Immunogenetics (2004) 56: 311–320
DOI 10.1007/s00251-004-0691-0

ORIGINAL PAPER

Nigel J. Burroughs · Rob J. de Boer · Can Keşmir

Discriminating self from nonself with short peptides from large proteomes

Received: 24 February 2004 / Revised: 26 May 2004 / Published online: 30 July 2004

Abstract We studied whether the peptides of nine amino acids (9-mers) that are typically used in MHC class I presentation are sufficiently unique for self:nonself discrimination. The human proteome contains 28,783 proteins, comprising 10^7 distinct 9-mers. Enumerating distinct 9-mers for a variety of microorganisms we found that the average overlap, i.e., the probability that a foreign peptide also occurs in the human self, is about 0.2%. This self:nonself overlap increased when shorter peptides were used, e.g., was 30% for 6-mers and 3% for 7-mers. Predicting all 9-mers that are expected to be cleaved by the immunoproteasome and to be translocated by TAP, we find that about 25% of the self and the nonself 9-mers are processed successfully. For the *HLA-A*0201* and *HLA-A*0204* alleles, we predicted which of the processed 9-mers from each proteome are expected to be presented on the MHC. Both alleles prefer to present processed 9-mers to nonprocessed 9-mers, and both have small preference to present foreign peptides. Because a number of amino acids from each 9-mer bind the MHC, and are therefore not exposed to the TCR, antigen presentation seems to involve a significant loss of information. Our results show that this is not the case because the HLA molecules are fairly specific. Removing the two anchor residues from each presented peptide, we find that the self:nonself overlap of these exposed 7-mers resembles that of 9-mers. Summarizing, the 9-mers used in MHC class I presentation tend to carry sufficient information to detect nonself peptides amongst self peptides.

Keywords Antigen presentation · Bioinformatics · MHC class I · Self:nonself discrimination · T-cell receptor specificity

N. J. Burroughs
Mathematics Institute, University of Warwick,
Coventry, UK

R. J. de Boer · C. Keşmir (✉)
Theoretical Biology/Bioinformatics, Utrecht University,
Padualaan 8,
3584 CH Utrecht, The Netherlands
e-mail: C.Kesmir@bio.uu.nl
Tel.: +31-30-2533637
Fax: +31-30-2513655

C. Keşmir
Center for Biological Sequence Analysis, BioCentrum-DTU,
Technical University of Denmark,
Lyngby, Denmark

Introduction

The ability of the immune system to discriminate self from nonself and thereby make appropriate immune responses to pathogens has been a subject of intense study and debate for over 50 years. The onset of the genomic era brings a new perspective on this issue and heralds a comparative approach, comparing host and pathogen epitopes. The essential problem is to understand how genomic differences between the host and a diverse array of pathogens can be utilized to detect the presence of an invading pathogen. Differences between genomes translate into differences between proteins, and therefore the essence of the detection problem is to extract sufficient information from expressed proteins as the basis of recognition. A good analogy is identifying the language of a scientific article; for languages using the same alphabet one only needs short words to distinguish languages. For proteins the issue is what length peptides (words) in the amino acid alphabet are required to accomplish discrimination between organisms.

Vertebrate immune systems process self and nonself proteins into peptide fragments consisting of 8–25 consecutive amino acids, which are presented to the T-cell repertoire by surface MHC molecules (Engelhard 1994; Germain and Margulies 1993). The typical length of peptides presented to $CD8^+$ T cells by MHC class I molecules is nine residues; the peptides presented to $CD4^+$T cells on class II molecules tend to be longer. Class I MHC molecules use at least two of the residues from a peptide of nine amino acids (9-mer) as "anchor residues" and "bury" these in their binding pockets. Thus,

Table 1 Proteomes, size of self, and overlaps with the human proteome. Non-redundant proteomes were downloaded from the EBI web site (see Materials and methods). The *columns* represent the pathogenicity for the human host, the GenBank accession number, the number of proteins included in the analysis, their total length in amino acids (*aa*), the number of 9-mers before and after processing by proteasome and translocation by TAP, their overlap with self when presented as exposed 7-mers, and the percentage of 9-mers presented by *A*0201*, and*A*0204*, respectively. Exposed 7-mers were created by removing the anchor residues at positions two and nine from the 9-mers. Because of the specificity of the MHC molecules the number of unique exposed 7-mers was very similar to the number of 9-mers (not shown)

