Immunogenetics (2004) 56: 405–419
DOI 10.1007/s00251-004-0709-7

ORIGINAL PAPER

Pedro A. Reche · John-Paul Glutting · Hong Zhang · Ellis L. Reinherz

Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles

Received: 3 May 2004 / Revised: 12 July 2004 / Published online: 3 September 2004

Abstract We introduced previously an on-line resource, RANKPEP that uses position specific scoring matrices (PSSMs) or profiles for the prediction of peptide-MHC class I (MHCI) binding as a basis for CD8 T-cell epitope identification. Here, using PSSMs that are structurally consistent with the binding mode of MHC class II (MHCII) ligands, we have extended RANKPEP to prediction of peptide-MHCII binding and anticipation of CD4 T-cell epitopes. Currently, 88 and 50 different MHCI and MHCII molecules, respectively, can be targeted for peptide binding predictions in RANKPEP. Because appropriate processing of antigenic peptides must occur prior to major histocompatibility complex (MHC) binding, cleavage site prediction methods are important adjuncts for T-cell epitope discovery. Given that the C-terminus of most MHCI-restricted epitopes results from proteasomal cleavage, we have modeled the cleavage site from known MHCI-restricted epitopes using statistical language models. The RANKPEP server now determines whether the C-terminus of any predicted MHCI ligand may result from such proteasomal cleavage. Also implemented is a variability masking function. This feature focuses prediction on conserved rather than highly variable protein segments encoded by infectious genomes, thereby offering identification of invariant T-cell epitopes to thwart mutation as an immune evasion mechanism.

Keywords Epitopes · Major histocompatibility complex · Prediction · Profile · Proteasome

P. A. Reche (✉) · J.-P. Glutting · H. Zhang · E. L. Reinherz
Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute,
44 Binney Street,
Boston, MA 02115, USA
e-mail: reche@research.dfci.harvard.edu
Tel.: +1-617-6323412
Fax: +1-617-6323351

P. A. Reche · E. L. Reinherz
Department of Medicine, Harvard Medical School,
44 Binney Street,
Boston, MA 02115, USA

Introduction

T-cells play a central role in the host adaptive immune defense by the recognition of foreign peptide antigens bound to cell-membrane expressed major histocompatibility complexes (MHC) via their T-cell receptors (TCR) (reviewed in Garcia et al. 1999; Margulies 1997; Wang and Reinherz 2001). Antigen presenting MHC molecules are extremely polymorphic (Reche and Reinherz 2003), and fall into two major classes, termed class I and class II (reviewed in Maenaka and Jones 1999; Stern and Wiley 1994). Antigens presented by MHC class I (MHCI) and MHC class II (MHCII) are recognized by two distinct sets of T-cells, CD8 and CD4 T-cells, respectively (reviewed in Wang and Reinherz 2001). Engaging both T-cell subsets is desirable for mounting a strong defensive immune response against cancer cells and pathogens.

T-cell immune responses are driven by antigenic epitopes whose identification is important for understanding disease pathogenesis and etiology as well as for vaccine design. Bona fide experimental identification of T-cell epitopes is costly and time consuming, requiring the synthesis of overlapping peptides spanning the entire length of a protein and necessitating complicated in vitro cellular assays for each peptide synthesized (Draenert et al. 2003). As a result, alternative computational approaches have been developed for the prediction of antigenic peptides. Since T-cells recognize antigenic peptides only in the context of MHC molecules (Zinkernagel and Doherty 1974), computer-aided methods for the anticipation of T-cell epitopes rely mostly on the prediction of peptide-MHC binding (Hammer 1995). Peptides bound to the same MHC are related by sequence similarity (Falk et al. 1991; Rammensee et al. 1995) so that peptide binding patterns (Sette et al. 1989) as well as motif matrices (De Groot et al. 1997; Rammensee et al. 1999) have also been used for the prediction of peptide-MHC binding. In this regard, we have recently introduced a more sophisticated matrix method involving the use of position-specific scoring matrices (PSSMs) or profiles (Gribskov et al. 1987) for the prediction of peptide-MHCI binding (Reche

et al. 2002). This resource is available online at our RANKPEP web server (http://www.mifoundation.org/Tools/rankpep.html). Profiles are derived from a set of aligned peptides known to bind to a given MHC. Correct alignment of peptides by structural or sequence similarity is essential for the proper prediction of peptide-MHC binding (Gribskov et al. 1987). Profiles of aligned peptides known to bind to MHCII molecules could also be used for the prediction of peptide-MHCII binding, and consequent anticipation of CD4 T-cell epitopes. However, alignment of MHCII ligands is quite challenging since peptides binding to a single MHCII molecule are extremely variable in length and share very limited sequence similarity (Barber and Parham 1993; Madden 1995; Stern and Wiley 1994). In response to this complexity, in this paper we describe the use of the motif discovery program MEME (Bailey and Elkan 1995) to generate alignment and profiles that are consistent with the binding mode of peptides to MHCII molecules. Fifty MHCII-specific profiles have been created, allowing the extension of the RANKPEP resource to the anticipation of CD4 T-cell epitopes. On average, the sensitivity of these MHCII-specific profiles is such that ~60% of known MHCII-restricted T-cell epitopes are found among the top 2% scoring peptides from their protein sources.

Anticipation of T-cell epitopes is heavily predicated on the prediction of peptide-MHC binding, yet prior to MHC binding, correct peptide processing must occur to liberate a peptide from its protein source. Processing of MHCII-restricted epitopes occurs in the endosomal compartment, being mediated by several endopeptidases in combination with amino-peptidases and carboxy-peptidases (Pieters 2000; Watts 2001). This complexity makes the identification of any pattern related with processing of class II restricted peptides difficult. In contrast, there is experimental evidence that the C-terminus of MHCI-restricted epitopes results from the selective proteolysis of cytosolic proteins mediated by the proteasome (Craiu et al. 1997). The proteasome thus plays a vital role in determining cytotoxic T-cell (CTL) epitopes, and consequently we have modeled the proteasomal cleavage site from MHCI-restricted peptides using statistical language models. Language models for proteasomal cleavage prediction (LMPCP) could properly recognize up to ~90% of the C-termini from a set of 554 known MHCI-restricted epitopes, independent of the training set. Prediction of those MHCI-peptide binders containing a C-terminus that is likely to be the result of proteasomal cleavage is now implemented by the RANKPEP web server, leading to refined CTL epitope predictions. Finally, because amino-acid sequence mutation offers a means for immune evasion exploited by some pathogenic organisms such us HIV, we implemented the RANKPEP web server with a feature to mask the sequence variability from a multiple sequence alignment (MSA) using the Shannon entropy measure (Shannon 1948) as a variability metric (Reche and Reinherz 2003; Stewart et al. 1997).

Materials and methods

Peptide and protein sequences

Sequences of peptides that bind to MHC molecules were collected from the MHCPEP database (Brusic et al. 1998b). All peptides in the MHCPEP database are binders, but their binding strength is reported as unknown, low, moderate, or high. In this work we have excluded MHC ligands that were labeled as low binders. Sequences of naturally restricted T-cell epitopes were collected from the SYFPEITHI database (Rammensee et al. 1999). Full-length sequences of proteins containing T-cell epitopes were isolated from the non-redundant database of GenBank (Benson et al. 2003) following a blast search (Altschul et al. 1997) with the relevant T-cell epitopes as the query.

Alignments and PSSMs of MHCI and MHCII-specific ligands

Block alignments of peptides binding to specific MHCI molecules were obtained as indicated elsewhere (Reche et al. 2002). Briefly, peptides were collected from MHCPEP according to their MHCI binding specificity, and subsequently grouped by their sequence length to create block alignments. Likewise, peptides binding to MHCII molecules were collected from the MHCPEP database, and divided into subsets according to their MHCII binding specificity. Peptides shorter than nine residues were not considered. Subsequently, motif block alignments of peptides binding to specific MHCII molecules were obtained using the motif discovery program MEME (Bailey and Elkan 1995), using the command *meme file.fasta -protein -mod oops -nmotifs 1-minsites 4-maxsites 300 -minw 9 -maxw 9 -evt 10,000*, where *file.fasta* corresponds to each of the MHCII-specific subsets of protein ligands in FASTA format; *-mod oops*, indicates that each sequence has a binding site; *-minsites 4 -maxsites 500*, indicates that the motif should contain between four and 500 sequences; *-min 9 -maxw 9*, indicates that the size of the motif is exactly nine; and finally *-evt 10000* is the expected threshold value for a sequence to be included in the motif. Collections of peptides binding to the same MHCII molecule usually contain overlapping peptides, and therefore block motifs yielded by MEME frequently contain repeated sequences corresponding to the overlapping regions of different peptides. Consequently, alignments were parsed to eliminate sequence redundancy, yielding block alignments of unique sequences.

Profiles were obtained from peptide alignments containing a minimum of five sequences using PROFILEWEIGHT (Thompson et al. 1994b) and the BLK2PSSM utility included in the BLIMPS package (Henikoff and Henikoff 1996; Henikoff et al. 1999). PROFILEWEIGHT uses a branch-proportional weighting method, whereas a position-based weighting method (Henikoff and Henikoff

1994) was applied to the PSSMs obtained with BLK2PSSM.

Scoring peptide-MHC binding using PSSMs and cross-validation tests

Peptide scores indicate the similarity (and hence binding potential) of the peptides to the set of aligned peptides known to bind to a given MHC molecule. Scores are obtained by aligning the PSSM with the protein segments, and adding up the appropriate profile coefficients matching the residue type and position in the protein segment. Scoring starts at the beginning of each sequence, and the PSSM is moved over the entire sequence one residue at a time.

