Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular Immunology 43 (2006) 2037–2044
Molecular Immunology
www.elsevier.com/locate/molimm

Class I T-cell epitope prediction: Improvements using a combination of proteasome cleavage, TAP affinity, and MHC binding

Irini A. Doytchinova, Darren R. Flower*

The Edward Jenner Institute for Vaccine Research, Compton RG20 7NN, UK

Received 27 September 2005; received in revised form 3 November 2005; accepted 23 December 2005
Available online 9 March 2006

Abstract

Cleavage by the proteasome is responsible for generating the C terminus of T-cell epitopes. Modeling the process of proteasome cleavage as part of a multi-step algorithm for T-cell epitope prediction will reduce the number of non-binders and increase the overall accuracy of the predictive algorithm. Quantitative matrix-based models for prediction of the proteasome cleavage sites in a protein were developed using a training set of 489 naturally processed T-cell epitopes (nonamer peptides) associated with HLA-A and HLA-B molecules. The models were validated using an external test set of 227 T-cell epitopes. The performance of the models was good, identifying 76% of the C-termini correctly. The best model of proteasome cleavage was incorporated as the first step in a three-step algorithm for T-cell epitope prediction, where subsequent steps predicted TAP affinity and MHC binding using previously derived models.

Keywords: Proteasome cleavage; Epitope; Additive method

1. Introduction

Major histocompatibility complex (MHC) molecules are highly polymorphic cell surface molecules that present peptidic ligands on the cell surface for inspection by T lymphocytes. MHC class I ligands are derived primarily from endogenously expressed proteins (Shastri et al., 2002) and usually are 8–12 amino acids long, although there is now evidence that longer peptides (13–15 amino acids) are presented by class I MHCs in many—possibly all—vertebrates including human, mouse, cattle, and horse (Probst-Kepper et al., 2004; Green et al., 2004). MHC class II ligands have a more variable length of 9–25 amino acids and are derived from exogenous proteins. Cross-presentation—an endogenous route acting via class II and an exogenous one acting via class I—is also important and now increasingly well understood (Ackerman and Cresswell, 2004). The main processing pathway for MHC class I ligands involves degradation of proteins by the proteasome, followed by transport of the products by the transporter associated with antigen processing (TAP) to the endoplasmic reticulum (ER), where peptides are bound to MHC class I molecules, and then presented on the cell surface by MHCs. The MHC class II processing pathway involves protein degradation by the lysosomal–endosomal apparatus, binding of peptides to MHC molecules, and subsequent transport of the complex to the cell surface.

There is much evidence to suggest that the proteasome is responsible for generating the C terminus but not the N terminus of the final presented peptide (Craiu et al., 1997; Mo et al., 1999; Serwold and Shastri, 1999; Cascio et al., 2001). The proteasome is a multimeric proteinase with three active sites: a site with trypsin-like activity (cleavage after basic residues), one with chymotrypsin-like activity (cleavage after hydrophobic residues), and another with peptidylglutamyl-peptide hydrolytic activity (cleavage after acidic residues) (Orlowski and Michaud, 1989; Djaballah et al., 1992; Orlowski et al., 1993). In addition, in vertebrates there are three γ-interferon-inducible subunits that replace the constitutive subunits (Tanaka and Kasahara, 1998) and assemble the immunoproteasome. The immunoproteasomes have an altered hierarchy of proteosomal cleavage, enhancing cleavage after basic and hydrophobic residues and inhibiting cleavage after acidic residues (Van den Eynde and Morel, 2001; Toes et al., 2001). This is in accord with the amino

* Corresponding author. Tel.: +44 1635 577954; fax: +44 1635 577901/577908.
E-mail address: darren.flower@jenner.ac.uk (D.R. Flower).

0161-5890/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.molimm.2005.12.013

acid preferences for binding to MHC class I molecules at the C terminus (Rammensee et al., 1995).

Several computer algorithms are currently available for the prediction of proteasomal cleavage. FragPredict (http://www.mpiib-berlin.mpg.de/MAPPP/cleavage.html) was the first published algorithm and is based on compilation of peptide cleavage data (Holzhutter et al., 1999; Holzhutter and Kloetzel, 2000). PAProC (http://www.paproc.de) is based on an evolutionary algorithm (Kuttler et al., 2000; Nussbaum et al., 2001). NetChop (http://www.cbs.dtu.dk/services/NetChop) is an artificial neural network (ANN)-based algorithm (Keşmir et al., 2002). All these programs are freely available via the Internet. Several combined models for predicting T-cell epitopes, which use the sequential nature of the dominant class I peptide presentation pathway to restrict the combinatorial explosion inherent in the process of T-cell epitope presentation, have appeared recently (Daniel et al., 1998; Brusic et al., 1999; Petrovsky and Brusic, 2004; Tenzer et al., 2005).

In the present study, the additive method (Doytchinova et al., 2002) was applied to a set of naturally processed T-cell epitopes to derive models capable of predicting peptide cleavage by the proteasome. The additive method is based on the assumption that each substituent—amino acids in our case—makes an additive and constant contribution to the biological activity regardless of substituent variation in the rest of the molecule. Possible interactions between substituents can be accounted for by cross-terms. The method was applied initially to predict the affinity of peptides binding to HLA-A*0201 molecules (Doytchinova et al., 2002) and then was extended to another 10 human MHC class I (Guan et al., 2003a,b; Doytchinova and Flower, 2003a,b), 3 murine class I (Hattotuwagama et al., 2004), 3 human class II (Doytchinova and Flower, 2003a,b) and 8 murine class II proteins (unpublished data). These predictive models are accessible free online using the MHCPred server via http://www.jenner.ac.uk/MHCPred (Guan et al., 2003a,b).