					Processed 9-mers			*HLA-A*0201*			*HLA-A*0204*		
Species	*P*	Acc. no.	Proteins	No. aa	9-mers	Overlap	9-mers	Overlap	%	9-mers	Overlap	%	9-mers	Overlap	%
Homo sapiens		NA	28,781	1.3×10^7	9,813,926	-	2,381,009	-	24.26	88,398	-	3.71	159,751	-	6.71
Chlamydia pneumoniae	*	AE002161	1,110	3.6×10^5	353,092	0.20	93,037	0.14	26.34	3,858	0.23	4.15	6,666	0.42	7.16
Escherichia. coli	*	AE005174	5,139	1.6×10^6	1,462,036	0.16	368,333	0.11	25.19	16,002	0.23	4.34	31,135	0.31	8.45
Helicobacter pylori	*	AE000511	1,555	4.9×10^5	473,659	0.18	128,795	0.11	27.19	5,320	0.19	4.13	8,519	0.26	6.61
Mycobacterium tuberculosis	*	AL123456	3,877	1.3×10^6	1,263,289	0.18	291,765	0.11	23.10	10,421	0.22	3.57	2,6095	0.40	8.94
Neisseria meningitidis	*	AL157959	2,039	5.8×10^5	555,207	0.27	139,132	0.19	25.06	5,481	0.40	3.94	10,957	0.42	7.88
Pseudomonas aeruginosa	*	AE04091	5,556	1.9×10^6	1,789,976	0.17	453,660	0.12	25.34	19,765	0.34	4.36	40,418	0.40	8.91
Rickettsia conorii	*	AE006914	1374	3.4×10^5	327,425	0.41	85,929	0.26	26.24	3,573	0.42	4.16	5,914	0.51	6.88
Salmonella typhi	*	NA	4,707	1.4×10^6	1,351,012	0.17	342,264	0.12	25.33	14,822	0.23	4.33	29,380	0.32	8.58
Staphylococcus aureus (N315)	*	NA	2,593	7.8×10^5	748,299	0.17	189,306	0.11	25.30	7,991	0.23	4.22	13,261	0.28	7.01
Streptococcus pneumoniae	*	AE005672	2,079	5.9×10^5	560,676	0.19	145,532	0.14	25.96	6,415	0.27	4.41	10,832	0.33	7.44
Vibrio cholerae	*	NA	3,785	1.2×10^6	1,115,633	0.17	287,829	0.13	25.80	12,775	0.24	4.44	24,365	0.32	8.47
Yersinia pestis	*	NA	3,898	1.2×10^6	1,195,283	0.18	303,791	0.12	25.42	13,702	0.29	4.51	26,478	0.29	8.72
Thermoplasma volcanium		BA000011	1,523	4.5×10^5	436,490	0.14	114,341	0.08	26.20	4,465	0.20	3.90	7,543	0.33	6.60
Bacillus subtilis		AL009126	4,099	1.2×10^6	1,175,995	0.16	303,783	0.11	25.83	12,509	0.21	4.12	22,438	0.29	7.39
Average						0.2±0.07		0.13±0.04	25.59±0.92		0.26±0.07	4.18±0.25		0.35±0.07	7.79±0.87
Dengue type 1	*	U88536	1[b]	3,392	3,383	0.00	878	0.00	25.95	35	0.00	3.99	63	0.00	7.18
Ebola	*	AF086833	9	5,493	4,946	0.00	1,249	0.00	25.77	52	0.00	4.16	82	0.00	6.57
Hepatitis A	*	M14707	1[b]	2,225	2,217	0.00	573	0.00	25.83	26	0.00	4.54	44	0.00	7.68
Hepatitis B	*	X5190	4[b]	1,613	1,579	0.00	486	0.00	30.78	24	0.00	4.94	39	0.00	8.02
Hepatitis C	*	AJ132997	1[b]	3,010	3,001	0.00	761	0.00	25.36	36	0.00	4.73	75	0.00	9.86
HIV-1	*	AJ006287	9	3,164	3,092	0.00	742	0.00	24.00	25	0.00	3.37	51	0.00	6.87
HTLVI-1	*	D13784	4	2115	2083	0.34	536	0.37	25.73	30	0.00	5.60	52	0.00	9.70
Influenza A segments 1–8	*	J02146[a]	10	4,449	4,366	0.07	1,093	0.09	25.03	43	2.33	3.93	67	1.49	6.13
Measles	*	K01711	8	5,456	4,849	0.00	1,228	0.00	25.32	50	0.00	4.07	93	0.00	7.57
Mumps	*	AB040874	8	4,977	4,765	0.02	1,183	0.00	24.83	43	0.00	3.63	96	0.00	8.11
Parvovirus H1	*	X01457	2	694	677	0.00	158	0.00	23.34	3	0.00	1.90	7	0.00	4.43
Polio virus 1	*	AJ132961	1[b]	2,209	2,200	0.05	539	0.00	24.50	20	0.00	3.71	32	0.00	5.94
Rabies	*	M31046	5	3,600	3,559	0.03	948	0.11	26.64	30	3.33	3.16	70	0.00	7.38
Respiratory syncytial virus	*	AF013254	11	4,540	4,451	0.00	1,144	0.00	25.70	47	0.00	4.11	79	1.27	6.91
Rubella	*	AF188704	2[b]	3,192	3,175	0.03	707	0.00	20.27	15	0.00	2.12	51	0.00	7.21
Sendai	*	M69046	6	4,808	4,759	0.13	1,188	0.08	24.96	39	0.00	3.28	88	0.00	7.41
Yellow fever	*	X03700	1[b]	3,411	3,402	0.00	858	0.00	25.22	52	0.00	6.06	76	0.00	8.86
Average						0.04±0.09		0.04±0.09	25.37±1.75		0.33±0.96	3.96±1.07		0.16±0.46	7.42±1.37

[a]The accession numbers for the other segments are: J02147, J02150, J02151, V00603, V01088, V01099, V01106
[b]At least one of the proteins is a polyprotein

a T cell can only utilize the information contained in the remaining amino acids to discriminate self from nonself. Although self:nonself discrimination is facilitated by several mechanisms, e.g., danger and innate signals during primary immune reactions (Matzinger 1994; Medzhitov and Janeway 2002), effector/memory cells will have to discriminate self from nonself peptides in the absence of such contextual signals during the effector/memory phase of the response. The information in each presented 9-mers should therefore suffice to let effector/memory T cells discriminate between self and nonself.

If an immunodominant foreign peptide is also a self peptide, T cells would either be expected to be tolerant to such a peptide or would cross-react with the self peptide, possibly leading to autoimmunity. The first question of this paper is whether short peptides consisting of nine amino acids, as used for MHC class I presentation, are sufficiently unique. Next we enumerate the number of distinct self peptides and foreign peptides that are expected to be processed by the immunoproteasome, translocated to the endoplasmic reticulum (ER), and presented by two MHC alleles. Finally, removing the anchor residues, we estimate the degree of overlap in the remaining residues that are exposed to the T-cell receptor.

