

Modeling the MHC-I pathway

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Biophysik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I

der Humboldt-Universität zu Berlin

von

Diplom-Physiker Björn Peters

geboren am 18.5.1973 in Hamburg

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I

Prof. Dr. Michael Linscheid

Gutachter:

1. Prof. Hermann-Georg Holzhütter

2. Prof. Reinhard Heinrich

3. Prof. Robert Tampe

eingereicht am: 24.02.2003

Tag der mündlichen Prüfung: 10.07.2003

Summary

A major task of the immune system is to identify cells that have been infected by a virus or that have mutated, and discriminate them from healthy cells. This duty is assigned to cytotoxic T-lymphocytes (CTL), which scan epitopes presented to them on cell surfaces derived from intracellular proteins through the MHC-I antigen processing pathway. The goal of this work is to provide computational methods that allow to predict which epitopes get presented from the large pool of peptide candidates contained in intracellular proteins. This is achieved by examining the selective influence of three major steps in the pathway: peptide generation by the proteasome, peptide transport into the ER by TAP, and binding of peptides to MHC-I molecules.

For peptide binding to MHC-I, a new algorithm is developed that combines a matrix-based method describing the contribution of individual residues to binding with pair coefficients describing pair-wise interactions between positions in a peptide. This approach outperforms several previously published prediction methods, and for the first time quantifies the impact of interactions in a peptide. The distribution of the pair coefficient values shows that interactions are not limited to amino acids in direct contact, but can also play a role over longer distances. Compared to the matrix entries, the pair-coefficients are rather small, explaining why methods completely ignoring interactions can nevertheless make good predictions.

Next, a novel algorithm is developed to predict the TAP affinities of peptides of any length. Longer peptides are important because several MHC-I epitopes are generated by N-terminal trimming of precursor peptides transported into the ER by TAP. As the true *in vivo* precursors of an epitope are not known, a generalized TAP score is established which averages across the scores of all precursors up to a certain length. The ability of this TAP score to discriminate between epitopes and random peptides shows that the influence of TAP is a consistent, strong pressure on the selection of MHC-I epitopes.

Using predicted TAP transport efficiencies as a filter prior to the prediction of MHC-I binding affinities, it is possible to further improve the already very high classification accuracy achieved using MHC-I affinity predictions alone. Such a 2-step prediction protocol failed when predictions of C-terminal proteasomal cleavages were combined with MHC-I affinity predictions. This disappointing result is thought to be caused by the lack of a sufficiently large set of quantitative and consistent experimental data on proteasomal cleavage rates, which are more difficult to measure and interpret than the affinity assays used to characterize peptide binding to TAP and MHC-I. Therefore, a new protocol for the evaluation of proteasomal digests is developed, which is applied to a series of experiments. This novel protocol addresses two problems: (1) Using mass-balance equations, a method is developed to quantify peptide amounts from MS-signals. (2) By introducing the first kinetic model of the 20S proteasome capable of providing a satisfactory quantitative description of the whole time course of product formation, cleavage probabilities can be extracted reliably from proteasomal *in vitro* digests.

Keywords: Prediction, Antigen Processing, MHC, TAP, Proteasome.

Zusammenfassung

Das Immunsystems muss gesunde Zellen von infizierten und Krebszellen unterscheiden können, um letztere selektiv zu bekämpfen. Dies ist die Aufgabe der CTL-Zellen, die dazu auf der Zelloberfläche präsentierte Peptide die aus intrazellulären Proteinen der jeweiligen Zelle stammen untersuchen. Diese präsentierten Peptide (Epitope) werden durch den MHC-I Antigenpräsentationsweg hergestellt. Das Ziel dieser Arbeit ist es Methoden zu entwickeln die Epitope aus der großen Zahl prinzipiell in Proteinen enthaltener Peptide herausuchen können. Dazu wird die Selektivität dreier wichtiger Komponenten des Präsentationsweges untersucht: Die Herstellung der Peptide durch das Proteasom, der Transport in das ER durch TAP, und das Binden von Peptiden an leere MHC-I Moleküle.

Zur sequenzbasierten Vorhersage der Bindung von Peptiden an MHC-I Moleküle wurde ein neuer Algorithmus entwickelt. Dieser kombiniert eine Matrix, welche die individuellen Beiträge einzelner Reste zur Bindung beschreibt, mit Paarkoeffizienten, die Wechselwirkungen zwischen verschiedenen Positionen im Peptid beschreiben. Dieser Ansatz macht bessere Vorhersagen als bisher publizierte Methoden, und quantifiziert erstmals den Einfluss von Wechselwirkungen innerhalb eines Peptids auf die Bindung. Die Verteilung der Werte der Paarkoeffizienten zeigt, dass sich Wechselwirkungen nicht auf benachbarte Aminosäuren beschränken. Im Vergleich zu den Matrixeinträgen sind die Werte der Paarkoeffizienten klein, was erklärt warum Vorhersagen die Wechselwirkungen komplett vernachlässigen trotzdem gut sein können.

Erstmals wurde ein Algorithmus zur Vorhersage der TAP-Transporteffizienz von Peptiden beliebiger Länge entwickelt. Das ist deshalb wichtig, da viele MHC-I Epitope als N-terminal verlängerte Prekursoren in das ER transportiert werden. Für die Vorhersage der Transportfähigkeit eines potentiellen Epitopes wird deshalb über die Transporteffizienz des Epitopes selbst und seiner Prekursoren gemittelt. Mit Hilfe dieser Definition von Transportfähigkeit wird gezeigt, dass TAP einen starken selektiven Einfluss auf die Auswahl von MHC-I Epitopen hat.

Indem man Peptide die als 'nicht-transportierbar' vorhergesagt werden als mögliche Epitope ausschließt, kann man die ohnehin schon hohe Qualität von MHC-I Bindungsvorhersagen weiter steigern. So eine zweistufige Vorhersage scheitert, wenn man statt des TAP Transports die Vorhersage der Generierbarkeit eines Epitopes durch das Proteasom als Filter verwenden möchte. Dieses schlechte Abschneiden der proteasomalen Schnittvorhersagen wird auf eine mangelhafte experimentelle Datenbasis zurückgeführt, da proteasomale Schnittraten schwieriger zu messen und interpretieren sind als die Affinitätsdaten für TAP und MHC-I. Um die experimentelle Datenbasis in Zukunft verbessern zu können, wurde ein neues experimentelles Protokoll entwickelt und an einer Reihe von Experimenten getestet. Dabei werden zwei Probleme behandelt: (1) Durch die Verwendung von Massenbilanzen werden MS-Signale in quantifizierte Peptidmengen umgerechnet. (2) Durch das erste kinetische Modell des Proteasomes das die Entstehung und den Abbau von Peptiden während eines Verdaus zufrieden stellend beschreiben kann, können aus den Verdaudaten Schnittraten bestimmt werden.

Schlagnworte: Vorhersage, Antigen Prozessierung, MHC, TAP, Proteasom.

Contents

1	<i>Introduction - the MHC-I pathway</i>	7
1.1	Structure and function of the main pathway components	9
2	<i>Peptide binding to MHC-I</i>	15
2.1	Overview of existing prediction methods	15
2.2	Experimental datasets	19
2.3	Obtaining predictions from published methods	20
2.4	Introducing the stabilized matrix method (SMM)	21
2.5	Evaluating prediction quality	24
2.6	Comparison of matrix based predictions: SMM, PM, BIMAS and SYFPEITHI	24
2.7	Comparison of general predictions: ANN, CART and the additive method	26
2.8	Extending SMM with pair coefficients	30
2.9	Distribution of pair coefficient values	34
2.10	Summary	36
3	<i>Peptide transport by TAP</i>	37
3.1	Published prediction methods of <i>in vitro</i> TAP affinity	37
3.2	Comparison of affinity predictions for 9-mers	38
3.3	Predictions of TAP affinities for longer peptides	41
3.4	Using TAP transport predictions for the identification of epitopes	43
3.5	Combining TAP transport predictions with predictions of MHC-I affinity	52
3.6	Confidence in the values of the free parameters α and L	54
3.7	Summary	54
4	<i>Peptide generation by the proteasome</i>	57
4.1	Evaluating published algorithms predicting proteasomal cleavage	57
4.2	Problems with evaluating experimental proteasome digests	61

4.3	Novel protocol of experimental evaluation	66
4.4	Application and testing of novel protocol	72
4.5	Differences between constitutive- and immuno-proteasomal digests	87
4.6	Summary	89
5	<i>Summary of main results and conclusions</i>	91
6	<i>Outlook</i>	97

References

Abbreviations

Acknowledgements

Lebenslauf

Publikationen

Erklärung

1 Introduction - the MHC-I pathway

A major task of the immune system is to identify cells that have been infected by a virus or that have mutated, and discriminate them from healthy cells. This duty is assigned to cytotoxic T-lymphocytes (CTL cells)¹. Since it is not possible to examine the entire contents of a cell without destroying it, the CTL cells rely on the inspected cells to exhibit a representative fraction of their content on the surface. This is realized by the MHC-I antigen procession and presentation pathway (Figure 1) which consists of the following steps: In the cytosol, proteins are degraded by the proteasome, some of them at the end of their useful lifetime, some of them (about 40%) directly after synthesis. Most of the peptide fragments generated by the proteasome are further degraded by other cytosolic proteases into single amino acids used for the synthesis of new proteins. Some of the peptides escape degradation and are transported into the endoplasmic reticulum (ER) by the membrane spanning transporter TAP. There the peptides can again be degraded by the recently identified aminopeptidase ERA(A)AP (Saric, et al., 2002; Serwold, et al., 2002; York, et al., 2002) or exported back into the cytosol, unless they are able to bind to an empty MHC-I molecule. Once a peptide binds, the MHC-I - peptide complex is transported to the cell surface, where it is presented to CTL cells. The presented peptides are called T-cell epitopes.

¹ If not explicitly cited otherwise, the information in this chapter is taken from three recent reviews: Kloetzel, P. M. (2001): Antigen processing by the proteasome, *Nat Rev Mol Cell Biol* 2 [3], pp. 179-87, Lankat-Buttgereit, B. and Tampe, R. (2002): The transporter associated with antigen processing: function and implications in human diseases, *Physiol Rev* 82 [1], pp. 187-204. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11773612, Shastri, N.; Schwab, S. and Serwold, T. (2002): Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules, *Annu Rev Immunol* 20, pp. 463-93.

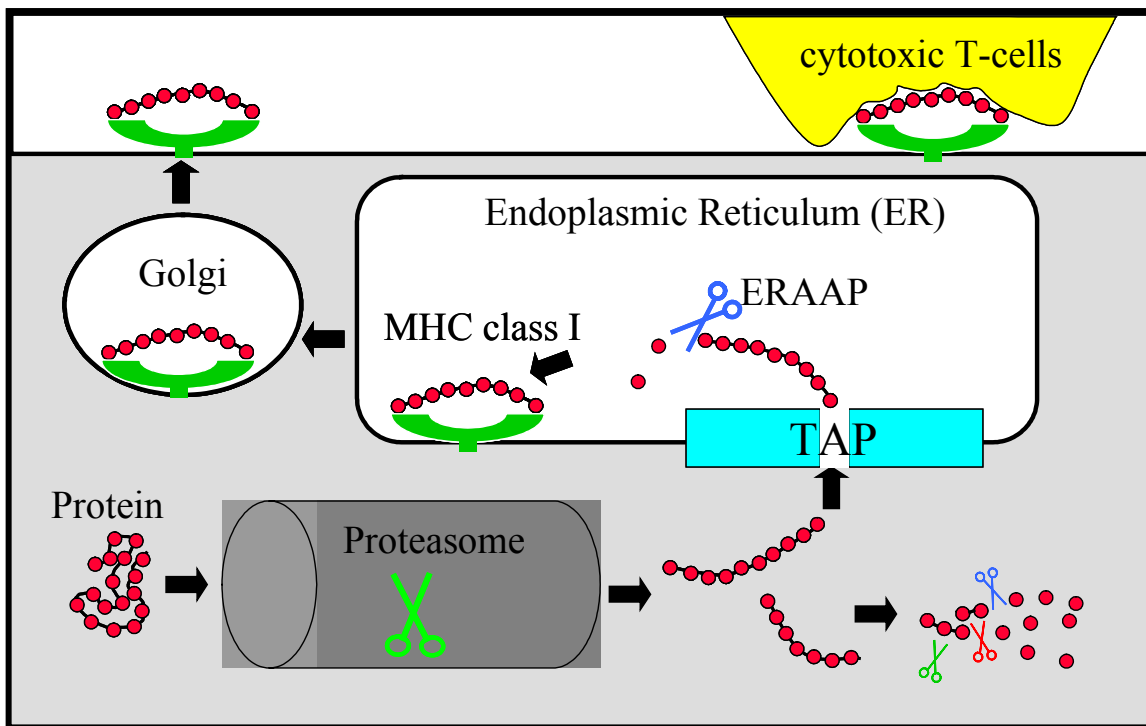


Figure 1: Schematic overview of the MHC-I antigen processing and presentation pathway

The CTL cells can discriminate between epitopes that are 'normally' presented to them, and those that are not. The definition of what is normal is made during development of the thymus: Of a large initial population of CTL cells, those reacting on any of the epitopes presented to them at this stage are deleted. Later in life, when detecting a foreign epitope, a CTL cell kills the inspected cell and secretes γ -interferon, which causes changes in the antigen procession of neighboring cells, some of which are described below.

The goal of this work is to provide computational methods that allow predicting which peptides from the large pool of candidates that in principle can be derived from intracellular proteins are presented as T-cell epitopes. Such prediction tools would be useful for several immunological applications including the intelligent design of peptide vaccines, i.e. predicting an epitope contained in a viral protein sequence which would be presented by cells infected with that virus, designing a vaccine containing this epitope in a less harmful context, and using this vaccine to train the immune system to illicit a strong response when encountering this epitope.

The approach to such a prediction taken in this work is to define the selective influences of three main agents in the pathway: Peptide generation by the proteasome (chapter 4), transport into the ER by TAP (chapter 3) and binding to MHC-I (chapter 2). Each of these steps is examined individually, resulting in algorithms that are able to predict the efficiency of each step for a given substrate. Combining the individual predictions leads to a model describing epitope selection of the entire pathway. While this is shown to work for TAP and MHC-I, predictions of proteasomal cleavage give inferior results, which is assumed to be the consequence of lesser quality experimental data. Therefore new ways of gathering and interpreting such data are introduced in chapter 4.

1.1 Structure and function of the main pathway components

In this section, an overview of the structure and biological function of the three main components of the MHC-1 pathway is given, which are examined in the rest of this work. Readers solely interested in the development of mathematical prediction algorithms can proceed directly to chapter 2.

1.1.1 The proteasome generates peptides by degrading proteins

Proteasomes are self-compartmentalizing multi-subunit protease complexes performing most of the non-lysosomal proteolysis in eukaryotic cells. All proteasomes isolated from eukaryotic cells until now contain the so-called 20S proteasome as the catalytic core. In Figure 2, the crystal structure of the 20S proteasome in yeast is shown. It depicts a cylindrical particle consisting of 28 subunits arranged in four heptameric rings. The two inner β -rings form the central cavity of the cylinder and harbor at their inner surface the proteolytic active sites. In eukaryotic 20S proteasomes, only three β subunits ($\beta 1$, $\beta 2$ and $\beta 5$) are active, with an N-terminal threonine as the catalytic residue. Each of these subunits has a distinct substrate preference, which is usually characterized by the rate in which it cleaves small fluorogenic peptides. Intriguingly, stimulation of cells by γ -interferon causes the replacement of the three active β -subunits $\beta 1$, $\beta 2$ and $\beta 5$ by their iso-forms $\beta 1i$, $\beta 2i$ and $\beta 5i$ in newly synthesized proteasomes. Because these subunits are induced by γ -interferon, signaling an infection in the vicinity of the cell, it is assumed that these new 'immuno-proteasomes' enhance the antigen procession capability of a cell. The immuno-

subunits do possess a distinct cleavage preference, but it is not exactly clear how this improves antigen processing.

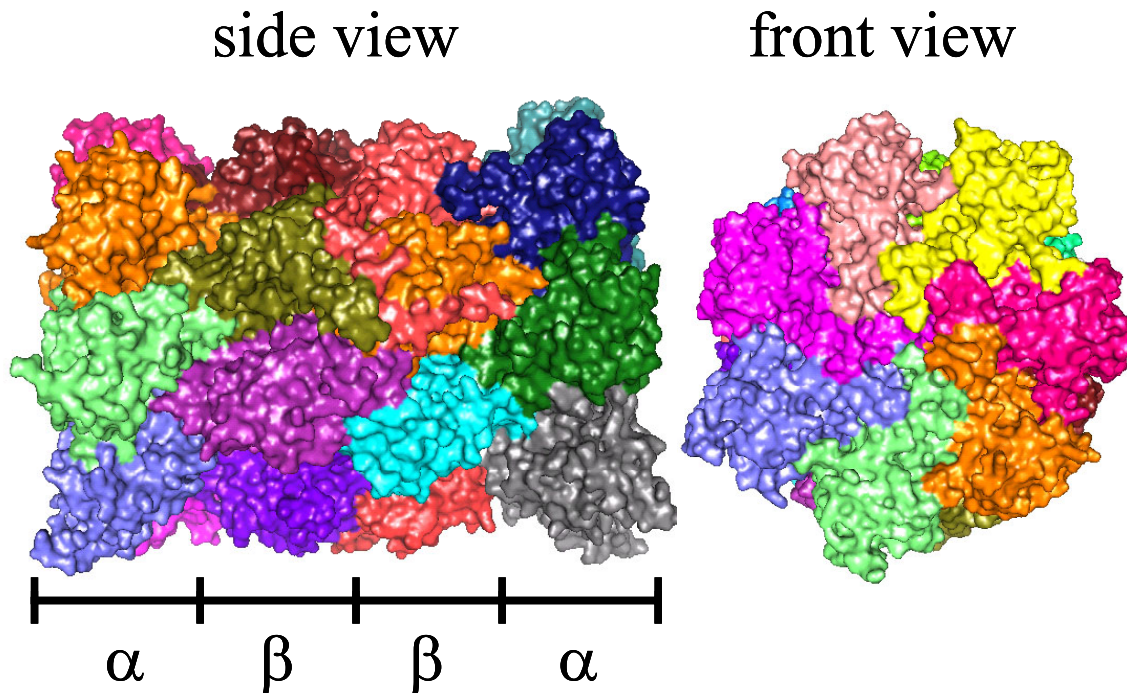


Figure 2: Structure of the 20S yeast proteasome published by (Groll, et al., 1997).

The α -subunits of the two outer rings form the boundary of a gated channel through which the traffic of incoming substrates and outgoing peptides is likely to proceed (Groll, et al., 2000; Kohler, et al., 2001). *In vivo*, the 20S core proteasome is usually found associated with 19S and / or 11S regulatory complexes. These regulators dock at the α rings and are believed to control the access to the core channel. The 19S regulators recognize proteins tagged with a poly-ubiquitin chain, which marks them for degradation. The 11S regulators are induced by γ -interferon which again makes it likely that their function enhances the antigen processing capability of a cell. In contrast to the 20S core alone, these 26S proteasomes need ATP to function.

Proteasomes are essential to life. Chromosomal deletions of each of the 14 yeast 20S proteasome genes are lethal (Heinemeyer, et al., 1991; Hilt and Wolf, 1995). Functional integrity of proteasomes has been demonstrated to be indispensable for a variety of cellular functions besides generation of antigenic peptides such as metabolic adaptation, cell differentiation, cell-cycle

control, stress response and removal of abnormal proteins (Hilt and Wolf, 1996). The role as supplier of antigenic peptides was presumably taken over by the proteasome during the evolution of the immune system because of its ancient property to cleave substrates into smaller peptides (Niedermann, et al., 1997).

Originally it was thought that peptides generated by the proteasome during normal protein turnover would be the only source of fragments for the MHC-I pathway. However it has been known for some time that around 40% of proteins are degraded by the proteasome within a minute of synthesis, which is thought to be a consequence of their inability to fold. Degradation of these defective ribosomal proteins (DRiPs) has been found to be a main source of antigenic peptides (Schubert, et al., 2000). This gives the immune system access to all proteins at the point of synthesis, independent of their lifetime and final location in the cell.

Apart from the proteasome, several other proteases have been implicated in the generation of antigenic peptides. Among these are the tripeptidyl peptidase II, furin and the thimet oligopeptidase. (Schwarz, et al., 2000). Their importance is not yet completely clear, but it can be assumed that because of their selective specificity, they can only play a role in the generation of a minority of observed antigenic peptides.

1.1.2 TAP transports peptides into the ER

TAP is a heterodimer consisting of TAP1 and TAP2, each of which contains transmembrane domains and an ATP binding motif. No crystal structure of TAP is currently available, but it is known from sequence homology analysis that TAP belongs to the super family of ATP-binding cassette transporters (ABC transporters). The TAP genes are coded in the MHC-II locus, and are up regulated after stimulation with γ -interferon (Ayalon, et al., 1998), which implicates the role of TAP in antigen procession.

The initial selective step of TAP transport is binding of the peptide, which involves both subunits of TAP. This is followed by a slow structural reorganization of the molecule, which is believed to trigger ATP hydrolysis and peptide translocation across the membrane (Neumann and Tampe, 1999). TAP specificity has been analyzed using combinatorial peptide libraries (Uebel, et al.,

1995) showing that the C-terminus and the three N-terminal residues of a peptide contribute most to binding to TAP. The optimal lengths of peptides for transport is 8-16 amino acids, but oligopeptides up to 40 residues in length have been shown to be transported (Momburg, et al., 1994; van Endert, et al., 1994).

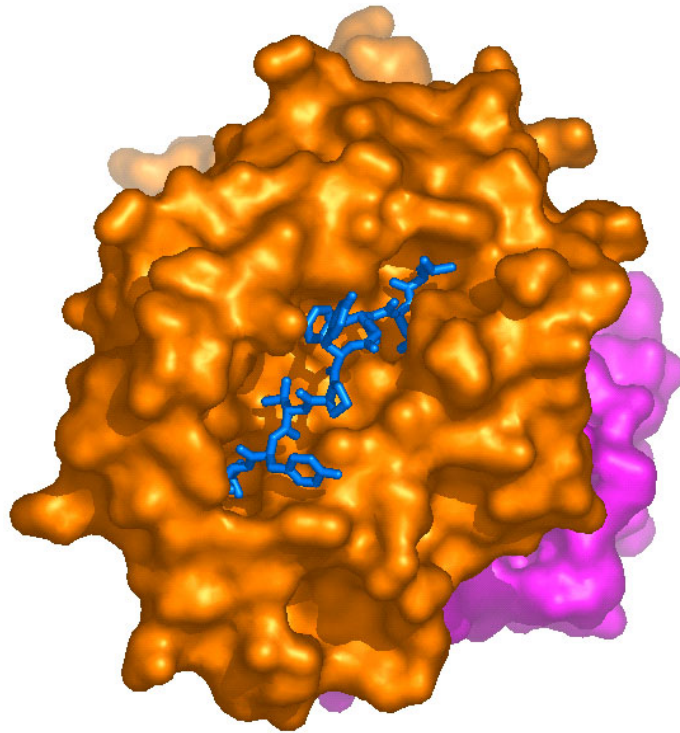


Figure 3: Epitope (blue) in the binding groove of the MHC-I α -chain (orange). Structure published by (Khan, et al., 2000).

1.1.3 MHC-I molecules present bound peptides on the cell surface

Loaded MHC-I molecules are heterotrimers consisting of the presented epitope bound to the polymorphic α -chain which is again bound to the invariant β 2 microglobulin. Figure 3 depicts a peptide in the binding pocket of the MHC-I molecule. Polymorphism in the α chain primarily involves residues in the binding pocket, giving rise to the large variety of binding specificities observed for different MHC-I alleles. Each human has up to six different MHC-I alleles, out of 980 different ones currently known (February 2003, www3.ebi.ac.uk/Services/imgt/hla/cgi-bin/statistics.cgi).

The assembled empty MHC-I molecules are associated with TAP, and the molecule tapasin acts as a bridge between the two. This places the empty MHC-I molecules close to the peptide source and retains them there until they are loaded with a peptide. The loaded MHC-I molecules leave the ER via the Golgi apparatus and the *trans*-Golgi network to the cell surface. Several hundred thousand copies of MHC-I molecules each containing a single epitope are presented at any time on the cell surface, where their epitopes are scanned by CTL cell receptors as shown in Figure 4.

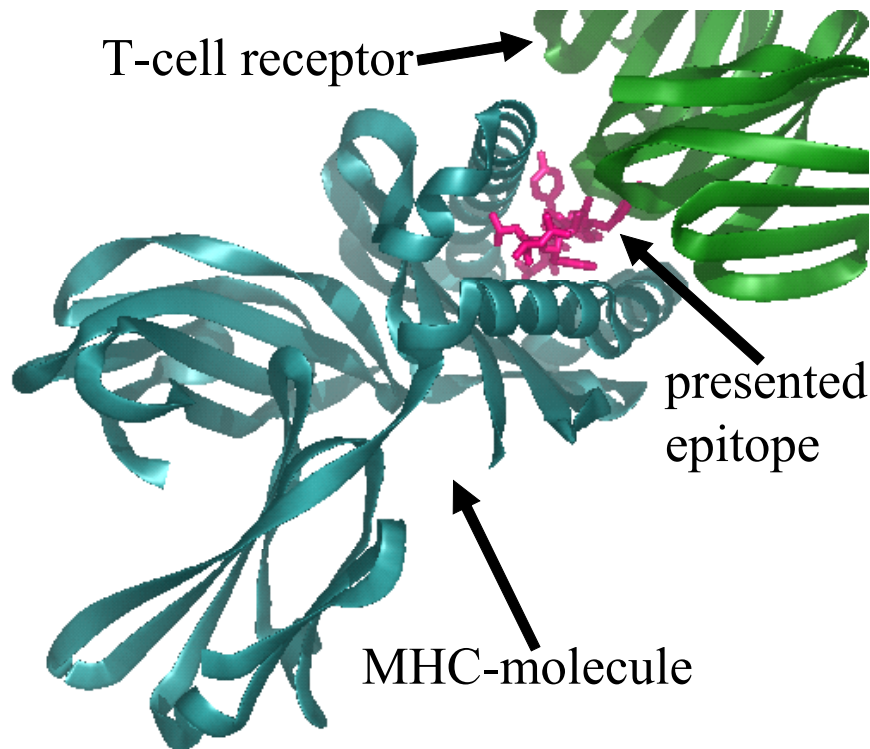


Figure 4: MHC-I bound epitope is scanned by T-cell receptor. Structure published in (Garboczi, et al., 1996).

2 Peptide binding to MHC-I

Binding to MHC-I is the most selective requirement for a peptide to become an epitope through the MHC-I pathway. This has directed most experimental and theoretical work aimed at MHC-I epitope prediction towards this step. In this chapter a new method to predict the affinity of a peptide to an MHC-I molecule is introduced and compared to existing prediction methods. Special interest is paid to the independent binding assumption, which states that each residue within a peptide contributes independently to the overall binding of the peptide. Several published prediction methods make this assumption, while others claim it to be invalid because it neglects interactions between peptide positions. Here, for the first time the violation of the independent binding assumption is quantified.

This chapter is organized as follows: The first section (2.1) gives an overview of existing prediction algorithms, followed by a description of the available datasets consisting of peptides with measured affinities, which are needed for training and testing of the algorithms (section 2.2). Section 2.3 explains how predictions from existing algorithms were obtained. This is followed by the introduction of the novel SMM prediction method (section 2.4), and a description of the means to compare the prediction quality of different methods (section 2.5). In section 2.6, all algorithms making the independent binding assumption are compared, including the novel SMM method. This is followed by a comparison of general prediction methods not making that assumption in section 2.7. In section 2.8, the SMM prediction method is extended to include pair coefficients describing the pair-wise interactions between peptide positions, thus dropping the independent binding assumption. Finally, the distribution of these quantified pair-wise interactions between peptide positions is analyzed (section 2.9).