Cross-validation tests to address the predictive performance of profile matrices were carried out using a ROC analysis (Swets 1988). The ROC curves were generated plotting the function SE versus 1-SP for various thresholds of top scoring peptides (0.5, 1, 2, 3, 5, 10, 20%), where SE, and SP represent the sensitivity and specificity of the predictions, respectively. The area under the ROC curve (*AUC*) provides a measure of overall prediction accuracy. SE and SP are calculated from Eqs. 1, 2

$$SE = TP/(TP + FN) \quad (1)$$

$$SP = TN/(TN + FP) \quad (2)$$

where TP are true positives (binders predicted as binders); FN are false negative (binders predicted as non-binders); TN are true negatives (non-binders predicted as non-binders) and FP are false positives (non-binders predicted as binders). For each of the MHC molecules in Table 1, known binders were divided into two distinct sets, a binding training set and a binding test set, with comparable numbers of peptides. PSSMs were derived from the training set using both PROFILEWEIGHT and BLK2PSSM as indicated elsewhere and then used to test whether the peptides in the binding test set were TP or FP at the mention thresholds. At any given threshold a test binding peptide was considered to be a TP if it was found among the predicted peptides at that threshold from a random protein of 1,000 (amino-acid composition after frequencies in the swissprot database) incorporating the tested peptide. The peptide was a FN if it was not among the predicted peptides. Calculation of SP requires having a set of experimentally determined non-binders. Unfortunately, because there are very few experimentally verified examples of peptides that do not bind to a particular MHC, we have followed an approach similar to that of Donnes and Elofsson (2002) to obtain a set of non-binder peptides. In any given protein, most of the peptides do not bind to the MHC molecule (90–98%), and consequently a randomly generated peptide could be considered as a non-binder. Thus, we have calculated the SP of the predictions from a set of randomly generated peptides (not identical to the binders). The number of non-binder

Table 1 MHC molecules targeted for peptide binding predictions

MHCI	Aln[a]	Proc[b]	MHCII	Aln[a]	Proc[b]
HLA-A2 (A*0201)	291	139	HLA-DQ2 (DQA1*0501×DQB1*0201)	31	15
HLA-A2 (A*0202)	20	15	HLA-DQ8 (DQA1*0301×DQB1*0302)	52	25
HLA-A2 (A*0204)	34	18	HLA-DP9 (DPA1*0201×DPB1*0901)	18	15
HLA-A2 (A*0205)	22	15	HLA-DR1 (DRB1*0101)	189	81
HLA-A2 (A*0206)	34	14	HLA-DR1 (DRB1*0102)	21	11
HLA-A3 (A*0301)	47	36	HLA-DR4 (DRB1*0401)	322	140
HLA-A11 (A*1101)	63	44	HLA-DR4 (DRB1*0402)	72	36
HLA-A24 (A*2402)	57	54	HLA-DR4 (DRB1*0405)	66	23
HLA-A33 (A*3301)	23	22	HLA-DR4 (DRB1*0404)	44	30
HLA-A68 (A*6801)	50	40	HLA-DR7 (DRB1*0701)	81	49
HLA-B7 (B*0702)	48	41	HLA-DR8 (DRB1*0801)	41	31
HLA-B27 (B*2703)	22	16	HLA-DR9 (DRB1*0901)	39	10
HLA-B27 (B*2704)	10	10	HLA-DR11 (DRB1*1101)	124	25
HLA-B27 (B*2705)	82	35	HLA-DR11 (DRB1*1104)	28	10
HLA-B35 (B*3501)	81	59	HLA-DR15 (DRB1*1501)	35	19
HLA-B51 (B*5101)	39	26	HLA-DR17 (DRB1*0301)	20	15
HLA-B51 (B*5102)	32	29	HLA-DR51 (DRB5*0101)	52	29
HLA-B51 (B*5103)	30	28	I-Ak	121	66
HLA-B53 (B*5301)	39	31	I-Ad	240	62
HLA-B54 (B*5401)	42	33	I-Ed	212	49
H-2Kb	84	18	I-Ek	226	58
H-2Qa-2a	22	21	I-Ag7	76	44
H-2Db	71	22	I-As	67	32
H-2Ld	64	14	I-Ab	97	53
H-2Kd	63	32			

[a]Number of ligands known to bind to the relevant MHC molecules included in the alignment.
[b]Peptides in the alignment identified as T-cell epitopes for which we retrieved protein sources. These peptides were targeted in the epitope prediction test using PSSMs derived from the relevant alignments.

peptides was double that of the peptides in the binding set. At each threshold a non-binder was considered a TN if it was not among the predicted peptides from a random protein of 1,000 amino acids incorporating the tested non-binder peptide. The non-binder peptide was considered an FP, if it was found among the predicted peptides.

Epitope prediction tests using PSSMs

Peptide-MHC binding prediction tests were carried out by determining the relative ranking of known MHC-restricted epitopes from their protein sources using the relevant profiles. Several thresholds of top scoring peptides were checked (0.5, 1, 2, 3, 5, 10, 20%). Peptides were considered to be predicted if they were among the top scoring peptides at the set threshold. MHC molecules targeted for peptide predictions are shown (Table 1). These MHC molecules were selected on the basis of alignments of known ligands that include at least ten sequences of MHCI-restricted or MHCII-restricted peptides, CD8 and CD4 T-cell epitopes, respectively. Moreover, the binding specificity of the ligands was known at the allelic level. MHCI-restricted peptides considered in these tests were all nonamers (9mers), and annotations about known CD8 and CD4 T-cell epitopes were taken from the SYFPEITHI database (Rammensee et al. 1999). Binding predictions of these epitopes to their MHC molecules were tested using PSSM that were derived from alignments with and without the epitope to be predicted. PSSMs for prediction tests were obtained using both PROFILEWEIGHT and BLK2PSSM as indicated elsewhere.

Prediction of proteasomal cleavage using statistical language models

The proteasomal cleavage site was modeled from a database consisting of 332 naturally MHCI-restricted epitope fragments and their C-terminal flanking regions using the SRI language modeling toolkit (SRILM) (Stolcke 2002). Selected epitopes were all restricted by human MHCI molecules (HLA I). To prevent biases towards any given *HLA I* allele, selected peptides included all peptides restricted by *HLA-C* (38) and *HLA-G* alleles (12), and a similar number of epitopes restricted by HLA-A and HLA-B binding peptides, 135 and 147, respectively. Moreover, no more than 5% of the selected epitopes were restricted by the same *HLA-A* or *HLA-B* allele. LMPCP were created using the SRILM NGRAM-COUNT utility over training sets derived from the above database. From this database different training sets of fragment size 10, 8, 6, and 4 were generated (length was fixed in all fragments of the same training set). Cutpoints in fragments were indicated by a *vertical line* ("|") after the C-terminal end of the epitope (P1 cleavage site), with fragments having an even number of residues on either side of the cleavage site. Representative fragments of training sets of ten and four residues will be EPRKL|VTQDL and KL|VT, respectively.

For each training set of a given fragment size (N), $N-2$ different LMPCPs were generated by changing the window size or order (i) of the model from $i=2$ to $i=N-1$. LMPCPs then varied with the length of the fragments in the training file and with the order i chosen to generate them, and for clarity will note them as LMPCP_N^i. Each LMPCP_N^i was tested using the SRLIM HIDDEN-NGRAM utility over test files containing peptide fragments of the same size (N) as the training file and using the same window size (i) as that of the LMPCP. HIDDEN-NGRAM is a word boundary tagger based on n-gram models (Stolcke 2002) which at a selected probability threshold indicates the cutpoints in the fragments of a testing file by inserting the cleavage marker "|" into a position determined by the LMPCP. Thirteen probability thresholds were tested for each LMPCP (0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70). Test files were derived from 554 human MHCI-restricted epitopes different from those in the training test but of the same size and nature (even number of residues around the cutpoint) as those used in the training sets and without the cleavage site indicated. Evaluation of the results was carried out by determining the percentage of correctly predicted cleavage sites (PCS) in the test files. In addition, a percentage of expected cleavage sites (ECS) was calculated for each model according to the equation

$$\mathrm{ECS} = 100 \times C/(N-1) \tag{3}$$

where C is the number of cutpoints per fragment yielded by a given model LMPCP_N^i when tested in a file of fragments size N. Conceptually, ECS would indicate the number of correctly PCS, if cleavage resulting from a given model was random. Thus, the above 23 different LMPCP_N^i [N=10, 6, 4; i=1–(N–2)] were tested for each probability threshold, and ranked by the difference between the PCS and ECS. In addition, each LMPCP was tested on files containing the full-length protein source of the 554 human MHCI-restricted peptide fragments used in the testing files, and the mean length of the fragments also was obtained.

Consensus sequence and sequence variability masking

Sequence variability is calculated from multiple amino acid sequence alignments as indicated by Reche and Reinherz (Reche and Reinherz 2003), using a variability metric (V) formally identical to the Shannon entropy equation (Shannon 1948). Briefly, V per site is given by

$$V = -\sum_{i=1}^{M} P_i \log_2 P_i \tag{4}$$

where P_i is the fraction of residues of amino acid type i, and M=20, the number of amino acid types. V ranges from 0 (total conservation, only one amino-acid type is present

at that position) to 4.322 (all 20 amino acids are equally represented in that position). Note that in order to achieve the maximum value V=4.3, at least 20 sequences are required. Gap symbols (–) are considered for deriving the consensus sequence but are not computed for the variability calculations. Given a sequence variability threshold V_t, a consensus sequence is generated from the sequence alignment as the most common amino acid for those positions with a $V \leq V_t$, whereas variable positions ($V > V_t$) are masked and represented in the consensus sequence with a *dot*. Segments with a position masked are not considered in the RANKPEP predictions of peptide-MHC binding.

Sequence logos

Peptide fragments of MHCI-restricted epitopes with their flanking regions were aligned centered around the C-terminal end (cleavage site) of the MHCI-restricted peptide, and sequence information was calculated for each position and displayed using a sequence logo (Schneider and Stephens 1990). In a sequence logo each of the residues present in a position of the sequence alignment is represented with a height that is proportional to its frequency, and the height of the entire stack is proportional to the total information content (R) in that position. Sequence information R per site was given by the following equation

$$R = 4.3 - V\,(\text{bits per position}) \qquad (5)$$

where 4.3 is the upper variability limit for 20 symbols, and V is the variability in that position (Eq. 4). Sequence information is given in bits.