Materials and methods

The nonredundant proteomes of the species listed in Table 1 were downloaded from the EBI web site (bacterial and eukaryotic proteomes were from http://www.ebi.ac.uk/proteome, while viral proteomes were from http://www.ebi.ac.uk/genomes/virus.html; downloads were made in June 2003). Peptides of the specified length (*n*-mers) were generated from each protein in the proteome using all positions as possible first positions; similar results were obtained using nonoverlapping peptides (data not shown). Peptides containing any altered (nonstandard) amino acids (B, X or Z) were removed from the analysis (these constituted less than 0.01% of the various proteomes). To check for possible mistakes in the databases, we analyzed all human proteins overlapping with nonself at the 25-mer or 40-mer level. All of these human proteins except two (Q9S459, Q14553) had a degree of homology with microorganism proteins that seemed reasonable for evolutionarily conserved proteins. Q9S459 is not a human, but a *Salmonella* protein. Q14553, which contains a large DNA insert from the hepatitis B virus (HBV), is found only in HBV-infected individuals. Both proteins were removed from the analysis.

Reliable prediction tools for MHC peptide binding are only available for a limited number of MHC alleles. For *HLA-A*0204*, a neural network predicting IC50 values is available at http://www.cbs.dtu.dk/services/netMHC (Buus et al. 2003). Peptides with a predicted IC50 of less than 500 nM were considered to be presented by *HLA-A*0204*. This neural network method is reasonably reliable because more than 85% of the experimentally verified good binders are correctly predicted (Buus et al. 2003). For *HLA-A*0201* the half-life of the peptide-MHC (pMHC) complex can be predicted with a matrix method (Parker et al. 1994) publicly available at http://wwwbimas.dcrt.nih.gov/molbio. Complexes having a half-life exceeding 60 s were considered to be good binders. There are similar matrices for other MHC alleles, but these are less reliable and only give a relative score rather than a true half-life of the peptide-MHC complex.

Proteasome predictions for producing peptides of length nine amino acids were performed using a neural network (Kesmir et al. 2002). For only three proteins in the human proteome were the predictions not available. We assume that a prediction above 0.5, as suggested by Kesmir and co-workers (2002), corresponds to a likely cleavage site. This neural network predicts the specificity of the immunoproteasome rather than the constitutive proteasome (Kesmir et al. 2002; Saxova et al. 2003). For TAP predictions we have implemented the method suggested by Peters and co-workers (2003). This method is highly accurate and was shown to increase the performance of MHC predictions when used as a prefilter (Peters et al. 2003). We used a threshold of 1 as the minimum score for being translocated. Peters and co-workers (2003) showed that only 1.5% of epitopes have a worse score than 1.

Results

The size of human self

We downloaded the human proteome from the EBI website (see Materials and methods), consisting of 28,783 proteins, with a total length of 1.3×10^7 amino acids. Enumerating the number of distinct 9-mers in the human proteome, we found that the "size of human self" is 9.8×10^6 distinct peptides. Surprisingly most of these peptides are unique, i.e., 76% of the 9-mers occur only once in this proteome, and 24% are repeated. Thus, the majority of the amino acids in the human proteome are starting points of a unique new 9-mer. Since the total peptide universe of 20^n possible peptides increases with increasing peptide length, n, the number of repeats in the proteome should decrease with the peptide length. The size of self indeed increases with the length of the peptides (see Fig. 1 A), but approaches saturation at peptides of six amino acids in length. Saturation occurs because the majority of the n-mers in the human proteome are unique above n=6. A similar saturation occurs for other organisms: the size of self of *Yersinia pestis* saturates at n=5–6, and that of the influenza virus saturates at n=3–4 because of its smaller genome size (see Fig. 1A and Table 1).

As a consequence of the saturation, the size of self at the n-mer level, when expressed as a fraction of the total peptide universe of 20^n possible peptides, decreases exponentially for n>5 (see Fig. 1B). A randomly made foreign peptide of more than seven amino acids is therefore unlikely to occur in the human self. Thus, one would expect very small overlaps between self and nonself at, e.g., the 9-mer level. Peptide usage is not random,