Most of the results presented in this chapter are taken from (Peters, et al., 2003).

2.1 Overview of existing prediction methods

When the first peptides binding to MHC-I molecules were sequenced (Rotzschke, et al., 1990), it soon became clear that each MHC-I molecule has its own narrow binding preference. Only peptides of a certain length were found, typically 8-10 amino acids long. Some positions in the

binding peptides called anchor positions could only be occupied by a few different amino acids. Apparently, peptides interact more strongly with the MHC-I molecule at these anchor positions, limiting the number of amino acids tolerated there. Using pool sequencing of peptides eluted from one type of MHC-I allele (Falk, et al., 1991), it could be characterized which amino acids are allowed at the anchor positions. Demanding compliance with this anchor motif was the first method to predict which peptides are potential MHC-I binders (Rotzschke, et al., 1991).

With more data available, it became apparent that there are several peptides binding to MHC-I that do not comply with the anchor motif or, more often, that peptides containing the motif did not bind (Jameson and Bevan, 1992). This led to extensions of the motif method, listing more and more good, bad or tolerated residues for binding at the different peptide positions. One deficiency of such a listing is that it does not say if a good residue at one position can compensate for a bad residue somewhere else. To achieve this, scores were assigned to each amino acid at each position. These can be summed up for a given peptide and the total score predicts if the peptide is likely to bind or not. This is called a matrix based prediction, because the scores for each peptide position and amino acid can be written in the form of a matrix like the one shown in Table 1.

2.1.1 Matrix prediction methods: BIMAS, SYFPEITHI and PM

The first scoring matrix for MHC-I binding was introduced by (Parker, et al., 1994), which is in the following called the BIMAS method. The experimental basis was a set of peptides with measured half-life dissociation constants for the HLA-A0201 allele. The matrix coefficients were determined mathematically by minimizing the distance between predicted matrix scores for the peptides and the logarithms of their measured half-life constants. This makes the matrix score a prediction of a peptides half-life of dissociation, and therefore of its strength of binding.

A major contribution to the prediction of epitopes was the establishment of the SYFPEITHI database (Rammensee, et al., 1999; Rammensee, et al., 1995). This database is a collection of sequences of naturally processed epitopes and peptides known to bind to MHC-I molecules, which serves as the basis for a scoring algorithm. To determine a scoring matrix for a given allele, the frequencies of amino acids in peptides binding to that allele are used to manually

determine scores for each peptide position. There is no experimental interpretation of the SYFPEITHI score, it simply reflects the agreement of a peptides sequence with that of previously known binders and epitopes. As the sequences of naturally processed epitopes also reflect the selectivity of other components in the antigen procession pathway, these matrices do not describe the pure binding specificity of an MHC-I allele, but it is assumed that the influence of other pathway agents is small compared to the selectivity of MHC-I binding.

Another matrix based prediction approach is called the polynomial method (PM) (Gulukota, et al., 1997). It has the advantage of being very easy to calculate from a set of peptides with known affinity values. Each matrix entry corresponding to a particular amino acid type at a particular peptide position is calculated as the mean affinity of all peptides containing that amino acid type at that position.

2.1.2 The independent binding assumption

A matrix based prediction necessarily assumes, that the contribution to binding of each residue in a peptide is independent of the other residues in the peptide. This is a daring assumption, known to be false for many other biological problems. However, matrix based algorithms work surprisingly well for MHC-I affinity predictions: Scoring matrices are usually calculated with little experimental data (typically hundreds of peptides) compared to the size of sequence space (for 9-mers: 20^9), and still give good predictions. This shows that the relationship between sequence and affinity can at least roughly be approximated by the independent binding assumption. The assumption is also supported by structural data showing that peptides bind in an extended conformation in the MHC-I groove, so that contacts between residues of the bound peptide are limited.

2.1.3 General prediction methods: ANN, CART and the additive method

Even if independent binding is a justified approximation, a method without this restriction should be able to make better predictions if interactions between peptide positions play a role at all. In the following, such methods are called general methods, in contrast to those making the independent binding assumption. The first general methods used to predict binding to MHC-I were artificial neural networks (ANN) (Gulukota, et al., 1997; Milik, et al., 1998). ANNS are a

biologically motivated approach to machine learning. An ANN consists of several layers of interconnected 'neurons', each of which transfers its input into an output according to some mathematical function. The free parameters of this function are determined by learning from a set of data where the correct output belonging to an input is known (see (Agatonovic-Kustrin and Beresford, 2000) for a review of ANN applications in biomedical sciences).

Another general prediction method are Classification and Regression Trees (CART). These are described in detail in (Breiman, et al., 1984), and were first used by (Segal, et al., 2001) to predict MHC-I binding. Briefly, a classification tree is built by introducing splits in a set of peptides with known binding capability according to what amino acid is at a certain position of a peptide. The splitting is repeated, leading to a tree shaped classification scheme (e.g. Figure 8). Each split is chosen so that it maximizes the homogeneity of the peptides in both daughter nodes. A perfectly homogenous node contains only binders or only non-binders. A very large tree is built first so that all nodes are perfectly homogenous. It is then pruned back to an optimal size determined by cross-validation. Each terminal node in the optimal tree is assigned a binding score, computed as the percentage of non-binders the node contains.

The third general prediction method used in this chapter is the additive method (Doytchinova, et al., 2002). It consists of a scoring matrix + coefficients describing pair-wise interactions between amino acids. The score of a peptide is calculated by adding those coefficients belonging to pairs of amino acids present in that peptide to its matrix score, as described below in equation (6). All interactions between neighboring and next-neighbor amino acids are considered. The matrix- and interaction coefficients are determined by a partial least square fit to a set of peptides with known affinities.

The prediction methods used in this chapter were chosen because they were freely accessible or reasonably easy to reproduce. Other classes of prediction methods based on binding data which are not treated here are Hidden Markov Models (Mamitsuka, 1998) and Support Vector Machines (Donnes and Elofsson, 2002). There are also prediction methods based on structural data of solved peptide-MHC-I complexes. These use threading or molecular modeling techniques to identify potential binders (Altuvia, et al., 1995; Schueler-Furman, et al., 2000).

2.2 Experimental datasets

Four non-overlapping sets of 9-mer peptides with known affinities to the HLA-A0201 allele are used in this chapter. All datasets are separated into 'binders' and 'non-binders'. The cutoff for making this separation is not critical as long as all the binders have a higher affinity than any of the non-binders. All sets have similar sequence composition: at the anchor positions 2 and 9, amino acids A, I, L, M, T and V are over-represented; at other positions, all 20 amino acids are roughly equally represented. For the SYFPEITHI-set, this reflects the occurrence of amino acids at these positions in naturally processed epitopes. For the *in vitro* datasets, where the choice of peptides is made by an experimentalist, this bias is also found because experimentalists want to include as many binders in their datasets as possible. While it would be better - from a mathematical viewpoint - to have binding information of randomly selected peptides, this would require far more measurements as only very few random peptides bind. As the described bias is present in all learning and test sets, no particular prediction method is favored.

Train-set: This set consists of 533 peptides with IC₅₀ values measuring their binding affinity to the HLA-A0201 molecule, as described in (Gulukota, et al., 1997). IC₅₀ values are measured in an assay in which both the peptide of interest and a reference peptide compete for binding to HLA-A0201. The IC₅₀ value of a peptide is the concentration at which it has suppressed 50% of the reference peptides from binding to MHC-I. The logarithm of the IC₅₀ value can be interpreted as the difference in binding free energy between a peptide and the reference.

Several peptides in the dataset are 'heavy non binders', i.e. their IC₅₀ value is too large to be measured. Using an IC₅₀ cutoff of 500 nM, the Train-set is split into 127 binders and 406 non-binders, which of course include the 'heavy non binders'. The cutoff lies within the intermediate-to low affinity range; peptides in this set with IC₅₀<50nM are considered to be high-affinity binders.

Blind-set: This set 173 of peptides with IC₅₀ values was measured with the same experimental setup as the Train-set. Using the IC₅₀ cutoff of 500 nM, the set is split in 67 binders and 108 non-binders.

BIMAS-set: This set of peptides was published by (Parker, et al., 1994). For 134 peptides, $\beta 2$ microglobulin dissociation half-lives have been measured. Four of the peptides overlap with the Train-set. By interpolating between their known IC₅₀ and half-life values, the IC₅₀=500nM cutoff in the Train-set is translated to a half-life cutoff of approximately 650 minutes. Using this cutoff and excluding the overlapping peptides, the BIMAS-set has 25 binders and 105 non-binders.

SYFPEITHI-set: 143 peptides binding to HLA-A0201 were taken from the SYFPEITHI database, most of which are naturally processed epitopes. All of these were classified as binders, even though there is no measured affinity available for them. To have non-binders in this set that definitely have lower affinities than these binders, the 59 heavy non-binders from the Blind-set were included.

2.3 Obtaining predictions from published methods

The BIMAS and SYFPEITHI predictions were obtained from web servers (BIMAS: http://bimas.dcrf.nih.gov/molbio/hla_bind/, SYFPEITHI: <http://www.uni-tuebingen.de/uni/kxi>). The training data of these methods contained the equally named test sets described above.

The PM, CART and ANN predictions were trained using the Train-set described above. For the CART method, two commercial software packages (SPSS and CART) were used, leading to an identical optimal classification tree shown in Figure 8. Switching from a classification tree used here, which needs binary experimental binding data for training, to a regression tree, which uses quantitative IC₅₀ values, improves the prediction performance only slightly.

The ANN was designed as described in (Gulukota, et al., 1997): A feed-forward neural network with three layers was built consisting of an input layer with 180 neurons, a hidden layer with 50 neurons and an output layer with one neuron. The Aspirin/MIGRAINES software package from the MITRE Corporation (<http://www.emsl.pnl.gov:2080>) was used to simulate the network.

For the Additive Method, the coefficient values determined in (Doytchinova, et al., 2002) were used. The training data used in that paper to determine the coefficient values was a subset of the Train-set described above.

2.4 Introducing the stabilized matrix method (SMM)

Scoring matrices quantify the contributions of individual residues in a peptide to binding. The matrix element $s_{i,a}$ corresponds to amino acid a at position i of the peptide. The total score S_k for a given peptide k with the amino acids $a_k(i)$ at positions i is then given by the summation:

$$S_k = s_0 + \sum_i s_{i,a_k(i)} \quad (1)$$

where s_0 is a constant offset.

In this section, the novel stabilized matrix method (SMM) is developed. It determines the values for $s_{i,a}$ and s_0 by minimizing the distance between predicted scores S_k and measured affinities for the peptides in the Train-set:

$$\Phi(\{s_{i,a}\}) = \sum_k \|S_k - \text{measured}_k\| \quad (2)$$

For a peptide with a measurable IC50 value, the norm in equation (2) has the form:

$$\|S_k - \text{measured}_k\| = (S_k - \ln(\text{IC50}_k))^2 \quad (3)$$

Several peptides in the Train-set have too low affinities to measure an IC50 value. For these 'heavy non-binders', the IC50 values are set equal to or greater than the largest experimentally measurable value, which was cutoff= $\ln(50,000)$ for the Train-set. Accordingly, for these peptides the norm in (2) is set to be

$$\|S_k - \text{measured}_k\| = \begin{cases} 0 & \text{if } S_k > \text{cutoff} \\ (\text{cutoff} - S_k)^2 & \text{if } S_k < \text{cutoff} \end{cases} \quad (4)$$

To avoid over-fitting, a second term is added to the minimization function in (2):

$$\Psi(\{s_{i,a}\}, \lambda) = \Phi(\{s_{i,a}\}) + \lambda \sum_{i,a} s_{i,a}^2 \quad (5)$$

By minimizing with a non-zero λ value, a tradeoff is introduced between optimally reproducing the experimental IC50 values (including their inevitable experimental error) and minimizing parameters $s_{i,a}$. This forces all parameters $s_{i,a}$ towards zero, which do not significantly lower the distance Φ . The optimal value for λ is determined by 10-fold cross-validation on the Train-set, i.e. splitting the total set of peptides into 10 subsets, establishing a scoring matrix using 9 of these subsets and making predictions with that matrix for the left out subset. Figure 5 depicts the sum of the distances between these predictions and the experimental values as a function of λ .

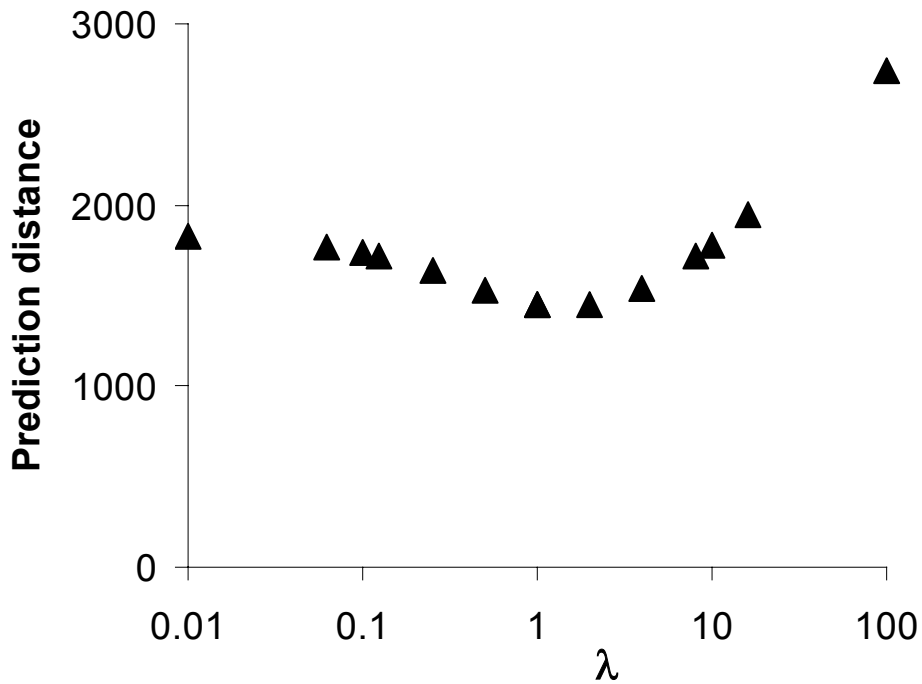


Figure 5: Cross validated distance as a function of λ

The optimal predictions were made with $\lambda_{\text{opt}}=1$. The resulting SMM scoring matrix is shown in Table 1. The lower a score, the better an amino acid is suited for binding at the given position.

A similar mathematical concept is used to solve 'inverse problems', where λ is called the regularization parameter. A short introduction to inverse problems is given in (Press, et al., 1992), chapter 18. To minimize equation (5), a commercial non-linear optimizer (Frontline Systems, 1999) using a generalized-reduced-gradient method is applied.

Table 1: SMM scoring matrix for binding to HLA-A0201

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.58	0.74	3.17	0.13	0.78	1.30	1.35	0.33	2.76
D	4.90	0.74	1.46	-1.50	-0.15	-0.14	1.50	1.06	2.76
N	1.47	0.74	2.03	-1.48	-0.36	-1.78	0.51	-0.76	2.76
C	0.93	0.74	0.96	-0.96	-0.82	-1.15	0.56	0.57	2.76
E	3.46	0.74	2.79	-1.24	0.40	-0.74	-0.17	0.39	2.76
Q	1.63	0.74	1.24	-0.34	-0.41	-0.52	0.61	0.36	2.76
G	0.50	0.74	1.08	-0.83	-0.64	-0.60	1.22	0.54	2.76
H	1.33	0.74	0.90	-1.44	-0.39	-0.60	-0.43	0.06	2.76
I	0.58	-0.42	1.54	1.67	-0.78	-1.40	-0.29	0.59	-0.50
L	0.29	-3.06	-0.38	-0.17	-0.03	-1.48	-0.01	0.26	0.50
K	-0.20	0.74	2.25	0.22	0.05	1.43	1.73	1.20	2.76
M	-1.46	-2.72	-1.02	0.36	-0.41	-1.43	-1.08	0.58	0.91
F	-1.24	0.74	0.13	-0.45	-1.84	-1.48	-1.64	-1.10	2.76
P	3.63	0.74	1.12	-0.67	0.93	-0.90	-1.60	-0.66	2.76
S	0.24	0.74	1.05	-0.25	-0.26	-0.26	0.18	0.08	2.76
T	0.82	0.17	1.70	-0.63	-0.75	-2.20	0.17	-0.00	1.06
W	0.38	0.97	-0.03	-2.38	-1.49	-1.07	-1.67	0.02	2.76
Y	-1.91	0.74	-0.89	-0.12	-1.74	-1.37	-2.34	-0.77	2.76
V	-0.09	-0.31	1.28	1.15	-0.44	-2.13	0.03	1.04	-0.81

Offset 10.14

2.5 Evaluating prediction quality

In this chapter, the predictions of diverse methods on equally diverse datasets have to be compared. To do this fairly, ROC curves are used (Bradley, 1997): For a given cut-off value which separates peptides by their predicted score into potential binders and non-binders, the two variables sensitivity (true positives / total positives) and 1-specificity (false positives / total negatives = false alarm rate) are calculated. By systematically varying the cut-off value from the lowest to the highest predicted score, a ROC curve like Figure 6 is plotted. Prediction quality is measured by the area under the curve (AUC), which is 0.5 for random predictions and 1.0 for perfect predictions. The AUC is equivalent to the probability that the score of a randomly chosen binder is higher than that of a randomly chosen non-binder. This measure has the advantage of not relying on a single arbitrarily chosen cut-off value for the prediction score, and can be equally applied to all datasets and prediction methods.

2.5.1 Statistical significance for differences in AUC

To assess if one prediction is significantly better than another, the set of peptides for which predictions are made was re-sampled. Using bootstrapping with replacement, 50 new datasets were generated with a constant ratio of positives to negatives. The difference in AUC for the two predictions on each new dataset is then calculated. One prediction is significantly better than another if the distribution of the differences in AUC values is significantly above zero, which is measured using a standard t-test with a p-value of 0.001.

2.6 Comparison of matrix based predictions: SMM, PM, BIMAS and SYFPEITHI

The four matrix based methods had to predict which of the peptides in the Blind-set are binders. The Blind-set is the only set truly 'blind' to all methods, i.e. none of peptides in this set were included in the training data of any of the prediction methods. Figure 6 depicts ROC curves for all predictions. It indicates that the performance ranks in the order of SMM>BIMAS>PM over almost the entire range. SYFPEITHI is the worst method for sensitivities above 0.42 and becomes the best for specificities above 0.97. This is due to the fact that SYFPEITHI predictions reflect other components of the antigen presentation pathway in addition to MHC-I binding,

leading to a decrease in sensitivity when predicting binding alone. The area under the ROC curve (AUC) gives a single number describing prediction accuracy of a method. Figure 6 translates to AUC values of 0.869 for SMM, 0.846 for BIMAS, 0.795 for PM and 0.745 for SYFPEITHI. The AUC values for the predictions of all methods on the other test sets (BIMAS-set and SYFPEITHI-set) are listed in Table 2. For both of these sets, SMM makes the best predictions of all truly blind methods.

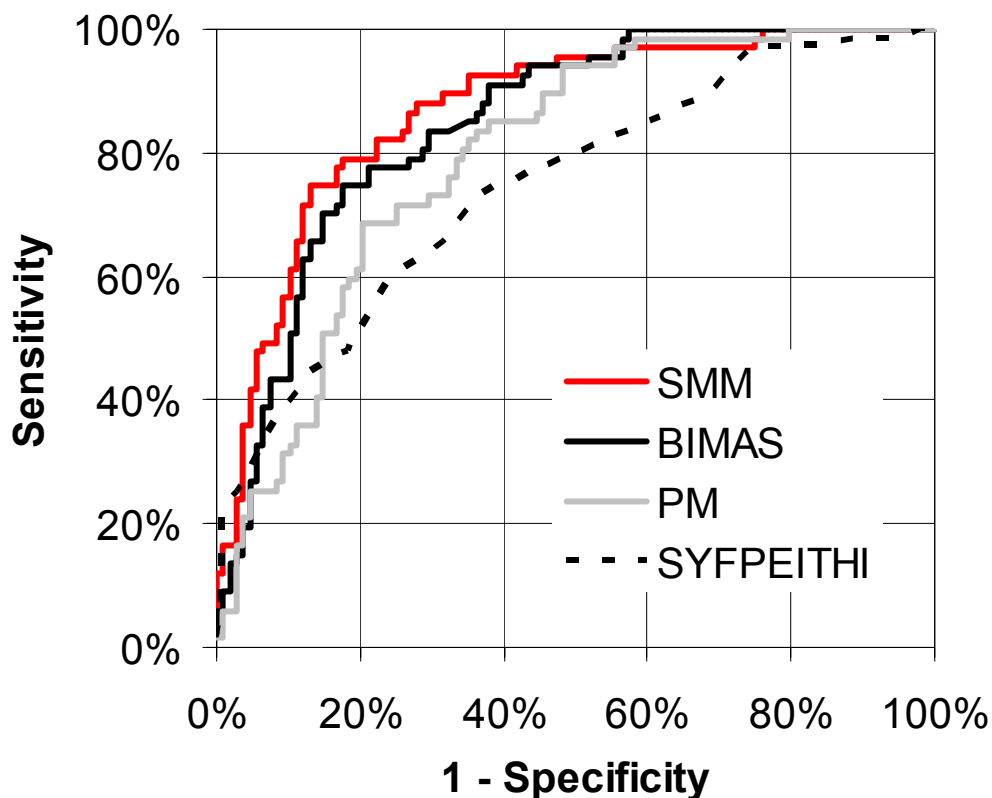


Figure 6: ROC curve for matrix based prediction methods on the Blind-set.

For each method, the cutoff was varied from the lowest to the highest predicted score of any peptide in the Blind-set. For each cutoff value, the sensitivity and specificity were calculated, and plotted in the graph.

Table 2: Comparison of prediction quality

Prediction Method	Independent Binding Assumption	AUC on Test Set		
		Blind	SYFPEITHI	BIMAS
SMM, $\lambda=1$	Yes	0.869	0.848	0.866
SMM, $\lambda=0$	Yes	0.856	0.846	0.865
BIMAS	Yes	0.846	0.829	(0.875)
PM	Yes	0.795	0.792	0.757
SYFPEITHI	Yes	0.745	(0.865)	0.754
SMM + pair coef.	No	0.873	0.852	0.869
ANN	No	0.796	0.788	0.762
Additive method	No	0.820	0.770	0.830
CART	No	0.708	0.620	0.539

Two elements make the SMM approach different from other matrix based methods. First, it incorporates the experimental information of heavy non-binders precisely into the distance defined in equation (2). Since only the lower bound of the IC50 values for heavy non-binders can be determined, previous approaches have either left them out entirely or tried to fit them exactly to the lower bound. Second, the regularization technique was used. With errors in experimental measurements, there can be multiple sets of matrix coefficients that can reproduce the experimental data within their range of the error. Choosing the set of coefficients that gives the minimum distance may mean to overfit the problem. By incorporating a regularization parameters (λ in equation 5), a set of coefficients is chosen that reproduces the experimental results reasonably while keeping the parameter values small. This effectively prevents overfitting. Table 2 indicates that the AUC values at $\lambda_{opt}=1$, are better than those at $\lambda=0$ for all test sets.

2.7 Comparison of general predictions: ANN, CART and the additive method

Figure 7 shows ROC curves for general prediction methods (not making the independent binding assumption) on the Blind-set. The corresponding AUC values are also listed in Table 2. ANN

and the additive method make consistently better prediction than the CART tree. None of these previously published general methods reaches the prediction quality of SMM or BIMAS which made the independent binding assumption.

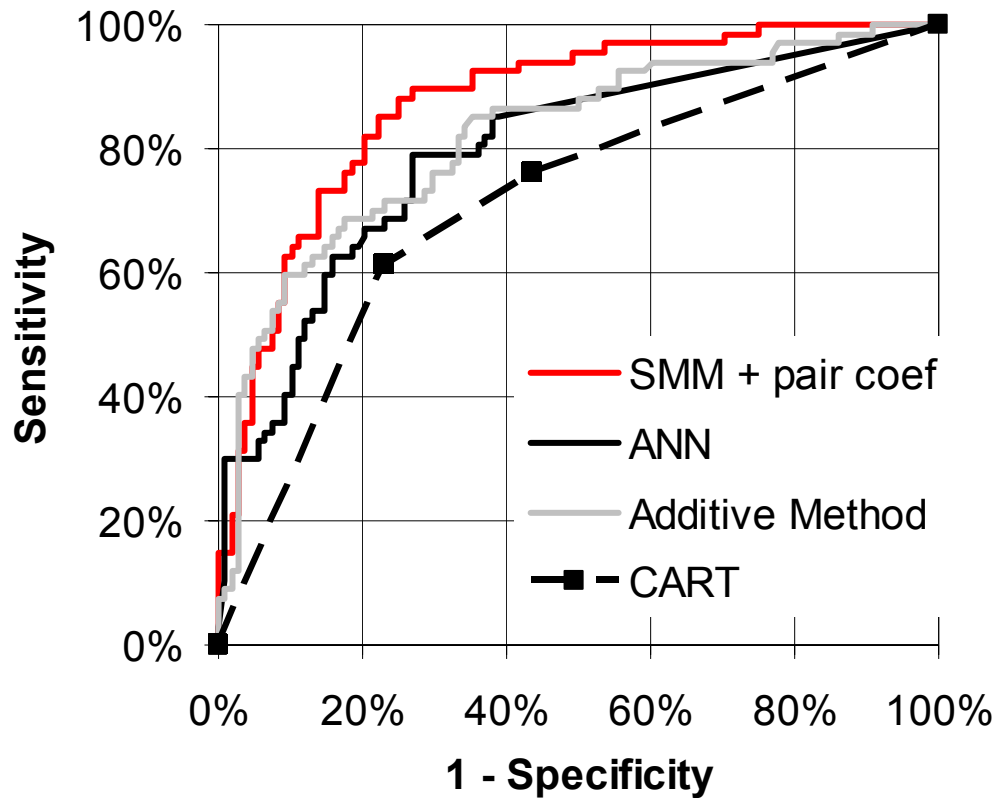


Figure 7: ROC curves for general prediction methods on the HLA-A2 Blind-set.

The Figure contains ROC curves described in the legend of Figure 6. Since there are only three terminal nodes in the CART tree, corresponding to three different scores, its ROC curve consists of only two non-trivial points.

At first glance, this is surprising, as the general methods should be able to describe all binding mechanisms, including the simple case of independent binding. Why does the more restrictive matrix approach perform better? This can best be seen for the CART algorithm. As shown in Figure 8A, CART suggests to split node 3, but not node 2. If this exactly described reality, peptides in node 3 would bind differently than peptides in node 2, signifying an interaction between positions 2 and 1 only for peptides with an L or M at position 2. In contrast, if the independent binding assumption is true and there are no interactions between positions 2 and 1, the split described for node 3 should also be applicable to node 2. Testing this on the Blind-set (Figure 8B) shows that transferring the split actually works, as the new nodes resulting from the transferred split are more homogenous than node 2. The CART algorithm cannot identify this split, because it can only use the peptides in node 2 to decide about splits at node 2, and there are only 11 binders left in that node. This shows that general methods simply require more training data to achieve the prediction quality of matrix-based methods, if the independent binding assumption holds to a high degree.

In case of the additive method and the ANN, lack of data has led to overfitting. The additive method has 1850 free parameters, the ANN architecture taken from (Gulukota, et al., 1997) more than 9000 neurons. This number of parameters cannot be determined reliably for the Train-set containing only 533 experimental data points. For the ANN, choosing a different architecture with less free parameters would probably have improved its prediction quality. For the additive method, the overfitting seems to affect mainly the interaction coefficients. When neglecting these coefficients which describe interactions between neighboring amino acids, and keeping only those compatible with the independent binding assumption, the prediction quality improves.