Results

Structure-based alignments of MHCI and MHCII ligands

MHC molecules, also known as human leucocyte antigens (HLAs) in humans, are highly polymorphic molecules imposing distinct chemical and physical constraints on their selective peptide binders which are related to each other by sequence similarity. PSSMs or profiles (Gribskov et al. 1987) created from a set of aligned peptide-MHC binders provide a means for the prediction of peptide binding to a given MHC molecule (Reche et al. 2002). However, for a PSSM to be a good predictor of peptide binding to MHC, peptides must be first aligned by structural and/or sequence similarity. MHCI and MHCII molecules bind peptides in similar yet distinct modes (Barber and Parham 1993; Madden 1995; Stern and Wiley 1994), and consequently PSSMs were derived differently for MHCI and MHCII ligands. MHCI ligands are of short length (8–11), as they are constrained into the MHCI peptide binding groove, with their N-terminal and C-terminal ends connected by a network of hydrogen bonds to conserved residues of the MHCI molecule (Madden 1995; Matsumura et al. 1992; Zhang et al. 1998) (Fig. 1a). While peptides bound to the same MHCI can differ by one or two amino acids in length from each other, proper

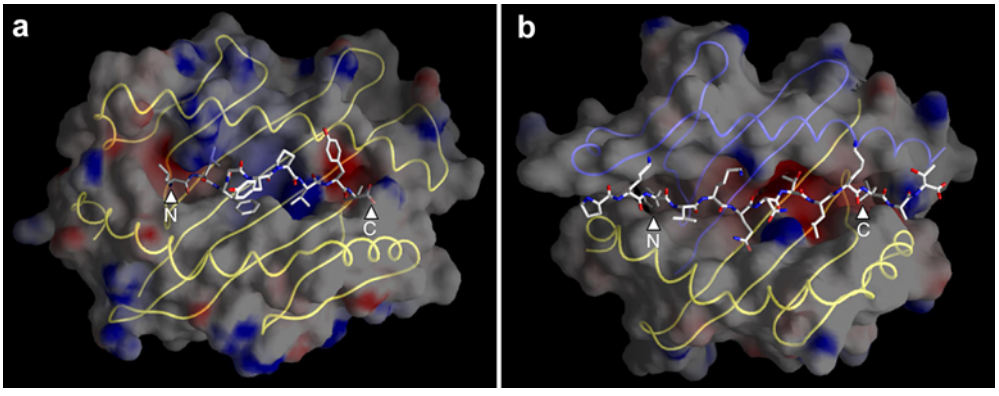

Fig. 1a, b Binding of peptide ligands to MHCI and MHCII molecules. The figure shows the top of the molecular surface of the antigen-presenting platform of representative human MHCI (**a**) and MHCII (**b**) molecules as viewed by the TCR. The MHCI molecule corresponds to HLA-A*0201 in complex with a peptide LLFGYPVYV from HTLV-1 TAX protein [PDB:1HHK (Madden et al. 1993)]. The MHCII molecule corresponds to HLA-DR1 in complex with peptide PKYVKQNTLKLAT from influenza hemagluttinin protein [PDB:1FYT (Hennecke et al. 2000)] The peptide binding platform of the MHCI molecule is composed of two anti-parallel α-helices sitting over a base of eight anti-parallel β-strands shown as worm representation under the molecular surface. Likewise, the peptide-binding platform of the MHCII molecule presents the same secondary features, but resulting from the association of two different polypeptide chains (an α1 chain in *blue* and a β1 chain in *yellow*). Peptides bound to these molecules are represented by *sticks* to highlight the contours of the binding groove. Note how the peptide binding groove of the MHCI molecule is closed, and peptides bind in a manner such that both the N-terminal and C-terminal ends of the peptide (indicated by *arrows*) are nested into the MHCI binding groove, restricting their lengths to 8–11 residues. In contrast, the peptide binding groove of the MHCII molecule is open, thereby imposing no limitation to the size of ligands, whose N-terminal and C-terminal ends can extend beyond the binding grove. The side chains of N-terminal and C-terminal end of the 9mer peptide core that fit into the MHCII binding groove are indicated. The molecular surface is colored by electrostatic potential (*blue* positively charged and *red* negatively charged). The figure was prepared using GRASP (Nicholls et al. 1991)

structural alignment of diverse peptides is best guaranteed if the peptides are of the same length (Reche et al. 2002). Accordingly, we have separated the peptides bound to a given MHCI molecule into subsets containing only peptides of the same length, and created separate profiles from ungapped block alignments. In contrast, the peptide binding groove of MHCII molecules is open, allowing both the N- and C-terminus of a peptide to extend beyond

Table 2 Performance of profiles for the prediction of peptide-MHC binding. Peptide binders (Aln column of Table 1) were randomly divided into a training and binding set and a ROC analysis was carried out to determine the *AUC* value (performance). Several ROC analyses were carried out with different training and binding sets, and the mean *AUC* value (*AUCm*) with its standard deviation is shown here. Also shown is the best *AUC* value obtained (*AUCb*). Sensitivities and specificities of the predictions with the *AUCb* at a 3% threshold are also given

MHCI	Matrix	SE/SP 3%	*AUCm*	*AUCb*	MHCII	SP/SE 3%	*AUCm*	*AUCb*
A*0201	PROFWG	0.90/0.86	0.79±0.06	0.86	HLA-DQ2	0.83/0.93	0.87±0.05	0.94
	BL2PSSM	0.95/0.84	0.85±0.03	0.89		0.85/0.94	0.88±0.06	0.95
A*0202	PROFWG	0.90/0.90	0.80±0.07	0.90	HLA-DQ8	0.73/0.93	0.70±0.06	0.78
	BL2PSSM	0.80/0.95	0.70±0.10	0.87		0.77/0.95	0.72±0.06	0.79
A*0204	PROFWG	0.82/0.94	0.68±0.09	0.79	HLA-DP9	0.84/0.93	0.80±0.06	0.91
	BL2PSSM	0.53/0.97	0.61±0.06	0.70		0.83/0.94	0.88±0.10	0.95
A*0205	PROFWG	0.73/0.95	0.67±0.07	0.79	DRB1*0101	0.72/0.89	0.74±0.05	0.80
	BL2PSSM	0.73/1.00	0.64±0.07	0.72		0.71/0.88	0.75±0.04	0.79
A*0206	PROFWG	0.94/1.00	0.85±0.05	0.94	DRB1*0102	0.68/0.94	0.72±0.05	0.80
	BL2PSSM	0.88/0.97	0.79±0.06	0.88		0.65/0.97	0.72±0.04	0.79
A*0301	PROFWG	0.83/0.96	0.77±0.03	0.84	DRB1*0401	0.66/0.83	0.68±0.01	0.69
	BL2PSSM	0.74/0.96	0.72±0.05	0.78		0.60/0.84	0.62±0.04	0.70
A*1101	PROFWG	0.97/0.90	0.88±0.03	0.94	DRB1*0402	0.73/0.93	0.70±0.07	0.80
	PROFWG	0.87/0.91	0.81±0.06	0.87		0.72/0.93	0.72±0.04	0.79
A*2402	PROFWG	0.93/0.95	0.72±0.05	0.81	DRB1*0404	0.73/0.91	0.70±0.06	0.79
	BL2PSSM	0.86/0.94	0.57±0.07	0.65		0.68/0.95	0.61±0.05	0.67
A*3301	PROFWG	0.82/0.91	0.61±0.10	0.80	DRB1*0405	0.75/0.98	0.76±0.05	0.82
	BL2PSSM	0.45/0.91	0.50±0.06	0.75		0.77/0.98	0.82±0.04	0.85
A*6801	PROFWG	0.88/0.95	0.76±0.05	0.86	DRB1*0701	0.70/0.92	0.71±0.02	0.74
	BL2PSSM	0.92/0.94	0.72±0.07	0.82		0.65/0.95	0.72±0.04	0.75
B*0702	PROFWG	0.92/0.98	0.87±0.05	0.94	DRB1*0801	0.55/0.96	0.52±0.07	0.64
	BL2PSSM	0.92/0.98	0.73±0.05	0.82		0.45/0.97	0.52±0.06	0.63
B*2703	PROFWG	0.91/0.95	0.80±0.04	0.86	DRB1*0901	0.80/0.92	0.80±0.06	0.90
	BL2PSSM	1.00/0.95	0.77±0.08	0.91		0.79/0.95	0.78±0.06	0.87
B*2704	PROFWG	0.80/0.90	0.61±0.13	0.81	DRB1*1101	0.49/0.92	0.57±0.04	0.65
	BL2PSSM	0.60/1.00	0.61±0.09	0.80		0.47/0.91	0.54±0.04	0.61
B*2705	PROFWG	0.88/0.91	0.84±0.03	0.88	DRB1*1104	0.93/0.98	0.91±0.04	0.96
	PROFWG	0.93/0.93	0.82±0.05	0.92		0.92/0.97	0.92±0.02	0.95
B*3501	PROFWG	0.90/0.92	0.82±0.04	0.89	DRB1*1501	0.62/0.94	0.61±0.08	0.72
	BL2PSSM	0.93/0.91	0.71±0.06	0.80		0.59/0.96	0.60±0.07	0.70
B*5101	PROFWG	0.95/0.95	0.72±0.08	0.83	DRB1*0301	0.55/0.95	0.54±0.09	0.67
	BL2PSSM	1.00/0.93	0.67±0.06	0.77		0.65/0.93	0.52±0.09	0.65
B*5102	PROFWG	0.88/0.98	0.68±0.05	0.76	DRB5*0101	0.79/0.96	0.83±0.03	0.86
	BL2PSSM	0.75/0.97	0.60±0.10	0.77		0.77/0.92	0.81±0.03	0.85
B*5103	PROFWG	0.87/0.97	0.66±0.07	0.77	I-A^{k}	0.65/0.90	0.64±0.04	0.71
	BL2PSSM	0.73/0.96	0.57±0.07	0.75		0.68/0.91	0.66±0.05	0.73
B*5301	PROFWG	1.00/0.97	0.94±0.04	0.97	I-A^{d}	0.63/0.89	0.71±0.02	0.74
	BL2PSSM	0.98/0.96	0.91±0.04	0.95		0.65/0.88	0.73±0.02	0.76
B*5401	PROFWG	0.90/0.98	0.87±0.03	0.91	I-E^{d}	0.81/0.87	0.89±0.02	0.91
	BL2PSSM	1.00/0.98	0.79±0.06	0.89		0.83/0.90	0.89±0.02	0.92
H-2 K^{b}	PROFWG	0.90/0.96	0.91±0.03	0.95	I-E^{k}	0.82/0.94	0.84±0.03	0.88
	BL2PSSM	0.90/0.96	0.90±0.03	0.93		0.80/0.97	0.84±0.03	0.88
H-2Qa-2a	PROFWG	1.00/0.99	0.96±0.02	0.99	I-Ag7	0.66/0.96	0.69±0.06	0.76
	BL2PSSM	1.00/0.99	0.96±0.02	0.99		0.65/0.99	0.63±0.06	0.72
H-2D^{b}	PROFWG	1.00/0.93	0.85±0.08	0.99	I-As	0.77/0.95	0.73±0.04	0.78
	BL2PSSM	0.77/0.95	0.79±0.05	0.85		0.76/0.98	0.72±0.08	0.81
H-2L^{d}	PROFWG	0.97/0.95	0.92±0.03	0.97	I-A^{b}	0.68/0.95	0.66±0.05	0.74
	BL2PSSM	1.00/0.94	0.92±0.04	0.97		0.67/0.96	0.67±0.05	0.75
H-2K^{d}	PROFWG	0.91/0.94	0.77±0.05	0.84				
	BL2PSSM	0.87/0.95	0.80±0.05	0.89				