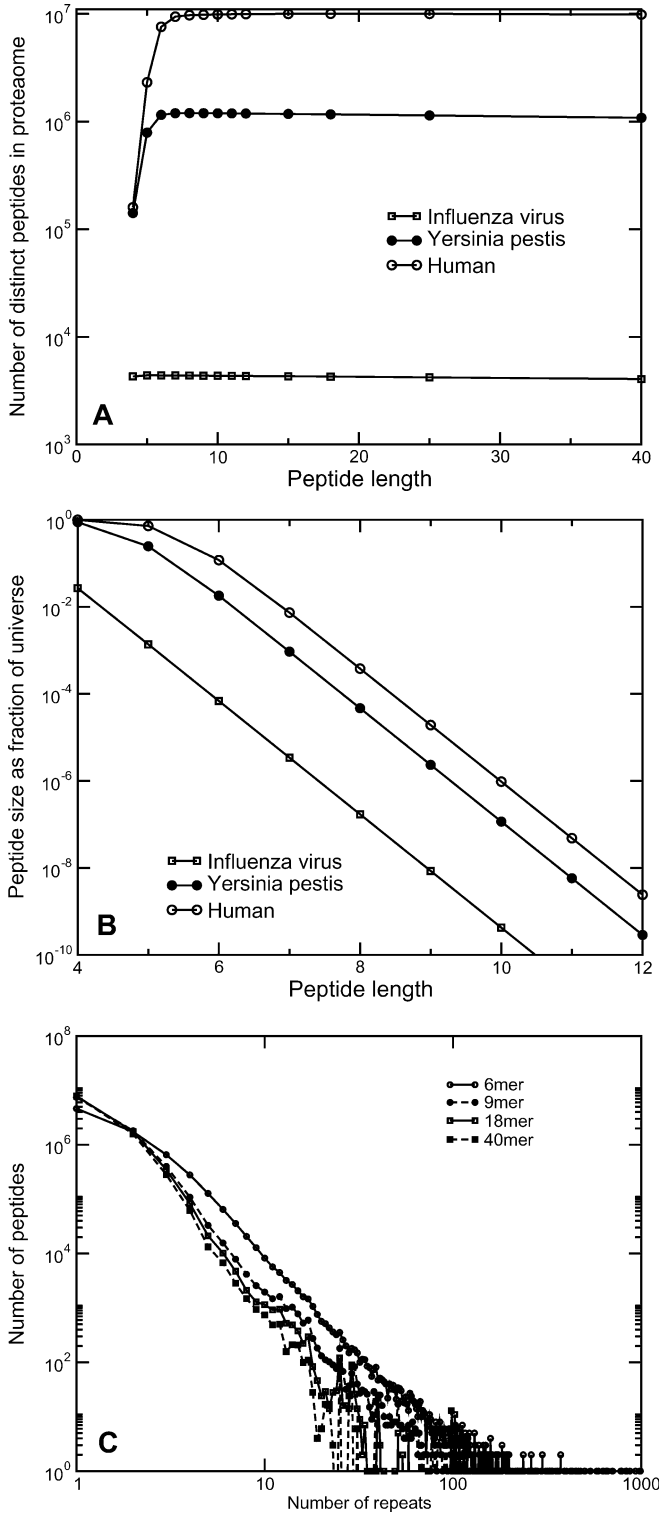

Fig. 1A–C Peptide statistics in whole proteomes. **A** The size of self for humans, *Yersinia pestis* and influenza A virus as a function of peptide length *n*. Self is defined as the number of distinct peptides occurring in the proteome of an organism. **B** Self size (as in **A**) as a fraction of all possible *n*-mers (a total of 20^n for 20 amino acids). **C** Number of peptides with a given number of repeats in the human proteome for 6-, 9-, 18- and 40-mers. The distributions follow a power-law scaling. This power-law behavior was observed in all the organisms analyzed, with higher levels of repeats observed in larger organisms (see also Fig. 2B, *open circles*). The average repeat frequency decreases with increasing peptide length, but is still significant at peptides of length 40

however; the number of peptides with a given repeat frequency follows a power law behavior, see Fig. 1C, which is reminiscent of power laws observed in gene family sizes (Huynen and Van Nimwegen 1998; Qian et al. 2001), and DNA sequences (Mantegna et al. 1995; Holste et al. 2001). Thus, although most *n*-mers are unique for $n \geq 6$, they are repeated much more frequently than what would be expected for a random peptide model. Finally, because the size of self for all organisms saturates by $n \geq 6$, the total information content of the realized *n*-mers in a protein would not increase when longer peptides are used.

Self:nonself overlaps

Detection of a foreign peptide relies upon the existence of peptides in the foreign organism that are not in the proteome of the host. Not all presented peptides from a microorganism need to be nonself, but the detection of a microorganism requires that at least one of the immunodominant peptides is not a self peptide. To study this, we measured the overlap between human self and foreign for a number of different viruses and bacteria (see Fig. 2 A and Table 1). The overlap is defined as the percentage of distinct peptides from a microorganism that also exist in the human proteome. The smaller the overlap the lower the probability that an immunodominant foreign peptide is also a self peptide.

The overlap between human self and foreign decreases with the length of the peptide (see Fig. 2A). Bacterial proteomes of different size have similar overlaps with human self (see Table 1). For peptides of length five and below, more than 80% of foreign peptides also occur in human self. The chance that a foreign peptide is also a self peptide drops below the 5% level when the peptide length exceeds six amino acids, e.g., decreases to an average of 3% at the 7-mer level. The sharpest decrease in the overlap between self and foreign peptides occurs when the peptide length increases from six to seven, and from seven to eight, amino acids (see Fig. 2A). Although the overlap between self and foreign becomes small for peptides of length seven and longer, it remains possible to find some microorganisms with a nonzero overlap for any reasonable peptide length (see Fig. 2A). Summarizing, if the immune system were to use all the information in a peptide, peptides of at least seven amino acids from a microorganism would occur in the human proteome with a maximum of 3% chance only.

For the 9-mers that are used in MHC class I presentation the probability that a randomly chosen foreign peptide is also present in the human self is about 0.2% (see Fig. 2A and Table 1). However, analyzing all human 9-mers we showed that the chance that a randomly chosen human 9-mer is repeated in the human self is 24%. Not unexpectedly, the differences between human and foreign 9-mers tend to be larger than those amongst human 9-