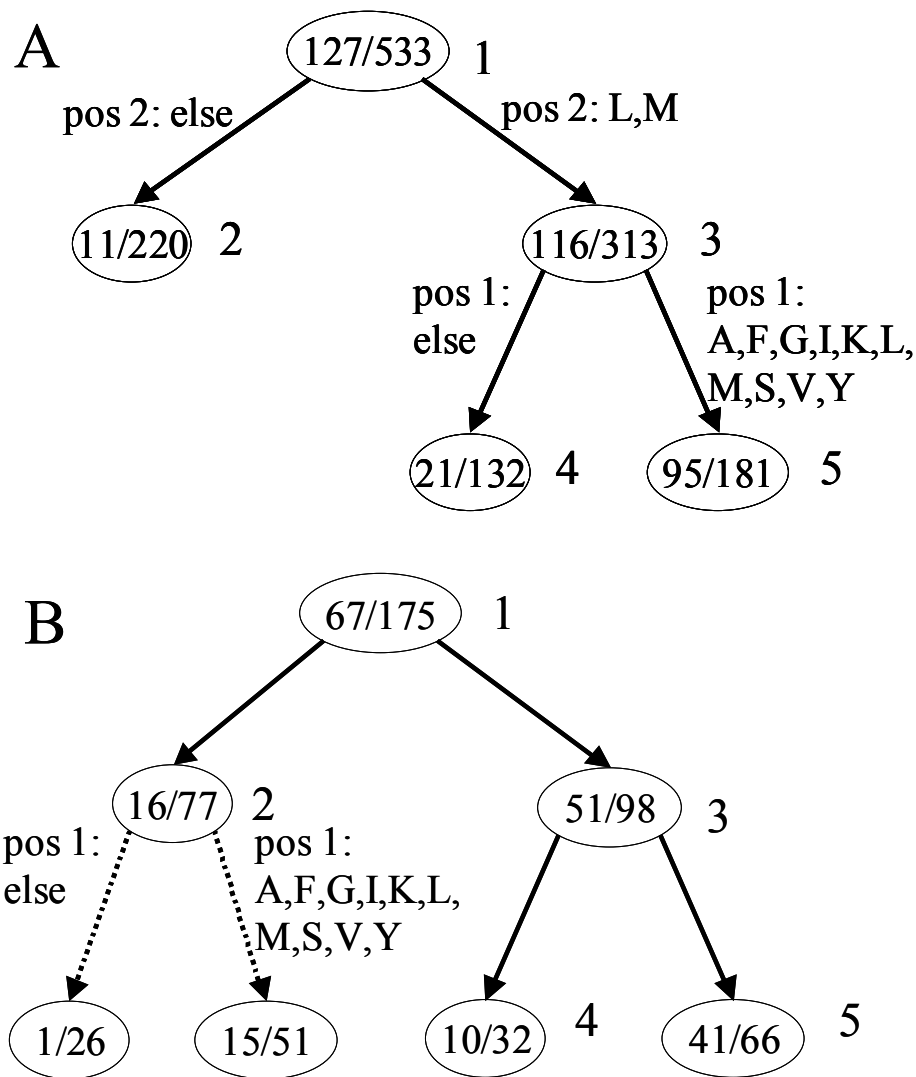


Figure 8: Classification tree for peptides binding to HLA-A0201

Each ellipse denotes a node corresponding to a set of peptides, with the first number indicating the number of binders, and the second number indicating the total number of peptides in the node. The splits in the nodes are symbolized by arrows, which lead to daughter nodes. To the right of each node is a reference number (1-5) used in the text. **(A)** the optimal tree generated for the Train-set. **(B)** the tree in (A) is used to classify peptides in the Blind-set. The additional split on the lower left with dotted arrows, which is taken from the split of node 3 in (A), improves the classification. This indicates that the tree in (A) was not optimal.

2.8 Extending SMM with pair coefficients

The results of the previous section show that given the limited data, general methods tend to perform worse than matrix based ones. In this section a general method that uses the matrix predictions as a starting point is developed. This is done by quantifying the contribution to binding of interactions of peptide positions with pair coefficients $s'_{i,a,i',a'}$. For example, coefficient $s'_{3A,7L}$ (**A***L**) describes the difference in binding between the following two scenarios: (1) an Alanine is at the third AND a Lysine at the seventh position of the peptide and (2) the sum of the average contributions of having an Alanine at the third position, described by matrix value s_{3A} (**A*****), and having a Lysine at the seventh position, described by s_{7L} (*****L**). The total score for a given peptide k with the amino acids $a_k(i)$ at positions i is then given by

$$S'_k = S_k + \sum_i \sum_{i'} s'_{i,a_k(i),i',a_k(i')} \quad (6)$$

where S_k is the matrix score defined in equation (1), and the sum includes all pairs of amino acids found in the peptide. For 9-mer peptides, this would result in $20*20*36 = 14400$ different pair coefficients. Since it is impossible to determine that many coefficients given the limited data, a two-stage selection process is applied: In the first step all coefficients for which fewer than $N_{\min}=10$ peptides exist in the Train-set are eliminated, as they lack sufficient experimental information. The optimal values for the remaining 269 pair coefficients can then be calculated by minimizing

$$\Psi'(\{s'_{i,a,i',a'}\}, \lambda') = \Phi'(\{s'_{i,a,i',a'}\}) + \lambda' \sum_{i,a,i',a'} s'^2_{i,a,i',a'} \quad (7)$$

where Φ' is the same as Φ in equation (2), except that scores S'_k are used instead of S_k . When optimizing the pair coefficients $s'_{i,a,i',a'}$ in equation (7), the matrix coefficients $s_{i,a}$ are frozen at their optimal value determined from equation (5).

In a second selection step, the Train-set is split into 10 equal-size non-overlapping subsets and 10 different optimal values for each pair coefficient are determined by leaving out one subset at a time and minimizing equation (7). If a pair coefficient contains both positive and negative

optimal values, it cannot be estimated reliably from the Train-set. Therefore, it is discarded. Setting all 145 discarded coefficients to zero, equation (7) is minimized to determine the values of the remaining 124 pair coefficients.

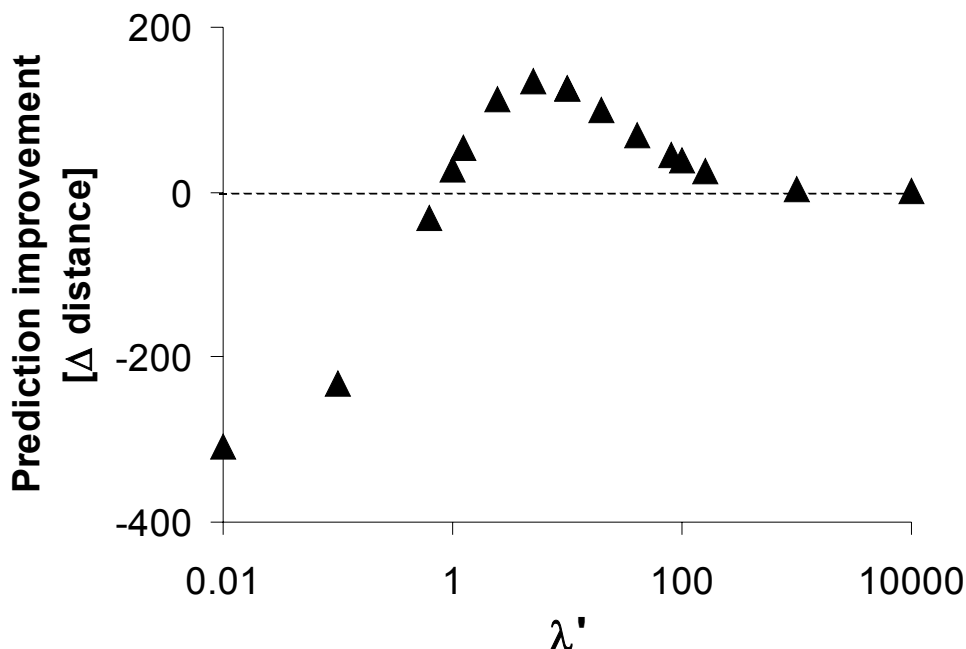


Figure 9: Using cross-validation to determine the optimal λ'

The distance between the predictions of SMM + pair coefficient and the experimental IC50 values of the Train-set is plotted in comparison to the distance achieved with the SMM matrix alone. The improvement in prediction quality is highest for $\lambda'_{\text{opt}} = 5$.

The optimal value for λ' is again determined using cross-validation, similar to λ in equation (5). The cross validated distance as calculated using (2) compared to that of the plain matrix predictions is shown in Figure 9 for various values of λ' . For very high values of λ' , the pair coefficients are forced to stay around zero, and the prediction accuracy is equal to that of the plain matrix approach. For $\lambda' = 0$, there is no restriction on the value of pair coefficients, leading to over-fitting, and the resulting performance is much worse than the plain matrix. In between,

there is a maximum in distance improvement at the optimal value $\lambda'_{opt}= 5$. The values of the pair coefficients are listed in Table 3.

Using the combined SMM matrix + pair coefficient method leads to an improvement in prediction quality over the plain SMM matrix on all test sets (Table 2). No other general method even reaches the SMM matrix predictions on one of the test sets. There are three main advantages of the novel matrix + pair coefficients approach: First of all, the pair coefficients are determined by systematic investigation of differences between the matrix predictions and the experimental values. As the matrix method is highly accurate, it is a better starting point than trying to determine both position contributions and position interactions all at once. Another novelty is that the interactions under investigations are limited to those with a sufficient amount of consistent training data. The third advantage is again the use of a regularization parameter (λ' in equation 7), which prevents the pair coefficients from overfitting the data. Its importance can be seen clearly in Figure 9: without it ($\lambda'=0$) the pair coefficients reduce prediction quality below that of the matrix alone.

Table 3: Pair coefficient values

Pos1	Pos2	value	Pos1	Pos2	value	Pos1	Pos2	value
5	S 9 L	-0.81	2	L 4 P	-0.16	2	V 6 A	0.18
2	V 5 A	-0.52	4	E 9 L	-0.15	8	P 9 L	0.19
2	L 4 Q	-0.43	2	L 8 L	-0.14	7	N 9 V	0.20
3	S 9 L	-0.42	1	A 7 A	-0.13	8	S 9 L	0.20
8	H 9 L	-0.41	1	A 9 V	-0.13	2	L 5 E	0.20
2	L 6 F	-0.41	4	P 9 L	-0.13	2	L 7 T	0.21
6	A 9 V	-0.39	2	L 7 P	-0.12	4	S 9 V	0.21
8	I 9 L	-0.36	8	A 9 V	-0.11	1	E 2 L	0.22
2	L 3 A	-0.36	5	A 9 V	-0.11	5	I 9 L	0.22
2	V 7 G	-0.36	6	V 9 L	-0.11	5	V 9 V	0.23
1	G 2 L	-0.34	4	A 9 V	-0.10	3	A 7 A	0.24
6	L 9 V	-0.34	4	K 9 V	-0.10	2	L 7 G	0.24
2	L 4 C	-0.33	6	S 9 L	-0.09	6	Q 9 L	0.25

Pos1	Pos2	value
5	T 9 V	-0.32
2	L 6 Y	-0.32
1	A 2 L	-0.31
2	L 3 L	-0.31
2	L 5 I	-0.31
2	L 3 Y	-0.29
2	V 5 V	-0.27
2	L 4 S	-0.26
4	K 9 L	-0.26
7	L 9 L	-0.26
6	P 9 L	-0.25
1	V 9 L	-0.25
2	L 6 V	-0.25
1	Q 9 L	-0.25
8	V 9 L	-0.24
2	L 7 S	-0.24
7	A 9 V	-0.22
7	P 9 L	-0.22
2	L 8 P	-0.21
1	L 9 V	-0.21
2	I 9 V	-0.20
6	A 7 A	-0.20
2	L 6 I	-0.18
8	G 9 L	-0.18
2	V 6 P	-0.17
1	T 2 V	-0.17
7	A 9 L	-0.17
2	L 6 A	-0.17
2	L 4 T	-0.17

Pos1	Pos2	value
2	L 4 K	-0.06
3	A 9 L	-0.06
1	A 6 A	-0.06
5	G 9 V	-0.01
5	A 7 A	0.03
5	A 8 A	0.04
2	V 9 L	0.04
3	V 9 V	0.07
2	L 6 H	0.09
7	A 8 A	0.10
5	L 9 A	0.10
1	I 9 V	0.11
2	L 5 R	0.12
4	K 5 A	0.13
2	L 4 E	0.13
2	L 6 G	0.13
7	T 9 L	0.14
3	A 6 A	0.14
3	A 4 K	0.14
2	V 4 S	0.15
7	V 9 V	0.15
1	D 9 L	0.15
7	G 9 L	0.16
1	L 9 L	0.16
2	V 5 L	0.17
2	L 3 P	0.17
2	L 8 G	0.18
1	C 2 L	0.18
1	A 8 A	0.18

Pos1	Pos2	value
6	P 9 V	0.26
5	E 9 L	0.28
2	L 3 S	0.28
3	A 5 A	0.28
2	L 4 G	0.29
2	V 3 R	0.30
7	S 9 V	0.30
5	Q 9 L	0.32
2	L 7 A	0.33
2	L 5 V	0.33
2	V 3 G	0.34
1	D 2 L	0.35
6	T 9 L	0.35
3	L 9 V	0.38
6	F 9 L	0.40
4	L 9 V	0.40
1	A 3 A	0.41
1	P 2 L	0.41
1	C 9 L	0.41
2	L 6 L	0.41
2	L 7 R	0.42
2	L 5 S	0.46
6	I 9 L	0.47
6	L 7 L	0.52
4	Q 9 L	0.56
8	D 9 L	0.67
5	S 9 V	0.74

Previously there was one published study comparing several methods to predict peptide binding to MHC-I (Yu, et al., 2002). There it was reported that the optimal choice of a prediction method depends on the number of peptides available for training: an ANN was outperformed by scoring matrices when the training data consisted of 234 peptides, while the ANN outperformed scoring matrices when trained on a set of over a thousand peptides. In principle this agrees with the results reported here.

The new scoring matrix + pair coefficients approach should work over a large range of training set sizes. If little data is available, few or no pair coefficients will meet the criteria for inclusion, and the method is reduced to the SMM matrix. With more training data available, more pair coefficients are included, thus adjusting the complexity of the method to the available training data.

2.9 Distribution of pair coefficient values

Another advantage of the pair coefficients over methods like ANNs is that the extracted rules for binding are easy to interpret. The values determined for the pair coefficients provide direct information about the MHC-I-peptide binding mechanism. Since peptides bind in an extended conformation, one would expect the absolute values of the pair coefficients to be lower if their associated amino acid positions are farther apart. In Figure 10, two quantities that reflect the influence of position distance are plotted: the percentage of pair coefficients discarded due to conflicting information and the average absolute value of retained pair coefficients at the distance. Distances 6 and 7 have the highest percentages of discarded coefficients and distance 7 has the lowest average value of retained coefficients, indicating weak or no interactions between positions at such distances. To a lesser extent, the levels of interaction at distances 2, 5, 6 and 8 are also weaker than those at distances 1, 3 and 4. This agrees with the expected trend of stronger interactions for closer positions, but to a much lesser extent than expected. Also, this study does not confirm that (i, i+2) neighbors influence each other more strongly than (i, i+1) neighbors, which was expected because the side chains of next-neighbor amino acids face in the same direction thus allowing for direct interactions. Taken together, this shows that interactions are not limited to amino acids in direct contact, but can also play a role over longer distances, probably through the conformation of the peptide back-bone.

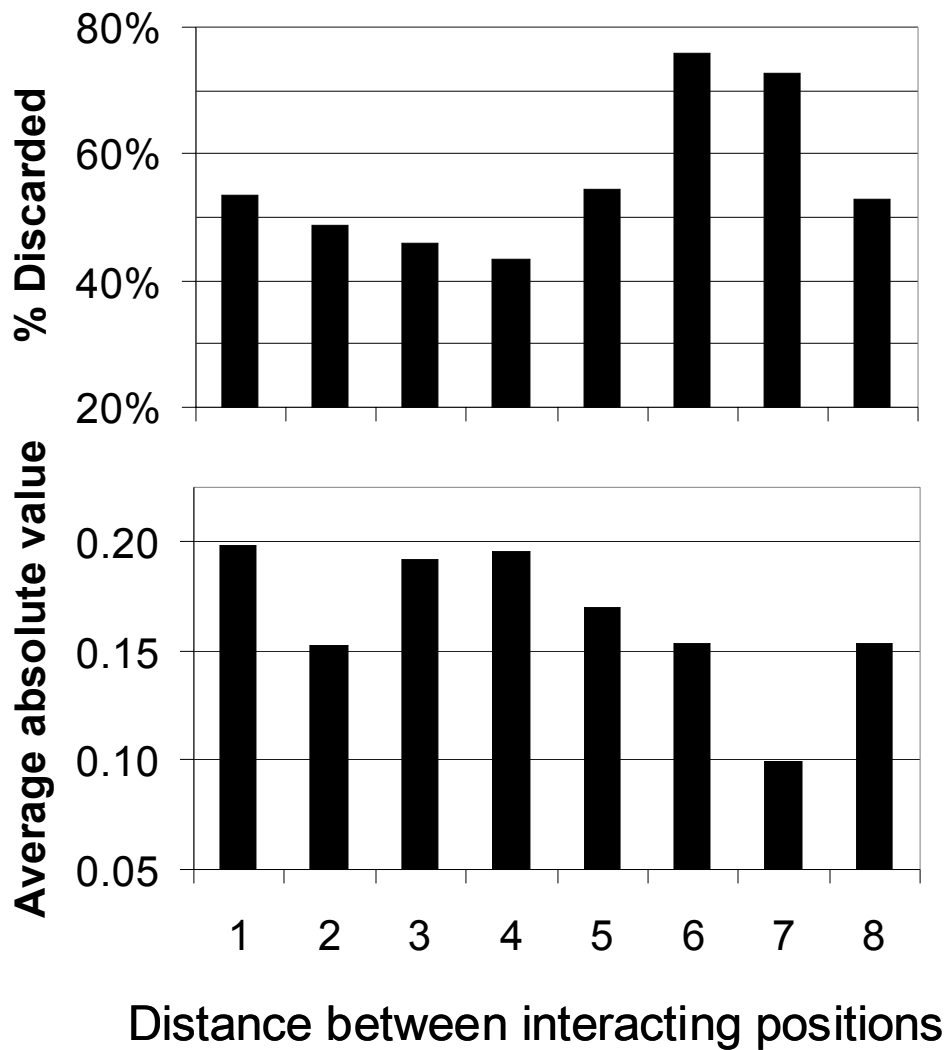


Figure 10: Distribution of pair coefficients.

The percentage of discarded pair coefficients (top) and the average values of retained pair coefficients (bottom) are plotted against the distance between two interacting positions.

While the presented results show that interactions between peptide positions do exist, the values of the pair coefficients are roughly an order of magnitude lower than the entries of the scoring matrix. This could be due to two reasons: the values for these coefficients are indeed small, or the noise level in the datasets is high – thus the coefficients are forced to low values by the regularization parameter λ . Since the plain SMM performs very well, it can be believed that the

true values of the pair coefficients are indeed significantly lower than the values of the matrix entries. The low impact of pair interactions compared to the contributions of individual residues explains why methods making the independent binding assumption can make good predictions at all.

2.10 Summary

A novel sequence based algorithm was introduced that predicts the affinity of peptides to MHC-I molecules. Its basis is a matrix based prediction (SMM) where the entries describe the contributions of individual residues in a peptide to binding. Determining the matrix entries by minimizing the distance between predicted scores and measured IC₅₀ values for a set of training peptides leads to significantly better predictions on three independent test sets than were obtained for a number of previously published methods.

The SMM matrix was combined with pair-coefficients describing interactions between peptide positions, which further improve prediction quality. The pair-coefficient values for the first time quantify the influence of these interactions on peptide binding. Compared to the values of the matrix entries, they are rather small, which explains why good predictions are possible without taking peptide interactions into account at all. The distribution of the coefficient values also shows, that interactions in a peptide are not limited to residues in direct contact, but can also play a role over longer distances, probably through the conformation of the peptide backbone.

3 Peptide transport by TAP

The TAP transporter is the main supplier of peptides binding to MHC-I molecules. As the TAP transport efficiency varies depending on the sequence of the transported peptide, the TAP preference influences the pool of peptides available for MHC-I binding. The importance of this influence *in vivo* is still subject to debate. Previous attempts to identify epitopes by an enhanced predicted TAP transport efficiency exhibited large allele specific differences. This has led to the conclusion that either MHC-I alleles are loaded by different degrees of TAP-independent transport (Brusic, et al., 1999), or that varying amounts of epitopes are transported as N-terminal prolonged precursors (Lauvau, et al., 1999). The second reasoning has been shown to be true for several epitopes (Goldberg, et al., 2002; Lauvau, et al., 1999) and receives further support by the identification of the protease ERA(A)P responsible for the N-terminal trimming of precursors in the ER (Saric, et al., 2002; Serwold, et al., 2002; York, et al., 2002). Motivated by this, a novel method to predict the effective transport of potential epitopes is developed in this chapter based on the predicted transportability of the epitopes themselves and their precursors.

This chapter consists of the following sections: First, an overview of existing methods to predict TAP affinities is given, which can be considered equivalent to predictions of TAP transport (section 3.1). The prediction quality of these methods is compared with a new SMM type scoring matrix established on a set of 9-mer peptides (section 3.2). In section 3.3 these predictions are generalized to be applicable to peptides of any length. This is the basis of a scoring algorithm to discriminate between presented epitopes and random sequences by their TAP transportability (section 3.4). Finally, it is shown in section 3.5 that epitope identification with combined predictions of TAP transportability and MHC-I affinity give better results than predictions using MHC-I affinity alone.

Most of the results presented in this chapter are taken from (Peters, et al., 2003).

3.1 Published prediction methods of *in vitro* TAP affinity

TAP transport rates can be determined experimentally using transport assays (Nijenhuis, et al., 1996; Wang, et al., 1998) where the transported peptides are trapped in the ER (e. g. by

glycolysation). However, these assays measure transport as well as further degradation in the ER, export from the ER and other side effects (Uebel and Tampe, 1999). Another experimental possibility is the use of *in vitro* affinity assays (Gubler, et al., 1998; Uebel, et al., 1997), in which the affinity has been shown to correspond closely to the transport rate of TAP (Gubler, et al., 1998). Affinity data is easier to measure and interpret, which allows to gather comparably large datasets, and is therefore the basis of this work. In the following, the correspondence of TAP transport and affinity is taken to be exact, which allows to equate predictions of TAP affinity with predictions of TAP transport.

To characterize the preference of TAP for 9-meric peptides, two scoring matrices were derived directly from experiments: The 'Ala-matrix' was constructed by using the peptide AAASAAAAY as a reference, and measuring IC50 values for the peptides possessing an exchanged amino acid at one of the 9 sequence positions (Daniel, et al., 1997; Gubler, et al., 1998). The 'Mix-matrix' was generated using libraries of 9-meric peptides $X_1X_2...Y...X_9$, where X_i stands for a mixture of different amino acids and Y is a specific amino acid occupying a fixed sequence position (Uebel, et al., 1997). These libraries compete in binding with the totally randomized peptide library $X_1X_2...X_9$. Similar to the scoring matrices derived in chapter 2, the entries in these matrices can be summed up to predict the affinity of any 9-meric peptide.

In (Daniel, et al., 1998), an ANN was trained on TAP binding data from a set of peptides. The ANN predictions were compared to those made by the Ala-matrix, and were shown to be slightly but significantly better.

3.2 Comparison of affinity predictions for 9-mers

The TAP affinity predictions of the two experimentally derived matrices described above were compared on a set of 430 peptides with measured IC50 values (Daniel, et al., 1998). The resulting scatter plots are depicted in Figure 11, showing that the Mix-matrix makes significantly better predictions than the Ala-matrix.

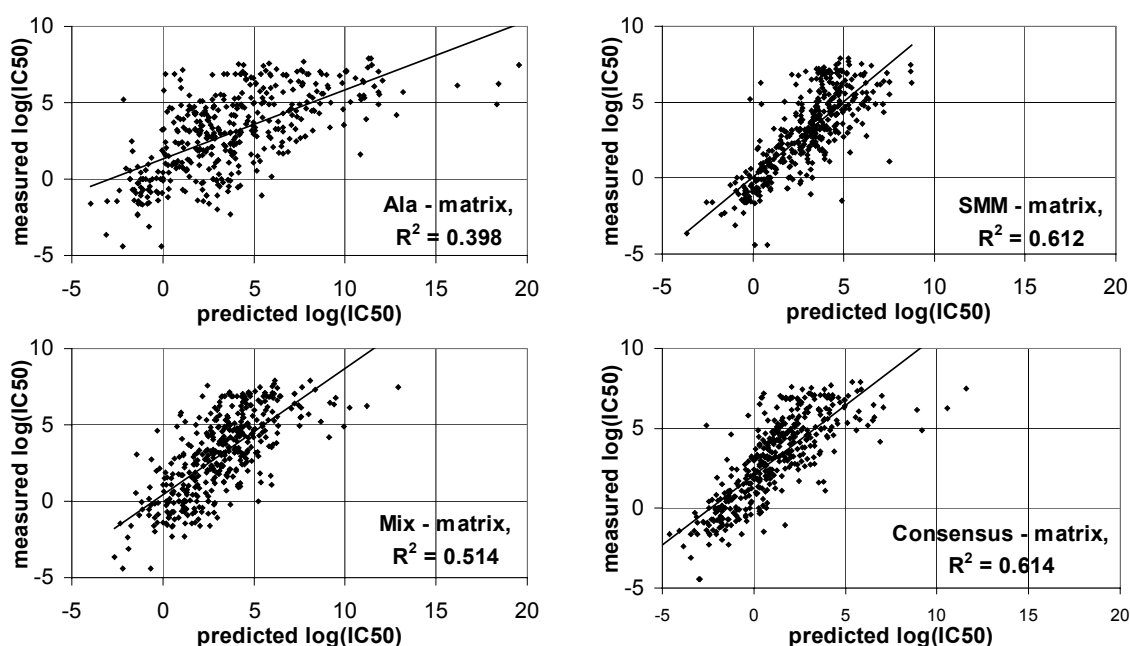


Figure 11: Comparison of predicted and measured *in vitro* TAP affinity values of 9-mer peptides

The scatterplots depict the observed $\log(\text{IC}_{50})$ values of 430 9-meric peptides versus predicted $\log(\text{IC}_{50})$ values using the scoring matrix indicated at the bottom right of each panel. The solid curves represent linear regression lines.

With the measured IC_{50} values of the 430 peptides, it is possible to establish an SMM matrix as described in section 2.4. To be able to compare the prediction quality of this method with that of the two matrices derived directly from experiments, five different SMM matrices were established each trained on a subset of the 430 peptides. For each of these 5 subsets, an optimal λ was determined by cross-validation as shown for one subset in Figure 12. Each of the five SMM matrices was then used to predict the IC_{50} values of the peptides not included in its training data. The resulting scatter plot is also depicted in Figure 11, which shows that the SMM matrix makes significantly better predictions than the Ala- or Mix-matrix.