the binding groove (Fig. 1b). Thus, peptides bound to MHCII molecules display a great variability in length (9–22 residues) even though only a peptide core of nine residues fits into the MHCII binding groove per se. The binding mode of this peptide core is conserved among the different peptide-MHCII complexes, providing the energy that anchors the peptide to the MHCII molecule (Wang and Reinherz 2001). An important contribution to the binding energy derives from a set of conserved hydrogen bonds between the backbone of the peptide core and conserved residues in the MHCII molecules (Barber and Parham 1993; Madden 1995; Stern and Wiley 1994). As a result, the peptide binding repertoire of MHCII molecules is broader than that of MHCI molecules, and consequently peptide-MHCII ligands share less sequence similarity than peptide-MHCI ligands. Poor amino acid sequence similarity between MHCII ligands together with their great variability in sequence length makes their alignment difficult and hampers the use of global alignment algorithms such as CLUSTALW (Thompson et al. 1994a). Since alignment of the MHCII ligands requires the identification of their binding core, we have turned to the motif discovery program MEME (Bailey and Elkan 1995). MEME uses an expectation maximization algorithm in combination with a priori information regarding the nature of the motif. The a priori information we used was consistent with described structural information about the binding of peptide to MHCII molecules, namely identification of a single preferred register of peptide binding to a given MHCII molecules whose length is nine residues (see "Materials and methods" for detail).

Performance of peptide-MHC binding predictions using profiles

A single measure of the accuracy of predictive models is provided by the *AUC* value, which results from plotting SE versus 1-SP at several thresholds (See Materials and methods). We have carried out such analysis for each of the MHC molecules shown in the Table 1, and the results are provided in Table 2. For cross-validation, ROC analysis was carried out using binding test sets containing different peptides than those used for profile generation (training sets). Calculated *AUC* values varied with the actual peptides in the training and binding sets, and consequently ROC analyses were carried out using ten different training and binding sets (generated by randomly dividing the peptide binders into two different sets), and the mean *AUC* value (*AUCm*) and standard deviation is shown the Table 2. Also in Table 2 is given the best *AUC* value (*AUCb*) obtained from the above ROC analysis, along with the SE and SP values at a 3% threshold. When using PSSMs we are computing the similarity of the peptide to a set of aligned peptides known to bind to the MHC molecule, and thus variation of the peptide-MHC binding predictions (given by the *AUC* standard deviation) is linked to the overall amino acid sequence similarity between the peptides in the training and test sets. Thus, *AUCb* must result from the division of the peptide binders into a training and a test set with the best overall sequence similarity. Note that when all peptide binders are included in the profiles, *AUCb* is likely to reflect the performance of the relevant profiles for the prediction of peptide-MHC binding.

Values of *AUC*=0.5 indicate random choice, while the accuracy of predictions is poor for values of *AUC*<0.7, good for values of *AUC*>0.8% and excellent for values of *AUC*>0.9 (Swets 1988). Following the above and considering the *AUCb* values, good peptide-MHCI binding predictions can be provided (PROFILEWEIGHT or BLKWPSSM profiles) to 21 of 25 tested MHCI molecules (Table 2, *AUCb*>0.8%) and excellent peptide-MHCI binding predictions to 12 MHCI molecules (Table 2, *AUCb*>90%). Peptide binding predictions to class II MHC molecules were not as good. Nevertheless, adequate peptide-MCHII binding predictions could be provided to 11 of 24 MHCII molecules, and excellent peptide binding predictions to five MHCII molecules (HLA-DQ2, HLA-DP9, DRB1*0901, DRB1*1104, I-Ed). Peptide binding predictions provided to DRB1*0801, DRB1*1101,

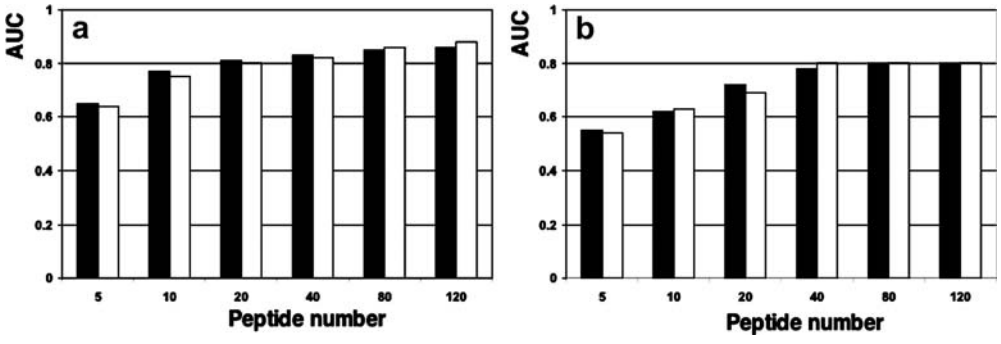

Fig. 2a, b Peptide-MHC binding predictions using PSSMs derived from alignments of different number of peptides. Performance of the peptide binding predictions to A*0201 (**a**) and DRB*0101 (**b**) computed as *AUC* values obtained from a ROC analysis where predictive profiles were derived from sets of 5, 10, 20, 40, 80 and 120 peptide binders and tested on the remaining peptide binders (Table 1). *AUC* values varied with the peptides included in the alignment (training set). The *AUC* values plotted in the figure correspond with the best value obtained after repeating the ROC analysis with 100 different training sets. Profiles were obtained using PROFILEWEIGTH (*black bars*) and BLK2PSSM (*white bars*)

DRB*0301 were poor (*AUCb*<0.7) but yet well above that of a random choice (*AUC*=0.5).

The effect of the size of the training set in the performance of the peptide binding predictions to A*0201 and DRB1*0101 is shown in Fig. 2. The results indicate that good peptide binding prediction (*AUC*>0.8) can be provided to A*0201 from matrices derived with as few as 20 peptides (Fig. 2a). On the other hand, a larger number of peptides (>40) are required to provide good peptide-binding predictions to the class II MHC molecule DRB1*0101 (Fig. 2b). It is known that MHCI ligands are more related by sequence similarity than MHCII ligands. Apparently, a large collection of MHCII ligands is needed to get a good representation of the class II peptide-binding motif. However, using this ROC analysis, a comparison of the performance of the peptide binding predictions across different MHC molecules with respect to the training set size should be approached with caution. For MHC molecules with only a few known peptide binders that are very similar to each other, the performance of the profiles would appear very good. On the other hand, if a large number of peptides are known to bind to the MHC molecule and their sequence diversity is high, the predictions might then not appear as adequate. Nevertheless, a profile derived from a large and diverse set of peptides is more likely to predict new binding peptides from a query protein than a profile derived from a few related peptides. Therefore, a more rigorous ROC analysis would need to be performed upon experimental determination of the binding of peptides predicted from a protein. A final cautionary note applies to the SP values. Due to the difficulty of finding experimentally determined non-binders, SP values were calculated from random peptides (see Materials and methods), and thus are quite high.

Prediction of MHCI-restricted and MHCII-restricted T-cell epitopes using profiles

Profiles of MHC ligands can be used in combination with the dynamic search algorithm to score and sort all peptides according to their binding potential to the relevant MHC molecule (see Materials and methods). Only peptides that bind to MHC with an affinity above a necessary threshold are able to elicit a T-cell response. Consequently, if PSSMs are good predictors of peptide-MHC binding, T-cell epitopes should be expected among the high scoring peptides within their protein sources. We checked the validity of this notion for the set MHC molecules shown in Table 1. All prediction tests were carried out using PSSMs that were generated using PROFILEWEIGHT (Thompson et al. 1994b), which uses branch-proportional sequence