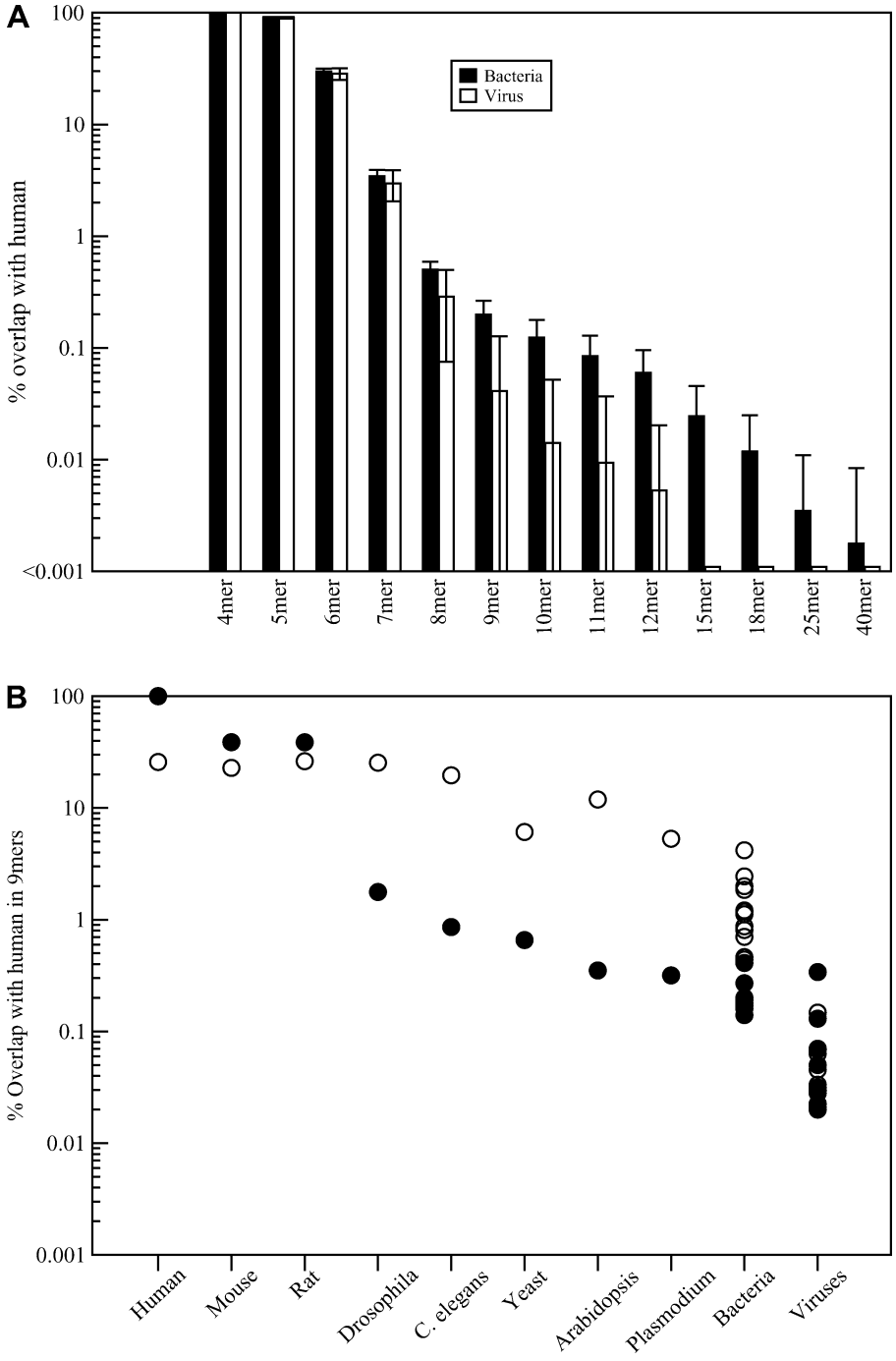

Fig. 2A, B Overlap with nonself. **A** The average percentage of distinct peptides in 17 viral and 14 bacterial proteomes that also occur in the human self as a function of the peptide length (*n*-mer). From each protein in the proteome, *n*-mers were generated using all sequential first positions. For each microorganism the percentage overlap between self and nonself was calculated by dividing the number of overlapping *n*-mers by the total number of distinct *n*-mers in the microorganism. The figure depicts the average of these percentages within each specified group, and the *error bars* show the standard deviation. **B** The self:nonself overlap between human 9-mers and 9-mers from a variety of species. Species are ordered by evolutionary distance as determined by Baldauf and co-workers (2000). The *closed circles* depict the overlaps; the *open circles* depict the fraction of 9-mers that is repeated within each proteome

mers. We therefore studied the self:nonself overlap between human and a number of other taxa as a function of the evolutionary distance (see Fig. 2B, *solid circles*). This indeed shows that the low self:nonself overlaps that we found for bacteria and viruses are due to the large evolutionary distance between human and these microorganisms. For instance, randomly selected 9-mers from close-by species like mouse or rat have a 40% chance of also occurring in the human self. The information content of 9-mers is therefore not sufficient to discriminate such closely related taxa, but tends to be sufficient for typical pathogens. The open circles in Fig. 2B depict the percentage of repeated 9-mers for each proteome, which was 24% in human, and remains close to 20% up to *Caenorhabditis elegans*. The organisms with small proteomes tend to have less repeats, probably because their genome is more compact. Microorganism overlaps follow similar patterns. Different microorganism species require different types of immune response (Janeway et al. 2001; Borghans and De Boer 2002). The immune system therefore also has to discriminate pathogen species from one another on the basis of short peptides. We computed overlap between all bacterial species in Table 1, where the "overlap" for each pair was defined as the number of

Table 2 Specificity of each step in antigen processing and presentation. To find the specificities for the nonself set, we combined all the bacterial and viral 9-mers and then considered only the unique ones to calculate ratios. *S* Self, *NS* nonself. The table shows two main results. First, the specificity of TAP, the immunoproteasome and MHC molecules are coevolved, because for each column the lowest fraction is obtained for the complete 9-mer set. Second, TAP and MHC molecules have a slight preference to translocate/present nonself 9-mers

	Proteasome		TAP		A*0201		A*0204	
Set	S %	NS %	S %	NS %	S %	NS %	S %	NS %
All 9-mers	33.9	33.6	58.5	61.7	1.4	1.7	2.6	3.6
Cleaved 9-mers			71.6	75.7	2.9	3.5	5.5	7
Translocated 9-mers					2.1	2.5	3.8	4.5
Processed 9-mers					3.7	4.2	6.7	8.1

overlapping peptides divided by the size of the smallest proteome. For 9-mers the overlaps were typically around 1%, with the exception of the *Escherichia coli* and *Salmonella typhi* pair, which overlapped in 35% of the 9-mers, of *Y. pestis* with *E. coli* and *S. typhi* at 14%, and of *Vibrio cholerae* with *E. coli*, *S. typhi* and *Y. pestis* at 4.5%. Because overlaps between unrelated bacterial species are typically around 1%, the immune system can reliably "remember" a pathogen by a single 9-mer: it is significantly unlikely that an immunodominant peptide of a subsequent unrelated pathogen will be identical to a previous one.

Processing: proteasomal cleavage and translocation by TAP

Only part of the 9-mers from a proteome are presented on the MHC. The 9-mers used in class I presentation have to be cleaved from endogenous proteins by the (immuno) proteasome, and have to be translocated to the ER by TAP molecules. To investigate whether these two processing steps contribute to self:foreign discrimination, we predicted all 9-mers expected to be processed and thus become available for class I presentation. The proteasomal cleavage predictions were made with a neural network predictor for the immunoproteasome specificity (Kesmir et al. 2002). The (immuno) proteasome is stochastic: there are many MHC ligands with possible internal cleavage sites (see, for example, Lucchiari-Hartz et al. 2000; Morel et al. 2000). The enzymes involved in N-terminal trimming of the peptides seem to have a broad specificity (Stoltze et al. 2000), and N-terminal trimming to a peptide length of nine amino acids is always possible. Therefore, we considered each predicted cleavage site in a protein to be C-terminal of a possible 9-mer and call these the "cleaved 9-mers" (see Table 2). In both human and microorganism proteomes the average cutting frequency was once every 3.1 positions (SD 0.01), with no difference between human and foreign proteomes (see Table 2). Because approximately 34% of the unique 9-mers were predicted cleavage products for human, bacteria, and viruses (see Table 2), we conclude that the immunoproteasome on its own fails to discriminate self from nonself.