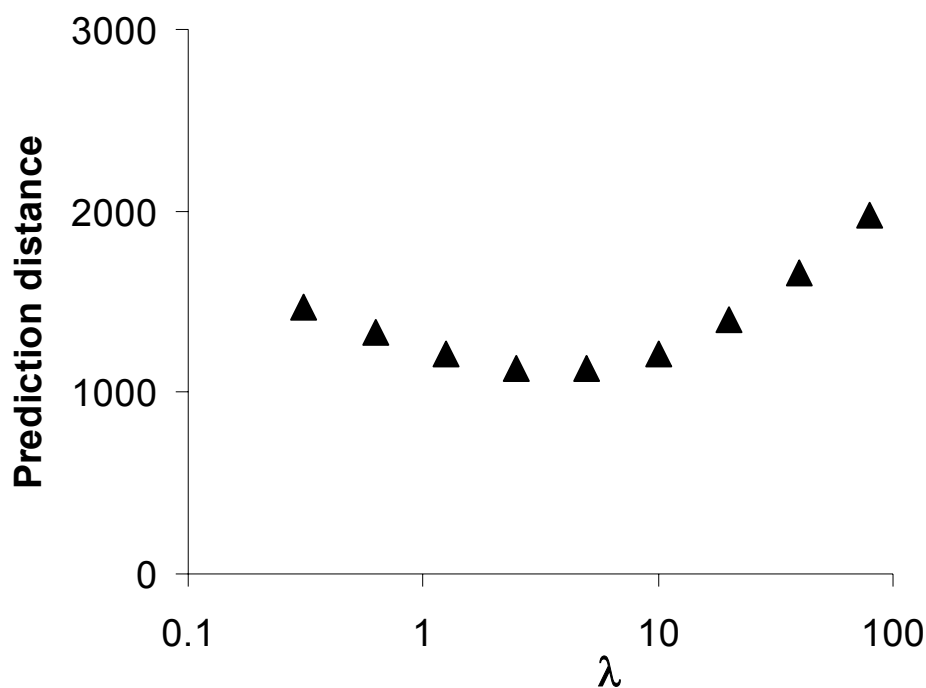


Figure 12: The Cross validated distance of measured and predicted TAP affinities is plotted as a function of λ

The distance between SMM matrix predictions and measured IC50 values in five-fold cross validation is plotted. The best predictions are made for $\lambda_{\text{opt}}=5$.

By averaging over the three scoring matrices, the 'consensus-matrix' (Table 4) is generated, which is expected to give better predictions than the individual matrices because their errors can partially compensate each other. A scatter plot for its predictions is also shown in Figure 11. As expected, the consensus matrix gives the best results although the SMM-matrix is only marginally worse. The ANN predictions from (Daniel, et al., 1998) were not available for a direct comparison. However, as they were only slightly better than those of the Ala matrix, which makes the worst predictions of the three individual matrices, it can be assumed that the consensus matrix predictions are at least as good as those made by the ANN.

Table 4: TAP consensus matrix

	(N1) Pos 1	(N2) Pos 2	(N3) Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	(C) Pos 9
A	-1.56	-0.25	-0.10	0.24	-0.10	0.17	0.27	-0.00	0.55
C	0.05	-0.01	-0.02	0.11	0.09	0.05	0.00	-0.13	0.00
D	1.37	1.42	1.83	-0.23	0.33	0.32	1.07	0.32	1.83
E	1.65	0.02	1.51	0.08	0.54	-0.13	0.64	0.44	1.58
F	1.03	-0.45	-1.05	-0.50	-0.26	0.08	-0.50	0.17	-2.52
G	0.28	1.14	1.70	0.45	0.66	0.12	1.41	-0.38	1.41
H	0.21	0.33	-0.23	-0.21	-0.11	-0.06	-0.19	0.39	0.55
I	-0.11	-0.49	-0.62	-0.09	-0.42	-0.75	-0.94	0.45	-0.52
K	-1.03	-0.41	0.09	-0.23	-0.08	-0.26	0.44	0.12	-0.45
L	-0.50	0.09	-0.11	0.11	-0.34	0.02	-0.73	0.01	-0.94
M	-0.38	-0.46	-0.58	-0.35	-0.26	0.30	-0.64	-0.11	-0.29
N	-1.43	0.69	1.01	0.38	0.49	-0.27	0.16	0.33	1.33
P	1.43	3.00	0.22	-0.04	-0.72	-0.13	-0.84	0.03	-0.09
Q	0.47	-0.97	0.39	0.15	0.15	-0.07	0.34	0.26	0.12
R	-1.34	-1.47	-0.42	-0.27	-0.32	-0.75	-0.09	-0.42	-1.47
S	-0.56	-0.34	0.11	0.27	0.45	0.31	0.87	-0.51	2.26
T	-0.12	-0.04	0.43	0.23	0.43	0.49	0.39	-0.46	0.72
V	-0.49	-0.50	-0.71	0.27	0.37	-0.02	-0.29	0.10	-0.30
W	0.54	-0.64	-1.65	-0.18	-0.78	0.31	-0.50	-0.63	-0.87
Y	0.50	-0.67	-1.80	-0.18	-0.13	0.28	-0.87	0.02	-2.91

3.3 Predictions of TAP affinities for longer peptides

It has been described in the literature that binding of peptides to TAP is mainly influenced by their C-terminal and three N-terminal residues (Daniel, et al., 1998; Uebel, et al., 1997; Uebel and Tampe, 1999; van Endert, et al., 1995). Motivated by this, a new scoring scheme to predict IC50 values of peptides with more than 9 residues is introduced, which neglects the influence of ‘inner’ residues: TAP affinities of peptides with arbitrary length are calculated by scoring only

the C-terminus and the three N-terminal residues using the four corresponding columns of the 9-mer matrix. Thus, for a peptide with the amino acid sequence N1, N2, N3, N4, ..., C the TAP score t is given by

$$t = mat_{1,N1} + mat_{2,N2} + mat_{3,N3} + mat_{9,C} \quad (8)$$

where mat_{i,X_i} denotes the score of residue X at sequence position i.

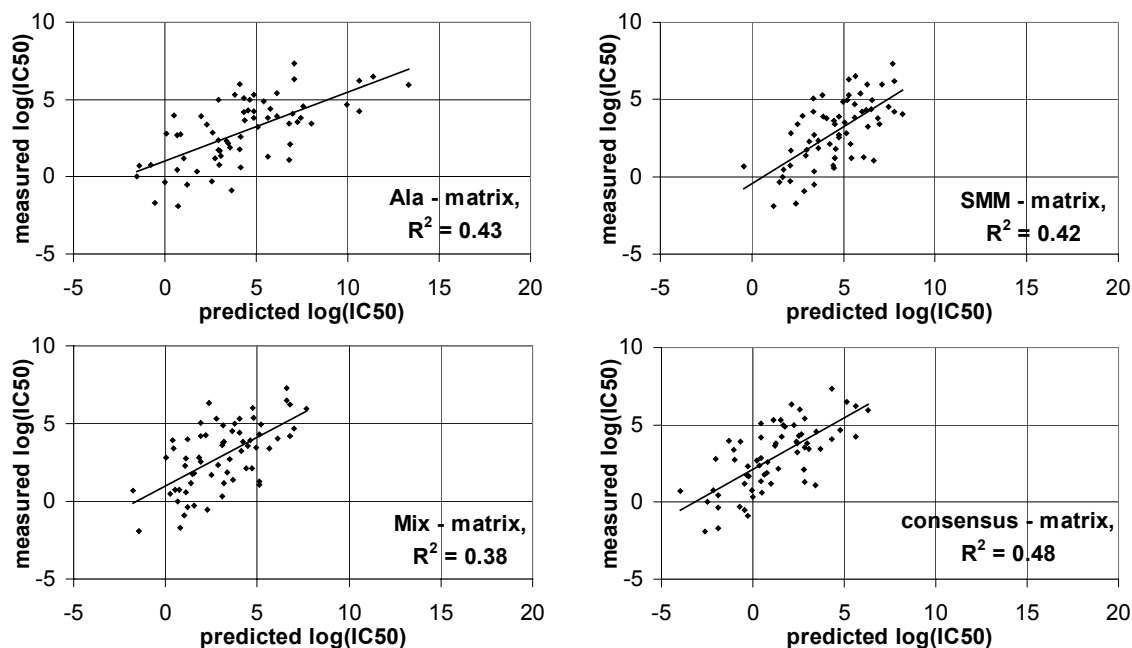


Figure 13: Comparison of predicted and measured *in vitro* TAP affinity values for peptides longer than 9 amino acids.

The scatterplots depict the observed $\log(\text{IC}_{50})$ values of 64 peptides versus theoretical $\log(\text{IC}_{50})$ values predicted using the scoring matrix indicated at the bottom right of each panel. The length distribution of the peptides was as follows: 36 10-mers, 18 11-mers, 6 12-mers, and one 13-, 15-, 16-, and 18-mer. The solid curves represent linear regression lines.

To test how well equation (8) predicts TAP affinities of peptides with more than 9 residues, it was applied to 64 peptides between 10 and 18 amino acids in length with measured affinities. As shown in Figure 13, the correlation between predicted and measured affinity values is lower than for the 9-mers, but still significant. The consensus matrix again provided higher correlation than all other matrices, so that it was used in all further applications.

3.4 Using TAP transport predictions for the identification of epitopes

To assess the selective role of TAP within the MHC-I presentation pathway, a test set of known naturally processed epitopes is needed. This is taken from the SYFPEITHI database (Rammensee, et al., 1999) and contains all known 9-meric epitopes that are presented naturally by any human MHC-I allele except those presented by HLA-A0201 (which are used later on), and for which the sequence of the source protein is available. MHC-I ligands, which are known to bind but which are not presented naturally are not included as well as epitopes derived from signal sequences. All other 9-mers contained in the protein sequences from which the epitopes originated are taken as random control peptides (=non-epitopes). In the following, this set of 203 epitopes and more than 60,000 random 9-mers is referred to as the HLA-X dataset.

To measure the prediction quality, again ROC curves and their integral (AUC) are used (section 2.5). First, the complete 9-mer consensus matrix is used to predict the TAP affinities of all 9-mers in the HLA-X dataset. These affinities are then used to separate epitopes from random 9-mers, resulting in the ROC curve plotted in Figure 14, curve (a), which corresponds to an AUC value of 0.702, indicating a relevant but not very good prediction.

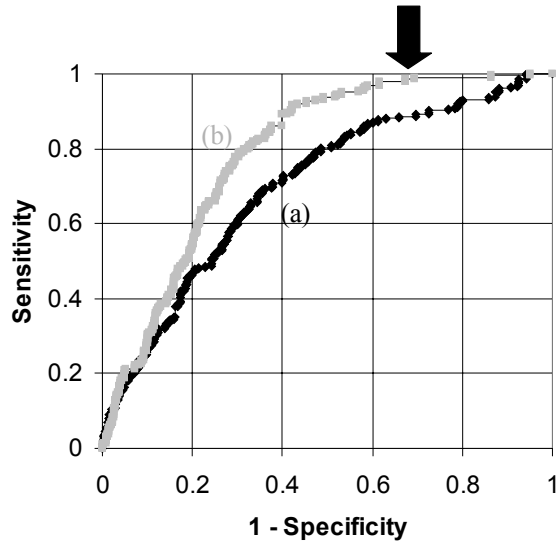


Figure 14: ROC curves for the HLA-X dataset.

Curve (a) was constructed using the entire consensus-matrix on the HLA-X dataset yielding an AUC value of 0.702. For curve (b) scoring equation (10) was used with $\alpha=0.2$ and $L=10$, giving AUC=0.791. The improvement is nearly completely in the high sensitivity region. The arrow indicates the point in curve (b) which corresponds to the sensitivity and specificity reached when choosing the cutoff=1, which is used later in the combined TAP and MHC-I predictions.

The same analysis was repeated but now including potential epitope precursors carrying N-terminal extensions. TAP affinities for N-terminal precursors of length 9, 10, ..., L were calculated for all epitopes and non-epitopes by means of equation (8). The TAP transport score of a potential 9-mer epitope is obtained by averaging over the TAP affinities of itself and its precursors up to a maximal length L:

$$\begin{aligned} \bar{t}_L &= \frac{1}{L-8} \sum_{l=9}^L t(\text{precursor}_l) \\ &= \text{mat}_{9,C} + \frac{1}{L-8} \sum_{l=9}^L \text{mat}_{1,N1} + \text{mat}_{2,N2} + \text{mat}_{3,N3} \end{aligned} \quad (9)$$

Note that all precursors contribute to the transport score with identical C-termini, while the N-terminal contributions are varying. Increasing successively the maximal number L of allowed N-terminal extensions and using the corresponding TAP transport scores to discriminate between epitopes and non-epitopes, the AUC values depicted in Figure 15, curve (a) are obtained. For $L=9$ (no N-terminal extension), equation (9) is equivalent to equation (8) and the AUC value amounts to 0.700, which is only marginally lower than the value 0.702 obtained when using the complete consensus matrix. This finding further justifies the usage of equation (8).

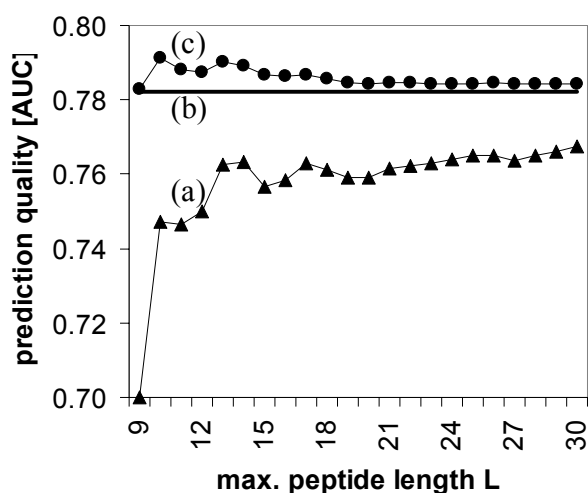


Figure 15: Prediction quality for the HLA-X dataset as a function of the maximal precursor length

Plotted is the prediction quality measured by the AUC of the TAP transport score for different predictions: (a) equal weight for N- and C-terminus (equation 9) (b) C-terminus score only (equation 10, $\alpha=0$) (c) optimal prediction with down-weighted N-terminus (equation 10, $\alpha=0.2$)

The AUC values improve significantly with increasing maximal precursor length L . This was not expected for L greater than 18, as the TAP transport efficiency for peptides exceeding this length has been shown to drop of significantly (van Endert, et al., 1994). Evidently, increasing step by step the possible length L of epitope precursors, the statistical average across their N-terminal scores will converge against a stable limit value thus rendering the influence of N-terminal

scoring less and less important for the prediction of TAP affinities. Hence in the limit $L \rightarrow$ infinity, only the C-terminus will account for differences in the TAP scores of different potential epitopes. To see how close this limit is, the AUC values were calculated using the C-terminus for scoring only (Figure 15, curve b). Surprisingly, the AUC value of 0.782 is higher than all AUC values obtained before. This finding raises the question whether the rise in AUC values seen with increasing length L of precursors does really reflect the usage of longer precursors in antigen production, or whether the N-terminal scores are just adding noise to the prediction, which is smoothed out with increasing L . To check this, the TAP transport scores of the N-terminal residues were weighted by a factor α :

$$\bar{t}_{L,\alpha} = mat_{9,C} + \frac{\alpha}{L-8} \sum_{l=9}^L mat_{1,N1} + mat_{2,N2} + mat_{3,N3} \quad (10)$$

In Figure 15, curve (a) corresponds to $\alpha = 1$ and curve (b) corresponds to $\alpha = 0$. If the increase in AUC values obtained with precursors $L > 9$ is only an artifact, one would expect the AUC for all values of L to grow monotonously when decreasing α from one to zero. If not, one would expect to find the optimal value of α somewhere between one and zero. The latter case is true: A maximum value of AUC was obtained for $\alpha = 0.2$ (curve(c) in Figure 15), which was significantly above the AUC value obtained when only scoring the C-terminus. Curve (b) in Figure 14 depicts the ROC obtained when choosing the optimal values $L = 10$ (i.e. one N-terminal extension) and $\alpha = 0.2$. Hence, predicting TAP affinities of N-terminally extended epitope precursors by down-weighting their N-terminal scores in comparison to their C-terminal scores significantly improves the discrimination between epitopes and non-epitopes. Possible explanations for the 'down-weighting' of the N-terminus will be analyzed below.

To exclude that the improvement in predictions obtained when choosing $\alpha < 1$ is a specific property of the HLA-X dataset, the same scoring procedure was applied to a completely independent set of mouse epitopes. This H2-X dataset was also extracted from the SYFPEITHI database following the same rules as those for the HLA-X dataset, but using mouse instead of human MHC-I alleles. Again it is tried to separate epitopes from random 9-mers using the predicted TAP transport efficiency (Figure 16), which is based on measurements of human TAP specificity. It has been shown that there are significant differences between the murine and

human TAP specificity (Momburg, et al., 1994), as human TAP translocates peptides with hydrophobic and basic C termini, whereas mouse TAP prefers only peptides with hydrophobic C termini. As expected, this results in generally lower AUC values than those for the HLA-X dataset. Nevertheless, qualitatively the three curves in Figure 16 (a)-(c) are related to each other in exactly the same way as those shown in Figure 15 for the HLA-X dataset: Using the scores for the N- and C-terminus with equal weights ($\alpha=1$) for the prediction of TAP affinities results in a worse discrimination between epitopes and non-epitopes than neglecting the N-terminus completely ($\alpha=0$). Again, a better prediction is achieved when the scores for the N-terminus are down-weighted with $\alpha=0.2$.

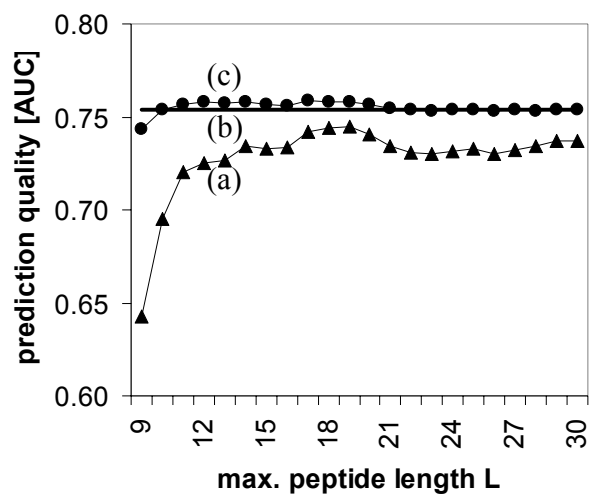


Figure 16: Prediction quality for the H2-X dataset as a function of the maximal precursor lengths

Plotted is the prediction quality measured by the AUC of the TAP transport score given in equation (10) for different predictions: (a) equal weight for N- and C-terminus ($\alpha=1$) (b) C-terminus score only ($\alpha=0$) (c) better prediction with down-weighted N-terminus ($\alpha=0.2$)

3.4.1 TAP transport predictions for individual MHC-I alleles

The calculations made in the previous section were repeated for individual MHC-I alleles that make up the HLA-X dataset to see how much the results vary. This analysis was restricted to those alleles for which at least 10 epitopes are present in the HLA-X dataset (Table 5). Epitopes presented by different allele subtypes were pooled in one set, for example the 'HLA-B27' set consists of epitopes listed in the SYFPEITHI database to be presented by HLA-B27 (unknown subtype) and the subtypes HLA-B2702, HLA-B2704 and HLA-B2705. While the binding preference of the allele subtypes can vary slightly, the datasets would otherwise be too small, especially as for many entries in the SYFPEITHI database the four digit code identifying the exact subtype is not given. The only exception is the HLA-A0201 set, for which only epitopes presented by this allele subtype are included.

First, it was studied how well the epitopes of each individual allele can be identified by TAP affinity scores computed without inclusion of possible precursors or down-weighting of the N-terminal residues (i.e. putting $L=9$ and $\alpha=1$ in equation (10)). The resulting AUC values (Table 5) show huge variations from 0.39 to 0.89. The differences in prediction quality for the individual alleles correspond very well with those reported in (Brusic, et al., 1999; Daniel, et al., 1998), where the alleles HLA-B27, -A3 and -A24 were classified as efficient for TAP loading (high AUC) and the alleles HLA-B07, B08 and A0201 were classified as inefficient for TAP loading (low AUC).

Repeating the AUC calculations with the optimal parameters $L=10$ and $\alpha=0.2$ obtained for the entire HLA-X dataset, the AUC values fall in a much narrower range between 0.71 and 0.88, i.e. a subdivision into TAP-efficient and TAP-inefficient alleles is no longer preserved. These results provide evidence that TAP plays an equally important role for peptide loading of all alleles considered. Intriguingly, some alleles such as HLA-B27 or HLA-A3 seem to be preferentially loaded with peptides directly imported from the cytosol whereas other alleles such as HLA-B35 or HLA-0201 are preferentially loaded with peptides entering the ER as N-terminally extended precursors where they are cut to final size.

Table 5: Individual alleles

	# Epitopes	AUC L=9, $\alpha=1$	AUC L=10, $\alpha=0.2$	Optimal α for L=10
HLA-B35	10	0.39	0.80	0.0
HLA-B07	11	0.43	0.71	0.0
HLA-B08	10	0.69	0.80	0.0
HLA-B44	11	0.78	0.88	0.0
HLA-A24	37	0.81	0.87	1.0
HLA-A3	11	0.82	0.75	1.2
HLA-B27	20	0.89	0.77	4.0
HLA-A0201	87	0.65	0.70	0.4

Finally, the optimal value of α for each individual allele was calculated when setting $L=10$. The resulting values vary between 0 and 4, showing that the optimal value of α is extremely allele specific: The better the C-terminal residues required for effective TAP transport agree with those C-terminal residues enabling effective MHC-I binding to the given allele, the lower the weight that has to be put on the N-terminal residues. The optimal value of $\alpha=0.2$ for the whole HLA-X dataset shows that, on the average, C-terminal amino acid motives required for effective TAP transport and MHC-I binding overlap stronger than the corresponding N-terminal motives. This is probably due to a stronger force for co-evolution on that motif, as the C-terminus undergoes no change from TAP transport to MHC-I binding, while the N-terminus can be trimmed.

3.4.2 Consequences of the uncertainty as to which N-terminally extended precursors are generated *in vivo*

Another explanation why better epitope predictions were achieved with $\alpha < 1$ is the uncertainty as to which epitope precursors are actually transported *in vivo* to liberate the definitive epitope in the ER by N-terminal trimming. Equation (10) is based on the unrealistic assumption that up to a critical length L all N-terminally prolonged precursors of an epitope are present in comparable abundance. Given that several precursor are not generated *in vivo*, their score for the N-terminus will 'dilute' that of the existent precursors. From the statistical point of view, this would favor to put a higher weight on the score of the C-terminus, or equivalently, to down-weight scores of the N-terminal residues.

To estimate the implications of precursor uncertainty for the choice of α , simplified simulations of the MHC-I pathway were performed: Using the protein sequences from which the epitopes of the HLA-X dataset originate, a set of m fragments per sequence obeying a log-normal length distribution is generated, as was observed for the cleavage products of the proteasome (Kisselev, et al., 1999). These m fragments per sequence are considered to be the pool of potential epitope precursors generated by the proteasome that contain a C-terminal 9-mer which can bind to an MHC-I molecule. Which of these fragments becomes an epitope is decided by their affinity to TAP, which is calculated using equation (8). The fragment with the highest affinity per sequence is chosen, defining with its last 9 down-stream residues an epitope. The other $m-1$ fragments are discarded. It is then tried to identify these artificially generated 9-mer epitopes among all other 9-mers contained in the protein sequences by applying the TAP transport score (equation 10) at varying values of α .

The highest AUC values in all simulations were indeed obtained when choosing $\alpha < 1$. Figure 17 shows the AUC values for such a simulated dataset. In this case the highest AUC value was obtained for $L=11$ and $\alpha=0.6$. Varying the width of the hypothetical length distribution in the simulations, the optimal α values were always between 0.6 - 0.9, i.e. larger than the value $\alpha=0.2$ yielding the best prediction of epitopes on real experimental datasets but always smaller than 1.

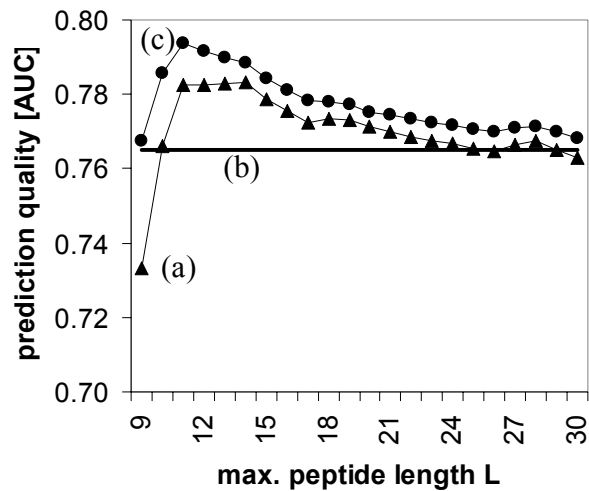


Figure 17: Prediction quality on a simulated dataset

Plotted is the prediction quality measured by the AUC of the TAP transport score given in equation 10 for different predictions: (a) equal weight for N- and C-terminus ($\alpha=1$) (b) C-terminus score only ($\alpha=0$) (c) optimal prediction with down-weighted N-terminus ($\alpha=0.6$)

There are three free parameters in the simulation: the number m of different fragments used to define a single epitope and the mean and standard deviation of the log-normal length distribution of peptides generated. The larger the value of m , the higher the selective power that TAP has in the pathway in comparison to the proteasome and the MHC-I molecules. By systematically increasing the value of m , it was found that with $m=10$ the AUC value on the basis of the TAP score for the C-terminus alone was close to those AUC values in Figure 15 and Figure 16 observed with real experimental data. The length dependence of the AUC values was in good concordance with that shown in Figure 15 and Figure 16 when choosing the mean of the log-normal length distribution in the range 9 – 11.

3.5 Combining TAP transport predictions with predictions of MHC-I affinity

It was tested whether the combination of predictions for two main steps of the presentation pathway, TAP transport and MHC-I binding, can improve the identification of epitopes. These calculations were performed on a set of 87 HLA-A0201 presented epitopes which had been omitted from the HLA-X dataset. For the prediction of peptide binding to HLA-A0201, the SMM scoring matrix developed in section 2.4 was used. On its own, this matrix already possesses a high capacity to identify the epitopes of the HLA-X dataset (AUC = 0.919, cf. Figure 18 curve (a)).

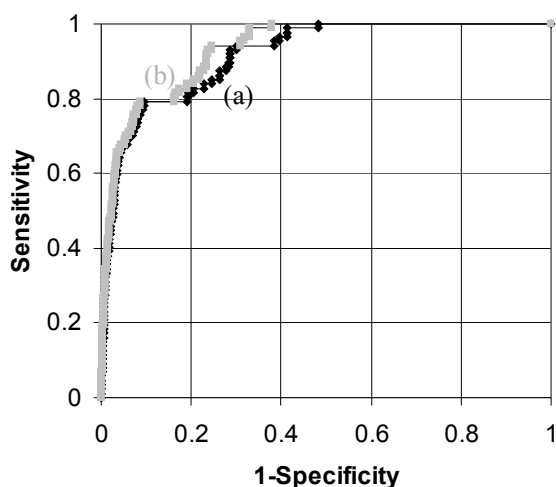


Figure 18: ROC curves for the combined TAP and MHC-I prediction on the HLA-A0201 dataset

The black curve (a, AUC=0.919) shows the (very high) level of the MHC-I prediction alone. The consistently better gray curve (b, AUC=0.932) is made by classifying all 9mers with a TAP transport score worse than 1 as not transported ($\alpha=0.2$, $L=10$), and limiting the MHC-I prediction to the transported peptides.

To combine predictions of MHC-I binding with predictions of TAP transport, first the TAP transport efficiency is calculated for all 9-mers contained in the source sequences of the HLA-A0201 epitopes using equation (10) with the parameters $L=10$ and $\alpha=0.2$. All 9-mers with TAP scores above the threshold value 1 are classified as not transportable and excluded from the set

of epitope candidates. This cutoff value was chosen by examining the ROC curve for the HLA-X dataset: Only 1.5% of epitopes but 32% of random 9-mers have a higher (=worse) TAP score (arrow in Figure 14). In the second step, the predicted MHC-I binding scores of the remaining peptides (having TAP scores < 1) were used to discriminate between epitopes and non-epitopes. Based on this two-step prediction protocol, the AUC value increases significantly to 0.932 (Figure 18, curve(b)). The improvement is largest in the high sensitivity region: Demanding 100% sensitivity, specificity is increased from 52% to 62% when using the combined prediction instead of MHC-I affinity prediction alone.

The same two-step prediction protocol was repeated for several mouse MHC-I alleles, using scoring matrices for the MHC-I affinity prediction that were measured by (Udaka, et al., 2000). Unfortunately, the number of epitopes available in the SYFPEITHI database per allele is small, ranging from 9 to 21 (Table 6). For three of the mouse alleles, the combined predictions gave better AUC values than MHC-I affinity predictions alone. For one allele (H2-Db), the combined prediction was worse. This shows that the combined MHC-I + TAP prediction using a human TAP matrix works for mouse epitopes, even though there are significant differences between the murine and human TAP specificity. This should improve significantly when using a scoring matrix based on experimental data for murine TAP.

Table 6: Combined TAP and MHC-I predictions

Dataset	# Epitopes	AUC values	
		MHC-I only	MHC-I + TAP
HLA-A201, 9-mers	87	0.919	0.932
H2-Kb, 8-mers	21	0.961	0.965
H2-Kb, 9-mers	9	0.855	0.879
H2-Db, 9-mers	20	0.971	0.949
H2-Ld, 9-mers	10	0.985	0.987

3.6 Confidence in the values of the free parameters α and L

There are two free parameters in the prediction of TAP transport scores: α and L . Throughout most of this chapter, the values $\alpha=0.2$ and $L=10$ were used, which were determined to be optimal for the HLA-X dataset containing epitopes from all human MHC-I alleles except HLA-A0201. These parameters show large variations when calculated for the individual alleles that make up the HLA-X dataset (Table 5), as they are heavily influenced by each individual alleles binding preference. The parameter values for the entire HLA-X dataset, which average out the individual alleles binding preferences, should reflect the true effect of TAP more accurately. The optimal parameter values calculated for the H2-X dataset ($\alpha_{\text{opt}}=0.02$ and $L_{\text{opt}}=11$), or the combined MHC-I and TAP predictions for HLA-A0201 ($\alpha_{\text{opt}}=0.6$ and $L_{\text{opt}}=18$), which should also reflect the true effect of TAP, are considerably different from those for the HLA-X set. However, the decrease in prediction quality when using $\alpha=0.2$ and $L=10$ instead of the optimal parameters for these datasets is quite small ($\Delta\text{AUC} < 0.006$) compared to the loss in prediction quality of ($\Delta\text{AUC} \sim 0.100$) when making predictions without down-weighting and neglecting precursors (i.e. $\alpha=1$, $L=9$). Apparently both parameters are meaningful, but their optimal values cannot be fixed within a narrow range. The usage of $\alpha=0.2$ and $L=10$ can therefore be recommended, even though, from a biological perspective, $L=10$ seems to be too small as longer precursors are known to be used *in vivo*.

3.7 Summary

In this chapter, a novel method to predict the TAP affinity of peptides of any length was introduced, which gave reasonably good predictions for peptides 9 - 18 residues long. This was used to assign an effective TAP transport score to a potential epitope, by averaging over the predicted TAP affinities of the epitope itself and its precursors. The ability of this score to discriminate between random 9-mers and presented epitopes improved when down-weighting the influence of the N-terminal residues. This was reasoned to be the consequence of the uncertainty which epitope precursors are present *in vivo* as well as possible co-evolution in the preference for the peptide C-terminus of TAP and the average MHC-I molecule.

Using the predicted TAP transport efficiency to identify naturally processed epitopes for individual MHC-I alleles showed that TAP does exert significant pressure on the epitope selection of all MHC-I alleles.

To combine TAP transport predictions with those of MHC-I affinity, all potential epitopes with a predicted TAP transport efficiency considered to be 'non-transportable' are eliminated. Using this as a filter prior to MHC-I affinity predictions improved the prediction quality considerably above that of the MHC-I affinity predictions alone.

4 Peptide generation by the proteasome

The proteasome degrades intracellular proteins to peptides between 3-30 amino acids in length. This pool of peptides is thought to be the main source of MHC-I epitopes or their N-terminally prolonged precursors. As demonstrated in the previous chapter, identification of epitopes can be improved by using predicted TAP transport efficiencies as a filter that rules out poorly transported peptides without notably reducing the number of true epitopes. The obvious next step is to check if the same strategy can be applied to filter out those epitope candidates that are unlikely to be generated by the proteasome.

This chapter starts with an evaluation of existing algorithms that predict proteasomal cleavage (section 4.1). Their predictions are poor, which is thought to be a consequence of the lesser quality of their training data, as proteasomal cleavage rates are inherently difficult to measure and interpret, which is discussed in section 4.2. To address this problem, a novel method to quantify proteasomal cleavage rates from time resolved experiments is introduced in section 4.3. This method is applied to a series of experiments analyzing the digestion of two polypeptide substrates by constitutive and immuno-type proteasomes (section 4.4). In the last section (4.5), the differences between digests with immuno- and constitutive proteasomes are discussed.

Most of the results reported in this chapter are taken from (Peters, et al., 2002; Peters, et al., 2003).

4.1 Evaluating published algorithms predicting proteasomal cleavage

Hitherto there are no indications that the C-terminus of proteasomal fragments undergoes further trimming along the MHC-I presentation pathway (Rock and Goldberg, 1999; Shastri, et al., 2002). Therefore, selecting potential epitopes and their N-terminally prolonged precursors by the probability that their C-terminus is generated by the proteasome should single out false epitope candidates without losing true epitope candidates. Currently, there exist three publicly available methods to predict proteasomal cleavage: NetChop (Kesmir, et al., 2002), PaProc (Nussbaum, et al., 2001) and FragPredict (Holzhutter, et al., 1999). All of these are trained on data from *in vitro* digests of proteins or oligopeptides.

4.1.1 NetChop

The NetChop algorithm (Kesmir, et al., 2002) is an artificial neural network trained on different sets of experimental data. Here, the 20S version of NetChop trained on *in vitro* digest of yeast enolase (Toes, et al., 2001) and bovine β -casein (Emmerich, et al., 2000) was used. The output of the algorithm for each possible cleavage site within a protein sequence is a continuous number indicating the likelihood of cleavage. Predictions were obtained online at www.cbs.dtu.dk/services/NetChop.

Alternative versions of NetChop are available that have been trained on collections of the flanking regions of known presented epitopes, which are thought to be cleavage sites of the proteasome. While this can be a valid approach, predictions trained this way can obviously not be used to evaluate the influence of the proteasome on epitope generation, as the proteasome is implied to be the source of all epitopes when using this kind of training data. Rather, the training data has to come directly from the proteasome itself, as is the case for proteasomal in-vitro digests.

4.1.2 PaProc

PaProc (Nussbaum, et al., 2001) is essentially a matrix based method combined with pair-coefficients describing the interaction between the residues P1 and P1' surrounding the cleavage site. The coefficient values were determined using an evolutionary algorithm. The training data consists mainly of an in-vitro digest of yeast enolase plus several polypeptides (Kuttler, et al., 2000). There are several implementations of the method based on different sets of experimental data. Here, the 'wild type III' method was used, which was trained on the largest dataset. PaProc is available online at www.paproc.de. Its output consists of 4 different discrete scores ('-', '+', '++' and '+++'), where '-' is designated to be 'non-cleavable'.

4.1.3 FragPredict

The FragPredict method (Holzhutter, et al., 1999) is not available online, but as a computer program distributed on request. It was the first published prediction method, trained on all in-vitro digests of polypeptides published at that time. It is capable not only of predicting cleavage

sites, but also to predict which fragments are formed from combinations of cleavages. To be comparable to the other methods, only the cleavage site prediction algorithm was used.

4.1.4 Identifying epitopes using proteasomal cleavage predictions

For each peptide, the predictions of its C-terminal cleavage were used to determine if it has the potential to become an epitope or not. Figure 19 depicts the ROC curves for the three cleavage prediction methods when applied to the HLA-X dataset described in section 3.4. According to the AUC values, the best discriminations between epitopes and random peptides were achieved with NetChop (AUC=0.61), closely followed by FragPredict (AUC=0.59), while PaProc (AUC=0.54) was significantly inferior to the other two prediction methods. Comparing the ROC curves of Figure 19 with those of Figure 14, it can be inferred that the discriminating power of existing prediction methods for proteasomal cleavage sites is far below that of TAP transport scoring developed in the previous chapter, let alone those for MHC-I affinity.

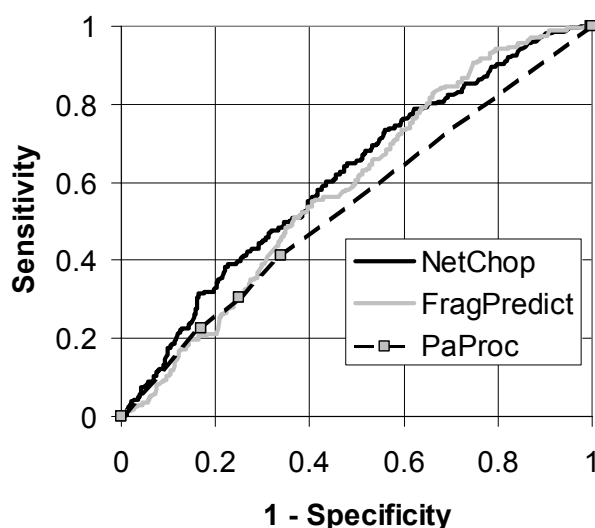


Figure 19: ROC curves for proteasomal cleavage predictions

For the three proteasomal cleavage prediction methods NetChop (AUC=0.61), FragPredict (AUC=0.59) and PaProc (AUC=0.54), the score for C-terminal cleavage is used to predict epitopes from the HLA-X dataset.

4.1.5 Combining proteasomal cleavage predictions with predictions of MHC-I affinity

Next, combined predictions of C-terminal proteasomal cleavage and MHC-I binding were tested, using the same two-step prediction protocol as described for TAP in section 3.5. For each of the three prediction methods of proteasomal cleavages, a cutoff value singling out peptides as 'not-generated' was chosen with a similar selective strength as the one used for TAP-transport, where 30% of the peptides were classified as 'not-transportable'. For PaProc, the fraction of omitted peptides was necessarily larger, as this method predicted about 60% of peptide bonds to have the lowest score ('-', not cleaved). The ROC curves for the combined predictions are shown in Figure 20; all of them indicating that the combined predictions are significantly worse than those based on predictions of MHC-I binding affinities alone.

Apparently, the 2-step prediction protocol used successfully to combine TAP and MHC-I prediction fails when predictions of C-terminal proteasomal cleavages were used as a filter. This disappointing result may have three different reasons. One is, that the selective power of the proteasome is weak as it generates nearly every possible peptide. Second, there might be other proteases serving as suppliers of antigenic peptides besides the proteasome. Finally, existing prediction algorithms of proteasomal cleavage sites might not be accurate enough. The last explanation seems most likely, because *in vitro* digests of epitope-containing model substrates by the proteasome provide with very few exceptions the epitope or one N-terminally prolonged precursor (Kessler, et al., 2001). The poor quality of prediction algorithms for proteasomal cleavage sites is also evidenced by contradictory results obtained when applying them to the same set of test protein sequences. Most likely, the poor prediction quality of proteasomal cleavages is mainly caused by the lack of a sufficiently large set of quantitative and consistent experimental data on cleavage rates, which are more difficult to measure and interpret than the affinity assays used to characterize peptide binding to TAP and MHC-I.

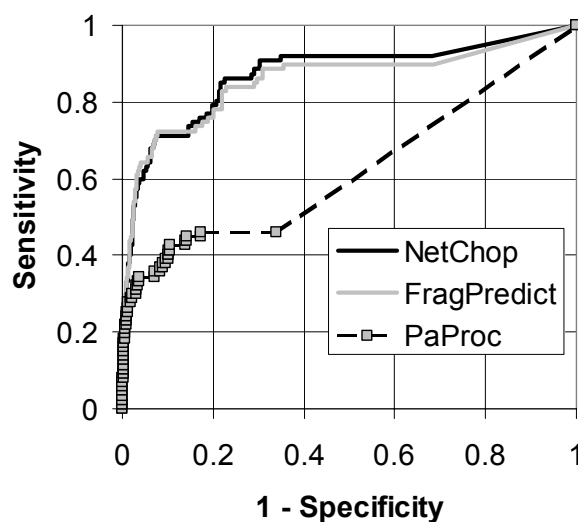


Figure 20: ROC curves for proteasomal cleavage + MHC-I binding predictions

Only peptides with a score for proteasomal cleavage of their C-Terminus better than a fixed cutoff are considered to be potential epitopes. They are assigned a score according to their predicted MHC-I binding affinity. The best results are obtained with NetChop (cutoff = 0.1, AUC=0.872) followed by FragPredict (cutoff = 0.5, AUC=0.858) and PaProc (cutoff = '+', AUC=0.623). All of these combined predictions are worse than using predicted MHC-I affinities alone (AUC = 0.919).

4.2 Problems with evaluating experimental proteasome digests

For an *in vitro* digest, proteasomes are incubated with a polypeptide or protein as a substrate. After a defined incubation time, the digest is stopped, and the generated mixture of peptide fragments is called the proteasomal digest of the substrate. To analyze these digests, usually Edman degradation or Mass Spectrometry (MS) are used. These methods are associated with different obstacles in the interpretation of results.

4.2.1 A single snapshot of a digest does not provide reliable cleavage rates

Using Edman degradation to analyze proteasomal digests, the peptide mixture is first separated using high performance liquid chromatography (HPLC). Ideally, each probe coming from the

HPLC should contain only one kind of peptide. The sequence and amount of each peptide can then be identified using Edman degradation. This is a reliable but time consuming method to produce quantified data, which has lead most experimentalists to limit the analysis of digests to a single incubation time, i.e. to analyze a snapshot of the fragment concentrations present in the digest at one time.

A naïve way of interpreting this snapshot is to divide the concentration of each generated fragment by the amount of depleted substrate and interpret these ratios as relative generation rates. This is not a valid interpretation, because proteasomal digests do not follow a simple substrate + enzyme \rightarrow substrate + product description. The proteasome can 're-process' its products, cutting them further into smaller fragments. While this re-processing may not play a significant role *in vivo*, where the products will either be degraded by other proteases or rescued from degradation by transport into the ER by TAP, it is unavoidable for *in vitro* experiments. Therefore, these relative generation rates would vary hugely depending on the incubation time, because longer fragments dominating at early times will later be cleaved into smaller fragments. This can also lead to misinterpretations of differences in the digests generated by different types of proteasomes. If two types differ only in their speed in which they degrade a substrate, the amounts of fragments generated can vary greatly after the same incubation time, even if their cleavage preference is completely identical (Figure 21)

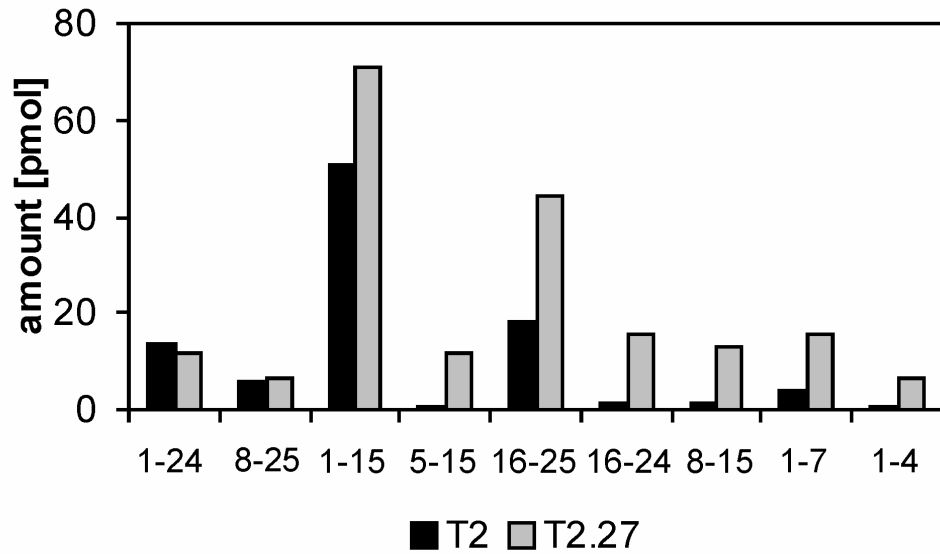


Figure 21: Different proteasome species with identical cleavage preference can produce large differences in individual fragment amounts

The data in the Figure stems from experiments described in section 4.4.1. The black and gray bars indicate the amount of nine pp89-25mer peptides produced by the T2 and T2.27 proteasome after 2h of incubation. The peptide amounts were assessed from the respective MS-signals by using calibration curves. The position of each peptide fragment in the sequence of the substrate is indicated on the x-axis. There are significant differences in the amount of the peptides 5-15, 8-15 or 16-24. Since the cleavage probabilities are unaltered (values given in Table 8), these differences result exclusively from the faster procession by the T2.27 proteasome and its tendency to re-process shorter peptides.

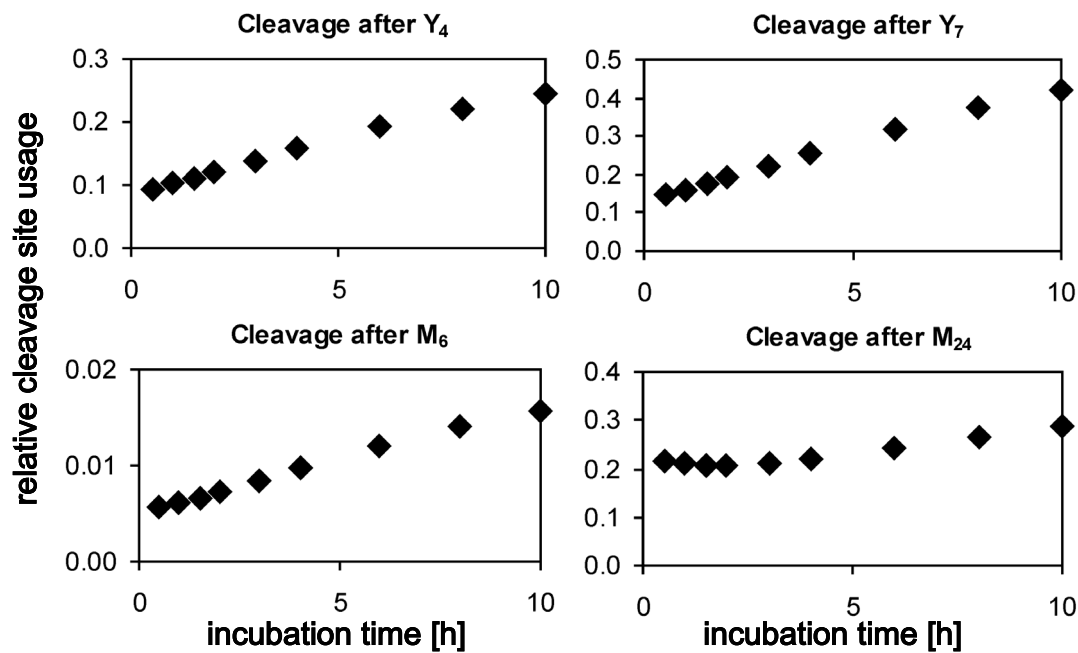


Figure 22: Re-processing of peptides makes the relative amounts of fragments associated with each cleavage site time dependent

The data used in this figure stems from the model fits described in 4.4.3, which are a noise-free set of peptide amount profiles. The four graphs depict the relative usage of the cleavage sites Y_4 , M_6 , Y_7 and M_{24} in the pp89-25mer at various time points of the simulated digestion experiment with T2.27. The relative usage of a cleavage site at a given time point is calculated by summing up the amounts of all peptides beginning or ending at that cleavage site, divided by the maximum sum found for any site at that time point (always after L_{15} in these experiments). If the relative usage of a cleavage site was equivalent to the cleavage probability in the substrate, it should be constant over time, as the cleavage probability is an intrinsic property of the substrate. As can be seen from the graphs, the relative usage is not constant over time, as re-processing of a fragment increases the usage of weaker cleavage sites that are still present in the fragments of the substrate.

A much better way to evaluate proteasomal digests is to sum up the amounts of fragments associated with each cleavage site, thereby assigning cleavage strengths, which are thought to be equivalent to cleavage site usage in the original substrate. While this is much better than to look at individual fragments, this definition of cleavage strengths also depends on the digestion time, as shown in Figure 22. This is due to the following reasons: (1) As the strongest cleavage sites are cut first, their number decreases faster than others, making it more likely that weaker cleavage sites are used when fragments are re-processed. (2) It is known that shorter peptides are less likely to be cleaved than longer peptides, making the cleavage site usage dependent on its surrounding sequence, which changes when fragments are re-processed.

4.2.2 MS-signals do not give quantified peptide amounts

As discussed in the previous section, experiments evaluating only one digestion time-point can only provide a snap-shop of the digest that cannot completely determine the mechanism of degradation of the proteasome. Using Edman degradation to analyze the digests, repeating an experiment for several different digestion times means lots of work. A much quicker method to analyze digest data is mass spectrometry (MS). Here, the peptides of the digest are again typically separated by HPLC and thereafter analyzed by MS. While this allows for a highly sensitive qualitative analysis of the digest (a list of peptides that were generated in a detectable amount after a certain incubation time), estimation of the quantities of the peptides is problematic. The intensity of the MS-signal is in principal related to the detected peptide amount, but several intrinsic properties of the peptides influence their ionization behavior and therefore the MS-signal. The presence of aromatic amino acids (Valero, et al., 1998), phosphate groups (Janek, et al., 2001), and charged side chains (Cohen and Chait, 1996) such as guanidino group of arginine (Krause, et al., 1999) as well as the peptide size (Olumee, et al., 1995) have been reported to influence the signal intensity. Hitherto there is no reliable theoretical approach enabling the calculation of the MS-signal intensity from the sequence of a given peptide. In principle, the problem to derive amount values from MS-signals can reasonably well be solved by synthesizing the observed peptides and measuring calibration curves for each of them, but this is also a rather time consuming work, especially for digests of long protein substrates in which a large number of observed peptides is produced.

4.3 Novel protocol of experimental evaluation

4.3.1 Determining peptide amounts from MS-signals

In this section, a much more efficient method to assess peptide amounts from MS-signals than the use of calibration curves is proposed. The basic idea is to use mass balance rules: At an arbitrary time point of the digest experiment, the amounts of all peptides having at least one sequence position in common must add up to the amount of the substrate at the beginning of experiment. Mathematically, this conservation rule can be stated as

$$\sum_{\{i:j \in f_i\}} a_i(t) = a_0 \quad (j=1, \dots, n) \quad (11)$$

where a_0 is the initial amount of the substrate of length n and $a_i(t)$ denotes the amount of peptide i at time t . The sum on the left-hand side of equation (11) includes all those peptides f_i that contain sequence position j . From the calibration curves shown in Figure 26, it can be inferred that the relationship between MS-signal and peptide amount can be roughly approximated by a linear function,

$$a_i = v_i s_i \quad (12)$$

where s_i denotes the MS-signal produced by peptide i and the signal conversion coefficient v_i is a characteristic constant determined by the physico-chemical properties of peptide i converting its MS-signal into the respective amount value. Demanding fulfillment of equation (11) for all sequence positions, one may estimate the scaling factors v_i by inserting relation (12) into the conservation equation (11):

$$\sum_{\{i:j \in f_i\}} v_i s_i(t_\alpha) = a_0 \quad (j=1, \dots, n; \alpha=1, \dots, m) \quad (13)$$

where the index α counts the number of discrete time points at which MS-signals for the peptides are available. Numerical values for the unknown conversion factors v_i can then be estimated by minimizing the violation of the $n \times m$ conservation conditions (13). Violation of these

conservation rules may result from three sources: First, measurements of the MS-signals are subject to random as well as systematic errors. Second, the true functional relationship between the signals and the amount of a peptide will certainly deviate from a simple linear one. Third, the set of detectable peptides will never be complete. In particular, short peptides (1-3 residues) are likely to escape from HPLC-MS analysis. The latter fact gives rise to a systematic loss of mass as more small peptides are formed during the time-course of the digest. Therefore it is reasonable to determine the unknown conversion factors by minimizing the violation of the conditions (13) between two successive time points of the experiments, i.e.

$$\Phi = \sum_{j=1}^n \sum_{\alpha=1}^{m-1} \left\| \frac{\sum_{\{i:j \in f_i\}} v_i s_i(t_{\alpha+1}) - \sum_{\{i:j \in f_i\}} v_i s_i(t_{\alpha})}{\sum_{\{i:j \in f_i\}} v_i s_i(t_{\alpha})} \right\| \rightarrow \text{MINIMUM!} \quad (14)$$

and choosing the distance metric in (14) as

$$\|x\| = \begin{cases} x^2 & \text{if } x < 0 \\ 5x^2 & \text{if } x > 0 \end{cases} \quad (15)$$

which punishes the unlikely 'gain' of peptides ($x > 0$) five times higher than their more likely 'loss' ($x < 0$).

When minimizing the functional (14) with respect to the unknown signal conversion coefficients v_i , one encounters the typical problem in regression analysis that the signal conversion coefficients of peptides with very small MS-signals are poorly determined because they can be largely varied without significant change of the functional Φ . Thus, to avoid unrealistic values of the calculated signal conversion coefficients, the minimization problem (14) is replaced by the constraint problem

$$\Phi + \lambda \Psi \rightarrow \text{MINIMUM!} \quad (16)$$

where the additional term

$$\Psi = \sum_i \left(\log \frac{v_i}{v_0} \right)^2 \quad (17)$$

measures the deviations of the v_i 's from a plausible reference value v_0 . This reference value v_0 was determined from a set of experimental calibration curves. Depending on the choice of the positive factor λ in (16), the minimization problem may become at the extreme either completely unconstrained ($\lambda \rightarrow 0$), or all signal conversion coefficients are forced to the reference value v_0 ($\lambda \rightarrow \infty$).

4.3.2 Kinetic modeling

In this section, a kinetic model of the proteasome is introduced which is supposed to serve as a mechanistic platform for the interpretation and comparison of kinetic data produced by *in vitro* digestion of model substrates. Proteasomal degradation comprises a multitude of distinct elementary processes, such as uptake of the substrate, transport through the interior of the proteasome, binding to the active sites, threonine-catalysed cleavage of peptide bonds under putative formation of covalent acyl-intermediates, hydrolytic liberation of these acyl-intermediates from the active-site threonine, and release of the products from the proteasome. As none of these elementary processes could be kinetically characterized so far, it makes no sense to incorporate them individually into a complex kinetic model containing a huge number of non-identifiable parameters. Instead, a simple kinetic model is established by lumping all elementary processes involved in the complete procession of a peptide into a single overall processing step. Compared with classical enzyme kinetics, the resulting proteasome model can be considered as a sort of Michaelis-Menten model expressing the most essential kinetic features in terms of a few phenomenological parameters which can be identified from the experimental data.

The time-dependent variation of the amount of peptides including the initial substrate is described by a system of linear kinetic equations,

$$\frac{d}{dt} \begin{pmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_N(t) \end{pmatrix} = \begin{pmatrix} -K_1 & 0 & 0 & \cdots & 0 \\ k_{21} & -K_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ k_{N1} & k_{N2} & k_{N3} & \cdots & 0 \end{pmatrix} \begin{pmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_N(t) \end{pmatrix} \quad (18)$$

Here k_{ij} is the rate constant with which peptide j is converted into peptide i per time unit and K_i is the total degradation rate of the i -th peptide. The peptides are labeled with decreasing lengths, a_1 being the substrate, so that $k_{ij} = 0$ for $i < j$ since cleavage always shortens a peptide.

In order to derive an explicit expression for the transition rates k_{ij} , two cardinal terms are introduced: the *procession rate* r_j of peptide j and the *cleavage probability* p_k of a cleavable peptide bond (= cleavage site). These two terms are explained in the following.

4.3.2.1 Procession rate

The procession rate is the rate (i.e. number of events per time unit) with which a peptide undergoes a procession cycle. A single procession cycle encompasses all events taking place between uptake of a peptide into the proteasome and release of all peptides derived from it. For peptide j with length L_j , it is put

$$r_j = \frac{r_{\max}}{1 + \left(\frac{L_0}{L_j}\right)^c} \quad (19)$$

where r_{\max} represents the maximum possible procession rate, L_0 represents a critical peptide length at which 50% of the maximum procession rate is reached, and the exponent $c > 0$ controls how sensitive the procession rate is to varying peptide lengths. This takes into account in a phenomenological manner that short peptides are degraded with lower turnover rates than longer peptides. A decelerated degradation with decreasing peptide length was observed for oligopeptides having up to 30 residues (Dolenc, et al., 1998), which is likely to be the maximum size of cleavage products. This type of length dependency can also explain why proteasomal

digests contain medium-size peptides which are not further degraded although they contain peptide-bonds which were cleavable in the original substrates.

4.3.2.2 Cleavage probability

A cleavage probability p_k is assigned to all cleavage sites k of the protein substrate, i.e. to those peptide bonds which need to be cleavable to explain the peptide pattern observed in the digest. The cleavage probability of all other peptide bonds is a priori put to zero. The assumption is made that multiple cleavages may occur independently and randomly during a processing cycle. This implies that there are as many different partitions, i.e. possible subdivisions of a given peptide into smaller pieces, as there are different combinations of possible cleavages. If the substrate contains n^* cleavage sites, there are 2^{n^*} such possible partitions, each of them occurring with a partition probability P_m ($m=0, \dots, 2^{n^*}-1$) that is determined by the cleavage probabilities of the individual cleavage sites (cf. Figure 23 for a simple example with $n^*=2$).

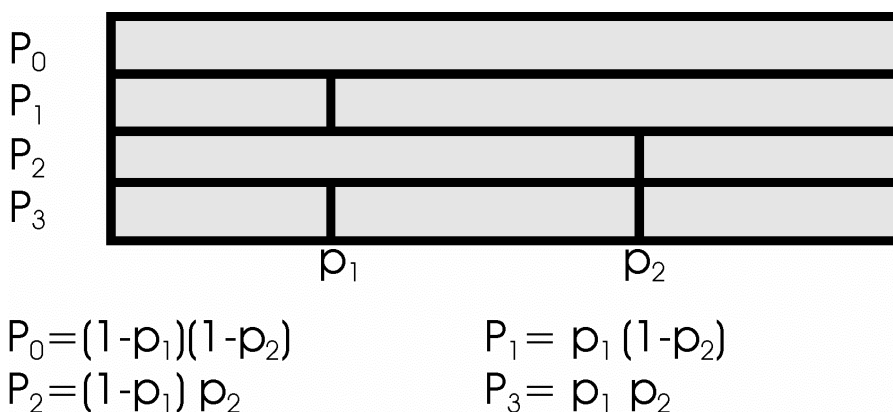


Figure 23: Possible partitions of a peptide containing 2 cleavage sites

Partition probabilities P_m are calculated by treating the individual cleavages as statistically independent events. For example: The probability P_2 to fractionize the substrate according to partition 2 is given by the probability p_2 for a cut to occur at cleavage site 2 times the probability $(1 - p_1)$ that cleavage site 1 is not cut.

As this model generates all peptides that can be produced by any combination of cleavage sites, there will usually be more peptides predicted in the model than observed in the experiment. These peptides are called hypothetical peptides.

4.3.2.3 Definition and estimation of rate constants

The rate constants k_{ij} in the equation system (18) are chosen as the procession rate for peptide j times the sum of the probability of all partitions in which peptide i is generated:

$$k_{ij} = r_j \sum_m P_m \quad (20)$$

Similarly, the coefficients K_i are given by

$$K_i = r_i \sum_{m>0} P_m \quad (21)$$

where the sum includes all partitions except P_0 , in which no cleavage occurs at all. For a given set of cleavage probabilities and procession rate parameters, the k_{ij} and K_i have explicit values for which the linear differential equation system (18) can be solved analytically yielding explicit mathematical formulas for the theoretical peptide amount profiles $a_i(t)$. Thus, numerical values for the unknown model parameters (r_{\max} , L_0 , c and p_k with $k=1, \dots, n^*$) can be determined by minimizing the distance between the theoretical peptide amount profiles and the observed ones. This minimization is performed using the following distance metric Δ :

$$\begin{aligned} a_{\text{sim}} < a_{\text{mid}} : \quad \delta &= \frac{a_{\text{mid}} - a_{\text{sim}}}{a_{\text{mid}} - a_{\text{min}}} \\ a_{\text{sim}} > a_{\text{mid}} : \quad \delta &= \frac{a_{\text{sim}} - a_{\text{mid}}}{a_{\text{max}} - a_{\text{mid}}} \end{aligned}$$

$$\Delta = \begin{cases} \delta & \text{if } \delta \leq 1 \quad \text{i.e. } a_{\text{sim}} \in [a_{\text{min}}, a_{\text{max}}] \\ 5\delta^2 - 4 & \text{if } \delta > 1 \quad \text{i.e. } a_{\text{sim}} \notin [a_{\text{min}}, a_{\text{max}}] \end{cases} \quad (22)$$

In (22) the symbols a_{mid} , a_{min} and a_{max} denote the mean, minimum and maximum peptide amount as derived from the measured MS-signal and a_{sim} denotes the simulated value predicted by the model. The distance metric Δ increases steeply (as a quadratic function) for values of a_{sim} lying outside of the experimental range $[a_{\text{min}}, a_{\text{max}}]$. The weighting factor 5 is somewhat arbitrary as long as it is greater than 1. Subtracting 4 ensures continuity of the distance at $\delta=1$. If a calibration curve was used to assess the amount of a peptide, the values for a_{mid} , a_{min} and a_{max} were taken directly from the calibration curve as described in Figure 26. If the mass balance method was used, the value for a_{mid} was determined using the signal conversion coefficient and putting $a_{\text{min}}=a_{\text{mid}}/2$ and $a_{\text{max}}=2 a_{\text{mid}}$.

To be consistent with the experiment, the hypothetical peptides found only in the model should have amounts below the quantification threshold. As discussed below, this threshold is about 5 pmol. To be on the safe side, the values $a_{\text{mid}} = a_{\text{min}} = 0$ and $a_{\text{max}} = 2$ pmol were chosen in the distance metric (22) for all hypothetical peptides.

4.4 Application and testing of novel protocol

4.4.1 Experimental setup

Time-dependent peptide profiles were obtained from degradation of the two model peptides: pp89 (a 25-mer derived from the IE pp89 of the *Murine Cytomegalovirus*) and LLO, a 27-mer representing a partial sequence region of listeriolysin O from *Listeria monocytogenes*). These two substrates were digested by a constitutive proteasome (T2) isolated from T2 cells lacking the gene region for $\beta 1i$ and $\beta 5i$ and by an immunoproteasome (T2.27) isolated from T2 cells transfected with $\beta 1i$ and $\beta 5i$ and characterized by an enhanced incorporation of the endogenous $\beta 2i$, the third immuno subunit.

4.4.1.1 Peptide synthesis

Peptides were synthesized by solid-phase methods on an automated Pioneer Peptide Synthesis System (PerSeptive Biosystems) using Fmoc chemistry. Peptide purity was confirmed by reversed-phase HPLC and MS. Syntheses were performed by Dr. P. Henklein (Charité, Berlin).

4.4.1.2 Purification and analysis of 20S proteasome complexes

Proteasomes were purified from the human lymphoblastic cell line T2 and T2 cells transfected with the proteasomal subunits LMP2 and LMP7 (T2.27) as previously described (Kuckelkorn, et al., 1995).

4.4.1.3 Peptide digestion assays.

20 µg of the pp89-25mer or LLO-27mer oligopeptide and 3.3 µg of purified proteasomes were incubated in 1 mL of assay buffer (20 mM Hepes / pH 7.8, 2 mM Mg(CH₃COO)₂, 1mM dithiothreitol) at 37°C for 0, 0.5, 1, 1.5, 2, 3, 4, 6, 8 and 10 h, stopped by adding 0.1 vol 1% trifluoroacetic acid then frozen at -20°C. For each time point, two samples of 30 µL digest were used independently for HPLC-MS-analysis. The resulting MS signal intensities were averaged. The experiments were carried out in duplicate.

4.4.1.4 HPLC-MS analysis

Samples (proteasomal digests and dissolved peptides for the calibrations curves) were separated by reversed-phase chromatography on a µRPC C2/C18 SC 2.1/10 column (Pharmacia Biotech) by linear gradient elution (eluent A, 0.05% trifluoroacetic acid in water; eluent B, 0.045% trifluoroacetic acid in 70% acetonitrile; flow rate, 73 µL/min). Analyses were performed online with an ion trap mass spectrometer (LCQ, Thermo-Finnigan) equipped with an electrospray ion source. As internal standard the peptide 9GPS (YPHF_MP_TN_LG_PS) was added to each sample. For calculation of the peak area the most intensive ion signal of the peptides was used.

4.4.2 Comparing theoretical and experimentally derived fragment amounts

4.4.2.1 Calibration curves

For calibration of the peptide amount with the MS signal, dilution series were prepared from stock solutions of 12 individual peptides found in the pp89-25mer. The peptide amounts were varied in a broad range between 1 and 500 pmol. Three types of dilution series were analyzed: In a first series of experiments, MS-signals were recorded for each individual peptide under isolated

conditions, i.e. without presence of other peptides. In order to assess the impact of collective effects, the MS-signals for peptide mixtures were also recorded in two further dilution series experiments, one with equimolar mixtures of 12 peptides, the other one with different molar ratios for three groups of peptides (group A : group B : group C = 10 : 5 : 1, see the last column in Figure 24 for the group assignments).

Peptide	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	group
1-25	R	L	M	Y	D	M	Y	P	H	F	M	P	T	N	L	G	P	S	E	K	R	V	W	M	S	A
1-24	[Bar]																								B	
5-25	[Bar]				[Bar]																				C	
5-22	[Bar]				[Bar]																		-			
8-25	[Bar]							[Bar]																		C
1-15	[Bar]															[Bar]										A
8-22	[Bar]							[Bar]															-			
4-15	[Bar]				[Bar]											[Bar]										-
5-15	[Bar]					[Bar]										[Bar]										B
16-25	[Bar]																[Bar]									A
7-15	[Bar]						[Bar]									[Bar]										C
16-24	[Bar]																[Bar]								B	
8-15	[Bar]							[Bar]								[Bar]										B
16-22	[Bar]																[Bar]									-
1-7	[Bar]						[Bar]																			B
5-10	[Bar]				[Bar]						[Bar]															-
1-5	[Bar]					[Bar]																				-
11-15	[Bar]										[Bar]															-
1-4	[Bar]				[Bar]																					B
12-15	[Bar]												[Bar]													-
1-3	[Bar]			[Bar]																						-
23-25	[Bar]																						[Bar]			-

Figure 24: Fragments of the pp89-25-mer

List of all fragments of the pp89-25mer detected in the proteasome digest. The second row contains the original 25-mer substrate, with those residues in bold print that are used as cleavage sites. For 12 peptides calibration curves were monitored, whereby the capital letter in the last column indicates the molar ratio with which the peptide was tested in a peptide mixture.

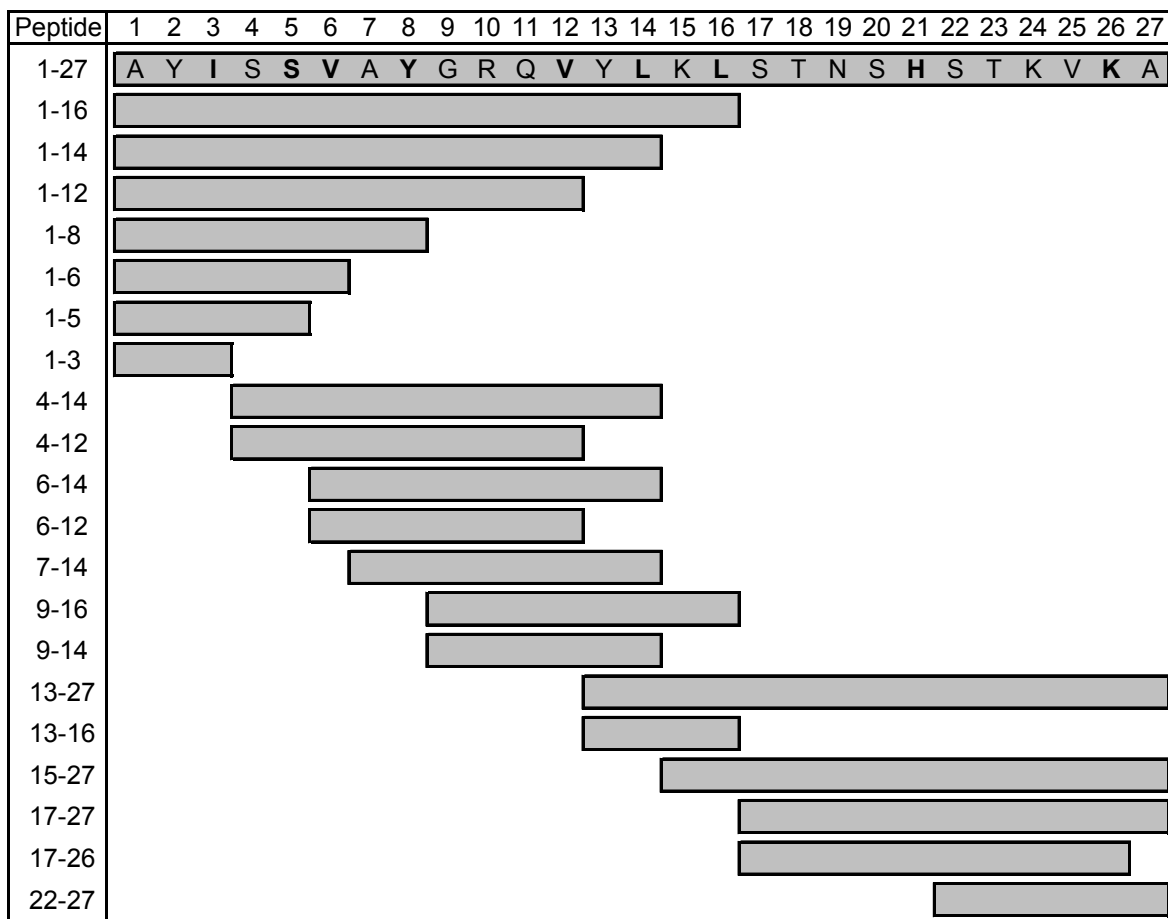


Figure 25: Fragments of the LLO-27mer

List of all fragments of the LLO-27mer detected in the proteasome digests. The second row contains the original substrate, with those residues in bold print that are used as cleavage sites.

A typical calibration curve obtained in this series of experiments is depicted in Figure 26. The average relative deviations of the recorded signals from the mean are listed in Table 7 for the set of 12 peptides at the various amounts tested. The major source of these deviations are systematic differences between the calibration curves recorded either under isolated conditions or in peptide mixtures. Compared with these systematic deviations the variations of MS-signals between repeat measurements carried out under identical conditions are small.

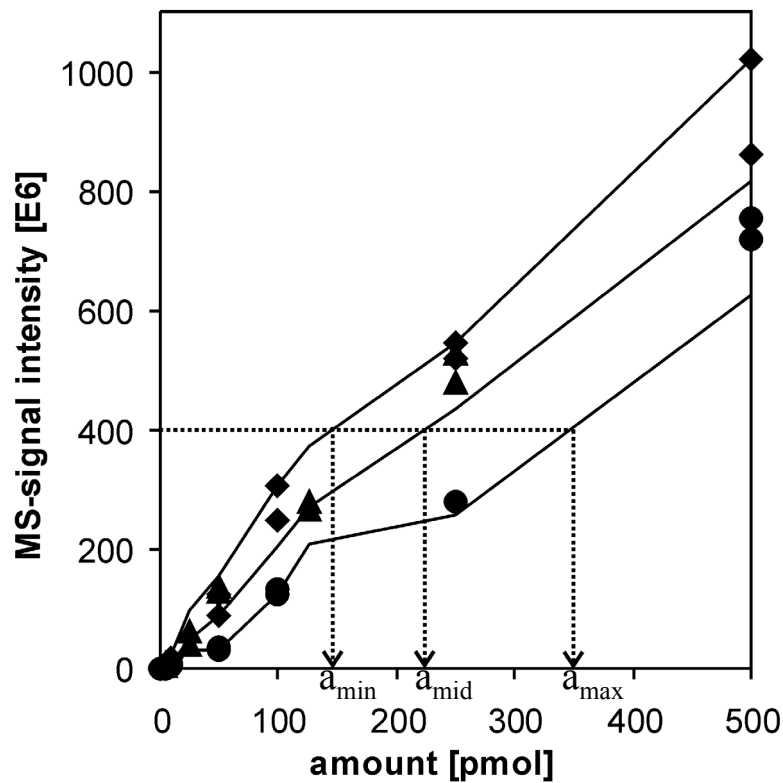


Figure 26: Calibration curve for the 11mer peptide 5-DMYPHFMPPTNL-15 contained in the pp89-25mer.

Three types of calibration curves were recorded, differing in the number and amount of peptides that were simultaneously present in one sample: ● only the 5-15 peptide present, ◆ all 12 analyzed peptides simultaneously present at the same amount, ▲ all 12 analyzed peptides simultaneously present but at different amounts. The solid lines interpolate linearly between the minimum, mean and maximum values recorded. In cases where the minimum or maximum recorded signal deviated from the mean by less than the average values given in Table 7, these average values were taken instead (see, for example, the minimum signal value at 500 pmol). The dotted lines indicate how a fixed MS-signal can be translated into an amount range. In this example, the MS-signal of 400×10^6 corresponds to a minimum amount of $a_{\min}=140$ pmol, mean amount of $a_{\text{mid}}=220$ pmol and a maximum amount of $a_{\max}=345$ pmol.

Table 7: Amount dependent signal deviations monitored in the calibration curves

Peptide Amount [pmol]	median relative difference max - mean	median relative difference min - mean
500	20%	-24%
250	26%	-40%
125	38%	-43%
100	40%	-43%
50	73%	-64%
25	93%	-66%
10	104%	-68%
5	143%	-80%
1	176%	-83%

Increasing the peptide amount by about two orders of magnitude from 1 pmol to 500 pmol, the lower and upper boundary for signal variations around the mean value decrease by about one order of magnitude to a level of about 20%. For low peptide amounts, the relative signal variations are very large. While all peptides produced a quantifiable MS-signal when added at an amount of 10 pmol or higher, no quantifiable MS-signal was detected in 6 out of 72 experiments at a peptide amount of 5 pmol. Diminishing the peptide amount further down to 1pmol, the number of experiments with unsuccessful peptide recovery increases to 15 out of 72. It was concluded that - under these experimental conditions - peptide amounts below 5 pmol do not necessarily produce quantifiable MS-signals.

4.4.2.2 *Assessment of peptide amounts from MS signals using the mass balance method*

For all peptides detected in the digests, time-dependent amount values were obtained from the measured MS-signals by applying the mass balance method described in section 4.3.1. Optimal values for the two control parameters λ and v_0 in (16) and (17) were determined in the following way: λ was continually increased starting with $\lambda=0$ until the maximum difference between the v_i 's was within a factor of 5. This yielded a value of $\lambda=1$ for the pp89-25mer digest and of $\lambda=0.1$ for the LLO-27mer digest. v_0 was determined by taking the logarithmic mean of the experimental minimum and maximum values; this yielded $v_0=1$.

Figure 27 shows the theoretically determined values of the signal conversion coefficients for the pp89-25mer digest in comparison to the experimental values assessed by means of calibration curves. Both methods are in good agreement: All calculated coefficients fall into the expected range, except for the peptide 1-4 which lies slightly above the experimental data.

4.4.3 Fitting the kinetic model to the experimental data

4.4.3.1 *Comparison of experimental and theoretical time-dependent amount profiles*

For both substrates, the time-courses of MS-signals were translated into time-courses of peptide amounts using the signal conversion coefficients calculated by means of the mass balance method. Fitting of the kinetic model to the time-dependent amount profiles was performed as described in section 4.3.2. As the partition probabilities P_m decline rapidly with increasing number of cleavages, only partitions including up to four cleavages for the pp89 25-mer and up to six cleavages for LLO 27-mer were considered in the calculation of the transition rates k_{ij} and K_i (cf. equations 20 and 21) in order to save computation time.

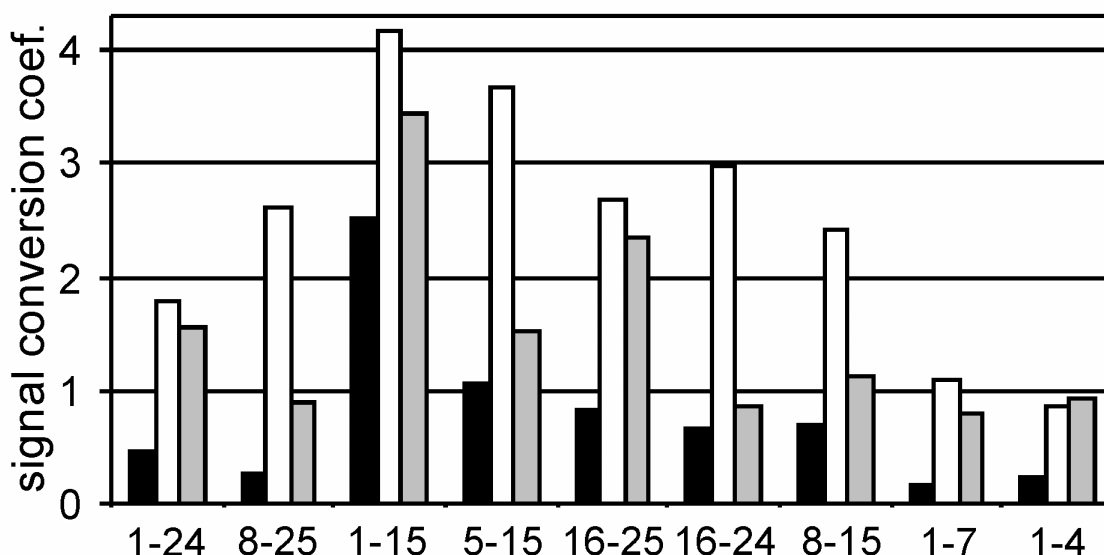


Figure 27: Comparison of experimentally and theoretically determined signal conversion coefficients for fragments derived from the pp89-25mer

Gray columns represent the signal conversion coefficients (cf. equation 12) obtained using the mass balance method. Black and white columns represent the experimental minimum and maximum for the coefficients. These boundaries were determined by picking the maximum signal of a given peptide measured in the time series of digest experiments and using the calibration curve to translate this signal into a minimum (a_{\min}) and maximum (a_{\max}) amount value (See arrows in Figure 26). Dividing the signal by a_{\max} and a_{\min} determines the experimental range for the value of the signal conversion coefficient.

Figure 28 depicts measured and calculated time-dependent amount profiles for the pp89 25-mer substrate and the 9 peptides generated with the highest abundance. The expected range of the experimental values as defined by the boundaries a_{\min} and a_{\max} is indicated by the gray shaded area. For both types of proteasomes, the calculated amount profiles of almost all peptides fall into the range expected from the experiment. This also holds true for the other 12 peptides found in this experiment (not shown). The calculated amount of all 31 hypothetical peptides predicted by the model but not detected in the experiment was below 2 pmol, i.e. remained below the reliable experimental quantification threshold.

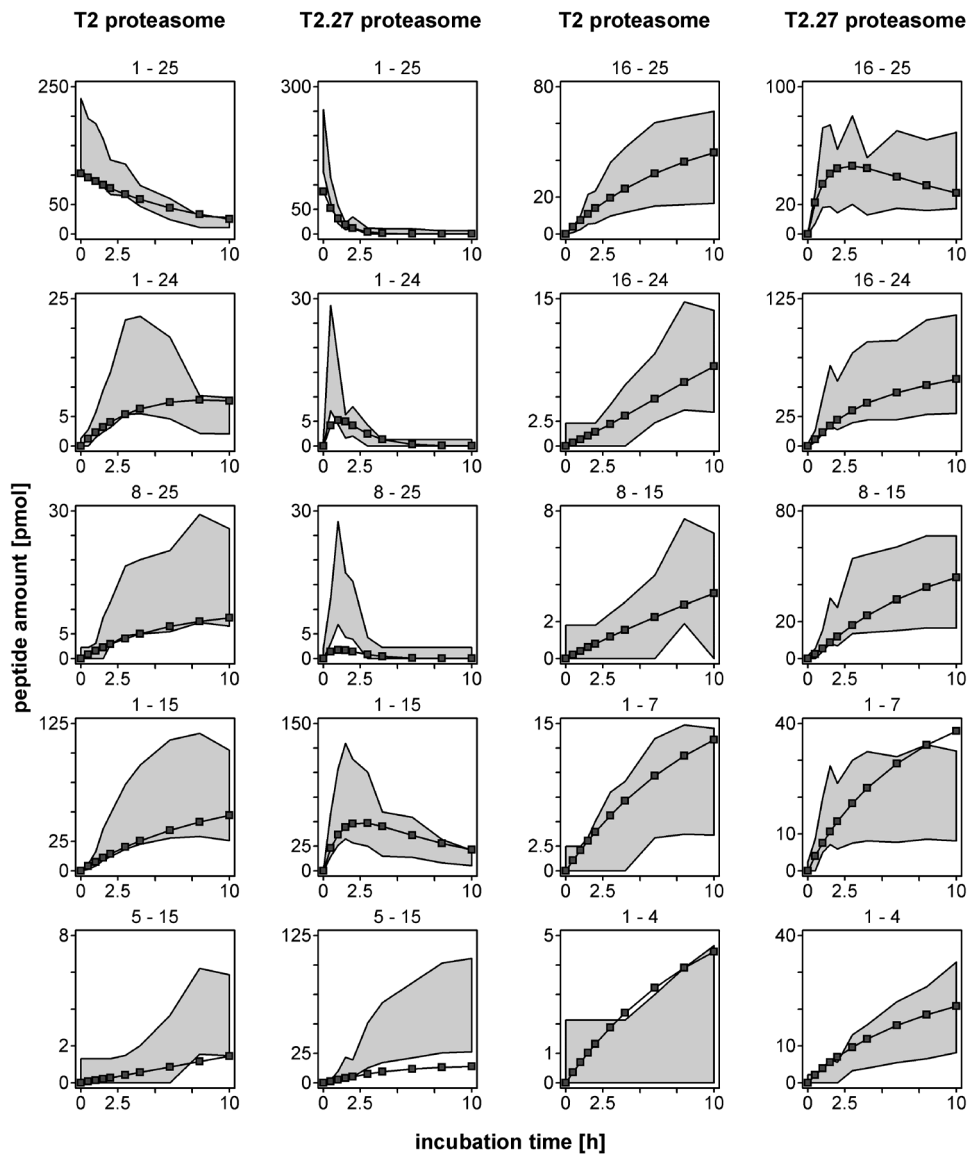


Figure 28: Time courses of peptide amounts for the pp89-25mer digests

Connected boxes represent the theoretical amount values predicted by the model. The shaded areas indicate the amount ranges determined from the MS-signals in the proteasome digests using the mass balance method. The caption of each graph names the position of the peptide in the sequence of the 25mer substrate. The x-axis indicates the incubation time in hours, and the y-axis the amount in units of pmol. Note the different scaling of the amount axis for the various peptides.