The additive method is a general method, which can be applied to any peptide–protein interaction. We applied this method to design new high affinity peptides binding to HLA-A*0201, generating several superbinders (Doytchinova et al., 2004a). Recently, the additive method was applied to derive a model for TAP binding affinity prediction (Doytchinova et al., 2004b). The model for proteasome cleavage prediction developed in this study, together with previously derived models for TAP and MHC binding prediction, were combined into a three-step algorithm for T-cell epitope prediction.

2. Materials and methods

2.1. Peptides

In order to develop additive models for proteasome cleavage prediction, a training set of 489 naturally processed T-cell epitopes (nonamer peptides) associated with HLA-A and HLA-B molecules was collected from our in-house database AntiJen (http://www.jenner.ac.uk/AntiJen) (Blythe et al., 2002; McSparron et al., 2003; Toseland et al., 2005). A test set of 231 peptides, as used by Saxova et al. (2003) to compare the performance of the available methods for proteasome cleavage prediction, was employed in our study for external validation. All common T-cell epitopes between the two sets were first excluded from the training set.

N4 N3 N2 N1 **P9 P8 P7 P6 P5 P4 P3 P2 P1** | P1' P2' P3' P4' P5'	Cleavage
N4 N3 N2 N1 **P9 P8 P7 P6 P5 P4**	0
N3 N2 N1 **P9 P8 P7 P6 P5 P4 P3**	0
N2 N1 **P9 P8 P7 P6 P5 P4 P3 P2**	0
N1 **P9 P8 P7 P6 P5 P4 P3 P2 P1** |	0
P9 P8 P7 P6 P5 P4 P3 P2 P1 | P1'	0
P8 P7 P6 P5 P4 P3 P2 P1 | P1' P2'	0
P7 P6 P5 P4 P3 P2 P1 | P1' P2' P3'	0
P6 P5 P4 P3 P2 P1 | P1' P2' P3' P4'	0
P5 P4 P3 P2 P1 | P1' P2' P3' P4' P5'	1

Fig. 1. Cleavage site presentation. Peptide positions are given in bold. The positions before the N terminus are denoted as "Nn", while the positions after the C terminus—as "Pn". The vertical line shows the cleavage site. When the C terminus of the epitope is located in the middle (position P1 of the decamer), the peptide is considered as positive and takes a value of 1, i.e. cleavage site present. The rest of the overlapped peptides are considered as negative and take the value 0 (cleavage site not present).

The epitopes were presented together with four flanking amino acids before the N terminus and five flanking residues after the C terminus (Fig. 1). Further, these parent 18aa peptides were broken into a set of overlapping decamers. The peptide which contained the C terminus of the epitope at position P1 of the decamer was considered as a positive example, i.e. the cleavage site was present. The rest of the overlapped peptides in each set were considered as negative examples (cleavage site not present). Thus, the initial training set of 489 epitopes generated 4370 decamers, 489 peptides of which had positive cleavages and 3881 peptides were negative.

As the peptides from the test set had a length of 8–12 residues, some of the parent peptides had lengths different from 18 amino acids. Four parent peptides with no flanking residues after the C terminus were excluded from the test set, since it was not possible to locate the cleavage site. Thus, the final test set included 227 epitopes. They generated 2100 decamers: 227 peptides were positive and 1873 peptides were negative.

2.2. Proteins

A set of proteins, containing recently published T-cell epitopes (since 2000), was collected from the AntiJen database (Blythe et al., 2002; McSparron et al., 2003) and used to test a three-step algorithm for T-cell epitope prediction, based on the additive method. The set included 32 epitopes, belonging to 21 proteins (Bourgault Villada et al., 2000; Gonzalez et al., 2000; Dannull et al., 2000; Geluk et al., 2000; Altfeld et al., 2001a,b; Bownds et al., 2001; Sharma et al., 2001; Rudolf et al., 2001; Peter et al., 2001; Maranon et al., 2001; Buslepp et al., 2001; Caccamo et al., 2002; Duraiswamy et al., 2003; Drexler et al., 2003; Tanaka et al., 2003; Kather et al., 2003; Jaye et al., 2003;

Terajima et al., 2003). The epitopes were all restricted by the HLA-A*0201 molecule and, to facilitate the calculation, all the proteins consist of less than 500 residues.

2.3. Additive method for proteasome cleavage prediction

For a set of decamers, the additive method generates a matrix with 200 (20 × 10) columns and a number of rows equal to the number of peptides. A term in the matrix is 1 when a certain amino acid exists at a certain position, and 0 when it is absent. A column containing the dependent variable (cleavage versus non-cleavage) is added and the matrix is solved by partial least squares (PLS) (Wold, 1995), as implemented in SYBYL 6.9 (Tripos Inc., 2004). Models including different positions next to the cleavage site were generated in order to assess the importance of the flanking residues. The prediction rate of T-cell epitopes versus non-T-cell epitopes was measured using receiver operating characteristic (ROC) curves (Bradley, 1997). Two variables (sensitivity (true T-cell epitopes/total T-cell epitopes) and 1-specificity (false T-cell epitopes/total non-T-cell epitopes)) were calculated at different cutoffs. The area under the curve (A_{ROC}) is a quantitative measure of the predictive ability and varies from 0.5 for a random prediction to 1.0 for a perfect prediction. The predictive ability of the models was assessed by leave-one-out cross-validation (LOO-CV) on the training set and by external validation on the test set.

2.4. Variable selection

In order to reduce the number of variables, two methods for variable selection were used: a genetic algorithm (GA) and stepwise regression, as implemented in the MDL QSAR Package (MDL Information Systems Inc., 2004). GA allows one to select a subset of the most significant predictors using two evolutionary operations: random mutation and genetic recombination (crossover) (Leardi et al., 1992). The performance of the algorithm was calibrated in terms of the size of the initial population, choice of parents, types of crossover and mutation, and fitness function. The best results were achieved with an initial population of size 32, tournament selection, uniform crossover, one-point mutation and Friedman's lack-of-fit scoring function with value 3 (Friedman, 1990). The regression equation was generated on the basis of the selected variables by ordinary multiple regression. The stepwise regression was used in a forward mode. Final models were assessed by ROC-statistics on the cross-validated training set and the external test set.

3. Results

3.1. Additive models for proteasome cleavage prediction

Additive models, which included different positions before and after the cleavage site, were used to assess the importance of flanking amino acids around the C terminus for accurate proteasome cleavage prediction. Peptide positions were denoted as shown in Fig. 1. Cross-terms were omitted as previous studies indicated that the contributions of the positions next to the cleavage site are additive (Altuvia and Margalit, 2000). Models including only variables selected using a GA or stepwise regression were also created. LOO-CV was used for the training set. The ability of the models to discriminate T-cell epitopes from non-T-cell epitopes was assessed by ROC-statistics on the training and test sets. True positive and true negative T-cell epitopes were calculated at different cutoffs and the areas (A_{ROC}) under the sensitivity/1-specificity curves are given in Table 1. The overall performance of the models is very good (all $A_{ROC} > 0.740$). Unsurprisingly, the predictions made on the training set by LOO-CV are slightly better than those made on the test set. Models which include selected variables (GA and stepwise) give significantly better predictions for the training set than for the test set. Models containing amino acids from both sides of the C terminus predict better than models which only include flanking positions before the cleavage site. The best performing models for the test set are models P2P1P1′P2′, P1P1′ and the stepwise model (Fig. 2). These models are given in Table 2. The greater

Table 1
ROC-statistics of the additive models for proteasome cleavage prediction

Model	A_{ROC}	
	Training set (LOO-CV)	Test set
P5P4P3P2P1P1′P2′P3′P4′P5′	0.777	0.753
P4P3P2P1P1′P2′P3′P4′	0.771	0.741
P3P2P1P1′P2′P3′	0.779	0.748
P2P1P1′P2′	0.773	0.761
P1P1′	0.772	0.759
P2P1P1′	0.776	0.753
P5P4P3P2P1	0.764	0.741
P4P3P2P1	0.764	0.741
P3P2P1	0.767	0.744
P2P1	0.754	0.741
P1	0.765	0.751
GA	0.838	0.748
Stepwise	0.817	0.763
PAProC		0.589
FragPredict		0.605
NetChop 2.0		0.771

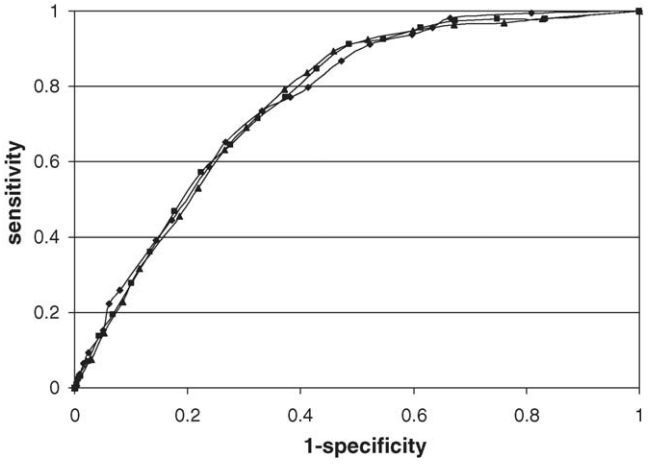

Fig. 2. ROC-statistics on the test set of the best performed additive models for proteasome cleavage prediction: P2P1P1′P2′ with $A_{ROC} = 0.761$ (squares), P1P1′ with $A_{ROC} = 0.759$ (triangles) and stepwise regression with $A_{ROC} = 0.763$ (circles).

Table 2
Additive models for proteasome cleavage prediction

	Model P1/P1′		Model P2P1/P1′P2′				Stepwise model									
	P1	P1′	P2	P1	P1′	P2′	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′	P5′
A	0.023	−0.012	0.003	0.025	−0.014	0.001					0.104					0.042
C	−0.031	0.053	0.017	−0.032	0.055	0.005	−0.076					0.129				
D	−0.064	0.011	−0.043	−0.065	0.012	−0.009				−0.059				0.053		0.052
E	−0.075	0.038	0.078	−0.075	0.037	0.029				0.102		0.056				
F	0.018	−0.029	−0.022	0.019	−0.030	−0.032	0.076				0.103			−0.050		
G	−0.096	0.035	0.007	−0.097	0.037	0.033			−0.045			0.050				
H	−0.049	0.009	0.049	−0.049	0.011	−0.026		0.063		0.075						
I	0.169	−0.046	−0.005	0.169	−0.045	−0.024					0.238					
K	0.061	0.031	0.010	0.060	0.030	0.006					0.160					
L	0.100	−0. 043	−0.015	0.096	−0.044	−0.004					0.192	−0.033		−0.031		−0.044
M	0.008	−0.020	0.008	0.008	−0.020	−0.013	−0.066				0.103					
N	−0.064	0.011	0.006	−0.065	0.012	0.021										
P	−0.109	−0.050	0.007	−0.110	−0.049	−0.009										
Q	−0.066	0.056	−0.013	−0.066	0.057	0.002						0.099				
R	0.039	0.022	−0.038	0.041	0.022	0.033			−0.060		0.141					
S	−0.092	−0.017	−0.007	−0.092	−0.015	0.013										0.053
T	−0.067	0.005	−0.001	−0.066	0.003	0.018										
Y	0.134	−0.039	−0.041	0.137	−0.042	−0.044					0.225		−0.054		−0.049	
W	−0.010	−0.011	0.010	−0.009	−0.010	−0.010					0.070			−0.109		
V	0.171	−0.006	−0.008	0.172	−0.007	0.009					0.242				−0.035	
Const.	0.101		0.104				0.013									

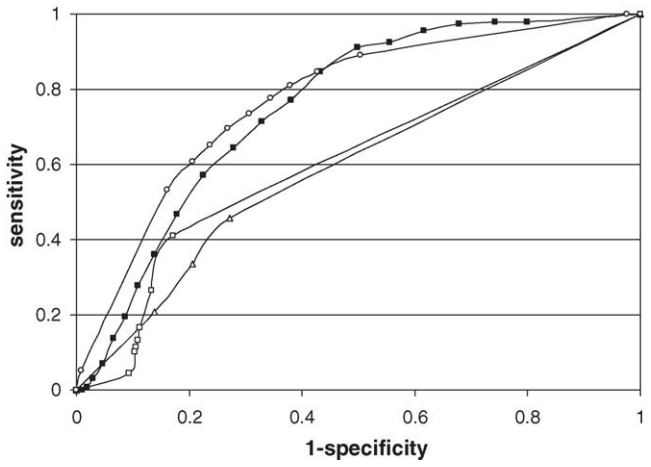

Fig. 3. ROC-statistics of the methods for proteasome cleavage prediction: additive model P2P1P1′P2′ with $A_{ROC} = 0.761$ (black squares), PAProc with $A_{ROC} = 0.589$ (white triangles), FragPred with $A_{ROC} = 0.605$ (white squares), and NetChop 2.0 with $A_{ROC} = 0.771$ (white circles).

the coefficient of an amino acid in the model, the greater is its effect on proteasome cleavage. Positive values increase the cleavage probability, negative ones reduce it.

3.2. Comparison with other algorithms for proteasome cleavage prediction

The test set used for external validation of the additive models was used for evaluation of the predictive algorithms PAProc (Kuttler et al., 2000; Nussbaum et al., 2001), FragPredict (Holzhutter et al., 1999; Holzhutter and Kloetzel, 2000) and NetChop 2.0 (Keşmir et al., 2002). Thresholds at 0, 1, 2 and 3 were selected for PaProc; 0, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 and >1.0 for FragPredict; 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 and >1.0 for NetChop. The sensitivity/1-specificity curves are given in Fig. 3. Good predictive ability close to that of the additive models was found for the NetChop algorithm ($A_{ROC} = 0.771$) and more moderate predictive ability for PAProc ($A_{ROC} = 0.589$) and FragPredict ($A_{ROC} = 0.605$).

3.3. Three-step algorithm for prediction of T-cell epitopes

A three-step algorithm based on the additive method was created to predict T-cell epitopes. Proteasome cleavage prediction was the first stage of the algorithm. Model P2P1P1′P2′ was used with a threshold value of 0.0, i.e. all negatively predicted cleavage sites were excluded. The next step of the algorithm was TAP affinity prediction. The previously derived additive model for TAP binding prediction (Doytchinova et al., 2004b) was used, with a cutoff set at 3.0 for TAP-independent alleles (HLA-A2, HLA-A23, HLA-B7 and HLA-B8) and a cutoff set at 5.0 for TAP-dependent ones (HLA-A1, HLA-A3, HLA-A11, HLA-A24, HLA-B15 and HLA-B27). The last step of the algorithm included the MHC binding affinity prediction models available on the MHCPred server.

The algorithm was tested on a set of 21 proteins, containing 32 recently published T-cell epitopes, presented by HLA-A*0201, and the predictions were compared with those made by other servers for T-cell epitopes predictions like SYFPEITHI (Rammensee et al., 1999) and BIMAS (Parker et al., 1994). SYFPEITHI (http://www.syfpeithi.de), a publicly available server, is based on the use of peptide binding motifs available in the literature to identify and score predicted epitopes. The threshold for SYFPEITHI was set to the top 20 scored peptides from each protein. Instructions for this server suggest that all naturally presented epitopes should be amongst the top-scoring 2% of peptides predicted, with an associated reliability of 80%. BIMAS (http://bimas.cit.nih.gov) estimates peptide binding affinities in terms of their half-life disassociation rates. The threshold for BIMAS was set to a half-life of 1 min. In terms of the "scores" returned by the additive algorithm, thresholds of 0.0 for the proteasome cleavage step, 3.0 for the TAP-binding step, and 5.3 for the MHC-binding step, were used in this study. Using the thresholds or cut-offs listed above, the additive algorithm predicted correctly all 32 T-cell epitopes, BIMAS 30 epitopes (94%), and SYFPEITHI 26 (81%).

4. Discussion

The identification of T-cell epitopes remains a critical step in the development of peptide-based vaccines (Luckey et al., 1998). The first step of such studies is usually in silico prediction of potential MHC binders from the sequence of a studied protein, followed by labor- , time- and resource-consuming experiments to verify the natural processing, presentation and T-cell recognition of the predicted peptides. As the veracity of initial in silico predictions improves, so subsequent "wet lab" work becomes faster, more efficient, and, ultimately, more successful. A wide range of computer-based algorithms have been developed to help predict T-cell epitopes (for reviews see Schirle et al., 2001; Golgberg et al., 2002; Flower, 2003).

The proteasome is the key enzyme responsible for the protein degradation in the cytosol and the generation of the C-termini of peptides presented by MHC class I molecules. Therefore, detailed knowledge on the specificity of protein degradation by the proteasome is crucial to T-cell epitope prediction. Previous studies on the positions flanking the cleavage site indicated that the amino acids in P3–P3′ have a strong influence on the cleavage site selection (Niedermann et al., 1996; Kuttler et al., 2000; Altuvia and Margalit, 2000), although cleavage-enhancement by Pro in P4 has been reported (Niedermann et al., 1996). Within the window of these six flanking amino acid residues, positions P1 and P1′ are the most significant. In accordance with these findings, the models derived in the present study show that P2, P1, P1′ and P2′ are the most influential positions. Cleavage appears after Val, Ile, Tyr, Leu, Lys, Arg, Ala, Phe and Met and/or before Gln, Cys, Glu, Gly, Lys, Arg, Asp, Asn, His and Thr. Arg and Lys make positive contributions at both positions, while Pro, Ser ad Trp contribute negatively at both. The preference for hydrophobic and basic amino acids at the C-termini in our models is compatible with previously reported results based on degradation experiments (Niedermann et al., 1996; Kuttler et al., 2000; Altuvia and Margalit, 2000). These preferences agree with the well-established requirements for binding to many MHC class

I alleles (Rammensee et al., 1995). Preferences for small (Cys, Gly), polar (Gln, Asn, Thr), positively (Lys, Arg, His) and negatively charged (Glu, Asp) amino acids at P1′ are also found in our models. These results, with the exception of negatively charged amino acids, are compatible with previously reported preferences of the P1′ position (Niedermann et al., 1996; Kuttler et al., 2000; Altuvia and Margalit, 2000). Surprisingly, positive contributions of the negatively charged aspartic and glutamic acids at P1′ position are seen in the present study.

Among the amino acids occupying position P2, Glu, His, Cys, Lys and Trp contribute positively and Asp, Tyr, Arg, Phe, Leu and Gln make negative contributions. At the P2′ position, Gly, Arg, Glu, Asn, Thr and Ser have positive coefficients, while Tyr, Phe, His, Ile, Met and Trp contribute negatively. The stepwise model suggests that certain amino acids at distant positions also contribute to cleavage. For example, Phe at P5 position and His at P4 position, as well as Asp at P3′ and P5′, and Ala and Ser at position P5′, make significant positive contributions. Negative contributions are made by Ala and Met at position P5, Gly and Arg at position P3, Trp, Phe and Leu at position P3′, Tyr and Val at position P4′ and Leu at position P5′.

A comparison between the three currently available web algorithms—PAProc, FragPredict and NetChop 2.0—and the additive models described here indicates that NetChop and the additive models predict almost equally well ($A_{ROC} = 0.771$ for the NetChop versus 0.763 for the stepwise model), followed by FragPredict ($A_{ROC} = 0.605$) and PAProc ($A_{ROC} = 0.589$). It appears that we are reaching the limits of current technology, in terms of the accuracy and universality realizable by proteasome prediction methods. The current paucity of proteasomal cleavage data, upon which these approaches depend, is clearly the limiting factor. It proved impossible to generate a reliable prediction method (unpublished data) based on the cleavage patterns apparent in the handful of proteins analyzed systematically for proteasome digested fragments (Nussbaum et al., 1998). Instead, we used data available from naturally processed epitopes and natural ligands eluted from the cell surface. One feature of such a dataset is that it contains peptides which have been generated through a complex combination of different proteases, which each exhibit a distinct specificity of cleavage. However, it is now well known (Craiu et al., 1997; Mo et al., 1999; Serwold and Shastri, 1999; Cascio et al., 2001), that such peptides retain, at the C terminus, a strong signal derived from proteasome cleavage. We thus exploit this observation in our work. Attempts to deconvolute N terminally cleavage patterns, derived from ERAAP and other proteases, are futile in the absence of explicit data on substrate specificity for such enzymes.

It is now clear that cleavage by the proteasome is only one event in antigen presentation: there are many more, and many of these are proteolytic. Analyses of peptide generation and T-cell epitopes expression in proteasome-inhibited cells suggest that cytoplasmic proteases other than proteasomes may also be involved in antigen processing pathway (Vinitsky et al., 1997; Luckey et al., 1998; Luckey et al., 2001). Tripeptidylpeptidase II (TPPII) was suggested to be a peptide supplier because of its ability to cleave peptides in vitro and its upregulation in cells surviving partial proteasome inhibition (Geier et al., 1999). Leucine aminopeptidase was found to generate antigenic peptides from N-terminally extended precursors (Beninga et al., 1998). Puromycin sensitive animopeptidase and bleomycin hydrolase were shown to trim N termini of synthetic peptides (Stoltze et al., 2000). Recently, an enzyme located in the lumen in ER and named ERAAP (ER aminopeptidase associated with antigen processing) (Serwold et al., 2002) or ERAP1 (ER aminopeptidase 1) (Saric et al., 2002; York et al., 2002), was proven to be responsible for the final trim of the N termini of peptides presented by MHC class I molecules. However, currently there is insufficient quantitative data about the role of these proteases to allow a precise bioinformatic evaluation of their impact on the antigen processing pathway.

The additive model for proteasome cleavage described here was incorporated in a three-step algorithm for T-cell epitope prediction. This approach will, in due course, be made available as a publicly accessible server for multi-step T-cell epitope prediction. The algorithm was tested on a set of newly reported T-cell epitopes and the predictions were compared with those made by the two best known servers for T-cell epitope prediction: SYPFEITHI and BIMAS. Of these three algorithms, the additive method performed best, predicting all 32 epitopes. This indicates that as we effectively model a more complete picture of the antigen presentation pathway, including other aspects of processing, such as TAP and proteasomal cleavage, the overall results, in terms of predicting epitopes, will be become ever more accurate. Such models, incorporating higher levels of complexity, thus represent a more efficient and more robust solution to this pivotal challenge: the successful in silico design of vaccines.

Acknowledgements

This work was supported by GlaxoSmithKline, Medical Research Council, Biotechnology and Biological Sciences Research Council, UK Department of Health and Medical Research Council at the Medical University in Sofia, Bulgaria.

References

Ackerman, A.L., Cresswell, P., 2004. Cellular mechanisms governing cross-presentation of exogenous antigens. Nat. Immunol. 5, 678–684.

Altfeld, M.A., Livingston, B., Reshamwala, N., Nguyen, P.T., Addo, M.M., Shea, A., Newman, M., Fikes, J., Sidney, J., Wentworth, P., Chesnut, R., Eldridge, R.L., Rosenberg, E.S., Robbins, G.K., Brander, C., Sax, P.E., Boswell, S., Flynn, T., Buchbinder, S., Goulder, P.J., Walker, B.D., Sette, A., Kalams, S.A., 2001a. Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif. J. Virol. 75, 1301–1311.

Altfeld, M., Addo, M.M., Eldridge, R.L., Yu, X.G., Thomas, S., Khatri, A., Strick, D., Phillips, M.N., Cohen, G.B., Islam, S.A., Kalams, S.A., Brander, C., Goulder, P.J., Rosenberg, E.S., Walker, B.D., 2001b. HIV study collaboration. Vpr is preferentially targeted by CTL during HIV-1 infection. J. Immunol. 167, 2743–2752.

Altuvia, Y., Margalit, H., 2000. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. J. Mol. Biol. 295, 879–890.

Beninga, J., Rock, K.L., Goldberg, A.L., 1998. Interferon-gamma can stimulate post-proteasomal trimming of the N terminus of an antigenic peptide by inducing leucine aminopeptidase. J. Biol. Chem. 273, 18734–18742.

Blythe, M.J., Doytchinova, I.A., Flower, D.R., 2002. JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics 18, 434–439.

Bourgault Villada, I., Beneton, N., Bony, C., Connan, F., Monsonego, J., Bianchi, A., Saiag, P., Levy, J.P., Guillet, J.G., Choppin, J., 2000. Identification in humans of HPV-16 E6 and E7 protein epitopes recognized by cytolytic T lymphocytes in association with HLA-B18 and determination of the HLA-B18-specific binding motif. Eur. J. Immunol. 30, 2281–2289.

Bownds, S., Tong-On, P., Rosenberg, S.A., Parkhurst, M., 2001. Induction of tumor-reactive cytotoxic T-lymphocytes using a peptide from NY-ESO-1 modified at the carboxy-terminus to enhance HLA-A2.1 binding affinity and stability in solution. J. Immunother. 24, 1–9.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.

Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J., Petrovsky, N., 1999. A neural network model approach to the study of human TAP transporter. In Silico Biol. 1, 109–121.

Buslepp, J., Zhao, R., Donnini, D., Loftus, D., Saad, M., Appella, E., Collins, E.J., 2001. T cell activity correlates with oligomeric peptide-major histocompatibility complex binding on T cell surface. J. Biol. Chem. 276, 47320–47328.

Caccamo, N., Milano, S., Di Sano, C., Cigna, D., Ivanyi, J., Krensky, A.M., Dieli, F., Salerno, A., 2002. Identification of epitopes of mycobacterium tuberculosis 16-kDa protein recognized by human leukocyte antigen-A*0201 CD8(+) T lymphocytes. J. Infect. Dis. 186, 991–998.

Cascio, P., Hilton, C., Kisselev, A.F., Rock, K.L., Goldberg, A.L., 2001. 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. EMBO J. 20, 2357–2366.

Craiu, A., Akopian, T., Goldberg, A., Rock, K.L., 1997. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. Proc. Natl. Acad. Sci. U.S.A. 94, 10850–10855.

Daniel, S., Brusic, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., Sinigaglia, F., Gallazzi, F., Hammer, J, van Endert, P.M., 1998. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J. Immunol. 161, 617–624.

Dannull, J., Diener, P.A., Prikler, L., Furstenberger, G., Cerny, T., Schmid, U., Ackermann, D.K., Groettrup, M., 2000. Prostate stem cell antigen is a promising candidate for immunotherapy of advanced prostate cancer. Cancer Res. 60, 5522–5528.

Djaballah, H., Harness, J.A., Savory, P.J., Rivett, A.J., 1992. Use of serine-protease inhibitors as probes for the different proteolytic activities of the rat liver multicatalitic proteinase complex. Eur. J. Biochem. 209, 629–634.

Doytchinova, I., Blythe, M.J., Flower, D.R., 2002. Additive method for the prediction of protein–peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. J. Proteome Res. 1, 263–273.

Doytchinova, I., Flower, D., 2003a. The HLA-A2 supermotif: a QSAR definition. Org. Biomol. Chem. 1, 2648–2654.

Doytchinova, I.A., Flower, D.R., 2003b. Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. Bioinformatics 19, 2263–2270.

Doytchinova, I.A., Walshe, V., Jones, N., Gloster, S., Borrow, P., Flower, D.R., 2004a. Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatics approach enabling prediction of superbinding peptides and anchorless epitopes. J. Immunol. 172, 7495–7502.

Doytchinova, I., Hemsley, S., Flower, D.R., 2004b. Transport associated with antigen processing preselection of peptides binding to MHC: a bioinformatic evaluation. J. Immunol. 173, 6813–6819.

Drexler, I., Staib, C., Kastenmuller, W., Stevanovic, S., Schmidt, B., Lemonnier, F.A., Rammensee, H.G., Busch, D.H., Bernhard, H., Erfle, V., Sutter, G., 2003. Identification of vaccinia virus epitope-specific HLA-A*0201-restricted T cells and comparative analysis of smallpox vaccines. Proc. Natl. Acad. Sci. U.S.A. 100, 217–222.

Duraiswamy, J., Sherritt, M., Thomson, S., Tellam, J., Cooper, L., Connolly, G., Bharadwaj, M., Khanna, R., 2003. Therapeutic LMP1 polyepitope vaccine for EBV-associated Hodgkin disease and nasopharyngeal carcinoma. Blood 101, 3150–3156.

Flower, D.R., 2003. Towards in silico prediction of immunogenic epitopes. Trends Immunol. 24, 667–674.

J. Friedman, Multivariate adaptive regression spline. Technical Report No. 102, Stanford University, Stanford, CA, 1990.

Geier, E., Pfeifer, G., Wilm, M., Lucchiari-Hartz, M., Baumeister, W., Eichmann, K., Niedermann, G., 1999. A giant protease with potential to substitute for some functions of the proteasome. Science 283, 978–981.

Geluk, A., van Meijgaarden, K.E., Franken, K.L., Drijfhout, J.W., D'Souza, S., Necker, A., Huygen, K., Ottenhoff, T.H., 2000. Identification of major epitopes of mycobacterium tuberculosis AG85B that are recognized by HLA-A*0201-restricted CD8+ T cells in HLA-transgenic mice and humans. J. Immunol. 165, 6463–6471.

Golgberg, A.L., Cascio, P., Saric, T., Rock, K.L., 2002. The importance of the proteasome and subsequent proteolitic steps in the generation of antigenic peptides. Mol. Immunol. 39, 147–164.

Gonzalez, J.M., Peter, K., Esposito, F., Nebie, I., Tiercy, J.M., Bonelo, A., Arevalo-Herrera, M., Valmori, D., Romero, P., Herrera, S., Corradin, G., Lopez, J.A., 2000. HLA-A*0201 restricted CD8+ T-lymphocyte responses to malaria: identification of new *Plasmodium falciparum* epitopes by IFN-gamma ELISPOT. Parasite Immunol. 22, 501–514.

Green, K.J., Miles, J.J., Tellam, J., van Zuylen, W.J., Connolly, G., Burrows, S.R., 2004. Potent T cell response to a class I-binding 13-mer viral epitope and the influence of HLA micropolymorphism in controlling epitope length. Eur. J. Immunol. 34, 2510–2519.

Guan, P., Doytchinova, I.A., Flower, D.R., 2003a. HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. Protein Eng. 16, 11–18.

Guan, P., Doytchinova, I.A., Zygouri, C., Flower, D.R., 2003b. MHCPred: a server for quantitative prediction of peptide—MHC binding. Nucleic Acids Res. 31, 3621–3624.

Hattotuwagama, C., Guan, P., Doytchinova, I.A., Flower, D.R., 2004. New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity. Org. Biomol. Chem. 2, 3274–3283.

Holzhutter, H.G., Frommel, C., Kloetzel, P.M., 1999. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20S proteasome. J. Mol. Biol. 286, 1251–1265.

Holzhutter, H.G., Kloetzel, P.M., 2000. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolutic fragments from oligomeric peptide substrates. Biophysics 79, 1196–1205.

Jaye, A., Herberts, C.A., Jallow, S., Atabani, S., Klein, M.R., Hoogerhout, P., Kidd, M., van Els, C.A., Whittle, H.C., 2003. Vigorous but short-term gamma interferon T-cell responses against a dominant HLA-A*02-restricted measles virus epitope in patients with measles. J. Virol. 77, 5014–5016.

Kather, A., Ferrara, A., Nonn, M., Schinz, M., Nieland, J., Schneider, A., Durst, M., Kaufmann, A.M., 2003. Identification of a naturally processed HLA-A*0201 HPV18 E7 T cell epitope by tumor cell mediated in vitro vaccination. Int. J. Cancer 104, 345–353.

Keşmir, C., Nussbaum, A.K., Schild, H., Detours, V., Brunak, S., 2002. Prediction of proteasome cleavage motifs by neural networks. Protein Eng. 15, 287–296.

Kuttler, C., Nussbaum, A.K., Dick, T.P., Rammensee, H.-G., Schild, H., Hadeler, K.-P., 2000. An algorithm for the prediction of proteasome cleavages. J. Mol. Biol. 298, 417–429.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemom. 6, 267–281.

Luckey, C.J., King, G.M., Marto, J.A., Venketeswaran, S., Maier, B.F., Crotzer, V.L., Colella, T.A., Shabanowitz, J., Hunt, D.F., Engelhard, V.H., 1998. Proteasomes can either generate or destroy MHC class I epitopes: evidence for nonproteosomal epitope generation in the cytosol. J. Immunol. 161, 112–121.

Luckey, C.J., Marto, J.A., Partridge, M., Hall, E., White, F.M., Lippolis, J.D., Shabanowitz, J., Hunt, D.F., Engelhard, V.H., 2001. Differences in the expression of human class I MHC alleles and their associated peptides in the presence of proteasome inhibitors. J. Immunol. 167, 1212–1221.

Maranon, C., Thomas, M.C., Planelles, L., Lopez, M.C., 2001. The immunization of A2/K(b) transgenic mice with the KMP11-HSP70 fusion

protein induces CTL response against human cells expressing the T. cruzi KMP11 antigen: identification of A2-restricted epitopes. Mol. Immunol. 38, 279–287.

McSparron, H., Blythe, M.J., Zygouri, C., Doytchinova, I.A., Flower, D.R., 2003. JenPep: a novel computational information resource for immunobiology and vaccinology. J. Chem. Inf. Comput. Sci. 43, 1276–1287.

MDL QSAR 2.2, 2004. MDL Information Systems Inc., San Leandro, CA.

Mo, X.Y., Cascio, P., Lemerise, K., Goldberg, A.L., Rock, K., 1999. Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides. J. Immunol. 163, 5851–5859.

Niedermann, G., King, G., Butz, S., Birsner, U., Grimm, R., Shabanowitz, J., Hunt, D.F., Eichmann, K., 1996. The proteolytic fragments generated by vertebrate proteasomes: structural relationships to major histocompatibility complex class I binding peptides. Proc. Natl. Acad. Sci. U.S.A. 93, 8572–8577.

Nussbaum, A.K., Dick, T.P., Keilholz, W., Schirle, M., Stevanovic, S., Dietz, K., Heinemever, W., Groll, M., Wolf, D.H., Huber, R., Rammensee, H.-G., Schild, H., 1998. Cleavage motifs of the yeast 20 S proteasome β subunits deduced from digests of enolase 1. Proc. Natl. Acad. Sci. U.S.A. 95, 12504–12509.

Nussbaum, A.K., Kuttler, C., Hadeler, K.-P., Rammensee, H.-G., Schild, H., 2001. PAProC: a prediction algorithm for proteosomal cleavage available on the WWW. Immunogenetics 53, 87–94.

Orlowski, M., Michaud, C., 1989. Pituitary multicatalytic proteinase complex. Specificity of components and aspects of proteolitic activity. Biochemistry 28, 9270–9278.

Orlowski, M., Cardozo, C., Michaud, C., 1993. Evidence for the presence of five distinct proteolytic components in the pituitary multicatalytic proteinase complex. Properties of two components cleaving bonds on the carboxy side of branched chain and small neutral amino acids. Biochemistry 32, 1563–1572.

Parker, K.C., Bednarek, M.A., Coligan, J.E., 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J. Immunol. 152, 163–175.

Peter, K., Men, Y., Pantaleo, G., Gander, B., Corradin, G., 2001. Induction of a cytotoxic T-cell response to HIV-1 proteins with short synthetic peptides and human compatible adjuvants. Vaccine 19, 4121–4129.

Petrovsky, N., Brusic, V., 2004. Virtual models of the HLA class I antigen processing pathway. Methods 34, 429–435.

Probst-Kepper, M., Hecht, H.J., Herrmann, H., Janke, V., Ocklenburg, F., Klempnauer, J., van den Eynde, B.J., Weiss, S., 2004. Conformational restraints and flexibility of 14-meric peptides in complex with HLA-B*3501. J. Immunol. 173, 5610–5616.

Rammensee, H.-G., Friede, T., Stevanović, S., 1995. MHC ligands and peptide motifs: first listing. Immunogenetics 41, 178–228.

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

Rudolf, M.P., Man, S., Melief, C.J., Sette, A., Kast, W.M., 2001. Human T-cell responses to HLA-A-restricted high binding affinity peptides of human papillomavirus type 18 proteins E6 and E7. Clin. Cancer Res. 7, 788s–795s.

Saric, T., Chang, S.-C., Hattori, A., York, I.A., Markant, S., Rock, K.L., Tsujimoto, M., Goldberg, A.L., 2002. An IFN-γ-induced aminopeptidase in the ER, ERAPI, trims precursors to MHC class I-presented peptides. Nat. Immunol. 3, 1169–1176.

Saxova, P., Buus, S., Brunak, S., Keşmir, C., 2003. Predicting proteasomal cleavage sites: a comparison of available methods. Int. Immunol. 15, 781–787.

Schirle, M., Weinschenk, T., Stevanović, S., 2001. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. J. Immunol. Methods 257, 1–16.

Serwold, T., Shastri, N., 1999. Specific proteolytic cleavages limit the diversity of the pool of peptides available to MHC class I molecules in living cells. J. Immunol. 162, 4712–4719.

Serwold, T., Gonzalez, F., Kim, J., Jacob, R., Shastri, N., 2002. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. Nature 419, 480–483.

Sharma, A.K., Kuhns, J.J., Yan, S., Friedline, R.H., Long, B., Tisch, R., Collins, E.J., 2001. Class I major histocompatibility complex anchor substitutions alter the conformation of T cell receptor contacts. J. Biol. Chem. 276, 21443–21449.

Shastri, N., Schwab, S., Serwold, T., 2002. Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules. Annu. Rev. Immunol. 20, 463.

Stoltze, L., Schirle, M., Schwarz, G., Schroeter, C., Thompson, M.W., Hersh, L.B., Kalbacher, H., Stevanović, S., Rammensee, H.-G., Schild, H., 2000. Two new proteases in the MHC class I processing pathway. Nat. Immunol. 1, 413–418.

SYBYL 6.9., 2004. Tripos Inc., St. Louis.

Tanaka, K., Kasahara, M., 1998. The MHC class I ligand-generating system: roles of immunoproteasomes and the interferon-γ-inducible proteasome activator PA28. Immunol. Rev. 163, 161–176.

Tanaka, Y., Amos, K.D., Fleming, T.P., Eberlein, T.J., Goedegebuure, P.S., 2003. Mammaglobin-A is a tumor-associated antigen in human breast carcinoma. Surgery 133, 74–80.

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M.M., Kloetzel, P.M., Rammensee, H.G., Schild, H., Holzhutter, H.G., 2005. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage. TAP transport and MHC class I binding. Cell Mol. Life Sci. 62, 1025–1037.

Terajima, M., Cruz, J., Raines, G., Kilpatrick, E.D., Kennedy, J.S., Rothman, A.L., Ennis, F.A., 2003. Quantitation of CD8+ T cell responses to newly identified HLA-A*0201-restricted T cell epitopes conserved among vaccinia and variola (smallpox) viruses. J. Exp. Med. 197, 927–932.

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., Schild, H., 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. J. Exp. Med. 194, 1–12.

Toseland S.P., Clayton D.J., McSparron H., Hemsley S.L., Blythe M.J., Paine K., Doytchinova I.A., Guan P., Hattotuwagama C.K., Flower D.R., 2005. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res., 1, 4 (online).

Van den Eynde, B.J., Morel, S., 2001. Differential processing of class-I-restricted epitopes by the standard proteasome and the immunoproteasome. Curr. Opin. Immunol. 13, 147–153.

Vinitsky, A., Anton, L.C., Snyder, H.L., Orlowski, M., Bennink, J.R., Yewdell, J.W., 1997. The generation of MHC class I-associated peptides is only partially inhibited by proteasome inhibitors: involvement of non-proteasomal cytosolic proteases in antigen processing? J. Immunol. 159, 554–564.

Wold, S., 1995. PLS for multivariate linear modeling. In: Van der Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design. VCH, Weinheim, pp. 195–218.

York, I.A., Chang, S.-C., Saric, T., Keys, J.A., Favreau, J.M., Goldberg, A.L., Rock, K.L., 2002. The ER aminopeptidase ERAPI enhances or limits antigen presentation by trimming epitopes to 8–9 residues. Nat. Immunol. 3, 1177–1184.