The second step in antigen processing is the translocation of the peptides into the ER by TAP molecules. To predict the TAP affinity, we implemented the weight matrix method suggested by Peters and co-workers (2003). TAP is less specific than the immunoproteasome; it translocates approximately 60% of all 9-mers, whereas the immunoproteasome cleaves only 34% of them (see Table 2). Importantly, the specificity of TAP and the immunoproteasome seems to have coevolved, because TAP translocates up to 76% of all "cleaved 9-mers", but only 60% of all enumerated 9-mers. In addition, TAP has a slight preference (5%) for translocating nonself 9-mers.

The other steps in antigen processing involve further degradation of the peptides in the cytoplasm by endopeptidases and N-terminal trimming in the ER by aminopeptidases (see, for example, Reits et al. 2004; Stoltze et al. 2000). Most of these enzymes seem to be unspecific, i.e., their effect on the antigen processing depends on how long the peptides are exposed to these enzymes. It is very hard to imagine that such unspecific enzymes allow for self:nonself discrimination. Therefore they are excluded from this analysis. For each organism in our analysis, we generated a "processed 9-mer" set by simply taking all cleaved 9-mers with a good TAP affinity. This set is 24% of all 9-mers in human and 25% of all 9-mers in microorganisms. The overlap between self and nonself decreases slightly when we use processed 9-mers. The proteasome uses the information in the flanking region of a 9-mer to make a cleavage (see, for example, Nussbaum et al. 1998; Beekman et al. 2000). A 9-mer occurring both in self and nonself proteomes will not be generated with the same efficiency due to possible differences in the flanking region of the 9-mer. Therefore, the processing by the immunoproteasome slightly decreases the overlaps, despite the fact that the proteasome fails to discriminate self from nonself. Moreover, TAP has a slight preference to present nonself, reducing self:nonself overlaps further.

Anchor residues

Once a 9-mer is presented on an MHC molecule, not all of the nine amino acids in a peptide are available for T-cell recognition because the so-called anchor residues are buried in the binding pockets within the groove of the MHC molecule. MHC molecules often bind distinct peptide subsets because they have different binding motifs

(Rammensee et al. 1999). Due to positive selection in the thymus, T cells tend to be restricted to a particular MHC allele in the host, and they typically fail to respond to peptides presented by the other MHC molecules in the host. Since T cells are ultimately the "detectors" discriminating self from nonself, it seems sufficient to compute overlaps at the level of the peptides presented by a single MHC allele. Because most MHC binding motifs have two anchor residues one would intuitively expect that the information available to a T cell corresponds to that of the 7-mers shown in Fig. 2.

In order to develop a better intuition for the effect of presentation on the self:nonself overlap, consider the set of processed 9-mers in the human proteome which partially overlaps with the set of processed 9-mers in any particular bacterial proteome. From each set select the (small) subset of 9-mers matching a particular MHC binding motif. If the binding motif were completely specific, i.e., if two particular amino acids were required for the two anchor residues, the self:nonself overlap at the level of the remaining "exposed 7-mers" would be equal to that of the original "presented 9-mers". On the other hand, if the binding motif were not specific, removing two degenerate anchor residues would increase the overlap, because 9-mers differing only at the anchor residues collapse onto identical 7-mers when these two residues are omitted. Summarizing, a fully specific binding motif is not expected to change the overlaps because all information is preserved, whereas with a hypothetical fully degenerate MHC binding motif the 9mer overlap of 0.2% would increase to the 7-mer overlap of 3.4% observed in Fig. 2. MHC molecules are neither completely degenerate nor completely specific. Therefore, at least part of the information of the two anchor residues is lost for self:nonself discrimination, increasing the peptide overlap. We studied this by predicting human and microorganism peptides that bind to a number of *HLA* alleles and then removed the anchor residues on these good binders.

Reliable prediction tools are publicly available (see Materials and methods) for two *HLA-A2* alleles. For *HLA-A*0201* we used a matrix method (Parker et al. 1994) and for *HLA-A*0204* a neural network (Buus et al. 2003). Using these methods, we enumerated all 9-mers that are predicted as good binders. Ignoring the primary anchor residues at positions 2 and 9 (Ruppert et al. 1993), we constructed the sets of exposed 7-mers, the sizes of which are listed in Table 1. On average the observed self:nonself overlap of the exposed 7-mers from the bacteria remains less than 0.4%, which is lower than the 8-mer overlap (see Fig. 2). Most viruses have an expected overlap of zero. Summarizing, the approximately 7% specificity of the MHC binding motif preserves most information of the two anchor residues, which allows the self:nonself overlaps of the exposed 7-mer to remain similar to that of the 9-mers.

For the human proteome we predict that the size of self is 8.8×10^4 and 1.6×10^5 9-mers on *HLA-A*020*1 and *HLA-A*0204*, respectively (see Table 1). For self peptides the specificity of these MHC molecules is 3.7% and 6.7%, respectively, whereas the presentation frequency of foreign 9-mers is 4.2% and 8.1% (see Table 1, 2), representing a bias for foreign of 12% and 17%, respectively. This bias is due to the differences in amino acid usage; a similar bias was obtained in a random peptide model using the amino acid frequencies in the human proteome and in bacterial proteomes (results not shown).