A similarly good correspondence between experiment and model was obtained for the LLO-27mer: Nearly all of the calculated time-dependent amount profiles for the 21 peptides detected in the digest (cf. Figure 25) were within the range of experimental uncertainty defined by the boundaries a_{\min} and a_{\max} , and the calculated maximal amount of all 26 hypothetical fragments was below the experimental quantification threshold of 2 pmol. For both substrates, the residuals between simulated and experimental time courses as measured by (22) are evenly spread across the entire time course, indicating random deviations between theoretical and experimental results.

Since the number of adjustable model parameters is small (pp89-25mer: 13 parameters / LLO-27mer: 12 parameters) compared with the number of data points (pp89-25mer: 220 experimentally observed data points + 310 hypothetical data points / LLO-25mer: 210 experimentally observed data points + 260 hypothetical data points), the good agreement between simulations and experimental data can be taken as a strong indication for the reliability of the model.

4.4.3.2 Assessing the variability of model parameters with a jack-knife procedure

To assess whether the numerical values for any model parameter differ significantly between the two types of proteasomes, it is necessary to relate the difference of the parameter values to their standard deviations. The standard deviation of a model parameter characterizes the expected range of its variability when determined from a set of independent repeat experiments. An alternative to carrying out new experiments is the so-called jack-knife procedure, which mimics the possible outcome of future experiments by replacing the original data base with a computer-generated artificial dataset. Such a jack-knife procedure was applied by omitting the measurements at 4 consecutive time points from the original dataset. Repeating this procedure and fitting the kinetic model to each dataset yields a collection of parameter estimates from which standard deviations can be assessed. The estimated numerical values for the model parameters and their jack-knife standard deviations are listed in Table 8 and Table 9, and are graphically displayed in Figure 29 and Figure 30. Note that the cleavage probabilities obtained in different fits cannot be directly compared because the model can provide equally good fits with either a high maximal procession rate r_{\max} combined with a low average level of the cleavage

probabilities p_i or, alternatively, with low r_{\max} and high p_i . Hence, to make cleavage probabilities comparable, they have to be related to the probability P_0 that no cleavage is made during procession of the substrate. The numerical estimates for the model parameters turn out to be fairly insensitive to large variations of the data base as produced in the jack-knife analysis.

Table 8: Estimated model parameters for the pp89-25mer digests

Parameter	T2		T2.27	
	mean	+/-	mean	+/-
$p_3 / (1-P_0)$	0.019	0.006	0.003	0.002
$p_4 / (1-P_0)$	0.054	0.007	0.053	0.006
$p_5 / (1-P_0)$	0.001	0.001	0.002	0.000
$p_6 / (1-P_0)$	0.015	0.007	0.004	0.003
$p_7 / (1-P_0)$	0.163	0.019	0.105	0.041
$p_{10} / (1-P_0)$	0.004	0.004	0.003	0.001
$p_{11} / (1-P_0)$	0.012	0.004	0.006	0.002
$p_{15} / (1-P_0)$	0.622	0.041	0.710	0.062
$p_{22} / (1-P_0)$	0.027	0.006	0.009	0.003
$p_{24} / (1-P_0)$	0.220	0.026	0.349	0.079
$p_0 = \prod (1 - p_i)$	0.65	0.07	0.43	0.13
r_{\max}	0.59	0.23	1.96	0.33
L_0	22.8	1.1	13.0	1.3
c	11.0	3.0	11.1	9.5

Table 9: Estimated model parameters for the LLO-27mer digests

Parameter	T2		T2.27	
	mean	+/-	mean	+/-
$p_3 / (1-P_0)$	0.05	0.01	0.23	0.03
$p_5 / (1-P_0)$	0.07	0.02	0.22	0.03
$p_6 / (1-P_0)$	0.16	0.01	0.14	0.03
$p_8 / (1-P_0)$	0.51	0.04	0.14	0.02
$p_{12} / (1-P_0)$	0.29	0.05	0.37	0.04
$p_{14} / (1-P_0)$	0.49	0.04	0.49	0.05
$p_{16} / (1-P_0)$	0.37	0.05	0.43	0.08
$p_{21} / (1-P_0)$	0.15	0.02	0.08	0.02
$p_{26} / (1-P_0)$	0.08	0.04	0.08	0.02
$p_0 = \prod (1 - p_i)$	0.11	0.04	0.12	0.09
r_{\max}	0.07	0.01	0.22	0.04
L_0	17.1	2.5	12.6	0.0
c	31.7	46.7	100.0	0.0

Inspection of the graphs in Figure 29 and Figure 30 reveals that for both substrates tested, the immunoproteasome possesses a significantly higher procession rate combined with a greater preference to process shorter peptides compared with the constitutive proteasome. The effects on cleavage probabilities associated with switching from the constitutive proteasome to the immunoproteasome are diverse for the two substrates. For the pp89-25mer (Figure 29, Table 8), there are no significant changes of cleavage probabilities at all. For the LLO-27mer (Figure 30, Table 9), the result of this analysis is a significant change of the cleavage probability at 4 of the 9

detectable cleavage sites: an increase of cleavage probabilities at the residues I₃ and S₅ and a decrease at Y₈ and H₂₁. These results are discussed in more detail in section 4.5.

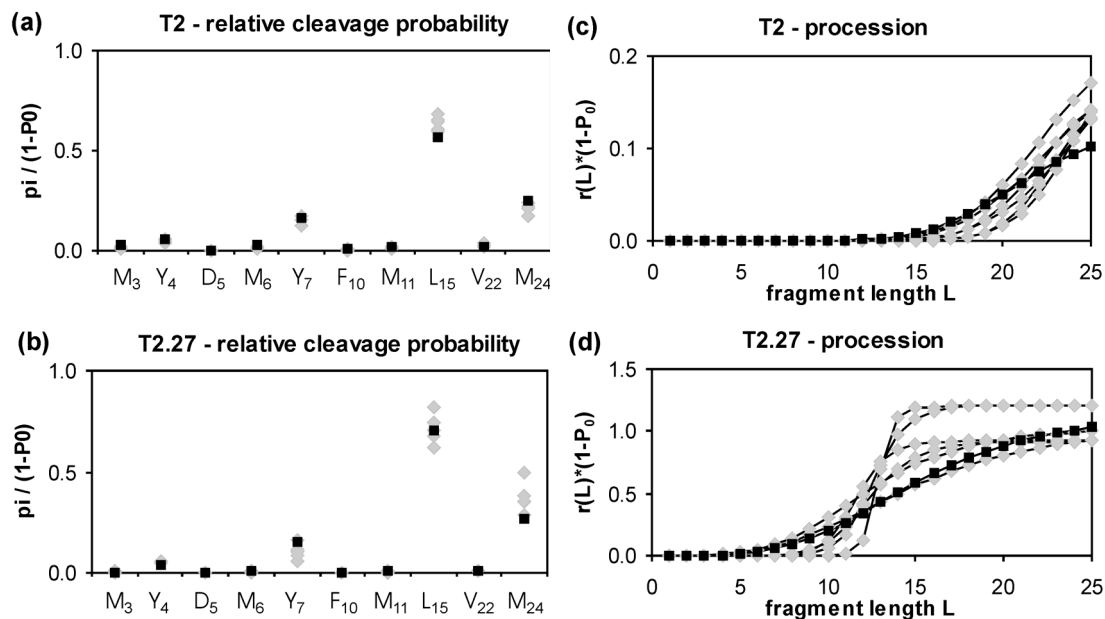


Figure 29: Variability of model parameters for the pp89-25mer digests assessed by a jack-knife procedure: Experimental peptide amounts determined by the mass balance method

(a)-(b): Cleavage probabilities normalized to $(1 - P_0)$. Each Figure shows the relative cleavage probabilities for 7 different fits: The black boxes indicate the cleavage probabilities obtained by fitting the model to the entire set of experimental data. The gray diamonds refer to parameter values obtained by fitting the model to reduced datasets where four 4 consecutive time points were left out from the experimental data. (c)-(d): The procession rate was calculated from the model parameters r_{\max} , L and c according to equation (19) and then multiplied by $(1 - P_0)$. Each Figure shows the results of 7 different fits detailed above.

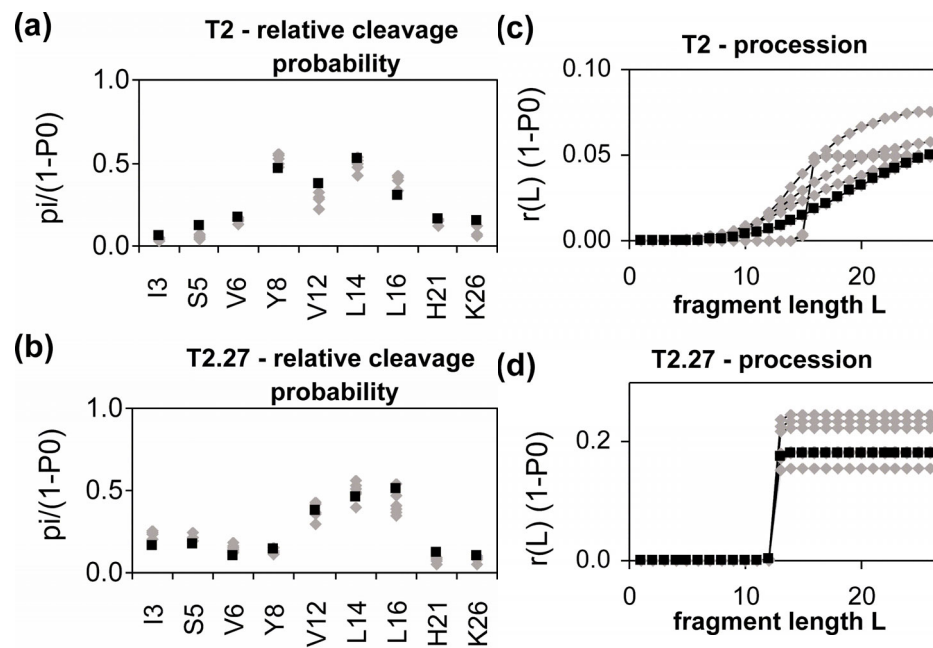


Figure 30: Variability of model parameters for the LLO-27mer digests assessed by a jack-knife procedure: Experimental peptide amounts determined by the mass balance method

The amount of all relevant peptides identified in the digests of the LLO-27mer were derived from MS-signals by using the mass balance method. Variability of model parameters was assessed by repeated model fitting to truncated datasets (see legend of Figure 29).

4.4.3.3 *Checking the equivalence of model computations based on either the mass balance method or experimental calibration curves*

In a second series of computations, estimation of the model parameters for the pp89-25mer digests was performed on the basis of time-dependent amount profiles which have been constructed by using the available calibration curves instead of using the novel mass balance method for the 12 peptides indicated in the last column of Figure 24. Again, robustness of the numerical estimates was assessed by a jack-knife procedure as outlined above (Figure 31). Importantly, mean values and variances of all model parameters do not significantly deviate from previous values obtained by fitting the model to time-dependent amount profiles derived

from MS signals by the mass balance method. This result underlines the finding in Figure 27 that the mass balance method enables reliable conversions of MS-signals into peptide amounts.

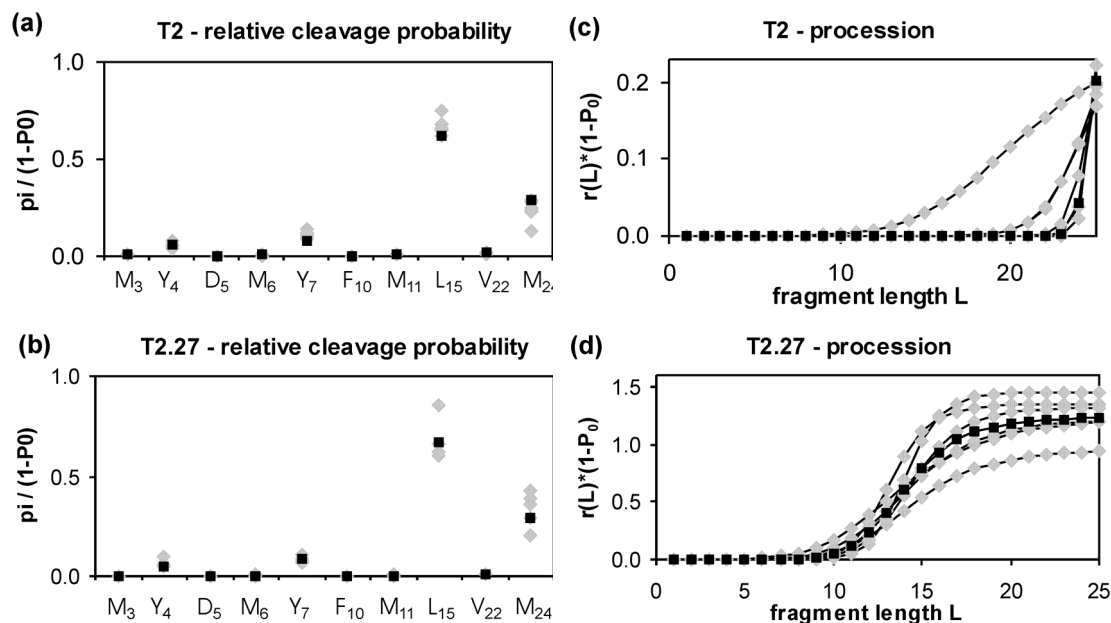


Figure 31: Variability of model parameters for the pp89-25mer digests assessed by a jack-knife procedure: Experimental peptide amounts determined by using calibration curves

Calibration curves were used to transform the MS-signals from the proteasome digests into peptide amounts, if available (see Figure 24). Using this dataset, the same seven fits as described in Figure 29 have been made to assess the variability of the model parameters. See the legend of Figure 29 for further details.

4.4.3.4 Adequate fitting of data requires a length-dependent procession rate

It was tested if a simpler version of the procession rate in equation 19 can produce similar quality time courses. To this end, constraints were imposed on the parameters L and c of the procession rate and the achieved quality of the fit was compared with the unconstrained one. The unconstrained fit yields a total residual distance of $\Delta=249$ ($\Delta=458$) for the pp89-25mer (LLO-27mer) digests according to the distance measure (22), when summed over both types of proteasome, all peptides and all time points. Setting $L_0=0$, i.e. neglecting the length-dependence by equating the procession rate to r_{\max} for all peptides, the total residual distance amounts to

$\Delta=427$ ($\Delta=605$), i.e. the quality of the fit decreases significantly. At the other extreme, setting $L_0 = (\text{substrate length} - 0.5)$ and $c=1000$ such that only the initial substrate undergoes procession with rate r_{\max} while re-processing of proteolytic fragments is prevented, the total residual distance amounts to $\Delta=456$ ($\Delta=549$) again indicating a clear drop in fit quality. These findings demonstrate that length restrictions in the procession of shorter peptides are an essential feature of proteasomal cleavage.

4.5 Differences between constitutive- and immuno-proteasomal digests

Comparing the model parameters determined for the digests by T2.2 and T2.27 proteasomes gives information about differences between them. First of all, both types of proteasome have a remarkably similar cleavage pattern. For the pp89-25mer, there is no significant difference in the determined cleavage probabilities at any cleavage site. This is an unexpected finding considering the large differences between the time-dependent product patterns produced by the two proteasome species. However, the theoretical analysis of the data demonstrates that these differences can be well accounted for by changes in the overall procession rate: Compared with the constitutive proteasome, the immunoproteasome works faster and accepts shorter peptides for re-procession. Since the kinetic model does not explicitly relate the procession rate to the various elementary steps involved in a procession cycle, it cannot be decided whether the higher procession rate of the immunoproteasome is due to an accelerated uptake and release of peptides or/and to a general increase in the catalytic capacity of its active sites. The finding that the immunoproteasome possesses a higher turnover rate than its constitutive counterpart is in agreement with previous observations (Boes, et al., 1994; Cardozo and Kohanski, 1998; Kuckelkorn, et al., 1995).

Using the substrate LLO-27mer, the differences in the overall procession rate of the constitutive proteasome and the immunoproteasome are very similar to those obtained for the pp89-25mer. In addition, there are significant alterations of the cleavage probabilities at four cleavage sites which in a concerted fashion give rise to an enhanced production of the epitope (VAYGRQVYL) by the immunoproteasome.

In summary, the results obtained with two different oligomeric substrates show that the kinetic effects associated with replacement of the constitutive proteasome by the immunoproteasome can be subdivided into a non-specific enhancement of the overall procession rate and peptide-bond specific alterations of cleavage probabilities. Since the latter effects are clearly restricted to a few cleavage sites it seems not very likely that the exchange of the active-site subunits by their interferon-inducible counterparts leads to a general stimulation of the trypsin-like and chymotrypsin-like activities accompanied by a depression of the peptidylglutamyl-peptide-hydrolyzing activity as postulated in several previous studies (Aki, et al., 1994; Boes, et al., 1994; Cardozo and Kohanski, 1998; Gaczynska, et al., 1996; Gaczynska, et al., 1993; Kuckelkorn, et al., 1995; Toes, et al., 2001). In particular, lacking changes of the cleavage probabilities at the three leucine residues present in the two substrates tested is hardly compatible with the common view (Groettrup, et al., 2001) that the immunoproteasome possesses a generally increased inclination for cleavages after certain categories of P1 residues (hydrophobic, branched chain, positively charged).

Recently Toes et al. (Toes, et al., 2001) have compared the fragment patterns of denaturated enolase-1 (436 amino acids) generated by constitutive and immunoproteasome. Only about 25% of the peptides produced by the immunoproteasome were also found in constitutive proteasome digests. Such a diversity in the peptide pools generated by either proteasomes was not seen here. For both oligomeric substrates, the two peptide pools detected in the digest were identical for both types of proteasome. The various peptides differed only in their amount which to a large extent could be explained by differences in the overall procession rate. The obvious inconsistency of the results reported here with those of Toes et al. is remarkable and may have two reasons. First, it is conceivable that the mechanisms by which the 20S proteasome degrades a denaturated long protein substrate and a relatively short (25 or 27 residues long) oligopeptide differ in that threading of a 436 long peptide chain through the proteasome may pose additional constraints on the accessibility of the active sites. Second, a moderate (2-5 fold) variation of cleavage probabilities as found for some cleavage sites of the LLO-27mer may amplify to larger variations (4 - 25 fold) of respective peptide amounts. Given that the abundance of a considerable portion of peptides derived from a long substrate is close to the detection threshold, such variations in peptide amounts could result in an apparent 'loss' or 'appearance' of peptides.

It has to be emphasized that the model in its present form was established to describe the degradation kinetics of oligopeptides as typically used in *in vitro* digests. Extension of this approach to kinetic experiments with long substrates will certainly require modifications of some basic assumptions, e.g. concerning the monotonous increase of the procession rate with peptide size or the statistical independence of cleavage combinations.

4.6 Summary

Existing algorithms describing protein degradation by the proteasome deliver poor results when used to identify epitopes by their predicted C-terminal cleavage. This is believed to be the consequence of the lesser quality of experimental data available for training of these prediction algorithms. To tackle this problem, a novel protocol to interpret proteasomal digests was developed. This protocol addresses two problems: (1) How to quantify the amounts of peptides present in a digest when only MS data is available, and (2) how to extract cleavage rates from a digest in which fragments are re-processed.

The conversion of MS-signals into peptide amounts is realized using mass balance equations and assuming a linear correlation between peptide amounts and their MS-signals. The amounts calculated with this approach are in good agreement with those determined using calibration curves. Problem (2) is addressed by developing a kinetic model of proteasomal digests. By fitting this model to the amount profiles from experimental digests, numerical values for cleavage rates are obtained, which are free parameters of the model. Comparing these fitted model parameters for digests made by constitutive and immuno-proteasomes shows that the differences in observed peptide amounts profiles can to a large extent be explained by an enhanced procession speed of the immuno-proteasome.

5 Summary of main results and conclusions

In the last chapters, the three main agents in the MHC-I pathway were examined with the goal to develop tools to predict their function in the antigen processing pathway. For peptide binding to MHC-I, a new prediction algorithm was developed. It combines a matrix-based method (SMM), which describes the contributions of individual residues to binding, with pair coefficients, which describe pair-wise interactions between positions in a peptide. This approach outperformed several previously published prediction methods, and for the first time quantified the impact of interactions in a peptide. The superiority of this approach is believed to be the consequence of three main novel features: (1) the use of a regularization parameter, which prevents the pair coefficients and the matrix entries from overfitting the data. (2) the pair coefficients are determined by systematic investigation of differences between the matrix predictions and the experimental values. As the matrix method is already highly accurate on its own, this is a better starting point than trying to determine both position contributions and position interactions all at once. (3) the interactions under investigations are limited to those with a sufficient amount of consistent training data.

The distribution of the pair coefficient values showed that interactions between adjacent peptide positions are somewhat stronger than those farther apart. However, this trend was seen to a much lesser extent than expected, signifying that interactions are not limited to neighboring amino acids in direct contact, but can also play a role over longer distances, probably through the conformation of the peptide back-bone. Compared to the SMM matrix entries, the pair-coefficients are rather small. This explains why methods completely ignoring interactions can still make good predictions.

Peptide affinities to TAP are considered to be closely related to their transport efficiencies. Therefore, the SMM matrix description developed to analyze peptide binding to MHC-I could also be applied to predict affinities of a set of 9-meric peptides to TAP. The SMM predictions were significantly better than those of two scoring matrices determined directly from experiments. Pair coefficients were not introduced here, to allow for the combination of all matrices into a single consensus matrix, which made the best overall predictions.

Using the experimental knowledge, that binding of a peptide to TAP involves mainly its C-terminus and three N-terminal residues, a 9-mer scoring matrix can be employed to predict the affinities of peptides of any length by taking only these residues into account. This was demonstrated to give good predictions of TAP affinities for peptides of size 10 to 18. Being able to predict TAP affinities of peptides longer than 9 amino acids (the typical epitope length) is important because it has become clear that several MHC-I epitopes are generated by N-terminal trimming of precursor peptides that are likely to be transported into the ER by TAP. As the true *in vivo* precursors of an epitope are not known, a generalized TAP score was established which averages across the scores of all precursors up to a certain length.

The highest prediction quality with this TAP score was achieved when the contribution of the N-terminal residues were down-weighted. It was reasoned on the basis of simulations and of results from scoring for individual MHC-I alleles, that this down-weighting partially reflects co-evolution of TAP and the average MHC-I allele as to the preference for certain C-terminal residues, as well as the uncertainty which epitope precursors are present *in vivo*. With this scoring method, the influence of TAP was found to be a consistent, strong pressure on the selection of MHC-I epitopes for all alleles. Using predicted TAP transport efficiencies as a filter prior to prediction of MHC-I binding affinities, it was possible to further improve the already very high classification accuracy achieved using MHC-I affinity predictions alone.

Such a two-step prediction protocol failed when predictions of C-terminal proteasomal cleavages were used as the filter, i.e. relying on MHC-I affinity predictions alone gave better results than combining them with proteasomal cleavage predictions. This disappointing result is thought to be caused by the lack of a sufficiently large set of quantitative and consistent experimental data on cleavage rates, which are more difficult to measure and interpret than the affinity assays used to characterize peptide binding to TAP and MHC-I. Therefore, in the last chapter a new protocol for the evaluation of proteasomal digests was developed, which was applied to a series of experiments. The first problem addressed in this protocol is the quantification of data from MS experiments. As the signal strength detected for a peptide depends not only on its amount but also on its chemical properties, additional information is needed to quantify a signal, which usually requires extra measurements in the form of calibration curves. To avoid these additional measurements, a novel method based on mass-balance equations was introduced which demands

that the total amount of peptides having one sequence position in common has to be conserved throughout the digest. This allowed for reasonable estimations of the peptide amounts from MS-signals in a digest.

Based on this quantified data, the first kinetic model of the 20S proteasome was developed which is capable of providing a satisfactory quantitative description of the whole time course of product formation measured in an *in vitro* digest. As known from conventional enzyme kinetics, the minimum ingredients to establish an enzyme-kinetic model are (1) the maximum activity characterizing the catalytic step of the enzyme under ideal working conditions (e.g. substrate saturation) and (2) the affinity characterizing the strength of interaction between enzyme and substrate. These two essential parameters have been incorporated into the proteasome model in terms of the parameters processing rate and peptide-bond cleavage probability. The crucial advantage of this model-based approach consists in the possibility of differentiating between non-specific changes of the procession rate and peptide-bond specific kinetic effects. Changes of the procession rate alone may lead to an increase or decrease in the amount of a specific peptide only if re-processing takes place - a typical situation under *in vitro* conditions. *In vivo*, re-processing of fragments is unlikely in view of the enormous amount of peptidase activity present in the cytosol. In this case, changes of the procession rate alone would result in a uniform increase or decrease of all fragments without affecting the relative proportions between them. Hence, a preponderance or repression of specific peptides (e.g. epitopes) over others can only be achieved by changes of the cleavage probability.

The analyzed proteasomal digests provide evidence that immuno-proteasomes have a consistently higher procession speed than the constitutive-proteasomes. The cleavage patterns for both types of proteasomes are rather similar: All cleavage sites are found to be used by both types of proteasome, and only a minority show significant changes in their probability of usage. However, the analysis of just two rather short model substrates does not allow for the generalization of these results. Also, many more substrates will have to be analyzed to have a sufficiently large training base to establish a new prediction algorithm of proteasomal cleavage.

Characterizing each element in the MHC-I pathway and combining predictions of their function is not the only possible approach towards a sequence based prediction of epitopes. It is also

possible to identify sequence motifs common to all epitopes presented by a specific MHC-I allele, as realized in the SYFPEITHI database (Rammensee, et al., 1999), and use this information for prediction. This approach does not differentiate between the influences of the proteasome, TAP or MHC-I on epitope selection, but has been shown to work well in practice. However, it has a principal drawback, as epitope sequences do not contain the full information used in the presentation pathway: The epitope may originate from a group of N-terminal prolonged precursors, generated by the proteasome, partially trimmed by cytosolic peptidases, transported by TAP into the ER and then cut to final size. These steps preceding binding to the MHC-I receptor will depend on sequence motifs in the flanking regions up- and downstream of the epitope, which are neglected when considering only the epitope sequences themselves. Hence, developing prediction algorithms for each individual step of the MHC-I presentation pathway and combining them should in principal be the superior approach. However, high quality experimental data for each step and advanced prediction techniques are needed to rival the prediction quality currently achieved by SYFPEITHI. Unfortunately, the predictive quality of the two approaches cannot be compared here, as there is no independent blind set available. SYFPEITHI is trained on the data used as test sets for the combined predictions developed in this work. For a neutral comparison, a significantly large set of newly identified naturally presented epitopes would be needed, or an older version of the SYFPEITHI prediction algorithm would have to be used and tested on more recently included epitopes. As a consequence, no conclusions about which method is currently better at identifying epitopes can be drawn here.

When applying an epitope prediction protocol that is based on algorithms for several individual steps of the MHC-I presentation pathway, it is of utmost importance that each prediction algorithm is trained on data containing only information on that specific step. For example, prediction methods that are supposed to predict MHC-I binding, but have been trained on data including epitope presentation, implicitly predict the effects of TAP and the proteasome. A combination of such an 'impure' MHC-I binding prediction with a prediction of TAP transport or proteasomal cleavage thus bears the risk of overestimating the role of TAP or the proteasome in the presentation pathway

The improvements achieved when including TAP transport of precursors into epitope predictions are in the high sensitivity regime of the ROC curve (cf. Figure 14). It is often argued that high

sensitivity of epitope predictions is of less practical relevance than having high specificity, i.e. to end up with a short list of high probability epitope candidates for a given protein sequence is all important. This view is wrong for two reasons: First, from the medical point of view, it can be equally interesting to identify **all** possible epitopes within a given protein sequence, requiring high sensitivity of the predictions. Secondly, when combining predictions for several steps of the MHC-I pathway whereby predictions of one step are used as a filter for the input to the next, it is very important to throw out as few true epitopes in each step as possible. Such a multi-step prediction protocol automatically increases specificity from one step to the next.

