding and non-binding peptides [see eq. (9)]. These weights can play a similar role as does the threshold value t [see eq. (13)] used to discriminate between binding and non-binding peptides. We studied the influence of different weights on recognition by monitoring the scoring function $f(\overrightarrow{x})$, eq. (3), for renumbered sequences $\overrightarrow{x}_n$ of the learning set $\mathbb{S}_{learn}$, which are ordered such that for subsequent sequences $\overrightarrow{x}_n$ and $\overrightarrow{x}_{n+1}$ we have $f(\overrightarrow{x}_n) < f(\overrightarrow{x}_{n+1})$. Thus, a scoring function is obtained whose value increases monotonously with sequence number n. The scoring function shown in Figure 1 is based on 200 binding and 200 non-binding peptides in the learning set $\mathbb{S}_{learn}$ to determine the parameters of the scoring function and its values. The crosses mark nonapeptides whose binding ability was determined incorrectly according to the value of the scoring function. The number of incorrect recognized binding (non-binding) sequences increases from 0 to 37 (decreases from 17 to 0) with decreasing weight w^+ for the binding peptides (Table 4). The ideal shape of the scoring function should be a step function with a function value of -1 for the first

200 non-binding peptides and +1 for the 200 binding peptides. For large weights w^+ of the binding peptides the positive step of the scoring function is very pronounced while the negative step is less distinct. The opposite is the case for small weights w^+.

Course of the Scoring Function. To study the behavior of the scoring function in more detail we employed a learning set of all available 538 binding peptides and added the same number of non-binding peptides at random from the set of 10,000 non-binding peptides. As in Figure 1 we renumbered the sequences to obtain monotonously increasing scoring functions. But, in this case we considered the binding and non-binding peptides separately yielding two branches $f^+(\vec{x})$ and $f^-(\vec{x})$ of the scoring function, respectively. The branches $f^-(\vec{x})$ of the non-binding peptides are located in the lower half, the branches $f^+(\vec{x})$ describing the binding peptides are located in the upper half of Figure 2. The fraction of non-binding peptides with $f^-(\vec{x}) < 0$ and of binding peptides with $f^+(\vec{x}) > 0$ are correctly recognized. In recognition mode (solid lines for LSM and dashed-dotted lines for SVM) the two different optimization methods (LSM and SVM) considered in this work yielded very similar results with a minor advantage for SVM in recognizing binding peptides, while LSM is marginally ahead in recognizing non-binding peptides. But, SVM seems to be superior in its ability to separate binding from non-binding peptides, since its scoring function is generally larger for binding peptides and smaller for non-binding peptides as compared to the corresponding LSM scoring function. The results obtained with LSM in prediction mode using the jackknife procedure (dashed lines) (leaving out one peptide in the learning mode, whose binding ability is predicted) yielded results that are very similar to the corresponding data obtained in recognition mode. Even with a rather small number of 50 binding and 50 non-binding peptides in the learning set, prediction of all 538 binding and non-binding peptides yields reasonable results (dotted lines).

Selecting Peptides from the Learning Set. The support vector machine has the ability to select a subset of data in feature space to optimize the performance. The least square optimization method does not directly offer such an option. However, after an LSM optimization is performed one can identify incorrectly recognized peptides and the peptides that are located in the twilight zone of vanishing values of the scoring function. The assumed binding ability of these peptides may have been wrongly assigned. This can particularly be the case for the randomly generated sequences from which we assumed that they are all non-binding. We had the option to eliminate these peptides in a second run of LSM optimization. To investigate this possibility, we started an LSM optimization with 300 binding and 5,000 non-binding peptides as learning set $\mathbb{S}_{learn}$ with $w^+ = 0.36$. In the prediction mode, we considered the remaining 238 binding and 5,000 non-binding peptides. Thus, in recognition mode 92.0% binding and 92.8% non-binding peptides were found. In the prediction mode 90.3% binding and 92.5% non-binding peptides were predicted correctly. In a second LSM run, we eliminated 31 non-binding peptides with $f(\vec{x}^-) > 0.7$ from $\mathbb{S}_{learn}$. With this choice, recognition was slightly reduced for the non-binding peptides yielding 92.5%, while it remained unchanged for the binding peptides. In the prediction mode we now observed that 91.6% of the binding and 92.2% of the non-binding peptides were predicted correctly, which is over all an improvement compared with the first LSM run.