Because reliable prediction methods are only available for human MHC molecules, we are not able to extend our analysis to other organisms. However, we think that our results would remain the same for other mammalians, such as the mouse, because (1) Sette and co-workers (2003) showed that mouse and human MHC molecules have similar specificities as a result of convergent evolution and (2) the overlap between mouse 9-mers and viral and bacterial 9-mers is as low as the overlap between human self and nonself (see Fig. 2B and results not shown).

Coevolution of the specificities in the class I presentation pathway

Yewdell and co-workers (2003) showed that the generation of MHC class I ligands from endogenous proteins is a highly inefficient process. Their rough estimates suggest that only 0.1% of specific peptides can reach the MHC molecule, others being destroyed by degradation. One way of increasing the efficiency of the processing pathway is to have similar specificities for TAP, immunoproteasome and MHC molecules, so that the products of the immunoproteasome would be good binders of TAP, and the peptides translocated to the ER tend to be good MHC binders. Recently, we have shown that the specificity of the immunoproteasome has coevolved with the human MHC class I molecules (Kesmir et al. 2003). The results presented in Table 2 confirm this, and add TAP into this coevolutionary relation. A larger fraction of the cleaved 9-mers have sufficient TAP and MHC affinities, and a larger fraction of the translocated 9-mers has sufficient affinity for the MHC. In combination, this increases the efficiency of presentation almost three-fold; an ordinary 9-mer has 1% chance of being presented, while 8% of processed 9mers are expected to be presented.

Discussion

The MHC class I restricted cellular immune response to a foreign pathogen typically focuses on a small number of immunodominant peptides of nine amino acids. We have shown that these few 9-mers sampled from the microorganism's proteome are likely to be unique to a $CD8^+$ T cell. They are not expected to be present in the host (human) proteome (the overlap is around 0.2%), nor are they expected to have occurred as an immunodominant peptide during a previous immune response to an unrelated microorganism. Thus, the information in a single 9-mer is sufficient to discriminate self from nonself and between different microorganisms. The binding of 9-mers to MHC molecules, and the recognition by T cells

invoke a loss of information because not all nine amino acids can be used. We have shown that there is hardly any loss of information due to antigen presentation because MHC molecules are fairly specific. For exposed 7-mers, the self:nonself overlap after presentation by one particular MHC allele is typically less than 0.5%. The overlaps would increase if T cells were not MHC-restricted to one (or a few) MHC alleles in the host, since one would have to sum the overlaps over all MHC alleles in the host. Thus, MHC restriction, i.e., positive selection, facilitates self:nonself discrimination. On the level of the exposed 7-mers presented by the MHC molecules there is enough information to allow for a reasonable degree of TCR degeneracy. Finally, we show that specificities of the molecules involved in class I antigen processing and presentation pathway have coevolved, and that this increases overall efficiency of the antigen presentation by three-fold.

A recent paper estimated the similarity of a few self and foreign proteins to large sets of self and nonself peptides (Ristori et al. 2000) and concluded that it is not possible to distinguish self from nonself by short peptides. The conflict between that study and our study lies in the definition of "distinguishing self from nonself". Ristori and co-workers (2000) focus on comparing the similarity of regions in proteins with self and nonself, and argue that short peptides are not discriminatory because almost every protein contains short sequences that are highly similar to self and to nonself. However, we think the relevant question is whether the few peptides from a pathogen that become immunodominant resemble any of the self-peptides, and have shown that single short foreign peptides can be discriminated from nonself. Moreover, most of the analysis by Ristori and co-workers (2000) was based upon a random peptide model. We agree that the amino-acid frequencies in the human proteome and foreign proteomes are too similar to allow for self:nonself discrimination under a random peptide model. Therefore, one needs to enumerate all human and nonself short peptides to find the overlap.

Several authors have suggested that many potential foreign epitopes are nonimmunogenic due to overlaps with self (Kourilsky and Claverie 1986; Ristori et al. 2000; Ohno 1992), which seems to contradict our results that self:nonself overlaps play a negligible role. Using an analysis of 4-mers, Kourilsky and Claverie (1986) observed a correlation between epitopes and regions of foreign proteins with low levels of overlap. We think this analysis is outdated because analyzing the complete proteomes that are currently available we found that almost every foreign 4-mer is present in the human self (see Fig. 2A). Moreover, the results of Kourilsky and Claverie (1986) were based on the only two epitopes that were by then known for the influenza nucleoprotein. Now four more epitopes have been identified in this protein (see http://www.syfpeithi.de); three of the new epitopes are in the high overlap regions defined by Kourilsky and Claverie (1986) and one is partially in the high overlap region. The paper by Ristori and co-workers (2000) discussed above also showed that epitopes tend to be located in regions of proteins where the similarity to foreign exceeded that to self. However, as argued above, while enumerating self:nonself overlaps for single (immunodominant) 9-mers, we have not been able to confirm their measures of similarity to self and nonself. Another study suggested that foreign T-cell epitopes are not the highest affinity MHC binders in the microorganism (Ohno 1992). We failed to reproduce the peptide-binding classification used by Ohno (1992) using current weight matrices for MHC binding predictions (HLA-B27, http://wwwbimas.dcrt.nih.gov/molbio, results not shown). These weight matrices are based on larger data sets than those used originally by Ohno (1992).

Two recent papers suggest that protein splicing plays a role in generating T-cell epitopes (Hanada et al. 2004; Vigneron et al. 2004). Via this mechanism, a T-cell epitope can be generated by combining two noncontiguous segments of a protein. In our analysis we have studied only the 9-mers generated from contiguous segments. If many T-cell epitopes were to be generated by the protein splicing, the size of self would increase. The effect on self:nonself overlaps is difficult to foresee. However, as long as protein splicing occurs infrequently, e.g., only in cancer cells (Hanada et al. 2004; Vigneron et al. 2004), its effect on our results would be small.