6 Outlook

Summarizing the attempts in this work to improve epitope identification by combining different prediction steps, it has to be concluded that currently the only reliable strategy is to filter out those peptides exhibiting poor TAP transport scores, and use MHC-I binding affinity predictions to identify epitopes among the transportable peptides. This algorithm is implemented on the publicly available website www.mhc-pathway.net. The website currently contains binding predictions for five different MHC-I alleles, which will be updated as more data becomes available. It is also planned to include more TAP scoring matrices describing its transport preference in different species.

The next step along this line is to include the proteasome, for which currently no prediction algorithms with sufficiently high reliability are available. Accurate prediction of proteasomal fragments would lead to a further improvement of TAP transport predictions which then - instead of considering all precursors up to length L as equally probable - can be restricted to those precursors actually generated. Eventually, this should also make the down-weighting of N-terminal residues in the TAP predictions obsolete, because there would be no uncertainty as to which precursors are generated, and co-evolution between peptide specificities of the proteasome, TAP and MHC-I would be included in the model. To establish a consistent database for proteasomal cleavage prediction, it is planned to apply the described novel evaluation protocol to a series of proteasomal digests with a large number of substrates and different types of proteasomes (e.g. constitutive / immuno proteasome, with and without the 11S and 19S regulators). The extracted cleavage probabilities can then be analyzed using the SMM framework established here for sequence based prediction of peptide affinities to MHC-I and TAP.

In principal, the SMM + pair coefficients algorithm can be applied to all problems that require the prediction of a property associated with a sequence. However, the approach is likely to be successful only when the assumption of independent additive contributions of each sequence positions to the property under investigation is a decent approximation. To test the SMM + pair coefficient approach on problems completely different from affinity experiments, it was applied to the identification of cis-prolines from their sequence environment (Lorenzen, et al.) and the

prediction of contacts between residues of membrane helices and either residues of other helices or lipids in the membrane itself (Hildebrand, et al.), both with positive preliminary results. The application, refinement and testing of the limits of this approach is another goal for the future.

References

- Agatonovic-Kustrin, S. and Beresford, R. (2000): Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, *J Pharm Biomed Anal* 22 [5], pp. 717-27. URL:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10815714
- Aki, M.; Shimbara, N.; Takashina, M.; Akiyama, K.; Kagawa, S.; Tamura, T.; Tanahashi, N.; Yoshimura, T.; Tanaka, K. and Ichihara, A. (1994): Interferon-gamma induces different subunit organizations and functional diversity of proteasomes, *J Biochem (Tokyo)* 115 [2], pp. 257-69.
- Altuvia, Y.; Schueler, O. and Margalit, H. (1995): Ranking potential binding peptides to MHC molecules by a computational threading approach, *J Mol Biol* 249 [2], pp. 244-50. URL:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7540211
- Ayalon, O.; Hughes, E. A.; Cresswell, P.; Lee, J.; O'Donnell, L.; Pardi, R. and Bender, J. R. (1998): Induction of transporter associated with antigen processing by interferon gamma confers endothelial cell cytoprotection against natural killer-mediated lysis, *Proc Natl Acad Sci U S A* 95 [5], pp. 2435-40. URL:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9482903
- 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 [3], pp. 901-9.
- Bradley, Andrew P (1997): The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition* 30 [7], pp. 1145-1159.
- Breiman, L.; Friedman, J.H.; Olshen, R.A. and Stone, C. J. (1984): *Classification and Regression Trees*, CRC Press.
- Brusic, V.; van Endert, P.; Zeleznikow, J.; Daniel, S.; Hammer, J. and Petrovsky, N. (1999): A neural network model approach to the study of human TAP transporter, *In Silico Biol* 1 [2], pp. 109-21.
- 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 [27], pp. 16764-70. URL:
<http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jbc.org/cgi/content/full/273/27/16764>
<http://www.jbc.org/cgi/content/full/273/27/16764>

- Cohen, S. L. and Chait, B. T. (1996): Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins, *Anal Chem* 68 [1], pp. 31-7.
- Daniel, S.; Brusic, V.; Caillat-Zucman, S.; Petrovsky, N.; Harrison, L.; Riganelli, D.; Sinigaglia, F.; Gallazzi, F.; Hammer, J. and van Endert, P. M. (1998): Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules, *J Immunol* 161 [2], pp. 617-24.
- Daniel, S.; Caillat-Zucman, S.; Hammer, J.; Bach, J. F. and van Endert, P. M. (1997): Absence of functional relevance of human transporter associated with antigen processing polymorphism for peptide selection, *J Immunol* 159 [5], pp. 2350-7.
- Dolenc, I.; Seemuller, E. and Baumeister, W. (1998): Decelerated degradation of short peptides by the 20S proteasome, *FEBS Lett* 434 [3], pp. 357-61.
- Donnes, P. and Elofsson, A. (2002): Prediction of MHC class I binding peptides, using SVMHC, *BMC Bioinformatics* 3 [1], p. 25. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12225620
- Doytchinova, Irini A. ; Blythe, Martin J. and Flower, Darren R. (2002): Additive Method for the Prediction of Protein-Peptide Binding Affinity. Application to the MHC Class I Molecule HLA-A*0201, *Journal of Proteome Research* 1 [3], pp. 263-272.
- Emmerich, N. P.; Nussbaum, A. K.; Stevanovic, S.; Priemer, M.; Toes, R. E.; Rammensee, H. G. and Schild, H. (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 [28], pp. 21140-21148.
- 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 [6324], pp. 290-6. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1709722
- Frontline Systems, Inc. (1999): Solver DLL, V3.5
- Gaczynska, M.; Goldberg, A. L.; Tanaka, K.; Hendil, K. B. and Rock, K. L. (1996): Proteasome subunits X and Y alter peptidase activities in opposite ways to the interferon-gamma-induced subunits LMP2 and LMP7, *J Biol Chem* 271 [29], pp. 17275-80. URL: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jbc.org/cgi/content/full/271/29/17275>
- Gaczynska, M.; Rock, K. L. and Goldberg, A. L. (1993): Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes [see comments] [published erratum appears in *Nature* 1995 Mar 16;374(6519):290], *Nature* 365 [6443], pp. 264-7. URL: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3ffield=medline_uid&search=8396732
- Garboczi, D. N.; Utz, U.; Ghosh, P.; Seth, A.; Kim, J.; VanTienhoven, E. A.; Biddison, W. E. and Wiley, D. C. (1996): Assembly, specific binding, and crystallization of a human TCR-alpha-beta with an antigenic Tax peptide from human T lymphotropic

virus type 1 and the class I MHC molecule HLA-A2, *J Immunol* 157 [12], pp. 5403-10. URL:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8955188

- Goldberg, A. L.; Cascio, P.; Saric, T. and Rock, K. L. (2002): The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides, *Mol Immunol* 39 [3-4], pp. 147-64.
- Groettrup, M.; Khan, S.; Schwarz, K. and Schmidtke, G. (2001): Interferon-gamma inducible exchanges of 20S proteasome active site subunits: Why?, *Biochimie* 83 [3-4], pp. 367-72.
- Groll, M.; Bajorek, M.; Kohler, A.; Moroder, L.; Rubin, D. M.; Huber, R.; Glickman, M. H. and Finley, D. (2000): A gated channel into the proteasome core particle, *Nat Struct Biol* 7 [11], pp. 1062-7.
- Groll, M.; Ditzel, L.; Lowe, J.; Stock, D.; Bochtler, M.; Bartunik, H. D. and Huber, R. (1997): Structure of 20S proteasome from yeast at 2.4 Å resolution, *Nature* 386 [6624], pp. 463-71.
- Gubler, B.; Daniel, S.; Armandola, E. A.; Hammer, J.; Caillat-Zucman, S. and van Endert, P. M. (1998): Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP, *Mol Immunol* 35 [8], pp. 427-33.
- Gulukota, K.; Sidney, J.; Sette, A. and DeLisi, C. (1997): Two complementary methods for predicting peptides binding major histocompatibility complex molecules, *J Mol Biol* 267 [5], pp. 1258-67.
- Heinemeyer, W.; Kleinschmidt, J. A.; Saidowsky, J.; Escher, C. and Wolf, D. H. (1991): Proteinase yscE, the yeast proteasome/multicatalytic-multifunctional proteinase: mutants unravel its function in stress induced proteolysis and uncover its necessity for cell survival, *Embo J* 10 [3], pp. 555-62.
- Hildebrand, Peter; Peters, B.; Goede, A.; Preissner, R and Frommel, C. Prediction of Contacts in Membrane Helices, manuscript in preparation.
- Hilt, W. and Wolf, D. H. (1995): Proteasomes of the yeast *S. cerevisiae*: genes, structure and functions, *Mol Biol Rep* 21 [1], pp. 3-10.
- Hilt, W. and Wolf, D. H. (1996): Proteasomes: destruction as a programme, *Trends Biochem Sci* 21 [3], pp. 96-102. URL:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8882582
- Holzthutter, 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 [4], pp. 1251-65. URL:
<http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.idealibrary.com/links/citation/0022-2836/286/1251>

- Jameson, S. C. and Bevan, M. J. (1992): Dissection of major histocompatibility complex (MHC) and T cell receptor contact residues in a Kb-restricted ovalbumin peptide and an assessment of the predictive power of MHC-binding motifs, *Eur J Immunol* 22 [10], pp. 2663-7. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1396971
- Janek, K.; Wenschuh, H.; Bienert, M. and Krause, E. (2001): Phosphopeptide analysis by positive and negative ion matrix-assisted laser desorption/ionization mass spectrometry, *Rapid Commun Mass Spectrom* 15 [17], pp. 1593-9.
- Kesmir, C.; Nussbaum, A. K.; Schild, H.; Detours, V. and Brunak, S. (2002): Prediction of proteasome cleavage motifs by neural networks, *Protein Eng* 15 [4], pp. 287-96.
- Kessler, J. H.; Beekman, N. J.; Bres-Vloemans, S. A.; Verdijk, P.; van Veelen, P. A.; Kloosterman-Joosten, A. M.; Vissers, D. C.; ten Bosch, G. J.; Kester, M. G.; Sijts, A.; Wouter Drijfhout, J.; Ossendorp, F.; Offringa, R. and Melief, C. J. (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 [1], pp. 73-88.
- Khan, A. R.; Baker, B. M.; Ghosh, P.; Biddison, W. E. and Wiley, D. C. (2000): The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site, *J Immunol* 164 [12], pp. 6398-405. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10843695
- Kisselev, A. F.; Akopian, T. N.; Woo, K. M. and Goldberg, A. L. (1999): The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation, *J Biol Chem* 274 [6], pp. 3363-71. URL: <http://www.jbc.org/cgi/content/full/274/6/3363>
- Kloetzel, P. M. (2001): Antigen processing by the proteasome, *Nat Rev Mol Cell Biol* 2 [3], pp. 179-87.
- Kohler, A.; Bajorek, M.; Groll, M.; Moroder, L.; Rubin, D. M.; Huber, R.; Glickman, M. H. and Finley, D. (2001): The substrate translocation channel of the proteasome, *Biochimie* 83 [3-4], pp. 325-32.
- Krause, E.; Wenschuh, H. and Jungblut, P. R. (1999): The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins, *Anal Chem* 71 [19], pp. 4160-5.
- Kuckelkorn, U.; Frenz, S.; Kraft, R.; Kostka, S.; Groettrup, M. and Kloetzel, P. M. (1995): Incorporation of major histocompatibility complex--encoded subunits LMP2 and LMP7 changes the quality of the 20S proteasome polypeptide processing products independent of interferon-gamma, *Eur J Immunol* 25 [9], pp. 2605-11.

- Kuttler, C.; Nussbaum, A. K.; Dick, T. P.; Rammensee, H. G.; Schild, H. and Hadelers, K. P. (2000): An algorithm for the prediction of proteasomal cleavages, *J Mol Biol* 298 [3], pp. 417-29. URL: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.idealibrary.com/links/citation/0022-2836/298/417>
- Lankat-Buttgereit, B. and Tampe, R. (2002): The transporter associated with antigen processing: function and implications in human diseases, *Physiol Rev* 82 [1], pp. 187-204. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11773612
- Lauvau, G.; Kakimi, K.; Niedermann, G.; Ostankovitch, M.; Yotnda, P.; Firat, H.; Chisari, F. V. and van Endert, P. M. (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 [9], pp. 1227-40.
- Lorenzen, S.; Peters, B.; Frommel, C. and Preissner, R Identification of Cis-prolines from their sequence environment, manuscript in preparation.
- Mamitsuka, H. (1998): Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models, *Proteins* 33 [4], pp. 460-74. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9849933
- Milik, M.; Sauer, D.; Brunmark, A. P.; Yuan, L.; Vitiello, A.; Jackson, M. R.; Peterson, P. A.; Skolnick, J. and Glass, C. A. (1998): Application of an artificial neural network to predict specific class I MHC binding peptide sequences, *Nat Biotechnol* 16 [8], pp. 753-6.
- Momburg, F.; Roelse, J.; Hammerling, G. J. and Neefjes, J. J. (1994): Peptide size selection by the major histocompatibility complex-encoded peptide transporter, *J Exp Med* 179 [5], pp. 1613-23. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8163941
- Momburg, F.; Roelse, J.; Howard, J. C.; Butcher, G. W.; Hammerling, G. J. and Neefjes, J. J. (1994): Selectivity of MHC-encoded peptide transporters from human, mouse and rat, *Nature* 367 [6464], pp. 648-51.
- Neumann, L. and Tampe, R. (1999): Kinetic analysis of peptide binding to the TAP transport complex: evidence for structural rearrangements induced by substrate binding, *J Mol Biol* 294 [5], pp. 1203-13. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10600378
- Niedermann, G.; Grimm, R.; Geier, E.; Maurer, M.; Realini, C.; Gartmann, C.; Soll, J.; Omura, S.; Rechsteiner, M. C.; Baumeister, W. and Eichmann, K. (1997): Potential immunocompetence of proteolytic fragments produced by proteasomes before evolution of the vertebrate immune system, *J Exp Med* 186 [2], pp. 209-20. URL: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jem.org/cgi/content/full/186/2/209>

- Nijenhuis, M.; Schmitt, S.; Armandola, E. A.; Obst, R.; Brunner, J. and Hammerling, G. J. (1996): Identification of a contact region for peptide on the TAP1 chain of the transporter associated with antigen processing, *J Immunol* 156 [6], pp. 2186-95.
- Nussbaum, A. K.; Kuttler, C.; Hadelers, K. P.; Rammensee, H. G. and Schild, H. (2001): PAMProC: a prediction algorithm for proteasomal cleavages available on the WWW, *Immunogenetics* 53 [2], pp. 87-94.
- Olumee, Z.; Sadeghi, M.; Tang, X. and Vertes, A. (1995): Amino Acid Composition and Wavelength Effects in Matrix-assisted Laser Desorption / Ionization, *Rapid Communications in Mass Spectrometry* 9, pp. 744-752.
- 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 [1], pp. 163-75.
- 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 [3], pp. 847-62.
- Peters, Björn; Bulik, Sascha; Tampe, Robert; van Endert, Peter M. and Holzhutter, Hermann-Georg (2003): Identifying MHC-I epitopes by predicting the TAP transport efficiency of epitope precursors, to be published in *Journal of Immunology*.
- Peters, Björn; Tong, Weiwei; Sidney, John; Sette, Alessandro and Weng, Zhiping (2003): Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules, to be published in *Bioinformatics*.
- Press, William H.; Teukolsky, Saul A.; Vetterling, William T. and Flannery, Brian P. (1992): *Numerical Recipes in C*, 2. ed., Cambridge University Press.
- Rammensee, H.; Bachmann, J.; Emmerich, N. P.; Bachor, O. A. and Stevanovic, S. (1999): SYFPEITHI: database for MHC ligands and peptide motifs, *Immunogenetics* 50 [3-4], pp. 213-9.
- Rammensee, H. G.; Friede, T. and Stevanoviic, S. (1995): MHC ligands and peptide motifs: first listing, *Immunogenetics* 41 [4], pp. 178-228. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7890324
- Rock, K. L. and Goldberg, A. L. (1999): Degradation of cell proteins and the generation of MHC class I-presented peptides, *Annu Rev Immunol* 17, pp. 739-79.
- Rotzschke, O.; Falk, K.; Deres, K.; Schild, H.; Norda, M.; Metzger, J.; Jung, G. and Rammensee, H. G. (1990): Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells, *Nature* 348 [6298], pp. 252-4. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1700304
- Rotzschke, O.; Falk, K.; Stevanovic, S.; Jung, G.; Walden, P. and Rammensee, H. G. (1991): Exact prediction of a natural T cell epitope, *Eur J Immunol* 21 [11], pp. 2891-4. URL:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1718764

- Saric, T.; Chang, S. C.; Hattori, A.; York, I. A.; Markant, S.; Rock, K. L.; Tsujimoto, M. and Goldberg, A. L. (2002): An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides, *Nat Immunol* 3 [12], pp. 1169-76.
- Schubert, U.; Anton, L. C.; Gibbs, J.; Norbury, C. C.; Yewdell, J. W. and Bennink, J. R. (2000): Rapid degradation of a large fraction of newly synthesized proteins by proteasomes, *Nature* 404 [6779], pp. 770-4. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10783891
- Schueler-Furman, O.; Altuvia, Y.; Sette, A. and Margalit, H. (2000): Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles, *Protein Sci* 9 [9], pp. 1838-46.
- Schwarz, K.; de Giuli, R.; Schmidtke, G.; Kostka, S.; van den Broek, M.; Kim, K. B.; Crews, C. M.; Kraft, R. and Groettrup, M. (2000): The selective proteasome inhibitors lactacystin and epoxomicin can be used to either up- or down-regulate antigen presentation at nontoxic doses, *J Immunol* 164 [12], pp. 6147-57. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10843664
- Segal, M. R.; Cummings, M. P. and Hubbard, A. E. (2001): Relating amino acid sequence to phenotype: analysis of peptide-binding data, *Biometrics* 57 [2], pp. 632-42.
- 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 [6906], pp. 480-3.
- Shastri, N.; Schwab, S. and Serwold, T. (2002): Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules, *Annu Rev Immunol* 20, pp. 463-93.
- Toes, R. E.; Nussbaum, A. K.; Degermann, S.; Schirle, M.; Emmerich, N. P.; Kraft, M.; Laplace, C.; Zwinderman, A.; Dick, T. P.; Muller, J.; Schonfisch, B.; Schmid, C.; Fehling, H. J.; Stevanovic, S.; Rammensee, H. G. and Schild, H. (2001): Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products, *J Exp Med* 194 [1], pp. 1-12. URL: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jem.org/cgi/content/abstract/194/1/1>
- Udaka, K.; Wiesmuller, K. H.; Kienle, S.; Jung, G.; Tamamura, H.; Yamagishi, H.; Okumura, K.; Walden, P.; Suto, T. and Kawasaki, T. (2000): An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries, *Immunogenetics* 51 [10], pp. 816-28.

- Uebel, S.; Kraas, W.; Kienle, S.; Wiesmuller, K. H.; Jung, G. and Tampe, R. (1997): Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries, *Proc Natl Acad Sci U S A* 94 [17], pp. 8976-81.
- Uebel, S.; Meyer, T. H.; Kraas, W.; Kienle, S.; Jung, G.; Wiesmuller, K. H. and Tampe, R. (1995): Requirements for peptide binding to the human transporter associated with antigen processing revealed by peptide scans and complex peptide libraries, *J Biol Chem* 270 [31], pp. 18512-6. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7543103
- Uebel, S. and Tampe, R. (1999): Specificity of the proteasome and the TAP transporter, *Curr Opin Immunol* 11 [2], pp. 203-8.
- Valero, M.-L.; Giralt, E. and Andreu, D. (1998): An Evaluation of Some Structural Determinants for Peptide Desorption in MALDI-TOF Mass Spectrometry, Ramage, R. and Roger, E., *Peptides* 1996 pp. 855-856, Mayflower Scientific Ltd., Kingswinford, UK.
- van Endert, P. M.; Riganelli, D.; Greco, G.; Fleischhauer, K.; Sidney, J.; Sette, A. and Bach, J. F. (1995): The peptide-binding motif for the human transporter associated with antigen processing, *J Exp Med* 182 [6], pp. 1883-95.
- van Endert, P. M.; Tampe, R.; Meyer, T. H.; Tisch, R.; Bach, J. F. and McDevitt, H. O. (1994): A sequential model for peptide binding and transport by the transporters associated with antigen processing, *Immunity* 1 [6], pp. 491-500.
- Wang, Y.; Guttoh, D. S. and Androlewicz, M. J. (1998): Peptide transport assay for TAP function, *Methods Enzymol* 292, pp. 745-53.
- York, I. A.; Chang, S. C.; Saric, T.; Keys, J. A.; Favreau, J. M.; Goldberg, A. L. and Rock, K. L. (2002): The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues, *Nat Immunol* 3 [12], pp. 1177-84.
- Yu, K.; Petrovsky, N.; Schonbach, C.; Koh, J. Y. and Brusica, V. (2002): Methods for prediction of peptide binding to MHC molecules: a comparative study, *Mol Med* 8 [3], pp. 137-48. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12142545

Abbreviations

ANN	Artificial Neural Network
AUC	Area Under Curve
CART	Classification and Regression Trees
CTL	Cytotoxic T-lymphocytes
ER	Endoplasmic Reticulum
ERA(A)P	ER Aminopeptidase associated with Antigen Processing
HLA	human major histocompatibility complex
HPLC	High Performance Liquid Chromatography
MHC-I	Major Histocompatibiliy Complex class I
MS	Mass Spectrometry
PM	Polynomial Method
ROC	Receiver Operating Characteristics
SMM	Stabilized Matrix Method
TAP	Transporter associated with Antigen Processing
T2	cell line containing constitutive proteasomes
T2.27	cell line containing immuno proteasomes

Acknowledgments

I want to express my gratitude to all those who have given me valuable scientific and personal support during this work. Most of all, I want to thank Prof. Holzhütter for introducing me to the field of antigen processing, which keeps on fascinating me. During my entire work, he always assisted me when I his needed expert advice and many ideas presented here were the results of our fruitful discussions.

I was privileged to be the first exchange student between the Bioinformatics program at Boston University and the graduate program "Dynamics and Evolution of Cellular and Macromolecular Processes" in Berlin, and I would like to thank Prof. Heinrich as the head of the Berlin graduate program and Prof. DeLisi from Boston University for making this possible and financing my stay.

During this stay from January to April of 2002, most of the work presented in the chapter on MHC-I binding was made. This was a joint project with my supervisor in Boston, Prof. Weng, W. Tong and the experimentalists that supplied us with the affinity data J. Sidney and A. Sette.

The chapter on TAP transport owes a great deal to Sascha Bulik, who is writing his diploma thesis on this subject, and to Prof. van Endert and Prof. Tampe who supplied us with TAP affinity data, and helped us with the resulting manuscript. I also want to thank K. Udaka for supplying us with affinity matrix data for several mouse MHC-I alleles.

The experimental analysis of proteasomal digests was carried out by several members of the Institute of Biochemistry at the Charite. This cooperation would not have been possible without the steady interest and support of Prof. Kloetzel and U. Kuckelkorn. The analysis of the dependencies between MS-signals and peptide amounts is based on discussions and much more experimental data than presented here from K. Jannek and T. Ruppert. Discussions with Prof. Kloetzel, U. Kuckelkorn and Prof. Dahmann helped form my understanding of proteasomal digests. The isolation of proteasomes and the accomplishment of the peptide digests was carried out by I. Drung, peptides were synthesized by P. Henklein, and B. Brecht, K. Textoris-Taube and B. Strehl helped with the analysis of MS data.

I also want to thank everyone in the group of Professor Frömmel for providing a good working atmosphere, in which I found many people willing to help me as a trained physicist learn basic biochemistry. Special thanks go to K. Rother for supplying me with the figures in the introduction.

For their personal support, I want to thank my parents, my girlfriend and good friends for keeping me in contact with the outside world.

Without them and the many people that I have not mentioned here by name, this work would not have been possible.

Lebenslauf

Name: Björn Peters
Geburtsdatum: 18. Mai 1973
Geburtsort: Hamburg
Nationalität: deutsch
Familienstand: ledig

1979 - 1993 Grundschole / Gymnasium Buckhorn, Hamburg
1989 - 1990 Highschool Abschluss, Reidsville, GA, USA
1993 - 1994 Zivildienst in der Schwerstbehindertenbetreuung, SPV Hamburg
1994 - 2000 Studium der Physik an der Universität Hamburg. Diplomarbeit am Institut für Laserphysik bei Prof. Toschek, Thema: "Räumliches Lochbrennen in einem Vielmodenlaser"
Juli - Aug. 1997 IAESTE Austausch am Institut für Kristallstrukturanalyse in Novi Sad, Jugoslawien
Juni - Sep. 1998 Praktikum im zentralen Managementnachwuchsprogramm von Bertelsmann bei Super-RTL.
2000 - 2003 Promotionsstudium der Biophysik an der Humboldt Universität Berlin
Juni 2001 ESMTB Summer School on the Biology and Mathematics of Cells
Oct. 2001 Participant at the OECD Expert Consultation in Berlin for TG 432 (In Vitro 3T3 NRU phototoxicity test)
Jan. - Apr. 2002 Forschungsaufenthalt an der Boston Universtiy, Bioinformatics Program bei Prof. DeLisi / Prof. Weng

Berlin, den 23.Februar 2002

Björn Peters

Publikationen

Peters, B., Hünkemeier, J., Baev, V.M., & Khanin, Y.I.(2001). Low-frequency dynamics of a Nd-doped glass laser. *Phys Rev A*, **64**, 023816.

Peters, B. Holzhütter, H.G (2002). *In vitro* Phototoxicity Testing: Development and Validation of a New Concentration Response Analysis Software and Biostatistical Analyses Related to the Use of Various Prediction Models. *ATLA*, **30**, 415-432.

Peters, B., Janek, K., Kuckelkorn, U. & Holzhütter, H. G. (2002). Assessment of proteasomal cleavage probabilities from kinetic analysis of time-dependent product formation. *J Mol Biol*, **318**(3), 847-62.

Peters, B., Tong, W., Sidney, J., Sette, A. & Weng, Z. (2003). Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules. *to be published in Bioinformatics*.

Peters, B., Bulik, S., Tampe, R., van Endert, P. M. & Holzhütter, H.-G. (2003). Identifying MHC-I epitopes by predicting the TAP transport efficiency of epitope precursors. *to be published in the Journal of Immunology*.

Manuskripte in Vorbereitung:

Hildebrand, P., Peters, B., Goede, A., Preissner, R. & Frommel, C. Prediction of Contacts in Membrane Helices.

Lorenzen, S., Peters, B., Frommel, C. & Preissner, R. Identification of Cis-prolines from their sequence environment.

Berlin, den 23.Februar 2002

Björn Peters

Erklärung

Die vorliegende Promotion habe ich selbstständig und ohne unerlaubte Hilfe angefertigt.

Ich besitze keinen entsprechenden Doktorgrad und habe mich anderwärts nicht um einen Doktorgrad beworben.

Die dem Promotionsverfahren zugrunde liegende Promotionsordnung ist mir bekannt.

Berlin, den 23. Februar 2002

Björn Peters