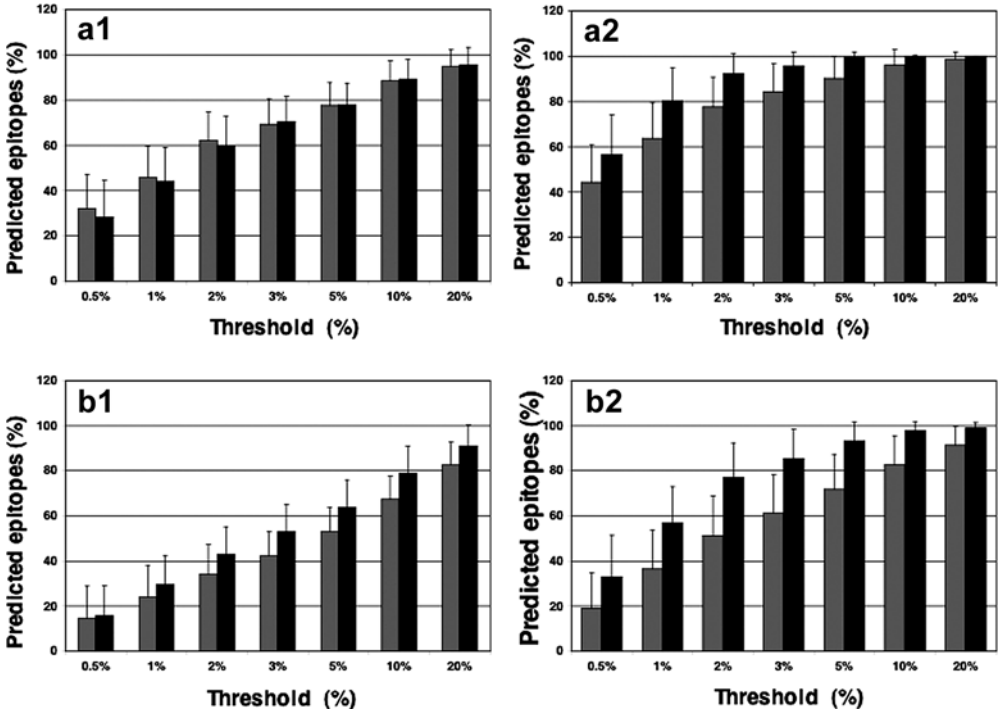

Fig. 3a, b Prediction of MHCI-restricted and MHCII-restricted epitopes using PSSMs. Prediction of peptide-MHC binding using PSSMs was evaluated for MHCI (**a1,2**) and MHCII (**b1,2**) molecules by scoring known MHC-restricted epitopes from their protein sources. MHC molecules targeted for peptide-MHC binding prediction are those shown in Table 1. Predictions were carried out at different thresholds (*abscissa*) [(percentage of top scoring peptides)]. A peptide is computed as predicted if it is found among the top scoring peptides (set by the threshold) from its protein source using the relevant PSSM, and the percentage of correctly predicted peptides is plotted in the figure (ordinate). Predictions were carried out using PSSMs derived from alignments that did (**a2, b2**) or did not (**a1, b1**) contain the peptide tested, and PSSMs were generated from these alignments using PROFILE-WEIGHT which uses a branch-proportional weighting method (*gray bars*) and BLK2PSSM under position-based weighting method (*black bars*). Given that several MHC molecules were targeted for peptide binding predictions (Table 1), plotted values correspond to the mean and standard deviation of percentage of properly predicted epitopes for all MHC molecules examined

weights, and BLK2PSSM in combination with position-based based weights (Henikoff and Henikoff 1994; Henikoff et al. 1999). The percentage of correctly predicted MHC-restricted epitopes at several thresholds of top scoring peptides (0.5, 1, 3, 5, 10, 20%) using the relevant PSSMs are shown in Fig. 3. Figure 3a illustrates predictions of MHCI-restricted epitopes (CD8 T-cell epitopes) using PSSMs generated from alignments without (Fig. 3a1) or with (Fig. 3a2) the epitope evaluated as a binder of a given MHC molecule. Likewise, Fig. 3b corresponds to predictions of MHCII-restricted epitopes (CD4 T-cell epitopes) using PSSMs generated from alignments without (Fig. 3b1) or with (Fig. 3b2) the targeted epitope. Results indicate that on average over 80% of CD8 T-cell epitopes were predicted at a 2% threshold (predictions with all peptides included in the alignment), whereas up to a 3–10% threshold was required to predict 80% of CD4 T-cell epitopes (predictions with all peptides included in alignment). Also, as reported elsewhere (Reche et al. 2002), for the prediction of MHCI-restricted epitopes we see no clear differences between results obtained from PSSMs derived with BLK2PSSM and PROFILEWEIGHT. When all peptides are included in the alignment, PSSM generated with BLK2PSMM gave better results (Fig. 3a2), but if the epitopes to be predicted are not included in the alignment, then PSSMSs generated from PROFILEWEIGHT gave slightly better results. On the other hand, for the prediction of MHCII-restricted epitopes PSSMs obtained with BLK2PSSM and a position-based weighting scheme gave better results independently of whether the epitopes to be tested were included in the alignment (Fig. 3b1,2). It is also important to note that there is some variability between the prediction tests obtained for the different MHC molecules targeted in this test (indicated by standard deviation in Fig. 3). Moreover, variability is greater in tests concerning the prediction of MHCII-restricted epitopes. This indicates that each PSSM has a different threshold of top scoring peptides at which known binders appear. Thus, for each PSSM we defined a PSSM-specific binding threshold (PSBT) as the score value that includes 85% of the peptides from which that PSSM was obtained.

Prediction of proteasomal cleavage using statistical language modeling tools

Statistical language modeling is the science of building probabilistic models from word strings. Language models including n-gram models are most frequently applied in speech recognition and natural language tagging (Rosenfeld 2000), but have also been applied to the sequence analysis and motif identification (Jimenez-Montano et al. 2002; Wu and Shivakumar 1994; Wu et al. 1996). Cleavage by the proteasome occurs at preferential sites within the protein, and sequence signals from antigenic peptides processed by the proteasome are especially conserved at position P1 of the cleavage site (the C-terminus of antigenic peptide) and its immediate flanking P1' residue (Altuvia and Margalit 2000). Prediction of proteasomal cleavage resembles the problem of language tagging (modeling the location of grammatic tags such as punctuation signs) and thus we have used the SRILM toolkit (Stolcke 2002) for statistical modeling of proteasomal cleavage sites. Training sets for statistical modeling of proteasomal cleavage were obtained from a database containing the C-terminus and flanking regions of 332 antigens restricted by human MHCI molecules (See

Table 3 Proteasomal cleavage prediction results using representative LMPCPs at different thresholds. LMPCPs were obtained using NGRAM-COUNT and tested using HIDDEN-NGRAM at different probability thresholds (*Pro*). *N*–2 LMPCP models were produced by varying the order of the model (*i*) from 2 to *N*–1, and the best model at each threshold is shown in this table. *N* size of the peptides fragments in the training and testing sets; *PCS* predicted cleavage sites; *ECS* expected cleavage sites calculated according to Eq. 3 (Materials and methods)

Model[a]	Frag. size (*N*)	Oder (i)	Pro[b]	PCS (%)	ECS (%)	PCS−ECS (%)	Mean size[c]
LMPCP$_{10}^{2}$(0.10)	10	2	0.10	87.6	39.2	48.4	2.84
LMPCP$_{2}^{8}$ (0.15)	8	2	0.15	80.3	37.4	42.9	2.95
LMPCP$_{6}^{6}$ (0.20)	6	6	0.20	73.9	37.2	36.7	2.98
LMPCP$_{6}^{6}$ (0.25)	6	6	0.25	66.7	31.9	34.8	3.43
LMPCP$_{4}^{2}$ (0.30)	4	2	0.30	82.1	47.6	34.5	2.48
LMPCP$_{4}^{2}$ (0.35)	4	2	0.35	78.6	43.1	35.5	3.07
LMPCP$_{4}^{2}$ (0.40)	4	2	0.40	75.7	40.0	35.7	3.31
LMPCP$_{4}^{2}$(0.45)	4	2	0.45	71.4	36.2	35.2	3.92
LMPCP$_{4}^{2}$ (0.50)	4	2	0.50	66.4	33.1	33.3	4.60
LMPCP$_{4}^{2}$ (0.55)	4	2	0.55	59.3	28.8	30.5	5.15
LMPCP$_{4}^{2}$ (0.60)	4	2	0.60	50.0	23.8	26.2	6.11
LMPCP$_{4}^{2}$ (0.65)	4	2	0.65	49.3	21.9	27.4	6.99
LMPCP$_{4}^{2}$(0.70)	4	2	0.70	47.9	20.5	27.4	7.77

[a]The models that were implemented in the RANKPEP web server are shown in *bold*
[b]Probability above which a cutpoint is predicted
[c]Mean size of the fragments yielded by the relevant LMPCP when tested over the full length proteins bearing the peptide fragments of the testing set

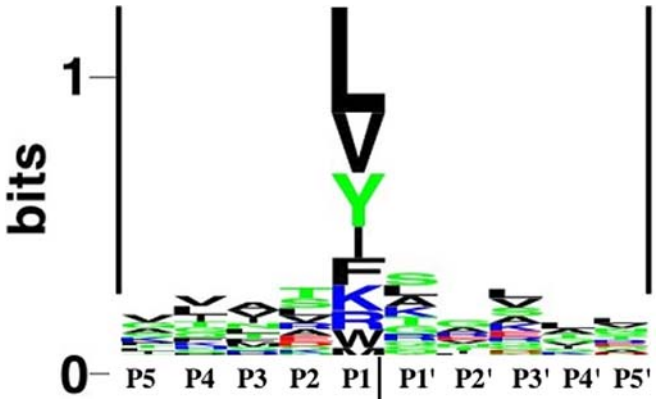

Fig. 4 Sequence logo of peptide fragments containing the C-terminal end of MHCI-restricted peptides. The sequence logo was built as indicated from a collection of peptide fragments containing the C-terminus of 332 human MHCI-restricted epitopes and their flanking regions (see Materials and methods). The C-terminus of each epitope corresponds to the P1 proteasomal cleavage site (shown by the *vertical bar*)

Materials and methods for details). Sequence information for this initial data set is shown in Fig. 4. As noted by others, P1 followed by P1′ are the positions with the largest information content (Altuvia and Margalit 2000; Kesmir et al. 2002). Significant sequence information is also found in the P3′ and P2 positions. LMPCP were based on n-gram statistics (Rosenfeld 2000) and created from training sets of peptide fragments of variable fragment length derived from the above database. LMPCP were tested at different cutpoint probabilities (0.1–0.70) using HIDDEN-NGRAM with testing files of peptide fragments not included in the training sets as well as with their entire protein sources. For each probability threshold, 23 different $LMPCP_N^i$ [N=10, 6, 4; i=1−(N−2)] were tested, where N is the fragment size of fragment in training and testing sets and i is the order of the LMPCP tested (See Materials and methods for details). The LMPCPs were ranked with regard to the percentage of correctly PCS minus the percentage of ECS (Eq. 3 in Materials and methods). The best LMPCP at each threshold is shown in Table 3 along with the indicated mean size of the fragment yield by the model. PCS varied with the relevant LMPCP, and was always under 50% if the cutpoint probability threshold was set above 0.7. Furthermore, LMPCPs from fragments of size four (two residues at each side of the cleavage site) were the best if the threshold cutpoint was above 0.3. PCS were under 50% if the models were generated from training sets containing fragment sizes longer than ten (for example, six residues to each side of the cleavage site) for all the cutpoint probability thresholds tested (0.1–0.7) (data not shown). With the exception of $LMPCP_6^6$ (0.25), PCS were above 70% only under soft cutting probability (0.1–0.4) (Table 3), indicating that the nature of the proteasome specificity is much less rigid than that of grammar tagging for which these languages models were originally intended. Yet, PCS by these models exceed the ECS by at least 30% (Table 3). The average size of the length of the peptide fragments produced by LMPCPs varied, with those providing a highest PCS yielding smaller fragments (around three residues) (Table 3).