A small overlap between self and nonself is favorable for several reasons. First, the risk of autoimmunity increases with each foreign epitope that overlaps with self. Second, the smaller the overlap between self and foreign the more peptides that remain as potential targets for detecting the presence of the pathogen. The latter is especially important for pathogens with small proteomes. For instance, for parvovirus we predict a very small number of peptides that can bind to the MHC alleles studied (see Table 1). These few foreign peptides have to be detected among the $>10^4$ self peptides that are expected to bind these *HLA-A2* alleles. Professional antigen-presenting cells have 10^4–10^6 surface MHC molecules, i.e., strong competition is expected among peptides to bind an MHC molecule. This implies that only the few highest binding viral peptides will be successfully presented.

The plant *Arabidopsis thaliana* has a proteome size similar to that of humans and shows similar degrees of overlap with the microorganisms analyzed here (not shown). Because *Arabidopsis* is not a host for the pathogens included in our study, the similarity in overlaps suggests that the 9-mer overlap levels are hardly influenced by host-pathogen evolution. However, one should not take this as evidence against molecular mimicry and/or host-pathogen coevolution. Mimicry could involve just one or a few immunodominant peptides, which is not detectable by our global proteome analysis. Indeed, we have shown elsewhere (Yusim et al. 2002) that several strains of HIV-1 have evolved regions in their proteins that escape proteasomal degradation and, as a consequence, harbor few T-cell epitopes. On the level of our complete genome analysis, such adaptations in specific areas of the pathogen genome would not be detected. Another

interesting example of host-pathogen coevolution is the fact that the two *HLA-A2* alleles included in our study have a small preference to present foreign peptides (see Table 1, 2). It will be interesting to explore whether other *HLA* alleles have evolved such biases and to study how and whether these affect the overlaps between self and nonself.

Acknowledgements We acknowledge the valuable input of Hugo van den Berg, José Borghans, Vera van Noort, and Paulien Hogeweg. C.K. was supported by the Bioinformatic Program of Netherlands organization for scientific research (NWO, 050.50.202).

References

Baldauf SL, Roger AJ, WenkSiefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977

Beekman NJ, Van Veelen PA, Van Hall T, Neisig A, Sijts A, Camps M, Kloetzel PM, Neefjes JJ, Melief C J, Ossendorp F (2000) Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site. J Immunol 164:1898–1905

Borghans JA, De Boer RJ (2002) Memorizing innate instructions requires a sufficiently specific adaptive immune system. Int Immunol 14:525–532

Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. Tissue Antigens 62:378–384

Engelhard VH (1994) Structure of peptides associated with class I and class II MHC molecules. Annu Rev Immunol 12:181–207

Germain RN, Margulies DH (1993) The biochemistry and cell biology of antigen processing and presentation. Annu Rev Immunol 11:403–450

Hanada K, Yewdell JW, Yang JC (2004) Immune recognition of a human renal cancer antigen through post-translational protein splicing. Nature 427:252–256

Holste D, Grosse I, Herzel H (2001) Statistical analysis of the DNA sequence of human chromosome 22. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 64:41917

Huynen MA, Van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589

Janeway CA, Travers P, Walport M, Shlomchik M (2001) Immunobiology. The immune system in health and disease, 5th edn. Garland, New York

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

Kesmir C, Van Noort V, De Boer RJ, Hogeweg P (2003) Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. Immunogenetics 55:437–449

Kourilsky P, Claverie JM (1986) The peptidic self model: a hypothesis on the molecular nature of the immunological self. Ann Inst Pasteur Immunol 137:3–21

Lucchiari-Hartz M, Van Endert PM, Lauvau G, Maier R, Meyerhans A, Mann D, Eichmann K, Niedermann G (2000) Cytotoxic T lymphocyte epitopes of HIVNef: generation of multiple definitive major histocompatibility complex class I ligands by proteasomes. J Exp Med 191:239–252

Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, Stanley HE (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 52:2939–2950

Matzinger P (1994) Tolerance, danger, and the extended family. Annu Rev Immunol 12:991–1045

Medzhitov R, Janeway Jr CA (2002) Decoding the patterns of self and nonself by the innate immune system. Science 296:298–300

Morel S, Levy F, BurletSchiltz O, Brasseur F, ProbstKepper M, Peitrequin AL, Monsarrat B, Van Velthoven R, Cerottini JC, Boon T, Gairin JE, Van den Eynde BJ (2000) Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. Immunity 12:107–117

Nussbaum AK, Dick TP, Keilholz W, Schirle M, Stevanovic S, Dietz K, Heinemeyer W, Groll M, Wolf DH, Huber R, Rammensee HG, Schild H (1998) Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. Proc Natl Acad Sci USA 95:12504–12509

Ohno S (1992) How cytotoxic T cells manage to discriminate nonself from self at the nonapeptide level. Proc Natl Acad Sci USA 89:4643–4647

Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide sidechains. J Immunol 152:163–175

Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 171:1741–1749

Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 313:673–681

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Reits E, Neijssen J, Herberts C, Benckhuijsen W, Janssen L, Drijfhout JW, Neefjes J (2004) A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation. Immunity 20:495–506

Ristori G, Salvetti M, Pesole G, Attimonelli M, Buttinelli C, Martin R, Riccio P (2000) Compositional bias and mimicry toward the nonself proteome in immunodominant T-cell epitopes of self and nonself antigens. FASEB J 14:431–438

Ruppert J, Sidney J, Celis E, Kubo RT, Grey M, Sette A (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. Cell 74:929–937

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

Sette A, Sidney J, Livingston BD, Dzuris JL, Crimi C, Walker CM, Southwood S, Collins EJ, Hughes AL (2003) Class I molecules with similar peptide binding specificities are the result of both common ancestry and convergent evolution. Immunogenetics 54:830–841

Stoltze L, Schirle M, Schwarz G, Schroter C, Thompson MW, Hersh L B, Kalbacher H, Stevanovic S, Rammensee HG, Schild H (2000) Two new proteases in the MHC class I processing pathway. Nat Immunol 1:413–418

Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, Van Der Bruggen P, Boon T, Van Den Eynde BJ (2004) An antigenic peptide produced by peptide splicing in the proteasome. