CMLS, Cell. Mol. Life Sci. 62 (2005) 1025–1037
1420-682X/05/091025-13
DOI 10.1007/s00018-005-4528-2

CMLS Cellular and Molecular Life Sciences

Research Article

Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding

S. Tenzer[a], B. Peters[b, d], S. Bulik[b], O. Schoor[c], C. Lemmel[c], M. M. Schatz[a], P.-M. Kloetzel[b], H.-G. Rammensee[c], H. Schild[a, *] and H.-G. Holzhütter[b]

[a] Institute of Immunology, Johannes Gutenberg University, Obere Zahlbacherstr. 67, 55131 Mainz (Germany),
[b] Institute for Biochemistry, University Medical School-Charite, Humboldt University, Monbijoustr. 2, 10117 Berlin (Germany), e-mail: schild@uni-mainz.de
[c] Institute for Cell Biology, Department of Immunology, University of Tübingen, Auf der Morgenstelle 15, 72076 Tübingen (Germany)
[d] Present address: La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, San Diego, California 92121 (USA)

Received 26 November 2004; received after revision 4 February 2005; accepted 4 March 2005

Abstract. Epitopes presented by major histocompatibility complex (MHC) class I molecules are selected by a multi-step process. Here we present the first computational prediction of this process based on in vitro experiments characterizing proteasomal cleavage, transport by the transporter associated with antigen processing (TAP) and MHC class I binding. Our novel prediction method for proteasomal cleavages outperforms existing methods when tested on in vitro cleavage data. The analysis of our predictions for a new dataset consisting of 390 endogenously processed MHC class I ligands from cells with known proteasome composition shows that the immunological advantage of switching from constitutive to immunoproteasomes is mainly to suppress the creation of peptides in the cytosol that TAP cannot transport. Furthermore, we show that proteasomes are unlikely to generate MHC class I ligands with a C-terminal lysine residue, suggesting processing of these ligands by a different protease that may be tripeptidyl-peptidase II (TPPII).

Key words. Proteasome; TAP; MHC; epitope prediction; antigen processing.

Most cells of the human body signal their internal status to the immune system by a specific repertoire of peptides derived from intracellular proteins, presented on the cell surface by major histocompatiblity complex (MHC) class I molecules [1] and screened by T cells. The proteasome, a multi-subunit protease complex representing about 1% of total cellular protein [2], has been implicated as the cardinal supplier of proteolytic fragments used for antigen presentation [3, 4]. Its active sites are found in three of the beta subunits of the 20S core particle [5], which is usually associated with the 19S cap to form the 26S proteasome. Upon stimulation with interferon (IFN)-γ, the three proteolytically active beta subunits of the constitutive proteasome X, Y and MB1 are exchanged for the so-called immunosubunits LMP2, LMP7 and MECL1 to form the immunoproteasome [6]. This is associated with alterations in proteasomal cleavage site preference [7–10], favoring the generation of the C-termini of some epitopes and abrogating the generation of others [11–15].

* Corresponding author.
S. Tenzer and B. Peters contributed equally to this work.

The primary proteasomal degradation products can be further processed by TPPII [16] and N-terminally trimmed by cytosolic aminopeptidases [17–19]. Some peptides escape complete degradation as they are transferred from the cytosol into the endoplasmic reticulum (ER) by a special transport system, the transporter associated with antigen presentation (TAP) [20]. Inside the ER, further N-terminal trimming may occur [21, 22] before the final products of the processing pathway, peptides of 8–12 amino acids (aa) in length, are loaded onto MHC class I molecules if they contain the correct binding motif. The MHC class I peptide complexes are then transported via the golgi to the cell surface. These naturally processed MHC class I peptide ligands (called 'MHC I ligands' in the following for brevity) are scanned by cytotoxic lymphocytes (CTLs) which will lyse the target cell if they are activated and recognize a foreign or 'altered self' peptide. MHC I ligands recognized by T cells in this way are called CTL epitopes.

For several areas in immunology, including the identification of CTL epitopes and vaccine design, reliable prediction of MHC I ligands is important [23]. In this paper, we describe such an algorithm encompassing predictions of peptide generation by the proteasome, transport by TAP and binding by MHC class I molecules, each of which is based on in vitro experiments characterizing these steps in isolation. In addition to the practical value of such a prediction, its performance also evaluates whether the underlying classical description of the MHC class I presentation pathway is sufficient to explain the selection of MHC I ligands found in vivo. We previously made a first step toward a combined prediction [24], showing that MHC I ligand identification could be improved when combining MHC-binding predictions with predictions of TAP transportability, if the transport of N-terminally prolonged precursor peptides was taken into account. However, there was little indication for a relevant impact of the proteasome on MHC I ligand generation, when applying proteasomal prediction algorithms available at that time. What remained unclear was if this was due to a lack of experimental proteasomal cleavage data, deficiencies in the selected prediction models or a less selective role of the proteasome in the pathway.

We have now developed a new matrix-based model for the differential prediction of c20S and i20S cleavages which we found to be superior to all previously published prediction methods for proteasomal cleavages. By combining its predictions with TAP transportability [24], we are able to predict the relative amounts of peptides generated from a given protein that are available in the ER for binding to MHC class I molecules. Finally, we combined this with existing predictions of peptide binding to MHC class I molecules.

As a test set, we identified by mass spectrometry a new set of 390 endogenously processed MHC I ligands, eluted from different renal cell carcinomas and cell lines with known proteasome composition. We evaluated how well these and other previously known MHC I ligands can be identified in their natural protein sequences using our combined predictions. This allows us to analyze if the selectivity of the proteasome, TAP [25] and MHC as characterized in in vitro assays is sufficient to explain the selection of the majority of MHC I ligands presented in vivo. More specifically, we could analyze the influence of the proteasomal subunit composition on the presented spectrum of MHC I ligands and the possible influence of other proteases such as TPPII [16, 26] on the generation of MHC I ligand C-termini.

The combined prediction model is available on http://www.mhc-pathway.net.

Materials and methods

Dataset of proteasomal cleavages

Three fully quantified in vitro digests of whole proteins [27–29] were used as training data. From these, 10-mers were generated consisting of the P6–P4' residues surrounding each potential cleavage site. Each cleavage site was associated with the corresponding amount of fragments generated by its usage. If the amount of fragments starting at or ending at a cleavage site differed, the higher value was used. For testing, in vitro digests of several peptides [30–32] and the Nef protein [33] were used, in which fragment amounts were not quantified. From these, 10-mers were generated as described above, which were classified as either cleavable or not cleavable.

Proteasomal cleavage predictions

For the prediction of proteasomal cleavages, we used a modified version of the SMM method described elsewhere [24, 34] using the quantified proteasomal cleavage data described above as training data. We applied the SMM algorithm as follows. For a sequence of 10 amino acids aa_{pos} (notation: 'aa' = amino acid and 'pos' = position in sequence window), a 20×10 matrix mat(aa,pos) can be used to predict the cleavage between positions 6 und 7 by summing up over matrix entries determined by the sequence [prediction = Σ_{pos} mat(aa_{pos}, pos)]. The matrix entries themselves are determined numerically by minimizing the distance Φ between the predicted scores for the 10-mer sequences in the training set and the logarithms of their associated measured amounts $\{\Phi=\Sigma_{\text{cleavage sites}}$ [log(measured amount) – prediction]$^2\}$. In cases of cleavage sites for which no fragments were found, the measured amount was set to the estimated detection threshold of 5 pmol. Any prediction for these cleavage sites equal to or below this threshold is considered perfect (distance = 0). To avoid over-fitting, a second term is added to the minimization function, punishing the deviation of matrix

entries from zero $\Psi = \Phi + \Sigma_{pos} \lambda_{pos} \Sigma_{aa} \text{mat}(aa_{pos}, pos)^2$. By minimizing this objective function with non-zero λ_{pos} values, a tradeoff is introduced between optimally reproducing the experimental values (including their inevitable experimental error) and minimizing the matrix entries $\text{mat}_{aa,pos}$. This in effect prefers 'simpler' solutions, in which matrix entries that do not significantly lower the distance Φ are kept at small values, and larger matrix values are kept at the same scale. Such a procedure is in general called regularization, but is also known as (local) Ridge regression or, in the context of neural networks, as weight decay.

In contrast to previous applications of the SMM method, the regularization parameter λ_{pos} is now position dependent. This is necessary because for proteasomal cleavages it is not clear how wide the sequence window around a potential cleavage site has to be to comprise all significant positions. The optimal values for the λ_{pos} are determined through fivefold cross-validation on the training set. Starting with λ_{pos} values optimal if the predictions were based on each column alone, the final values are determined by searching for a minimal cross-validated distance as a function of the λ_{pos} through steepest descent. This procedure is repeated 20 times, each time using different random splitting of the experimental dataset into training and test data for the cross-validation. The final prediction matrix is given as the mean of the 20 independently obtained matrices. For the outer sequence positions, the optimal λ_{pos} values were high, forcing the corresponding matrix values toward zero. This demonstrates that no reliable information about the influence of amino acids at these outer positions could be extracted.

MHC binding and TAP transport predictions

For the HLA-A*02-binding predictions, an SMM matrix [34] was used. Its values were transformed from natural logarithmic values to $\log_{10}$ values which are used throughout this paper ($\log=\log_{10}$). For all other MHC-binding predictions, previously published ARB matrices were used [35–39]. Their matrix entries were log transformed. If a residue at a given position abrogates binding (ARB=0), its log(ARB) value is set to –3. For HLA-A*01, two ARB matrices were published for different anchor residues at P2. We used only the D,E at P2 residue matrix and excluded other ligands. The TAP predictions were taken from Peters et al. [24], with α=0.2 and the maximal precursor length L set to MHC I ligand length +1. As with the HLA-A*02 matrix, the TAP matrix values were transformed to $\log_{10}$ values. To make comparison and combination of matrix scores easier, the TAP and HLA-A*02 matrices were multiplied by –1, so that a higher score always corresponds to a better cleavage/transport/binding prediction.

Statistical significance for differences in area-under-the-curve values

We used receiver-operator-characteristic (ROC) curves to analyze predictions and the area under the ROC curve (AUC) as a measure of prediction quality, as explained in the text and in Bradley [40]. To assess if one prediction is significantly better than another, we resampled the dataset for which predictions are made. Using bootstrapping with replacement, 50 new datasets were generated with a constant ratio of positives to negatives. We then calculated AUC values for the two predictions for each new dataset. One prediction is called significantly better than another if a paired two-tailed t test shows the two distributions of AUC values to be significantly different with a p value of 0.001.

Dataset of MHC ligands

We extracted all naturally presented MHC class I ligands of 8–12 aa in length that are derived from human proteins from the SYFPEITHI database [25]. The corresponding source proteins were identified using a BLAST search for 'short, nearly exact matches' and the protein sequences were extracted from the GenBank database. If there was no exact match for an MHC I ligand in a protein sequence or if the protein sequence contained undetermined residues, the MHC I ligand was discarded. For all ligands, the flanking sequences and the corresponding LocusLink IDs of the source proteins are given in supplementary table S3 (http://www.mhc-pathway.net/supplement). This dataset is referred to as SYF-human.

Identification of HLA ligands from solid tumors by liquid chromatography-mass spectrometry

Natural HLA ligand pools from solid tumors were analyzed by a reversed-phase Ultimate liquid chromatography (LC) system (Dionex, Amsterdam, Netherlands), coupled to a hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer (Q-TOF; Micromass, Manchester, UK) equipped with a micro-ESI source. Samples were loaded onto a C_{18} pre-column for concentration and desalting. After loading, the pre-column was placed in line for separation by a fused-silica microcapillary column (75 μm internal diameter × 250 mm) packed with 5 μm C_{18} reversed-phase material (Dionex). Solvent A was 4 mM ammonium acetate/water. Solvent B was 2 mM ammonium acetate in 80% acetonitrile/water. Both solvents were adjusted to pH 3.0 with formic acid. A binary gradient of 15–50% B within 120 min was performed, applying a flow rate of 200 μl/min reduced to approximately 300 nl/min by the Ultimate split-system. A gold-coated glass capillary (PicoTip; New Objective, Cambridge, Mass.) was used for introduction into the micro-ESI source. In tandem-mass spectromtry (MS) experiments, sequence information was obtained by interpretation of fragment spectra using computer-assisted database-

searching tools. Between 30 and 50% of the recorded spectra could be attributed to peptides and their source proteins by database-assisted identification techniques.

RNA sources

Renal cell carcinoma (RCC) samples were obtained from the Department of Urology, University of Tübingen. The local ethical committee approved this study, and informed consent was obtained from the patients. All patients had histologically confirmed RCC. Tissue samples were dissected, shock-frozen and ground by mortar and pestle under liquid nitrogen. Total RNA was prepared using TRIzol (Invitrogen, Karlsruhe, Germany) followed by clean-up with RNeasy (Qiagen, Hilden, Germany). Quality and quantity were confirmed on the Agilent 2100 Bioanalyzer (Agilent, Waldbronn, Germany) using the RNA 6000 Pico LabChip Kit (Agilent).

High-density oligonucleotide microarray analysis

Double-stranded DNA was synthesized from 5 μg of total RNA using SuperScript RTII (Invitrogen) and the primer (MWG Biotech, Ebersberg, Germany) as given by the Affymetrix manual (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). In vitro transcription using the BioArray High Yield RNA Transcript Labeling Kit (ENZO Diagnostics, Farmingdale, N. Y.), fragmentation, hybridization on Affymetrix HG-U133A GeneChips (Affymetrix, Santa Clara, Calif.), and staining with streptavidin-phycoerythrin and biotinylated anti-streptavidin antibody (Molecular Probes, Leiden, The Netherlands) followed the manufacturer's protocols (Affymetrix). The Affymetrix GeneArray Scanner was used, and data were analyzed with the Microarray Analysis Suite 5.0 software. For normalization, 100 housekeeping genes provided by Affymetrix were used. Pairwise array comparisons were calculated using the autologous healthy kidney sample as baseline. Significance of differential expression was judged by the 'change' values given by the statistical algorithms implemented in the Microarray Analysis Suite 5.0 software. Probesets reflecting expression of the proteasomal subunits were selected according to the annotations provided by Affymetrix.

Immunoblotting

One microgram of purified proteasomes or 50 μg of tumor lysates was separated by 12% SDS-PAGE by standard techniques and transferred to nitrocellulose membranes (Amersham Pharmacia Biosciences, Freiburg, Germany) with a semidry transfer system. Membranes were blocked overnight in blocking buffer (TBS containing 4% bovine serum albumin and 0.1% Tween-20). Human LMP7 was detected using a rabbit polyclonal antiserum (1:10,000, PW8200; Affiniti Research Products, Exeter, UK); LMP2 was detected using a rabbit polyclonal antiserum (1:10,000, PW8205; Affiniti Research Products); MECL1 was detected by a rabbit polyclonal antiserum (1:50,000, PW8350; Affiniti Research Products). All blots were washed with TBS-Tween (TBS containing 0.1% Tween-20) and then incubated with goat-anti-rabbit-HRP (1:5000, Dianova, Hamburg, Germany) and specific bands detected by chemiluminescence (Western Lightning; PerkinElmer, Wellesley, Mass).

New set of MHC class I ligands established

We identified 79 MHC class I ligands from RCC75, 95 ligands from RCC98 and 123 ligands from RCC68. In the case of RCC75 and RCC98, most of the ligands identified by mass spectrometry could be attributed to MHC alleles expressed by the patients using binding motifs derived from the literature. In the case of RCC68, a unique assignment was not possible, as the MHC class I alleles expressed by this patient showed overlapping binding preferences. The expression of proteasomal subunits is known for these tumors: RCC75 expresses only constitutive beta subunits, whereas RCC68 and RCC98 express high levels of immunosubunits. This dataset was extended with 23 MHC class I ligands from cell lines (LCL 721 and LCL 721.221) expressing mainly immunoproteasomes and 70 MHC class I ligands from cell lines (JY, MGAR) expressing only constitutive proteasomal beta subunits as verified by Western blotting (see supplementary table S4, http://www.mhc-pathway.net/supplement).

In vitro digests of peptides

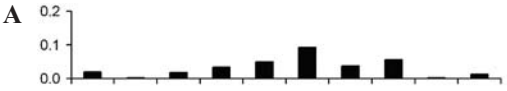
Purification of 20S proteasomes, in vitro degradation of the peptides and separation and analysis of cleavage products were performed as described elsewhere [27]. Peptides [10 nmol (μg)] were incubated for 6 h with 2 μg immunoproteasomes or for 6 h with 2 μg of constitutive proteasomes in digestion buffer (20 mM Tris-HCl, pH 7.6, 10 mM NaCl, 10 mM KCl, 2 mM $MgCl_2$, 0.5 mM dithiothreitol) and the reaction stopped by freezing the reaction mixture at –80ºC.

Results

Generation of novel proteasomal cleavage prediction

We have developed a matrix-based prediction method of proteasomal cleavages that we call ProteaSMM trained on data from in vitro proteasomal digests of yeast enolase-1 [29] and casein [28]. In essence, the prediction method assigns scores to each amino acid located in a 10-residue window around the scissile bond. The sum of these scores is the predicted usage of the cleavage site. We established two scoring matrices using different sets of training data. Both sets contain cleavages from the casein digest made with the constitutive proteasome, one (ProteaSMM-c) also contains cleavages in enolase obtained with constitutive proteasomes (c20S), and the other (ProteaSMM-i) con-

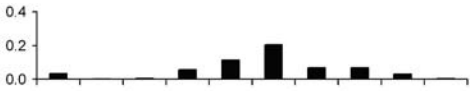
A



	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'
A	-0.01	0.00	0.03	-0.02	-0.02	0.05	**0.11**	-0.04	0.00	-0.01
C	0.00	0.00	0.00	0.00	0.04	-0.02	-0.01	-0.01	0.00	0.00
D	0.03	0.00	-0.03	-0.06	-0.06	**0.19**	-0.04	**0.14**	0.00	-0.01
E	0.01	0.00	0.00	0.01	-0.02	0.03	-0.04	-0.02	0.00	-0.02
F	-0.02	0.00	0.02	0.00	0.03	0.03	0.02	0.03	0.00	0.01
G	-0.03	0.00	0.01	-0.05	0.00	-0.08	-0.03	0.00	0.00	-0.02
H	0.00	0.00	-0.01	0.01	0.04	-0.04	-0.01	0.02	0.00	-0.01
I	0.02	0.00	0.02	0.00	-0.02	-0.01	-0.04	0.06	0.00	0.01
K	-0.04	0.00	-0.03	0.05	0.00	-0.16	0.00	-0.13	0.00	0.02
L	-0.01	0.00	0.00	0.05	-0.05	**0.17**	-0.03	0.10	0.00	0.00
M	0.00	0.00	0.01	-0.03	0.00	0.03	0.04	0.02	0.00	0.00
N	-0.01	0.00	-0.01	-0.01	0.00	-0.08	0.00	-0.02	0.00	-0.01
P	0.04	0.00	0.04	-0.01	-0.13	-0.17	-0.01	-0.05	0.00	0.01
Q	-0.02	0.00	-0.02	0.01	0.02	-0.07	-0.04	-0.02	0.00	0.02
R	0.00	0.00	-0.01	0.02	0.03	0.01	0.04	-0.03	0.00	0.00
S	0.02	0.00	-0.02	-0.07	0.01	-0.02	0.02	-0.06	0.00	0.00
T	0.02	0.00	0.00	0.00	0.06	-0.05	0.00	0.01	0.00	0.00
V	0.00	0.00	0.00	0.06	**0.11**	0.05	-0.01	-0.01	0.00	0.00
W	0.00	0.00	0.00	0.01	-0.03	0.06	0.01	0.00	0.00	-0.01
Y	0.01	0.00	0.01	0.02	-0.02	0.08	0.03	0.01	0.00	0.01

Offset: 0.94

B



	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'
A	-0.02	0.01	0.01	-0.06	0.00	0.06	**0.14**	0.02	0.02	0.01
C	-0.01	0.00	0.00	0.03	0.10	0.00	-0.01	-0.04	-0.01	0.00
D	0.08	0.00	-0.02	-0.05	-0.22	-0.04	-0.07	**0.16**	-0.01	0.01
E	-0.05	0.00	0.00	0.02	0.01	-0.14	-0.02	-0.03	0.01	-0.01
F	-0.04	0.00	0.00	-0.05	0.06	**0.27**	0.00	-0.01	-0.02	0.00
G	0.02	0.00	0.00	-0.10	-0.02	-0.21	0.02	0.00	0.05	0.00
H	0.04	0.00	0.00	-0.01	0.06	-0.05	-0.01	0.07	0.02	-0.01
I	0.01	-0.01	0.01	0.07	0.05	0.10	-0.11	**0.12**	0.01	0.00
K	-0.05	0.00	-0.01	0.10	0.02	-0.17	0.01	-0.12	-0.04	0.00
L	0.04	0.00	0.00	0.02	-0.06	**0.43**	-0.13	0.03	0.06	-0.01
M	-0.02	0.00	0.01	-0.04	0.06	-0.03	0.09	0.02	-0.03	0.00
N	0.00	0.00	0.00	0.01	0.02	-0.12	0.08	0.01	0.06	0.00
P	0.03	0.00	0.01	0.02	-0.34	-0.38	-0.08	-0.03	-0.03	0.00
Q	-0.05	0.00	0.00	0.02	0.04	-0.13	-0.05	-0.02	-0.02	0.00
R	0.00	0.00	0.00	0.01	0.05	-0.01	0.08	-0.11	-0.03	0.00
S	-0.02	0.00	-0.01	-0.10	0.02	-0.07	0.06	-0.06	-0.03	0.00
T	0.03	0.00	0.00	-0.03	**0.13**	-0.14	-0.01	0.08	-0.02	0.00
V	-0.02	0.00	0.00	**0.12**	**0.16**	-0.01	-0.03	-0.02	0.00	0.00
W	0.02	0.00	0.00	0.00	-0.07	**0.36**	-0.02	-0.05	-0.02	0.00
Y	0.00	0.00	0.00	0.03	-0.07	**0.29**	0.05	-0.02	0.02	0.00

Offset: 1.04

Figure 1. Scoring matrices for constitutive (*A*) and immuno-type (*B*) proteasomes. Each matrix column belongs to a sequence position close to a potential cleavage site between P1 and P1′. To predict the usage of this cleavage site, the matrix values determined by the residues surrounding the cleavage site have to be summed and the offset value added. The resulting score estimates the logarithm of the total amount of fragments generated by the cleavage site. Residues significantly enhancing the cleavage probability are in bold, residues suppressing a cleavage are shaded in gray. The bars at the top of the matrices depict the standard deviation of the scores for the corresponding sequence position, indicating the importance of the position for cleavage.

tains the enolase digest with immunoproteasomes (i20S). It would obviously be preferable to use pure training sets of just immuno- or constitutive proteasomal cleavages, but having a similar-sized dataset for each algorithm is also important as is the possibili*y* to compare fairly our predictions to other methods that were generated using the same data.

The ProteaSMM-i and -c prediction matrices are shown in figure 1. The relative importance of a given sequence position for the cleavage preference of the proteasome can be assessed by the variance of its corresponding scores for the 20 possible amino acids (see bar diagram above the scoring matrices in fig. 1). For both the constitutive and the immunoproteasome, the sequence positions P4–P2′ in close proximity to the scissile bond determine most of the matrix prediction, with the P1 position having by far the highest influence. It is important to note that our method only extracts highly evident information from the training data, thereby suppressing the influence of inevitable noise arising from both the finite amount of training data and the experimental error associated with it. More distant positions may, therefore, have an influence on the usage of a cleavage site, but this influence is not captured in the matrix because it is not significant enough in the training data.

Comparison to existing algorithms predicting proteasomal cleavages

There are three published methods predicting proteasomal cleavages trained on in vitro data: FragPredict [41, 42], PaProC 1.0 [43, 44] and NetChop 20S [45]. Also, an as yet unpublished version of PaProC exists that differentiates between immuno-type cleavages (PaProC 2i) and constitutive-type cleavages (PaProC 2c). All PaProC versions and NetChop 20S were trained on data from in vitro proteasomal digests of yeast enolase-1 [29] and casein [28], whereas the FragPredict method was trained on a limited set of peptide digests. To allow for a fair comparison of the prediction methods, we retrained FragPredict using the enolase and casein data.

An alternative approach to proteasomal cleavage predictions not using in vitro digests is made by NetChop C2.0, which bases its predictions on the analysis of C-termini of known MHC I ligands.

To compare the quality of the above prediction methods, we collected several sets of sequences for which in vitro digests were described in the literature or performed in our laboratory (see supplementary table S1) (http://www.mhc-pathway.net/supplement). For each of these test sets, the cleavage predictions were used to classify peptide bonds as cleavable or not cleavable, depending on the predicted score being above or below a given cutoff. Comparing the predictions with the experimental results, one can calculate the rate of true-positive and false-positive predictions. By varying the classification cutoff, one can calculate a ROC curve like that shown in figure 2. The AUC can be used as a measure of prediction quality [40]: its value tends to 1.0 for a perfect prediction method and to 0.5 for random predictions. Table 1 contains AUC values for all test sets and prediction methods, and notes what type of proteasome was used in each digest. For an overall comparison of prediction performance, we com-

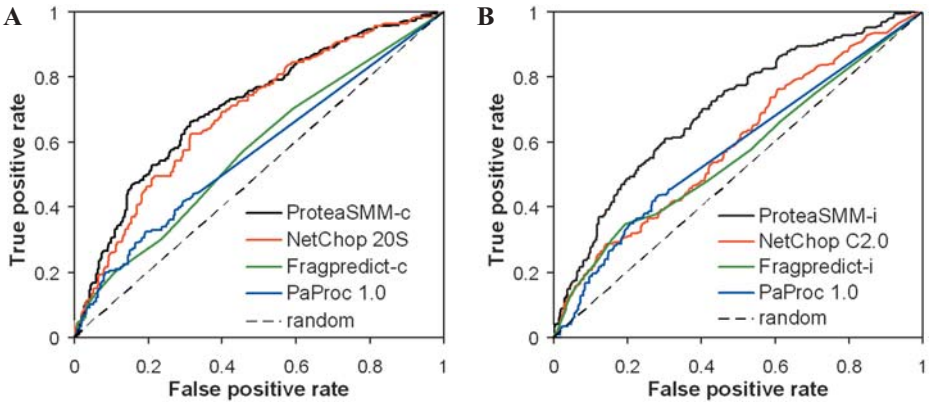


Figure 2. Prediction of in vitro proteasomal cleavages. ROC curves evaluating the ability to predict cleavages found in in vitro digests. For cleavage data from constitutive proteasomes (*A*), the ProteaSMM-c method (AUC=0.71) makes the best predictions closely followed by NetChop 20S (AUC=0.69). Contrary to our expectations, the best results for the constitutive dataset are obtained by ProteaSMM-i (AUC=0.73, ROC curve not shown). For cleavage data derived from digests using immunoproteasomes (*B*), the ProteaSMM-i method again clearly makes the best predictions (AUC= 0.70).

bined all digests generated with constitutive proteasomes in one pool and immunoproteasomal data in the other. For each pool, the prediction models that are trained on the corresponding proteasome type are used to draw the ROC curves in figure 2. The statistical analysis shows that for all except one dataset, the ProteaSMM-i method outperforms the ProteaSMM-c method. Unexpectedly, this includes several test sets derived from constitutive proteasomes, for which the ProteaSMM-c method is expected to do better. We can only speculate on the reason for this finding. One possible explanation is the presence of three major cleavage specificities (tryptic, chymotryptic and PGPH-like) in the constitutive proteasome compared to only two major specificities (tryptic and chymotryptic) in the immunoproteasome. The two shared cleavage specificities might be better simulated in the ProteaSMM-i algorithm due to the lesser degree of complexity of the training data.

It is generally accepted that the vast majority of the C termini of MHC I ligands are generated by the proteasome whereas the N termini may result from trimming of N-terminally extended precursor peptides. Therefore, the efficiency with which an MHC I ligand is generated by the proteasome can be estimated by the predicted cleavage strength after the residue forming its C-terminus. Figure 3 contains gray ROC curves, in which the predicted proteasomal generation efficiencies are used to identify known MHC I ligands within their protein source sequences. This test set comprises all MHC I ligands available to us (SYF-human, RCC-c20S and RCC-i20S datasets) and is described in Materials and methods. The AUC value for predictions on the basis of the ProteaSMM-i matrix

Table 1. Comparison of proteasomal cleavage predictions using in vitro cleavage data from proteins and peptides.

Digest substrate	Proteasome type in digest		Prediction method								
			NetChop		FragPredict		PaProc			ProteaSMM	
	c20S	i20S	20S	C2.0	-c	-i	1	2c	2i	-c	-i
Peptide mix	X		0.71	0.65	0.54	0.56	0.63	0.61	0.60	0.72	**0.74**
Peptide mix		X	0.74	0.67	0.55	0.63	0.62	0.56	0.58	0.74	**0.80**
Peptides ssx2	X		0.78	0.70	0.58	0.69	0.56	0.60	0.67	**0.81**	**0.81**
Peptides prame		X	0.79	0.67	0.68	0.65	0.61	0.57	0.56	0.80	**0.82**
Prion	X		0.59	0.57	0.60	0.60	0.52	0.60	0.54	0.64	**0.68**
Prion		X	0.60	0.50	0.55	0.51	0.53	0.63	0.52	0.63	**0.65**
nef	X	X	0.63	0.62	0.60	0.61	0.52	0.53	0.57	0.67	**0.69**
Prediction of MHC I ligand C-termini			0.69	(0.83)*	0.67	0.68	0.60	0.64	0.58	0.67	**0.76**

*As the NetChop C2.0 prediction was derived from such C-termini and not from in vitro cleavage data, its AUC value cannot be compared with the other predictions.The area under the ROC curves are indicated for each test dataset and each prediction method. The training sets of all methods including ProteaSMM-c and ProteaSMM-i did not include any data in the test sets. The values for the best prediction of each dataset are printed in bold. The lowest row does not contain in vitro digest data, but C-termini of known epitopes from the complete epitope dataset described in Materials and methods.

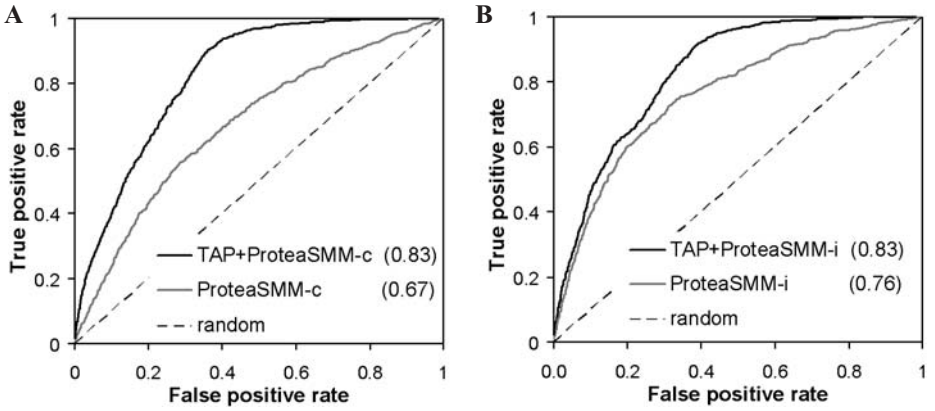


Figure 3. Identification of MHC I ligands by predicting their antigen-processing efficiency. ROC curves are used to measure the quality of proteasomal cleavage predictions (gray line) and the combined TAP+proteasome predictions (black line) in identifying MHC I ligands in their protein source sequences. The evaluations for the constitutive (*A*) and, for the immunoproteasomal (*B*) cleavage algorithm are depicted. The complete MHC I ligand dataset is used.

amounts to 0.76 which is much higher than the AUC value for predictions with the ProteaSMM-c matrix (0.67), or any other proteasomal cleavage prediction methods based on in vitro digests (compare the lowest row of table 1). All AUC values are significantly above 0.5 expected for random predictions, demonstrating that on average the proteasome generates MHC I ligands (or N-terminally extended precursors of them) with higher efficiency than other peptides.

The set of source protein sequences used is very likely to contain more MHC I ligands than known to us. We treat all peptides contained in these proteins not known to be MHC I ligands as negatives. Does this make the above analysis invalid? Not if one can assume that the unknown ligands are not systematically different from the known ligands. In that case, the rate of ligands identified by a given prediction at a given cutoff (true-positive rate) would remain unchanged if the additional ligands were known. The rate of false-positives at each cutoff would even be reduced. Therefore, treating all unknown peptides as non-ligands will likely make our epitope identification look worse than it is. Importantly, this should affect all prediction methods equally, so that any comparison in prediction quality should remain valid.

Combining proteasome and TAP predictions

The affinity of a peptide to TAP $IC50_{binding}$ has been shown to be proportional to its transport efficiency $IC50_{transport}$ [46]. Therefore, the amount of peptide transported into the ER is proportional to the amount generated in the cytosol divided by the $IC50_{TAP}$ value, if transport is competitive. The score predicted by ProteaSMM for a cleavage site is proportional to the logarithm of the amount of peptides generated by its usage. We combine this with our predictions of TAP transport for the potential MHC I ligand and its N-terminal precursors generated by this cleavage, which gives logarithms of TAP IC50 values as an output. By adding the two values for a potential MHC I ligand, one essentially predicts the logarithm of the total amount of ligand and its N-terminally prolonged precursors present in the ER. Using these predicted amounts in the ER to identify MHC I ligands within their natural protein sequences gives the black ROC curves in figure 3 with an AUC value of 0.83. This is significantly better than the ROC curves of either proteasome prediction or the TAP prediction alone (compare table 2). In combination with TAP, the difference between the ProteaSMM-c and ProteaSMM-i predictions is small.

Influence of proteasome subunit composition on the selection of MHC I ligands

To better assess the impact of proteasomal subunits on MHC I ligand generation, we established a high-quality dataset of MHC I ligands extracted from RCCs with known proteasome composition. RCC75 expresses only constitutive beta subunits, whereas RCC68 and RCC98

Table 2. AUC values for combined proteasome and TAP predictions

MHC I ligand dataset	ProteaSMM-i	ProteaSMM-c	TAP only	TAP +	
				ProteaSMM-i	ProteaSMM-c
RCC-i	0.785	0.668	0.844	0.862	0.857
RCC-c	0.755	0.684	0.818	0.831	0.825
Complete	0.761	0.674	0.814	0.831	0.827

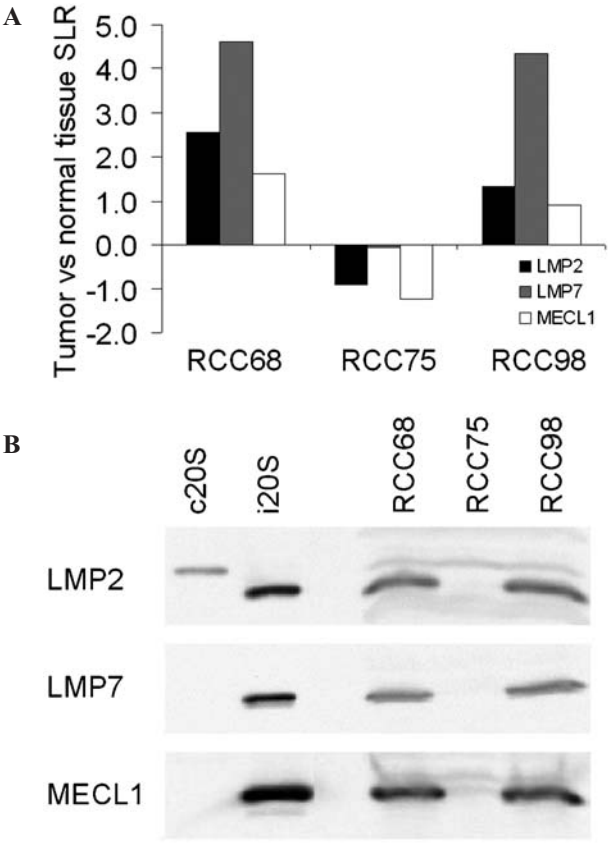


Figure 4. Expression of proteasomal immunosubunits in RCC-68, RCC-75 und RCC-98 as determined by gene chip analysis (*A*) and Western blotting (*B*). (*A*) Signal-$\log_2$ ratios (tumor vs normal tissue SLR) of the mRNA expression levels of the immunosubunits LMP2 (black bars), LMP7 (gray bars) and MECL1 (white bars) in tumor vs normal tissue as detected by Affymetrix gene chip analysis. In normal renal tissue, no expression of the immunosubunits was detectable. (*B*) Expression of the immunosubunits in tumor lysates as detected by Western blotting against the immuno subunits LMP2, LMP7 and MECL1. Purified c20S and i20S proteasomes were loaded as controls.

tumors express high levels of immunosubunits as determined by gene expression analysis (see fig. 4A). The presence or absence of immunoproteasomes in the tumor tissue was additionally verified by immunoblotting against the immunosubunits LMP2, LMP7 and MECL1 (see fig. 4B). To extend this dataset, we also included 93 MHC class I ligands from cell lines with known proteasome composition. The complete set contains 149 non-redundant ligands extracted from cells with constitutive proteasomes (RCC-c) and 241 ligands extracted from cells expressing an immuno-type proteasome (RCC-i). Table 2 compares the predictions of ProteaSMM-c and -i for these datasets. Similar to the results reported for the complete MHC I ligand dataset above, ProteaSMM-i predictions are substantially better for both RCC datasets. This is surprising at first, given the absence of immunoproteasomes from the cells in which the MHC I ligands of the RCC-c dataset were derived. However, once the ProteaSMM predictions are combined with TAP, the differences in prediction quality between the constitutive and immuno predictions again vanish. This suggests that which MHC I ligands are present on the cell surface in large amounts does not depend very much on the type of proteasome present in the cell, because the TAP selectivity filters out peptides that can be produced exclusively by constitutive proteasomes, but not the immuno-type.

To test this hypothesis more directly, we predicted TAP transport efficiencies for peptides found in the proteasomal digests of yeast enolase-I. We analyzed three different pools of peptide: (i) peptides with a C-terminus that is exclusively generated by c20S proteasomes, (ii) peptides with a C-terminus that is generated by both c20S and i20S proteasomes and (iii) peptides with a C-terminus that is exclusively generated by i20S proteasomes. When looking at the mean predicted TAP transport efficiencies of peptides associated with these C-termini, pools B and C show similar values, whereas pool A shows a markedly reduced TAP transport efficiency (A: –0.41±0.06; B: 0.00±0.06; C: 0.09±0.06). This clearly supports the above hypothesis that the immunoproteasome avoids generating peptides that are unfit for TAP transport.

Increasing the training-set size for proteasomal cleavage predictions

Throughout this work, we used only the data derived from proteasomal digests of enolase and casein to train our cleavage predictions, which makes them comparable to previously published methods. To further improve predictions, we also included cleavage data from the prion digests [27] in the training dataset. This allows us to train predictions purely on immunoproteasomal digests (prion and enolase) or on only constitutive digests (prion, enolase and casein). As hoped, this significantly improves the identification of MHC I ligands using the proteasome predictions alone from AUC=0.674 to AUC=0.723 for ProteaSMM-c and from AUC=0.761 to AUC=0.777 for ProteaSMM-i on the complete MHC I ligands dataset. In combination with the TAP predictions, however, there was no significant improvement to the results reported above when including the prion data (from 0.827 to 0.828 for ProteaSMM-c and from 0.831 to 0.829 for ProteaSMM-i). Including the prion data in the training set also leaves less data to test the in vitro cleavage predictions. Evaluating predictions on the remaining in vitro cleavage data gives ambiguous results. For the peptide digests, the predictions including the prion cleavage data in their training sets perform significantly worse than those without, while predictions for the only remaining protein digest (Nef) improve marginally. The prediction based on the immunoproteasomal digests still outperforms the constitutive predictions, though to a lesser extent.

Interpreting differences in prediction quality of MHC I ligands for different proteasomal prediction methods is immensely difficult, since proteasomal cleavage is only one factor in their generation, and arguably the least influential. An improvement in ligand identification for a new prediction can therefore occur even if the function of the proteasome itself is modeled less acurately, but overlaps better with other processing steps, such as TAP transport. In summary, without demonstrating that in vitro cleavages are also better predicted, it is safer to use the original smaller training set.

Comparison with NetChop C2.0

The NetChop C2.0 prediction method is based on analyzing the C-termini of known naturally presented ligands. These C-termini not only reflect the selectivity of the proteasome but also that of TAP and the average C-terminal-binding preference of MHC molecules. Therefore, the predictions of NetChop C2.0 also do not purely mirror the specificity of the proteasome. This is reflected in table 1, where the NetChop C2.0 predictions are inferior in their prediction of proteasomal cleavages to the NetChop 20S predictions or either ProteaSMM method. However, the NetChop C2.0 prediction is very good at identifying C-termini of MHC I ligands, reaching an AUC value of 0.833 on the complete MHC I ligand dataset. This is the same level of prediction quality reached with either the ProteaSMM-c + TAP or ProteaSMM-i + TAP predictions. The fact that the NetChop C2.0 predictions are not significantly better than the combined proteasome + TAP predictions shows that the latter are sufficient to explain the generation of the majority of MHC I ligand C-termini.

Table 3. Comparison of predicted proteasomal cleavage and TAP transport efficiencies of C-terminal amino acids with their frequency in MHC ligands.

	ProteaSMM-i	ProteaSMM-c	Mean	TAP	Number of epitopes (n = 925)
L	0.43	0.17	0.30	0.41	262
W	0.36	0.06	0.21	0.38	20
Y	0.29	0.08	0.19	1.26	174
F	0.27	0.03	0.15	1.09	96
D	–0.04	0.19	0.07	**–0.80**	2
A	0.06	0.05	0.06	**–0.24**	22
I	0.10	–0.01	0.04	0.22	54
V	–0.01	0.05	0.02	0.13	118
M	–0.03	0.03	0.00	0.13	20
R	–0.01	0.01	0.00	0.64	46
C	0.00	–0.02	**–0.01**	0.00	2
S	–0.07	–0.02	**–0.05**	**–0.98**	3
H	–0.05	–0.04	**–0.05**	**–0.24**	4
E	–0.14	0.03	**–0.06**	**–0.68**	2
N	–0.12	–0.08	**–0.10**	**–0.58**	0
T	–0.14	–0.05	**–0.10**	**–0.31**	3
Q	–0.13	–0.07	**–0.10**	**–0.05**	4
G	–0.21	–0.08	**–0.15**	**–0.61**	1
K	–0.17	–0.16	**–0.17**	0.20	***86***
P	–0.38	–0.17	**–0.28**	0.04	6

Amino acids are sorted by their mean proteasomal cleavage score. Values indicating below average cleavage by proteasomes or transport by TAP are in bold. Lysine (K) residues occur much more frequently at C-termini of MHC ligands than expected from their proteasomal cleavage score.

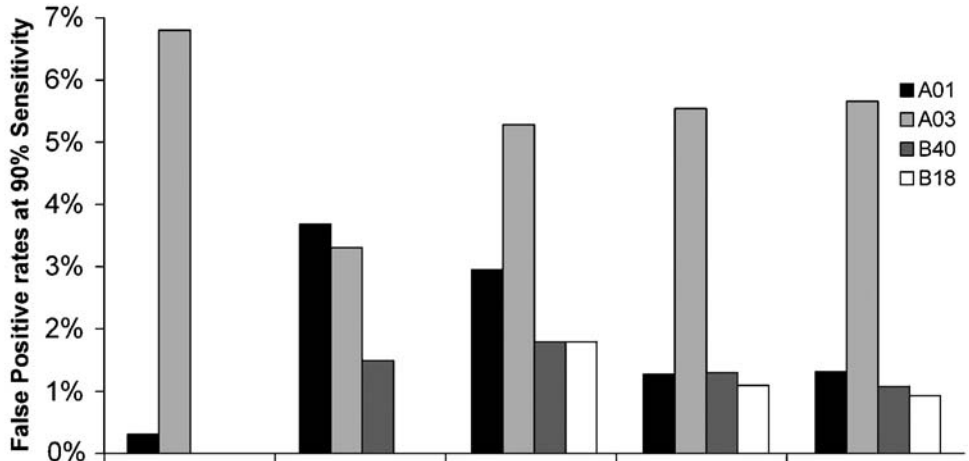
MHC-I ligands with a C-terminal lysine (P1K ligands)

By far the most influential residue for the proteasome and TAP prediction is the C-terminus of a potential MHC I ligand. Table 3 shows how often an amino acid is found at the C-terminus of an MHC I ligand compared with the prediction score that this C-terminus receives in the proteasome and TAP predictions. In general, C-termini that are predicted to be favorably cleaved by proteasomes and not poorly transported by TAP are found in high numbers at the C-termini of MHC I ligands. However, one exception is obvious: nearly 10% of known human MHC I ligands have a lysine at the C-terminus (P1K ligands), even though they are not predicted to be preferably cleaved by either constitutive or the immunoproteasomal algorithm. TAP does not specifically favor or restrict the presentation of these P1K ligands. Taken together, a protease other than the proteasome is likely involved in the generation of the C-termini of P1K ligands.

Further evidence for the involvement of an additional protease could be found when analyzing the P1′ position of MHC I ligands, which is only seen by the protease generating the C-termini of the ligands. When comparing the amino acid distribution at position P1′ for P1K ligands with that of other MHC I ligands (supplementary fig. 1 (http://www.mhc-pathway.net/supplement/)), we found a significantly higher frequency of F, V, L, M and I for P1K ligands than in the P1′ position of all other MHC I ligands, whereas amino acids like G, Q, S and T were found in significantly lower frequencies for P1K ligands. This is in conflict with the proteasomal cleavage motif, which tries to avoid residues L, I and V in the P1′ position. We also analyzed the P2′ and P3′ residues associated with P1K ligands and found smaller, but relevant differences to all other MHC I ligands (data not shown).

Combination with MHC-binding prediction

To predict the binding affinity of peptides to MHC class I molecules, we used ARB affinity matrices determined by in vitro binding studies, which have been published by the Sette group for a large number of MHC alleles [35–39]. We normalized these matrices so that they approximately give logarithms of IC50 values as an output, which can then be added to the score of the proteasome and TAP predictions. To evaluate these predictions, we first selected alleles for which an affinity matrix is available and for which at least 10 MHC I ligands are contained in the SYF-human dataset. The average AUC value



AUC	SYFPEITHI	BIMAS	MHC only	MHC+TAP +ProteaSMM-c	MHC+TAP +ProteaSMM-i
HLA-A01	**0.9989**	0.9934	0.9913	0.9959	0.9962
HLA-A03	0.9743	0.9774	**0.9785**	0.9769	0.9757
HLA-B40	-	0.9933	0.9930	0.9949	**0.9954**
HLA-B18	-	-	0.9949	0.9970	**0.9974**

Figure 5. AUC values and the corresponding false-positive rates of different MHC I ligand prediction algorithms at a sensitivity of 90%. The RCC dataset is used to compare the quality of the different algorithms. The AUC values for the best prediction of each dataset are printed in bold.

of the MHC affinity predictions alone is 0.965. This already very high prediction quality is consistently improved for every allele when adding the proteasome and TAP predictions, to an average AUC of 0.971 and 0.970 for the ProteaSMM-c and ProtaSMM-i predictions, respectively. Supplementary table S2 (http://www.mhc-pathway.net/supplement/) gives AUC values for individual alleles.

To compare fairly the prediction quality of our combined prediction with existing methods to identify MHC I ligands such as SYFPEITHI [25] or BIMAS [47], the above SYF-human ligand dataset cannot be used, as knowledge of these MHC I ligands was used in the establishment of those methods. Instead, we took the new RCC dataset containing only previously unknown MHC I ligands. The comparison in figure 5 shows that our combined prediction has a quality similar to that of SYFPEITHI and BIMAS. Interestingly, the HLA-A*03 dataset is the only one predicted better with the MHC prediction alone. This can be explained by the fact that many MHC I ligands presented by HLA-A*03 are P1K ligands, where the C-terminus is very likely not generated by the proteasome. For the other three cases, the combination with ProteaSMM-i outperforms that of the ProteaSMM-c prediction slightly but significantly.

Alternative scheme to combine prediction scores

As an alternative to adding the prediction scores for proteasome, TAP and MHC as we did above, one could establish score thresholds for each prediction, and demand that a potential ligand has a prediction better than the threshold for each step. As an example, we chose thresholds corresponding to a 99% true-positive rate for the TAP and proteasome predictions on the complete MHC I ligand dataset. The corresponding false-positive rates were 97.0% for the ProteaSMM-c, 93.9% for ProteaSMM-i and 66.4% for TAP. Combining proteasome and TAP predictions resulted in a reduction of false positives to 64.0% for the ProteaSMM-c and 62.6% for the ProteaSMM-i predictions. This demonstrates that this is a valid alternative approach, which can be extended in a similar manner to the combination with the MHC-binding predictions (data not shown). However, we did not find a threshold combination that lead to consistently better overall results than the addition of scores.

Practical comparison of prediction methods

AUC values are a powerful tool to compare prediction efficiencies. However, the improvement in AUC values achieved by a combined prediction seems to be only marginal if one is not familiar with the evaluation of these scores. For example, the AUC value for the prediction of HLA-A*01 ligands changes from 0.991 to 0.996 when combining MHC-binding predictions with those of TAP and proteasome. These figures become more meaningful when translated into the number of peptides required to be analyzed for the positive identification of MHC I ligands (see fig. 5). If a researcher wants to find all HLA-

A*01 MHC I ligands in a given protein with a reliability of above 90%, he/she has to synthesize the top 3.0% of scored peptides when relying on the MHC-binding prediction (AUC = 0.991). When changing to the combined prediction (AUC = 0.996), only the top 1.2% of scored peptides need to be tested to identify all MHC I ligands with the same probability of 90%. This means that the number of peptides to be analyzed decreases from 30 to 12 for a virtual protein with the length of 1000 amino acids.

Discussion

For most MHC I ligands, proteasomal generation of the correct C-terminus is the first step in antigen processing [3]. Here, we have developed algorithms predicting proteasomal cleavages allowing a discrete analysis of constitutive versus immunoproteasomal function. These predictions outperform other publicly available algorithms when compared on their ability to predict in vitro cleavages of small peptides, whole-protein digests or the C-termini of MHC I ligands. The latter is particularly important, as the poor correlation between previous predictions of proteasomal cleavage and observed MHC I ligand C-termini had raised doubts about whether proteasomes are truly selective in the MHC class I antigen presentation pathway [24].

By combining the proteasomal cleavage predictions with those for TAP transport, we essentially predict the pool of peptides available for MHC binding in the ER. This allows one to screen protein sequences for potential MHC I ligands independently of the type of MHC allele with a higher efficiency than either TAP or proteasome predictions alone. Such a screening can be used to disqualify very unlikely candidates (the bottom 50% of scored peptides contain only 3% of all MHC I ligands) or to come up with a shortlist of high-quality candidates (the top 1.3% of scored peptides contain 10% of all MHC I ligands). If an MHC-binding prediction is available, the rate of false positives drops significantly (averaging over all alleles in this study, the top 0.06% scored peptides contain 10% of all ligands), but obviously only for ligands of that allele. This makes the combined proteasome and TAP prediction especially valuable to identify ligands for MHC alleles with poorly characterized binding specificity, but also to identify a set of peptides likely to contain ligands independent of MHC haplotype.

The observed agreement of our predictions with the distribution of MHC I ligands shows that either the impact of antigen processing by non-classical pathways of antigen generation [48] is small or the selectivity of non-classical processing is similar to that of the standard proteasome + TAP model. One proposed alternative pathway for the generation of MHC I ligand C-termini involves TPPII [26]. This protease seems to be mainly responsible for the N-terminal trimming of primary proteasomal degradation products [16], but obviously also generates C-termini of new potential MHC I ligands when the trimming products are long enough to contain MHC-I ligands. As the only known MHC I ligand produced exclusively by TPPII has a lysine at the C-terminus [49], we investigated if evidence could be found that more of these P1K ligands are produced by TPPII rather than proteasomes. Two observations make this very likely. First, the number of P1K ligands in the entire ligand set is much higher than expected. According to our proteasome predictions, very few P1K ligands are expected, because a lysine residue at the P1 position is unfavorable for the necessary cleavage between P1 and P1′. Second, the distribution of residues at the P1′ position after P1K ligands shows an increase of hydrophobic residues (L, V and I), which is not seen for other MHC I ligands. As hydrophobic residues at P1′ are also unfavorable for proteasomal cleavage between P1 and P1′, a different protease is likely responsible for the cleavages generating the C-terminus of P1K ligands.

One main goal of our work was to analyze the role of proteasomal subunit composition in MHC I ligand generation. Several in vitro digests have shown differences in the cleavage patterns of constitutive and immunoproteasomes [27, 29], but there is also a large overlap in their cleavage preferences. Analyzing the effect of the proteasome on MHC I ligand generation is complicated by the fact that for most MHC I ligands we do not know if they were generated by constitutive or immunoproteasomes. We therefore established two large datasets of previously unknown MHC I ligands with defined proteasome composition (either constitutive or immuno) in the source tissue. Surprisingly, predicting C-termini of MHC I ligands is always more successful when using the prediction method based on data from the immunoproteasome. The reason for this becomes clear when combining the proteasomal prediction with TAP, which levels the observed differences in cleavage prediction for the two proteasome species. We interpret this as follows. The constitutive proteasome has a broader specificity, producing several peptides that do not match the specificity of human TAP, above all those having a negatively charged amino acid (D, E) at the C-terminus. Using a combined prediction, these are efficiently filtered out by the TAP prediction method. The immunoproteasome produces a more limited range of peptides which match the TAP specificity. From this perspective, the immunoproteasome has not evolved that much to change the pattern of peptides on the cell surface, but simply to generate potential MHC I ligands with a higher efficiency.

Combining the proteasome and TAP predictions with MHC-binding predictions derived from affinity matrices, we have established the first complete pathway model based solely on the in vitro characterization of each com-

ponent. The quality of predictions of this combined model is equal or better than those made using BIMAS, which is based on MHC binding or SYFPEITHI, which was developed on the basis of known presented MHC I ligands. The quality of our prediction shows that our classical model of the MHC pathway is capable of explaining the selection of most MHC I ligands. We have set up a website (http://www.mhc-pathway.net) on which the presentation of MHC class I ligands for a large number of MHC class I alleles (A01, A02, A03, A11, A24, A31, A33, A68, B07, B18, B35, B40, B44, B45, B51, B53, B54 for 9-mers and A01, A24, B18, B40, B44, B45 for 10-mers) can be predicted. These predictions will be extended and improved as more and better binding matrices become available. A similar approach for MHC class II ligand prediction is conceivable, once enough experimental data characterizing the individual steps in MHC class II ligand processing are available.

Acknowledgements. This work was supported by the EU, QLQ2-CT-1999-00713 and the NIH to H.-G. R., the Deutsche Forschungsgemeinschaft (SFB 490-B7, SFB 510-C1 and priority program 1045) to H. S. and the Deutsche Forschungsgemeinschaft, priority program 1045 to H.-G. H. We thank Prof. A. Stenzl for the RCC samples.