Web implementation

Predictions of peptide-MHCI binding using PSSMs are available on-line from the Molecular Immunology Foundation web server hosted by Dana-Farber Cancer Institute (http://www.mifoundation.org/Tools/rankpep.html). The server consists of a set of python and perl scripts that handle the input, combine the prediction of peptide-MHC binding and proteasomal cleavage and serve the output over the Internet. The interface to the server, shown in Fig. 5a, is divided in six major sections: PSSMs, INPUT, THRESHOLD, PROTEASOME CLEAVAGE, and ADVANCED OPTIONS. The PSSM section includes a selection of 88 MHCI-specific (81 human and seven mouse), and 50 MHCII-specific (38 human and 12 mouse) PSSMs for the prediction of peptide binding. Alternatively, the users can input their own PSSMs for the prediction of peptide-MHC binding. Optional PSSMs included in the server for the prediction of MHCI-peptide binding are those obtained using PROFILEWEIGHT (Thompson et al. 1994b), whereas PSSMs for the prediction of peptide binders to MHCII are those obtained using BLK2PSSM (Henikoff et al. 1999) with position-based weights (Henikoff and Henikoff 1994)(see Results). The INPUT query for the prediction of peptide binders to MHC molecules can be sequence(s) in FASTA format or an MSA. If an MSA is entered, the server creates a consensus sequence in which the variable positions are masked (see Materials and methods), and prediction of peptide-MHC binding is restricted to the conserved regions. Default variability for masking is 1.0 (positions with a variability above 1.0 will be masked). Roughly, a position in the alignment with a variability under 1.0 is either occupied by a prevalent residue (around 90% of the residues are identical) or by two different residues equally represented (~50% each). The variability threshold can be set to other values in the ADVANCED OPTIONS section. Values must range between 0 and 4.3 to be consistent with Eq. 4 (Materials and methods). Using the THRESHOLD options, peptides can be sorted by the number of top scoring peptides or the percentage of top scoring peptides. Filtering the sorted peptides by molecular weight can also be done using the ADVANCED OPTIONS section. Models for the prediction of proteasomal cleavage are selected in the PROTEASOME CLEAVAGE section. The current models (highlighted in Table 3) include $LMPCP_{10}^2$ (0.1) (option 1), $LMPCP_4^2$ (0.45) (option 2), and $LMPCP_2^4$ (0.7) (option 3). The RANKPEP result page (Fig. 5b) displays the PSSM selected, the optimum sequence (consensus) for that PSSM, i.e., the sequence that gives the highest score, and a list of peptides whose number is determined in the THRESHOLD section and ordered by score. For every sorted peptide, the server also outputs its molecular weight, and its relative score in percentage to that of optimum score. Peptides whose

scores are equal or greater than the PSBT score will be highlighted in red and, when predicting MHCI-binding peptides, those containing a C-terminal end predicted to be the result of proteasomal cleavage are shown in violet (Fig. 5b).

Discussion

Prediction of peptide-MHC binding using profiles: selection of PSSMs and thresholds

PSSMs or profiles are useful for representing sequence motifs. Indeed, popular databases such as BLOCK (Henikoff et al. 1999), PROSITE (Hofmann et al. 1999), and IMPALA (Schaffer et al. 1999) rely on motif profiles for the functional classification of new sequences via their similarity to these profiles. Similarly, PSSMs of peptides known to bind to MHC can be used for the identification/prediction of peptide-MHC binders. We first applied this idea to the prediction of MHCI-peptide binding (Reche et al. 2002). Here we have extended the method to the prediction of peptide-MHCII binding. PSSMs for the prediction of both MHCI and MHCII binding are now available at the RANKPEP web site (http://www.mifoundation.org/Tools/rankpep.html). PSSMs for the prediction of peptide-MHCI binding have been derived from ungapped alignments of peptides of the same length. Since MHCI molecules can bind peptides between eight and 11 residues in length, several PSSMs might be available at the RANKPEP web server for the independent prediction of 8mer, 9mer, 10mer or 11mer peptide binders to a given MHCI molecule. Most of the known MHCI-restricted peptides are 9mers (~90%)(data not shown), and therefore, in the absence of a certain preference for a given size, we suggest selecting PSSMs for the prediction of 9mer peptide binders. PSSMs for the peptide-MHCII binders always target the 9mer core of the peptide binders.

Overall, PSSMs for prediction of peptide-MHCI binding and MHCI-restricted epitopes are more sensitive than those used for the prediction of peptide-MHCII binding and MHCII-restricted epitopes. Consequently, a higher threshold is required to predict a similar percentage of epitopes (Table 2, Fig. 3) This result does not necessarily indicate that PSSMs specific for the prediction of MHCII were derived from incorrect alignments, but rather could reflect the greater structurally inherent peptide binding promiscuity of MHCII molecules (see Results). Indeed,

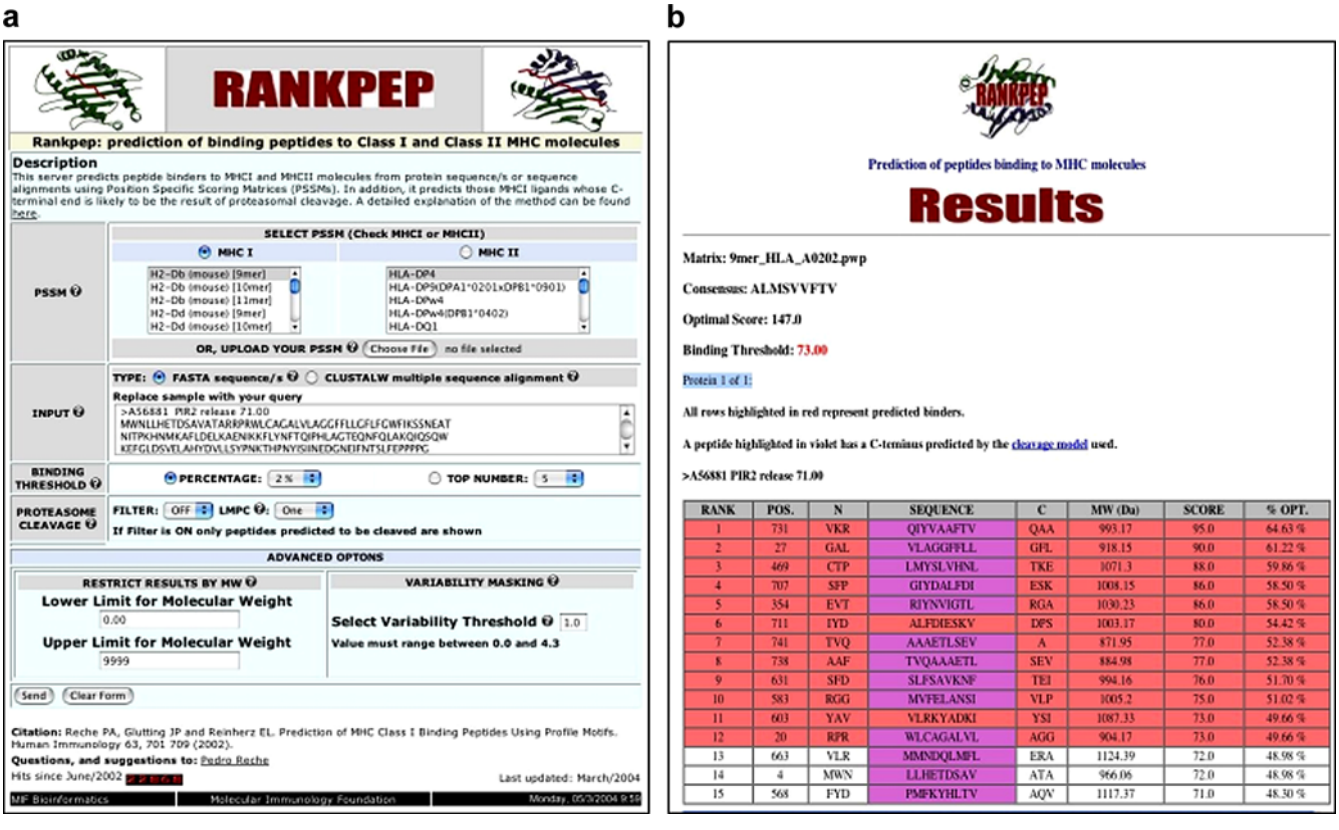

Fig. 5a, b The RANKPEP web server. **a** RANKPEP input page. The page is divided into several sections: PSSM for the selection of MHCI-specific and MHCII-specific matrices; INPUT; THRESHOLD, and PROTEASOME CLEAVAGE as discussed in Results; ADVANCE OPTIONS include filtering the peptide results by the molecular weight (MW) of peptides, and selection of a variability threshold (0–4.3) to mask sequence variability from inputs in the form of multiple sequence alignment. **b** RANKPEP result page. The page lists a number of peptides from the query given at a selected threshold. Also indicated in the result page is the PSSM selected and the binding threshold of the PSSM. Peptides whose scores are above the PSBT are shown in *red*. Peptides shown in *violet* contain a C-terminal residue that is predicted to be the result of proteasomal cleavage. If the proteasomal cleavage filter is checked ON, only *violet* peptides will be shown. Proteasomal cleavage options are only applied to the prediction of MHCI-restricted peptides

from the available crystal structures of peptide-MHCII complexes, we determined that the peptide binding core fitting onto the MHC groove was properly defined in the relevant alignments from which the PSSMs were derived (data not shown). PSSMs in RANKPEP are associated with a specific binding threshold above which sorted peptides are highlighted in the results page (Fig. 5b). Since PSBT is variable with regard to the sequence space diversity of the aligned peptides, both overestimation and underestimation of the number of peptides that are predicted to bind to a given MHCI molecule can occur. Therefore, following the results summarized in Table 2, Fig. 3, we recommend using a 2–3% binding threshold of top scoring peptides for the prediction of MHCI-restricted peptides, and a 4–6% threshold for the prediction of MHCII-restricted peptides.

Prediction of peptide-MHC binding using profiles: comparison with other methods

Prediction of peptide-MHC binding is important for the anticipation of T-cell epitopes and so determination of peptides that can bind to MHC molecules has been approached by a large array of methods including sequence patterns (Sette et al. 1989), motif-matrices (De Groot et al. 1997; Rammensee et al. 1999), quantitative matrices (QM) (Guan et al. 2003; Hammer et al. 1994; Parker et al. 1994; Stryhn et al. 1996; Udaka et al. 2000), virtual quantitative matrices (VQM) (Raddrizzani and Hammer 2000; Sturniolo et al. 1999), artificial neural networks (ANN) (Adams and Koziol 1995; Brusic et al. 1998a; Gulukota et al. 1997; Honeyman et al. 1998); hidden Markov motifs (HMM) (Mamitsuka 1998; Udaka et al. 2002); structural peptide threading (SPT) (Altuvia et al. 1997; Schueler-Furman et al. 2000; Swain et al. 2001), support vector machine (SVM) algorithms (Donnes and Elofsson 2002; Zhao et al. 2003) and stepwise discriminant analysis meta-algorithm (SDA) (Mallios 1999). QM and VQM methods are derived from actual binding experiments, whereas SPT is an entirely computer-based method that relies on the evaluation of peptide fit into the binding groove, and despite its great potential is currently still under development. On the other hand, techniques such as sequence patterns, motif-matrices, ANN, HMM, SVM, and SDA algorithms rely on the analysis of the sequences of peptides that are experimentally known to bind. Prediction of peptide-MHC binding using PSSMs lies within the motif-matrices methods, although in previous methods the matrices coefficients were adjusted either manually (Rammensee et al. 1999) or were not specified (De Groot et al. 1997). In any case, motif-matrices are a more accurate predictor of peptide-MHC binding than simple single sequence patterns (Reche et al. 2002). Most of these methods have been applied to both the prediction of peptide-MCHI and peptide-MHCII binding, and as occurs with the use of PSSMs, the success of the predictions seems to be greater for the prediction of peptide-MHCI binding than peptide-MHCII binding. In terms of accuracy (a balance between sensitivity and specificity) ANN and HMM have been reported to be best predictors of peptide-MHC binding, perhaps because they can model binding interferences, positive or negative, between the side chains of the peptides. Other methods assume independent binding of each side chain. Nevertheless, independent binding is generally the case, as supported by experimental evidence (Parker et al. 1994; Sturniolo et al. 1999). Indeed, in a recent study of the independent binding assumption for binding of peptides epitopes to MHCI molecules, there was only marginal improvement when sidechain pair interactions were introduced into the motif-matrix predictor (Peters et al. 2003). Furthermore, the accuracy of our profiles, given by the *AUC* value (Table 2), is similar to that reported by ANN and HMM methods. However, an objective comparison between these methods should be done upon experimental determination of the binding of peptides predicted from a protein query. It is in fact revealing that in practicum, only 30–50% of predicted peptides from query proteins turn out to be significant binders, independently of the method used. In the absence of such experimental testing, a rigorous computer-based comparison of the various peptide-MHC binding prediction is not straightforward, as the various methods have been training with different set of data and the results are dictated by the chosen test peptides. Thus, in this paper we have also determined whether known T-cell epitopes can be predicted from their protein sources using realistic thresholds. In this scenario, we find that using PSSMs around 80% of the known MHCI-restricted and MHCII-restricted epitopes appear among the top 3% and 5% scoring peptides, respectively (Fig. 3a,b).

Examples of online web servers for the predictions of peptide-MHC binding are available for most of the methods discussed above (see Table 1 in Guan et al. 2003). However, all these sites are for the prediction of peptide binding to either MHCI or MHCII molecules alone. The exception is the SYFPEITHI web site (Rammensee et al. 1999) that contain matrices for the prediction of peptide-MHCI and peptide-MHCII binding, but therein, the number of MHCII molecules that can be targeted for peptide prediction is very limited. The RANKPEP web site contains the largest set of predictors for the anticipation of peptide-MHCI and peptide-MHCII binding (88 and 50 PSSMs for targeting peptide binding predictions to independent MHCI and MHCII molecules, respectively).

Antigen processing: prediction of proteasomal cleavage using statistical language models

Antigen processing occurs prior to MHC binding, thus determining the pool of peptides that can become T-cell epitopes. CD8 T-cell epitopes, and MHCI-restricted peptides in general, derive from protein fragments generated by the protease activity of the proteasome. Protein fragments thus generated are substrates for amino-

peptidases that destroy most of the fragments. Nevertheless, a few peptides ranging between eight and 15 residues in length are translocated to the endoplasmic reticulum (ER) by the TAP transporter, where they can either be destroyed by an additional amino-peptidase or be rescued by binding to MHCI molecules (Pamer and Cresswell 1998; Rammensee 2002; Serwold et al. 2002). Thus, the N-terminus of any class I restricted peptide is shaped by activity of several amino-peptidases with the resulting loss of information, whereas the C-terminus is the result of the original proteasomal cleavage (Craiu et al. 1997). The proteasome is a multi-enzyme complex, whose catalytic subunits, and resulting specificity, change in the presence of IFN-γ. The form in the absence of IFN-γ is referred to as the constitutive proteasome, whereas the form in the presence of IFN-γ is known as the immunoproteasome (Fruh and Yang 1999; Toes et al. 2001). Although subject to debate, it is believed that the immunoproteasome is responsible for the generation of CD8 T-cell eptiopes (Chen et al. 2001; Toes et al. 2001; van Hall et al. 2000). Therefore, to increase the immunological relevance of our study we have modeled the specificity of the proteasome from a set of known CD8 T-cell epitopes and their flanking regions using statistical language models. Three of these models (LMPCP) [(Table 3)] are now implemented in the RANKPEP web site to predict whether the C-terminus of a given peptide might result from proteasome activity. The default model [model one: $LMPCP_{10}^{2}$ (0.1)] predicted about 85% of the cleavage sites (Table 3), providing the largest increase of PCS over the ECS of all tested models (48.4%). In a genome-wide characterization of CD8 T-cell epitopes from influenza virus in mouse, Zhong et al. (Zhong et al. 2003) proved that the combined use of this LMPCP can reduce the list of peptide-MHCI binders by ~30% without compromising the number of peptides that are true T-cell epitopes. The average fragment length yielded by $LMPCP_{10}^{2}$ (0.1) is, however, much smaller (~3 residues) than that experimentally determined for the proteasome (of 7–9 residues) (Kisselev et al. 1999; Toes et al. 2001), suggesting that the specificity of this particular LMPCP is rather low (many false positives). The smaller fragment size yield by $LMPCP_{10}^{2}$ (0.1) could also reflect the clustering and consequent overlap of epitopes observed within protein regions [(Meister et al. 1995); our own unpublished observations from the HIV CTL database in Los Alamos; url: http://www.hiv.lanl.gov/]. Nevertheless, to anticipate the possibility of rather low specificity of model $LMPCP_{10}^{2}$ (0.1), RANKPEP also provides two additional models, $LMPCP_{4}^{2}$ (0.45) and $LMPCP_{4}^{2}$ (0.7), which although less sensitive (Table 2), produce larger fragments and thereby are expected to be more specific. In particular, $LMPCP_{4}^{2}$ (0.7) yields peptide fragments with an average size (~8 residues) that is consistent with that thought to be generated by the proteasome (Table 3). However, it is important to note that LMPCP are not meant to predict proteasome fragmentation patterns, but to indicate whether the C-terminus of a peptide can result from proteasomal cleavage. Finally, there is an extra benefit of combining LMPCP with peptide-MHCI binding prediction using PSSMs. It is known that the C-terminal position of the peptide is always an anchor residue (Fig. 4). Note that prediction of peptide-MHCI using PSSMs assumes an independent contribution of each residue, and there are occasions in which top ranking peptides may contain a C-terminus that is not likely to be an anchor residue. In this scenario, the coupled usage of LMPCP will help to discard those peptides, thereby improving the MHCI-binding prediction. Moreover, some valuable information about the TAP transport of peptides into the ER may also have been incorporated into our LMPCP. TAP transport is essential for epitope generation, and our LMPCP models were trained using known epitopes.

A similar approach for modeling the proteasome cleavage site from known MHCI-restricted T-cell epitopes using ANN has already been reported (Kesmir et al. 2002). Relative to the language model herein, ANN produced larger fragments (~9 residues) more consistent with those thought to be generated by the proteasome. However, the training set consisted of peptides of up to 18 residues (9 residues at each side of the C-terminus end), and therefore the result would be biased toward the average length of the epitope, as there is significant sequence information with regard to the background along the entire length of the epitope. Other approaches for modeling the proteasome cleavage site include the analysis of the fragmentation patterns of a given protein with purified constitutive proteasome cores (Holzhutter and Kloetzel 2000; Kesmir et al. 2002; Kuttler et al. 2000). The biological relevance of the data generated from these important studies might be limited due to the fact the degradation was mediated by the proteasome rather than the immunoproteasome.

Processing of MHCII-restricted ligands relies mainly on exogenous proteins that are directed to the endosomal compartment, where they are degraded by the action of several endo-peptidases as well as by amino-peptidases and carboxy-peptidases (Pieters 2000; Watts 2001). This processing complexity together with the fact that MHCII molecules bind peptides of different yet overlapping lengths, makes the generation of models for the prediction of MHCII-antigen processing difficult. Nevertheless, recent reports indicate the existence of conserved regions flanking the core CD4 T-cell epitopes that are related to antigen processing rather than peptide-MHC interaction (Sant'Angelo et al. 2002). Moreover, some reports argue that they may contribute to immunogeneicity as well (Carson et al. 1997). Thus, when anticipating MHCII-restricted T-cell epitopes using RANKPEP, we suggest considering peptides consisting of the predicted 9mer binding cores plus the three most proximal amino acids flanking their N-terminal and C-terminal ends.

Conclusions

Engagement of both CD8 and CD4 T-cells is desirable for mounting a strong defensive immune response against

cancer cells and pathogens. Since antigen processing and presentation by MHCI and MHCII molecules differ, prediction of T-cell epitopes requires the development of bioinformatics tools that are able to cope with this complexity. To this end, our RANKPEP server represents a powerful tool that allows: (1) the prediction of peptide binding to MHCI and MHCII molecules using motif profiles; (2) greater specificity of CD8 T-cell epitope identification through combined proteasomal cleavage site prediction; and (3) prediction of conserved epitopes from MSAs. Finally, RANKPEP is a versatile and flexible web server, providing many sorting options and the possibility of using custom built matrices for prediction of peptide-MHC binding.

Acknowledgments This manuscript was supported by NIH grant AI50900 and the Molecular Immunology Foundation. We wish to acknowledge the insightful comments and corrections provided by Drs Esther Lafuente, Robert Mallis, and Weimin Zhong.

References

Adams HP, Koziol JA (1995) Prediction of binding to MHC class I molecules. J Immunol Methods 185:181–190

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

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

Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H (1997) A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. Hum Immunol 58:1–11

Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3:21–29

Barber LD, Parham P (1993) Peptide binding to major histocompatibility complex molecules. Annu Rev Cell Biol 9:163–206

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. Nucleic Acids Res 31:23–27

Brusic V, Rudy G, Honeyman JH, Harrison LC (1998a) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neuronal network. Bioinformatics 14:121–130

Brusic V, Rudy G, Kyne AP, Harrison LC (1998b) MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 26:368–371

Carson RT, Vignali KM, Woodland DL, Vignali DA (1997) T-cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. Immunity 7:387–399

Chen W, Norbury CC, Cho Y, Yewdell JW, Bennink JR (2001) Immunoproteasomes shape immunodominance hierarchies of antiviral CD8(+) T-cells at the levels of T-cell repertoire and presentation of viral antigens. J Exp Med 193:1319–1326

Craiu A, Akopian T, Goldberg A, Rock KL (1997) Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. Proc Natl Acad Sci USA 94:10850–10855

De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G (1997) An interactive web site providing major histocompatibility ligand predictions: application to HIV research and AIDS. AIDS Res Hum Retroviruses 13:529–531

Donnes P, Elofsson A (2002) Prediction of MHC class I binding peptides, using SVMHC. BMC Bioinform 529–531:25

Draenert R, Altfeld M, Brander C, Basgoz N, Corcoran C, Wurcel AG, Stone DR, Kalams SA, Trocha A, Addo MM, Goulder PJ, Walker BD (2003) Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T-cell responses. J Immunol Methods 275:19–29

Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296

Fruh K, Yang Y (1999) Antigen presentation by MHC class I and its regulation by interferon gamma. Curr Opin Immunol 11:76–81

Garcia KC, Teyton L, Wilson IA (1999) Structural basis of T-cell recognition. Annu Rev Immunol 17:369–397

Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355–4358

Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCPred: a server for quantitative prediction of peptide-MHC binding. Nucleic Acids Res 31:3621–3624

Gulukota K, Sidney J, Sette A, DeLisi C (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 267:1258–1267

van Hall T, Sijts A, Camps M, Offringa R, Melief C, Kloetzel PM, Ossendorp F (2000) Differential influence on cytotoxic T lymphocyte epitope presentation by controlled expression of either proteasome immunosubunits or PA28. J Exp Med 192:483–494

Hammer J (1995) New methods to predict MHC-binding sequences within protein antigens. Curr Opin Immunol 7:263–269

Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F (1994) Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. J Exp Med 267:1258–1267

Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243:574–578

Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci 12:135–143

Henikoff S, Henikoff JG, Pietrokovski S (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 15:471–479

Hennecke J, Carfi A, Wiley DC (2000) Structure of a covalently stabilized complex of a human alphabeta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. EMBO J 19:5611–5624

Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Res 27:215–219

Holzhutter HG, Kloetzel PM (2000) A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. Biophys J 79:1196–1205

Honeyman MC, Brusic V, Stone NL, Harrison LC (1998) Neural network-based prediction of candidate T-cell epitopes. Nat Biotechnol 16:966–969

Jimenez-Montano MA, Ebeling W, Pohl T, Rapp PE (2002) Entropy and complexity of finite sequences as fluctuating quantities. Biosystems 64:23–32

Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S (2002) Prediction of proteasome cleavage motifs by neural networks. Protein Eng 15:287–296

Kisselev AF, Akopian TN, Woo KM, Goldberg AL (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 274:3363–3371

Kuttler C, Nussbaum AK, Dick TP, Rammensee HG, Schild H, Hadeler KP (2000) An algorithm for the prediction of proteasomal cleavages. J Mol Biol 298:417–429

Madden DR (1995) The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 13:587–622

Madden DR, Garboczi DN, Wiley DC (1993) The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. Cell 75:693–708
Maenaka K, Jones EY (1999) MHC superfamily structure and the immune system. Curr Opin Struct Biol 9:745–753
Mallios RR (1999) Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 15:432–439
Mamitsuka H (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins 33:460–474
Margulies DH (1997) Interactions of TCRs with MHC-peptide complexes: a quantitative basis for mechanistic models. Curr Opin Immunol 9:390–395
Matsumura M, Fremont D, Peterson PA, Wilson IA (1992) Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science 257:927–934
Meister GE, Roberts CG, Berzofsky JA, De Groot AS (1995) Two novel T-cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. Vaccine 13:581–591
Nicholls A, Sharp K, Honig B (1991) Protein folding and association insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11:281–296
Pamer E, Cresswell P (1998) Mechanisms of MHC class I—restricted antigen processing. Annu Rev Immunol 16:323–358
Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. J Immunol 152:163–175
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. Bioinformatics 19:1765–1772
Pieters J (2000) MHC class II-restricted antigen processing and presentation. Adv Immunol 75:159–208
Raddrizzani L, Hammer J (2000) Epitope scanning using virtual matrix-based algorithms. Brief Bioinform 1:179–189
Rammensee HG (2002) Survival of the fitters. Nature 419:443–445
Rammensee HG, Friede T, Stevanoviic S (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41:178–228
Rammensee HG, Bachmann J, Emmerich NPN, Bacho OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219
Reche PA, Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. J Mol Biol 331:623–641
Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63:701–709
Rosenfeld R (2000) Two decades of statistical language modeling: where do we go from here? Proc IEEE 88:1–11
Sant'Angelo DB, Robinson E, Janeway CA Jr, Denzin LK (2002) Recognition of core and flanking amino acids of MHC class II-bound peptides by the T-cell receptor. Eur J Immunol 32:2510–2520
Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011
Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097–6100
Schueler-Furman O, Altuvia Y, Sette A, Margalit H (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 9:1838–1846
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
Sette A, Buus S, Appella E, Smith JA, Chesnut R, Miles C, Colon SM, Grey HM (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. Proc Natl Acad Sci USA 86:3296–3300
Shannon CE (1948) The mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
Stern LJ, Wiley DC (1994) Antigen peptide binding by class I and class II histocompatibility proteins. Structure 2:245–251
Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, Litwin S (1997) A Shannon entropy analysis of immunoglobulin and T-cell receptor. Mol Immunol 34:1067–1082
Stolcke A (2002) SRILM—an extensible language modeling toolkit. In: Ohala TMNJJ, Derwing BL, Hodge MM, Wiebe GE (eds) Proceedings of the International Conference of Spoken Language Processing. Center for Spoken Language Research, Boulder, pp 901–904
Stryhn A, Pederson LO, Romme T, Holm A, Buus S (1996) Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. Eur J Immunol 26:1911–1918
Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Sinigaglia F, Hammer J (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nature Biotech 17:555–561
Swain MT, Brooks AJ, Kemp GJL (2001) An automated approach to modelling class II MHC alleles and predicting peptide binding. Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering. IEEE Computer Society, New York, pp 81–88
Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293
Thompson JD, Higgins DG, Gibson TJ (1994a) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weigh matrix choice. Nucleic Acids Res 2:4673–4680
Thompson JD, Higgins DG, Gibson TJ (1994b) Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput Appl Biosci 10:19–29
Toes RE, Nussbaum AK, Degermann S, Schirle M, Emmerich NP, Kraft M, Laplace C, Zwinderman A, Dick TP, Muller J, Schonfisch B, Schmid C, Fehling HJ, Stevanovic S, Rammensee HG, Schild H (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. J Exp Med 194:1–12
Udaka K, Wiesmuller KH, Kienle S, Jung G, Tamamura H, Yamigishi H, Okumura K, Walden P, Suto T, Kawasaki T (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. Immunogenetics 51:816–828
Udaka K, Mamitsuka H, Nakaseko Y, Abe N (2002) Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides. J Immunol 169:5744–5753
Wang J-H, Reinherz E (2001) Structural basis of T-cell recognition of peptides bound to MHC molecules. Mol Immunol 38:1039–1049
Watts C (2001) Antigen processing in the endocytic compartment. Curr Opin Immunol 13:26–31
Wu C, Shivakumar S (1994) Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. Nucleic Acids Res 22:4291–4299
Wu CH, Zhao S, Chen HL, Lo CJ, McLarty J (1996) Motif identification neural design for rapid and sensitive protein family search. Comput Appl Biosci 12:109–118
Zhang C, Anderson A, DeLisi C (1998) Structural principles that govern the peptide-binding motifs of class I MHC molecules. J Mol Biol 281:929–947
Zhao Y, Pinilla C, Valmori D, Martin R, Simon R (2003) Application of support vector machines for T-cell epitopes prediction. Bioinformatics 19:1978–1984
Zhong W, Reche PA, Lai CC, Reinhold B, Reinherz EL (2003) Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. J Biol Chem 278:45135–45144
Zinkernagel RM, Doherty PC (1974) Restriction of in vitro T-cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature 248:701–702