Interestingly, the SVM optimization yielded here only 50.6% correctly recognized and 48.7% correctly predicted binding peptides, while the non-binding peptides were found with 99.5% in recognition and prediction mode. It is known since recently that SVM fails if an unbalanced data set is used for learning size of the set of binding as appears in the present application with 300 binding and 5,000 non-binding nonapeptides [22, 32] and remedies are complicated. A simple procedure is to artificially increase the minority set of binding peptides in the SVM procedure by considering 16 identical copies of the original set of 300 binding peptides to balance the number of binding and non-binding peptides considered. In this case, we obtained with the SVM correct recognition of 96.0% binding and 92.0% non-binding peptides and correct prediction of 93.0% binding and 91.0% non-binding peptides. Note that in the LSM optimization a balanced consideration of binding and non-binding peptides in the

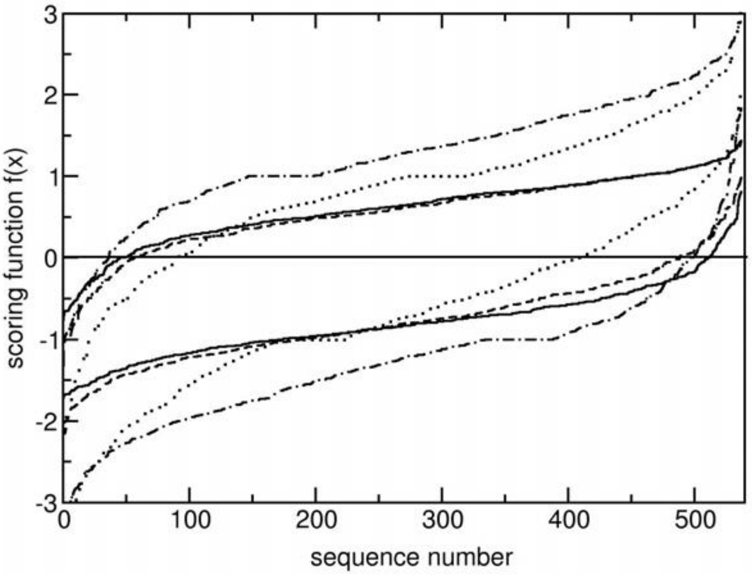

Figure 2: Course of the scoring function $f(\overrightarrow{x})$, eq. (3), monitored separately for binding and non-binding peptides. A weighting of $w^+ = 0.36$ and $w^- = 0.64$ was used for the least square optimization method (LSM). In contrast to Figure 1 binding and non-binding peptides were renumbered separately and the resulting courses of the scoring functions were displayed separately. From the pair of scoring functions displayed in the same line style the lower $f^-(\overrightarrow{x})$ refers to non-binding peptides and the upper $f^+(\overrightarrow{x})$ to binding peptides. For the learning set all available 538 binding peptides and the same number of non-binding peptides chosen from the set of 10,000 random peptides were considered. The dashed-dotted lines display results of recognition using SVM. The pair of solid lines describe the same as before but using LSM, which is also used for all other data displayed in this figure. The pair of dashed lines displayed results in prediction mode obtained with the jackknife method. The dotted lines display results in prediction mode using for learning only 50 binding and 50 non-binding peptides that were randomly chosen.

learning mode is achieved directly using the weighting factors w^+ and w^- to evaluate the averages, eq. (9). In conclusion one can say that with a well tuned least square optimization the same quality of predicting of peptides can be achieved as with the SVM optimization.

Quality Control. A sensitivity selectivity plot as described in the method section can be used as quality control for recognition and prediction (see Figure 3). The area under the function sensitivity(specificity), eq. (13), provides an overall measure of quality. The area is 0.9926 for recognition (dashed line) and 0.9559 for prediction (solid line) using the LSM optimization and 0.9613 for prediction (dotted line) using SVM optimization. Here, SVM shows again its superiority being slightly stronger in its ability to discriminate between binding and non-binding peptides.

A more reliable measure of prediction quality is obtained by performing a statistical survey of learning and predicting. For this purpose, we considered randomly chosen sets of peptides for learning and predicting considering the LSM optimization, where we generated at random 400 different learning sets $\mathbb{S}_{learn}$ and disjoint predicting sets $\mathbb{S}_{predict}$ from the total number of 538 binding peptides and 10,000 non-binding peptides to determine the parameters of the scoring function. Table 5 shows the results for small learning sets of 50 binding and 50 non-binding peptides, which yield perfect results for recognition, since the learning sets are so small, but exhibit rather modest results in the prediction mode with an average success below 80% and a large variance of about 20%. With a much larger learning set of 300 binding and 5,000 non-binding peptides the average fraction of correctly recognized peptides diminishes being now at 95% and 91% for binding and non-binding nonapeptides, respectively. At the same time, the average prediction quality improves considerably being now close to 90% for

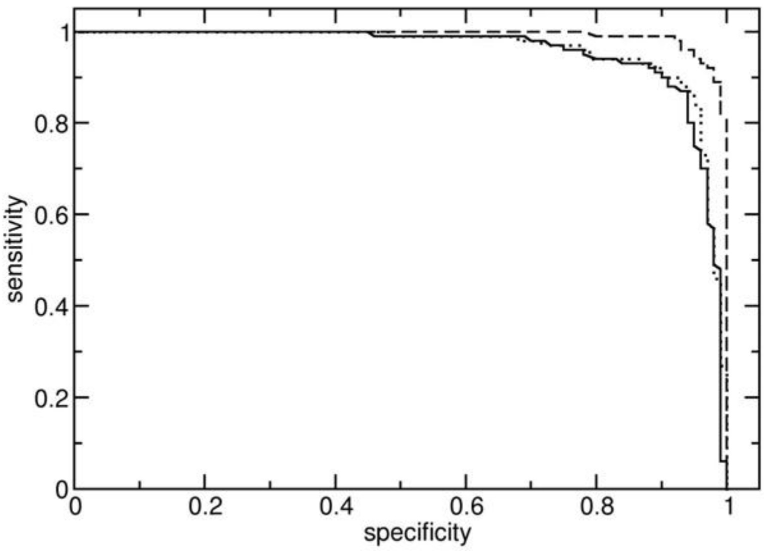

Figure 3: Sensitivity specificity plot for two different, disjoint sets $\mathbb{S}_{learn}$ and $\mathbb{S}_{predict}$ of 269 binding and 269 non-binding peptides for the learning and predicting mode, respectively, using the same data as for Table 3. Dashed line: learning mode, solid line: predicting mode with LSM optimisation. Dotted line: SVM in predicting mode. The area under the curves are a measure for the prediction and recognition quality. They adopt the values of 0.9926 for recognition and 0.9559 for prediction using LSM and 0.9613 for prediction using SVM.

binding and non-binding peptides. The variance of these averages is now much smaller due to the larger data base of binding and non-binding peptides used for the determination of the parameters of the scoring function (Table 5).

It is astonishing that the present method based on a simple linear scoring function can yield a prediction accuracy that is surpassing results of equivalent approaches [9, 14, 23, 31] and coming close to results of more powerful procedures like hidden Markov models [5, 20, 30]. However, there is one important difference between the present approach and neural network and hidden Markov model in that the latter can cope with binding peptides of variable length while the former can deal with aligned peptides of a fixed length only.

Table 5: Recognition and prediction statistics of binding for different learning sets of peptides[a].

	size of learning sets binding/non-binding	50/50[b]	size of learning sets binding/non-binding	300/5,000[b]
	binding peptides	non-binding peptides	binding peptides	non-binding peptides
recognition[c]	100% ±0.0%	100% ±0.0%	94.8% ±0.8%	91.3% ±0.1%
prediction[d]	78.8% ±21.3%	73.0% ±19.5%	90.0% ±2.9%	90.8% ±0.2%

[a] The learning sets are generated at random 400 times using least square optimization (LSM) with weighting factors of $w^{+} = 0.45$ and $w^{-} = 0.55$.

[b] Number of binding and non-binding peptides.

[c] Recognition mode: probing recognition probability of the different learning sets of peptides.

[d] Prediction mode: probing prediction probability of randomly chosen 238 binding and 5,000 non-binding peptides that are disjoint from the learning set.

4 Conclusions

We have generalized a least square optimization method to predict peptide binding at the class I major histocompatibility complex using a general linear scoring function. A new weighting procedure allows to treat asymmetric data sets with a small number of binding and a large number of non-binding peptides in a balanced way maintaining the prediction quality, which is in the case far better than the results from the support vector machine (SVM). However, the apparent deficiency of SVM can probably be repaired by generalizing existing programs solving the SVM problem. But, even for a symmetric data set the prediction quality of LSM comes very close to the SVM results. Further generalizations of the LSM may possess the potential to reach or surpass the prediction quality of other prediction schemes.

Acknowledgments

We are grateful for financial support from the Deutsche Forschungsgemeinschaft: Sfb498, GRK80/2, GRK268, GRK788/1, Forschergruppe 475.

References

[1] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne P.E., The protein data bank, *Nucleic Acids Research*, 28:235–242, 2000.

[2] Bouvier, M. and Wiley, D.C., Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules, *Science*, 265:398–402, 1994.

[3] Bradley, A.P., The use of the area under the roc curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30:1145–1159, 1997.

[4] Brooks, D.J., Fresco, J.R., Lesk, A.M., and Singh, M., Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code, *Molecular Biology and Evolution*, 19:1645–1655, 2002.

[5] Brusic, V., Petrovsky, N., Zhang, G., and Bajic, V.B., Prediction of promiscuous peptides that bind HLA class I molecules, *Immunology and Cell Biology*, 80:280–285, 2002.

[6] Brusic, V., Rudy, G., and Harrsison, L.C., MHCPEP, a database of MHC-binding peptides: Update 1997, *Nucleic Acids Research*, 26:368–371, 1998.

[7] Brusic, V., Rudy, G., Honeyman, M., Hammer, J., and Harrison, L., Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, *Bioinformatics*, 14:121–130, 1998.

[8] Burges, C.J.C., A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 121–167, 1998.

[9] Donnes, P. and Elofsson, A., Prediction of MHC class I binding peptides, using SVMHC. BMC, *Bioinformatics*, 3:25, 2002.

[10] Falk, K., Rotzschke, O., Stefanovic, S., Jung, G., and Rammensee, H.G., Allele-specific motifs revealed by sequencing of self peptides eluted from MHC molecules, *Nature*, 351:290–296, 1991.

[11] Garboczi, D.N., Ghosh, P., Utz, U., Fan, Q.R., Biddison, W.E., and Wiley, D.C., Structure of the complex between human T-cell receptor, viral peptide and HLA-A2, *Nature*, 384:134–141, 1996.

[12] Garcia, K.C., Degano, M., Pease, L.R., Huang, M., Peterson, P.A., Leyton, L., and Wilson, I.A., Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen, *Science*, 279:1166–1172, 1998.

[13] Guilloux, Y., Lucas, S., Brichard, V.G., Van Pel, A., Viret, C., De Plaen, E., Brasseur, F., Lethe, B., Jotereau, F., and Boon, T., A peptide recognized by human cytolytic T lymphocytes on HLA-A2 melanomas is encoded by an intron sequence of the N-acetylglucosaminyltransferase V gene, *J. Exp. Med.*, 183:1173–1183, 1996. .

[14] Gulukota, K., Sidney, J., Sette, A., and DeLisi, C., Two complementary methods for predicting peptides binding major histocompatibility complex molecules, *Journal of Molecular Biology*, 267:1258–1267, 1997.

[15] Hearst, M.A., Scholkopf, B., Dumais, S., Osuna, E., and Platt, J., Trends and controversies - support vector machines, *IEEE Intelligent Systems*, 13:18–28, 1998.

[16] Henderson, R.A., Michel, H., Sakaguchi, K., Shabanowitz, J., Apella, E., Hunt, D.F., and Engelhard, V.H., HLA-A2.1 associated peptides from a mutant cell line. A second pathway of antigen presentation, *Science*, 255:1264–1266, 1992.

[17] Hoerl, A.E. and Kennard, R.W., Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12:55–67, 1970.

[18] Joachims, T., *Making large-Scale SVM Learning Practical*, MIT Press, 1999.

[19] Lanzavecchia, A., Reid, P.A., and Watts, C., Irreversible association of peptides with class II MHC molecules in living cells, *Nature*, 357:249–252, 1992.

[20] Mamitsuka, H., Predicting peptides that bind to MHC molecules using supervised learning of hidden markov models, *Proteins*, 33:460–474, 1998.

[21] Matthews, B., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta.*, 405:442–451, 1975.

[22] Musicant, D.R., Kumar, V., and Ozgur, A., Optimizing F-measure with support vector machines, *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, Russell, I. and Haller, S., editors, AAAI Press:356–360, 2003.

[23] Peters, B., Bulik, S., Tampe, R., van Endert, P.M., and Holzhutter, H.G., Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors, *Journal of Immunology*, 171:1741–1749, 2003.

[24] Rammensee, H.G., Bachmann, J., Emmerich, N.N., Bachor, O.A., and Stevanovic, S., SYFPEITHI: Database for MHC ligands and peptide motifs, *Immunogenetics*, 50:213–219, 1999

[25] Regner, M., Claesson, M.H., Bregenholt, S., and Ropke, M., An improved method for the detection of peptide-induced upregulation of HLA-A2 Molecules on TAP-deficient T2 cells, *Exp. Clin. Immunogenet*, 13:30–35, 1996.

[26] Rosenfeld, R., Zheng, Q., Vajda, S., and Delisi, C., Computing the structure of bound peptides - application to antigen recognition by class-I major histocompatibility complex receptors, *Journal of Molecular Biology*, 234:515–521, 1994.

[27] Rotzschke, O., Falk, K., Stevanovic, S., Jung, G., and Rammensee, H.G., Peptide motifs of closely related HLA class-I molecules encompass substantial differences, *European Journal of Immunology*, 22:2453–2456, 1992.

[28] Stolze, L., Nussbaum, A.K., Sijts, A., Emmerich, N.P., Kloetzel, P.M., and Schild, H., The function of the proteasome system in MHC class I antigen processing, *Immunol. Today*, 21:317–319, 2000.

[29] Tortorella, D., Gewurz, B.E., Furman, M.H., Schust, D.J., and Ploegh, H.L., Viral subversion of the immune system, *Annu. Rev. Immunol.*, 18:861–926, 2000.

[30] Udaka, K., Mamitsuka, H., Nakaseko, Y., and Abe, N., Prediction of MHC class I binding peptides by a query learning algorithm based on hidden Markov models, *Journal of Biological Physics*, 28:183–194, 2002.

[31] Yu, K., Petrovsky, N., Schonbach, C., Koh, J.L.Y., and Brusic, V., Methods for prediction of peptide binding to MHC molecules: A comparative study, *Molecular Medicine*, 8:137–148, 2002.

[32] Wu, G., Chang and Edward, Y., Adaptive feature-space conformal transformation for imbalanced-data learning, *Proceedings of the Twentieth International Conference on Machine Learning*, Washington DC, 816–823, 2003.

Appendix

Table 6: List of considered binding nonapeptides as described in text.

0	AAAKAAAAV	AAGIGIIQI	AAGIGILTV	AAPTPAAPA	ACDPHSGHF	ACRTVALTA	AFHHVAREL	AIAKAAAAV	AIMDKNIIL	AIVDKVPSV
10	ALAAVVTEV	ALACAAAAV	ALADAAAAV	ALADGVQKV	ALAKAAAAA	ALAKAAAAI	ALAKAAAAL	ALAKAAAAM	ALAKAAAAN	ALAKAAAAR
20	ALAKAAAAT	ALAKAAAAV	ALAKAAAEV	ALAKAAAFV	ALAKAAAGV	ALAKAAALV	ALAKAAAPV	ALAKAAEAV	ALAKAALAV	ALAKAANAV
30	ALAKAAPAV	ALAKAAYAV	ALAKAGAAV	ALAKAIAAV	ALAKAPAAV	ALAKARAAV	ALAKAYAAV	ALAKEAAAV	ALAKGAAAV	ALAKLAAAV
40	ALAKNAAAV	ALAKYAAAV	ALANGIEEV	ALAPAAAAV	ALASHLIEA	ALATAAAAV	ALAVAAAAV	ALCRWGLLL	ALEKAAAAV	ALFDGDPHL
50	ALFGALFLA	ALFKAAAAV	ALFPQLVIL	ALGLGLLPV	ALGRNSFEV	ALIHHNTHL	ALKKAAAAV	ALLNIKVKL	ALLPPINIL	ALMDKSLHV
60	ALMKAAAAV	ALMPLYACI	ALNELLQHV	ALNKMFCQL	ALNKMFYKL	ALNKMLCQL	ALQDSGLEV	ALQPGTALL	ALSDHHIYL	ALSDLEITL
70	ALSKAAAAV	ALSNLEVKL	ALSRKVAEL	ALSTGLIHL	ALWDIETGQ	ALWGFFPVL	ALWNLHGQA	ALYVDSLFF	AMAIHKQSQ	AMAKAAAAV
80	AMFQDPQER	ATAKAAAAV	AVAKAAAAV	AVFDRKSDA	AVGIGIAVV	AVVPFIVSV	AVVPFLVSV	CINGVCWTV	CLGGLITMV	CLGGLLTMV
90	CLTKWMILA	CLTSTVQLV	DLERKVESL	DLFGIWSKV	DLMGYIPLV	DLMLSPDDI	DLVHFASPL	DPKVKQWPL	DVASVIVTK	EAAGIGILT
100	ELIRVEGNL	ELTLGEFLK	ELVSEFSRM	ELVSEVSKV	EMFRELNEA	EVAPPLLFV	FAFRDLCIV	FIAGNSAYE	FIASNGVKL	FIDSYICQV
110	FIYAGSLSA	FKNIVTPRT	FLAKAAAAV	FLCKQYLNL	FLDEFMEGV	FLDGNELTL	FLDGNEMTL	FLDQVPFSV	FLEPGPVTA	FLFDGSPTY
120	FLGAAGSTM	FLGENISNF	FLGGTPVCL	FLGGTTVCL	FLKEPVHGV	FLLDKKIGV	FLLLADARV	FLLPSFAPD	FLLSLGIHL	FLLTRILTI
130	FLLWATAEA	FLPSDFFPS	FLQSRPEPT	FLTPKKLQC	FLWAIMHTE	FLWEFPHDL	FLWGPRALV	FLWGPRAYA	FLWTLEGDV	FLYCYFALV
140	FLYEAVPQL	FLYERVPQL	FLYGALLLA	FMFDLAAEL	FMFESPWNV	FMLDWFPTI	FTDQVPFSV	GAGIGVAVL	GAGIGVLTA	GELGFVFTL
150	GIAGGLALL	GIGIGVLAA	GIGILTVIL	GILGFVFTL	GILGFVFTM	GILGFVFTV	GILTVILGV	GIVPFIVSV	GIVPFLVSV	GLAPPQHEI
160	GLAPPQHLI	GLCTLVAML	GLDVLTAKV	GLHCYEQLV	GLIEKNIEL	GLIMVLSFL	GLLGFVFTL	GLLGNVSTV	GLLGTLVQL	GLLGWSPQA
170	GLPVEYLQV	GLQDCTMLV	GLRDLAVAV	GLSEFTEYL	GLSPTVWLS	GLSRYVARL	GLVPFIVSV	GLVPFLVSV	GLYDGMEHL	GLYPGLIWL
180	GLYSSTVPV	GMLGFVFTL	GMNCRPILT	GMNERPILT	GMNKRPILT	GMNRHPILT	GMNRRPILT	GQLGFVFTL	GTLGFVFTL	GTLGIVCPI
190	GTLSKIFKL	GVALQTMKQ	GVLGFVFTL	GVLVGVALI	HEIRVEGNL	HLEGKVILV	HLESLFTAV	HLGNVKYLV	HLIDYLVTS	HLIKVEGNL
200	HLIRVEGNL	HLLVGSSGL	HLSLRGLPV	HLSTAFARV	HLYQGCQVV	HLYSHPIIL	HMTEVVRHC	HMTEVVRRC	IAGIGILAI	IIDQVPFSV
210	IISAVVGIL	IISCTCPTV	IISLWDQSL	ILAGYGAGV	ILAKFLHWL	ILAPPVVKL	ILAQVPFSV	ILDQKINEV	ILDQVPFSV	ILDTGTIQL
220	ILFEPVHGV	ILFGHENRV	ILGFVFTLT	ILHNGAYSL	ILKEPVHGV	ILKEYVHGV	ILKSPVHGV	ILLLCLIFL	ILMEHIHKL	ILMQVPFSV
230	ILSPFMPLL	ILSPLTKGI	ILSQVPFSV	ILTVILGVL	ILWEPVHGV	ILYEPVHGV	IMDKNIILK	IMDQVPFSV	IMIGVLVGV	ITAQVPFSV
240	ITDQVPFSV	ITFQVPFSV	ITMQVPFSV	ITSQVPFSV	ITWQVPFSV	ITYQVPFSV	IVGAETFYV	KACDPHSGH	KARDPHSGH	KASEKIFYV
250	KIFGSLAFL	KILSVFFLA	KINEPVIII	KINEPVIIL	KINEPVILI	KINEPVILL	KINEPVLII	KINEPVLIL	KINEPVLLI	KINEPVLLL
260	KKREEAPSL	KLAEYVAKV	KLAKAAAAV	KLFCQLAKT	KLGEFYNQM	KLHLYSHPI	KLIANNTRV	KLLEPVLLL	KLLPENNVL	KLNEILWSI
270	KLNEPVIII	KLNEPVIIL	KLNEPVILI	KLNEPVILL	KLNEPVLII	KLNEPVLIL	KLNEPVLLI	KLNEPVLLL	KLPAQFYIL	KLPQLCTEL
280	KLTPLCVTL	KLTSLCNTV	KLVALGINA	KLVANNTRL	KMFCQLAKT	KMFYQLAKT	KMVELVHFL	KTWGQYWQV	KVAELVHFL	KVLEYVIKV
290	KYLATASTM	LAGIGLIAA	LERPGGNEI	LIVIGILIL	LLAQFTSAI	LLARNSFEV	LLCLIFLLV	LLDFVRFMG	LLDGTATLR	LLDVPTAAV
300	LLFAGVQCQ	LLFDRPMHV	LLFGYPVYV	LLGANSFEV	LLGATCMFV	LLGRASFEV	LLGRDSFEV	LLGRNAFEV	LLGRNSAEV	LLGRNSEEM
310	LLGRNSFAV	LLGRNSFEM	LLGRNSFEV	LLGRRSFEV	LLIENVASL	LLLCLIFLL	LLLLTVLTV	LLMDCSGSI	LLMGTLGIV	LLNATAIAV
320	LLNATDIAV	LLPENNVLS	LLQYWSQEL	LLSSNLSWL	LLSVPLLLG	LLTEVETYV	LLWAARPRL	LLWFHISCL	LLWKGEGAV	LLWTLVVLL
330	LLYDWDFGL	LMAQEALAF	LMIIPLINV	LMWAKIGPV	LQTTIHDII	LTVILGVLL	LVVLGLLAV	MDHARHGFL	MIMVKCWMI	MLDLQPETT
340	MLGTHTMEV	MLLALLYCL	MLLAVLYCL	MLLSVPLLL	MLMAQEALA	MLWEGFTYI	MMQDIDFYL	MMRKLAILS	MMWYWGPSL	MVDGTLLLL
350	MVDGTTLLL	NLGPWIQQV	NLLPKLHIV	NLQSLTNLL	NLSWLSLDV	NLTISDVSV	NLVPMVATV	NMFCQLAKT	NMFTPYIGV	PILTIITLE