1 Falk K., Rotzschke O., Stevanovic S., Jung G. and Rammensee H. G. (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature **351:** 290–296
2 Coux O., Tanaka K. and Goldberg A. L. (1996) Structure and functions of the 20S and 26S proteasomes. Annu. Rev. Biochem. **65:** 801–847
3 Kloetzel P. M. (2001) Antigen processing by the proteasome. Nat. Rev. Mol. Cell Biol. **2:** 179–187
4 Rock K. L., York I. A., Saric T. and Goldberg A. L. (2002) Protein degradation and the generation of MHC class I-presented peptides. Adv. Immunol. **80:** 1–70
5 Groll M., Ditzel L., Lowe J., Stock D., Bochtler M., Bartunik H. D. et al. (1997) Structure of 20S proteasome from yeast at 2.4 Å resolution. Nature **386:** 463–471
6 Groettrup M., Ruppert T., Kuehn L., Seeger M., Standera S., Koszinowski U. et al. (1995) The interferon-gamma-inducible 11 S regulator (PA28) and the LMP2/LMP7 subunits govern the peptide production by the 20 S proteasome in vitro. J. Biol. Chem. **270:** 23808–23815
7 Eleuteri A. M., Kohanski R. A., Cardozo C. and Orlowski M. (1997) Bovine spleen multicatalytic proteinase complex (proteasome): replacement of X, Y, and Z subunits by LMP7, LMP2, and MECL1 and changes in properties and specificity. J. Biol. Chem. **272:** 11824–11831
8 Boes B., Hengel H., Ruppert T., Multhaup G., Koszinowski U. H. and Kloetzel P. M. (1994) Interferon gamma stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes. J. Exp. Med. **179:** 901–909
9 Gaczynska M., Rock K. L. and Goldberg A. L. (1993) Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. Nature **365:** 264–267
10 Cardozo C. and Kohanski R. A. (1998) Altered properties of the branched chain amino acid-preferring activity contribute to increased cleavages after branched chain residues by the 'immunoproteasome'. J. Biol. Chem. **273:** 16764–16770
11 Sijts A. J., Ruppert T., Rehermann B., Schmidt M., Koszinowski U. and Kloetzel P. M. (2000) Efficient generation of a hepatitis B virus cytotoxic T lymphocyte epitope requires the structural features of immunoproteasomes. J. Exp. Med. **191:** 503–514
12 Schwarz K., Broek M. van den, Kostka S., Kraft R., Soza A., Schmidtke G. et al. (2000) Overexpression of the proteasome subunits LMP2, LMP7, and MECL-1, but not PA28 alpha/beta, enhances the presentation of an immunodominant lymphocytic choriomeningitis virus T cell epitope. J.Immunol. **165:** 768–778
13 Morel S., Levy F., Burlet-Schiltz O., Brasseur F., Probst-Kepper M., Peitrequin A. L. et al. (2000) Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. Immunity **12:** 107–117
14 Sijts A. J., Standera S., Toes R. E., Ruppert T., Beekman N. J., Veelen P. A. van et al. (2000) MHC class I antigen processing of an adenovirus CTL epitope is linked to the levels of immunoproteasomes in infected cells. J.Immunol. **164:** 4500–4506
15 Hall T. van, Sijts A., Camps M., Offringa R., Melief C., Kloetzel P. M. et al. (2000) Differential influence on cytotoxic T lymphocyte epitope presentation by controlled expression of either proteasome immunosubunits or PA28. J. Exp. Med. **192:** 483–494
16 Reits E., Neijssen J., Herberts C., Benckhuijsen W., Janssen L., Drijfhout J. W. et al. (2004) A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation. Immunity **20:** 495–506
17 Beninga J., Rock K. L. and Goldberg A. L. (1998) Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase. J. Biol. Chem. **273:** 18734–18742
18 Levy F., Burri L., Morel S., Peitrequin A. L., Levy N., Bachi A. et al. (2002) The final N-terminal trimming of a subaminoterminal proline-containing HLA class I-restricted antigenic peptide in the cytosol is mediated by two peptidases. J. Immunol. **169:** 4161–4171
19 Stoltze L., Schirle M., Schwarz G., Schroter C., Thompson M. W., Hersh L. B. et al. (2000) Two new proteases in the MHC class I processing pathway. Nat. Immunol. **1:** 413–418
20 Lauvau G., Kakimi K., Niedermann G., Ostankovitch M., Yotnda P., Firat H. et al. (1999) Human transporters associated with antigen processing (TAPs) select epitope precursor peptides for processing in the endoplasmic reticulum and presentation to T cells. J. Exp. Med. **190:** 1227–1239
21 Saric T., Chang S. C., Hattori A., York I. A., Markant S., Rock K. L. et al. (2002) An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. Nat. Immunol. **3:** 1169–1176
22 Serwold T., Gonzalez F., Kim J., Jacob R., and Shastri N. (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. Nature **419:** 480–483
23 Nussbaum A. K., Kuttler C., Tenzer S. and Schild H. (2003) Using the World Wide Web for predicting CTL epitopes. Curr.Opin.Immunol. **15:** 69–74
24 Peters B., Bulik S., Tampe R., Endert P. M. van and Holzhutter H. G. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J. Immunol. **171:** 1741–1749
25 Rammensee H. G., Bachmann J., Emmerich N. P. N., Bachor O. A. and Stevanovic S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics **50:** 213–219
26 Yewdell J. W. and Princiotta M. F. (2004) Proteasomes get by with lots of help from their friends. Immunity **20:** 362–363
27 Tenzer S., Stoltze L., Schonfisch B., Dengjel J., Muller M., Stevanovic S. et al. (2004) Quantitative analysis of prion-protein degradation by constitutive and immuno-20S pro-

teasomes indicates differences correlated with disease susceptibility. J. Immunol. **172:** 1083–1091
28 Emmerich N. P. N., Nussbaum A. K., Stevanovic S., Priemer M., Toes R. E. M., Rammensee H. G. et al. (2000) The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate. J. Biol. Chem. **275:** 21140–21148
29 Toes R. E. M., Nussbaum A. K., Degermann S., Schirle M., Emmerich N. P. N., Kraft M. et al. (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. J. Exp. Med. **194:** 1–12
30 Kessler J. H., Beekman N. J., Bres-Vloemans S. A., Verdijk P., Veelen P. A. van, Kloosterman-Joosten A. M. et al. (2001) Efficient identification of novel HLA-A(*)0201-presented cytotoxic T lymphocyte epitopes in the widely expressed tumor antigen PRAME by proteasome-mediated digestion analysis. J. Exp. Med. **193:** 73–88
31 Ayyoub M., Stevanovic S., Sahin U., Guillaume P., Servis C., Rimoldi D. et al. (2002) Proteasome-assisted identification of a SSX-2-derived epitope recognized by tumor-reactive CTL infiltrating metastatic melanoma. J.I mmunol. **168:** 1717–1722
32 Peters B., Janek K., Kuckelkorn U. and Holzhutter H. G. (2002) Assessment of proteasomal cleavage probabilities from kinetic analysis of time-dependent product formation. J. Mol. Biol. **318:** 847–862
33 Lucchiari-Hartz M., Lindo V., Hitziger N., Gaedicke S., Saveanu L., Endert P. M. van et al. (2003) Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef. Proc. Natl. Acad. Sci. USA **100:** 7755–7760
34 Peters B., Tong W. W., Sidney J., Sette A. and Weng Z. P. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-1 molecules. Bioinformatics **19:** 1765–1772
35 Kondo A., Sidney J., Southwood S.,Guercio M. F. del, Appella E., Sakamoto H. et al. (1997) Two distinct HLA-A*0101-specific submotifs illustrate alternative peptide binding modes. Immunogenetics **45:** 249–258
36 Kondo A., Sidney J., Southwood S.,Guercio M. F. del, Appella E., Sakamoto H. et al. (1995) Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class I molecules. J. Immunol. **155:** 4307–4312
37 Sidney J., Southwood S., Pasquetto V. and Sette A. (2003) Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. J.Immunol. **171:** 5964–5974
38 Sidney J., Southwood S., Guercio M. F. del, Grey H. M., Chesnut R. W., Kubo R. T. et al. (1996) Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. J. Immunol. **157:** 3480–3490
39 Sidney J., Grey H. M., Southwood S., Celis E., Wentworth P. A., Guercio M. F. del et al. (1996) Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. Hum. Immunol. **45:** 79–93
40 Bradley A. P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30:** 1145–1159
41 Holzhutter H. G., Frommel C. and Kloetzel P. M. (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. J. Mol. Biol. **286:** 1251–1265
42 Holzhutter H. G. and Kloetzel P. M. (2000) A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. Biophys.J. **79:** 1196–1205
43 Kuttler C., Nussbaum A. K., Dick T. P., Rammensee H. G., Schild H. and Hadeler K. P. (2000) An algorithm for the prediction of proteasomal cleavages. J. Mol. Biol. **298:** 417–429
44 Nussbaum A. K., Kuttler C., Hadeler K. P., Rammensee H. G. and Schild H. (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. Immunogenetics **53:** 87–94
45 Kesmir C., Nussbaum A. K., Schild H., Detours V. and Brunak S. (2002) Prediction of proteasome cleavage motifs by neural networks. Protein Eng. **15:** 287–296
46 Endert P. M. van, Riganelli D., Greco G., Fleischhauer K., Sidney J., Sette A. et al. (1995) The peptide-binding motif for the human transporter associated with antigen processing. J. Exp. Med. **182:** 1883–1895
47 Parker K. C., Bednarek M. A. and Coligan J. E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J. Immunol. **152:** 163–175
48 Gromme M. and Neefjes J. (2002) Antigen degradation or presentation by MHC class I molecules via classical and non-classical pathways. Mol. Immunol. **39:** 181–202
49 Seifert U., Maranon C., Shmueli A., Desoutter J. F., Wesoloski L., Janek K. et al. (2003) An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope. Nat. Immunol. **4:** 375–379