360	PLDGEYFTL	PLKQHFQIV	PLLPIFFCL	PLQPEQLQV	PLSSSVPSQ	PLTFGWCYK	PLTSIISAV	QAGIGILLA	QIRGRERFE	QLAKTCPVQ
370	QLFHLCLII	QLIDKVWQL	QLQARILAV	QLSLLMWIT	QMFCQLAKT	QMVTTTNPL	QVCERIPTI	RGPGRAFVT	RIIPRHLQL	RILGAVAKV
380	RLCVQSTHV	RLDSYVRSL	RLGFLHSGT	RLGRNSFEV	RLIRVEGNL	RLLDYVVNI	RLLQETELV	RLMKQDFSV	RLNMFTPYI	RLPKDFRIL
390	RLPRIFCSC	RLSSNSRIL	RLTRFLSRV	RLVTLKDIV	RMFPNAPYL	RMGAVTTEV	RMPEAAPPV	RMYSPISIL	RTLDKVLEV	RTQDENPVV
400	RVIEVLQRA	SAHKGFKGV	SIIVRALEV	SILVRALEV	SIPSGGIGV	SIPSGGLGV	SLADTNSLA	SLAGGIIGV	SLDDYNHLV	SLDQSVVEL
410	SLFEGIDFY	SLFNTVATL	SLFPGKLEV	SLGGLLTMV	SLHVGTQCA	SLIGHLQTL	SLINVGLIS	SLIVRALEV	SLKKNSRSL	SLLAPGAKQ
420	SLLGGDVVS	SLLGLLVEV	SLLLELEEV	SLLMWITQC	SLLPAIVEL	SLLPPDALV	SLLPPTALV	SLLQHLIGL	SLLVRALEV	SLMAFTAAV
430	SLPDFGISY	SLPSGGIGV	SLPSGGLGV	SLRELGSGL	SLSEKTVLL	SLSRFSWGA	SLVIVTTFV	SLWGQPAEA	SLYADSPSV	SLYAVSPSV
440	SLYGGTTTI	SLYNTIAVL	SLYNTVATL	SLYSFPEPE	SMVGNWAKV	STAPPAHGV	STAPPHVNV	STNRQSGRQ	STPPPGTRV	SVASTITGV
450	SVFAGVVGV	SVRDRLARL	SVYDFFVWL	TGAPVTYST	TITDQVPFS	TIWVDPYEV	TLDDLIAAV	TLDSQVMSL	TLEEITGYL	TLFIGSHVV
460	TLGIVCPIC	TLHEYMLDL	TLIEDILGV	TLIKIQHTL	TLNAWVKVV	TLSKIFKLG	TLSPGKNGV	TLWVDPYEV	TMDHARHGF	TTAEEAAGI
470	TVILGVLLL	VDGIGILTI	VILGVLLLI	VISNDVCAQ	VIYQYMDDL	VKTDGNPPE	VLAGLLGNV	VLAKAAAAV	VLATLVLLL	VLDGLDVLL
480	VLEETSVML	VLEWRFDSR	VLF RB RG	VLS SDFRI	VLHDDLLEA	VLLCESTAV	VLLDQ ML	VLPDVFIRC	VLQ G LL	VLQ ASLAV
490	VLSPLPSQA	VLVKSPNHV	VLYRYGSFS	VMAGVGSPY	VMAPRTLVL	VMNILLQYV	VVHFFKNIV	VVLGVVFGI	VYDGREHTV	WILRGTSFV
500	WLDQVPFSV	WLEPGPVTA	WLNEILWSI	WLSLLVPFV	WLWYIKIFI	WTDQVPFSV	YIGEVLVSV	YLATASTMD	YLDNGVVFV	YLDPAQQNL
510	YLDQVPFSV	YLEPGPVTA	YLEPGPVTI	YLEPGPVTL	YLEPGPVTV	YLGEVIVSV	YLGEVLVSV	YLKEPVHGV	YLKKIKNSL	YLKKIQNSL
520	YLKTIQNSL	YLLEMLWRL	YLLPAIVEL	YLLPAIVHI	YLLPRRGPR	YLNKIQNSL	YLQLVFGIE	YLSGANLNL	YLVAYQATV	YLVSFGVWI
530	YLVTRHADV	YMDDVVLGA	YMDGTMSQV	YMLDLQPET	YMNGTMSQV	YTAFTIPSI	YTDQVPFSV	YVDPVITSI		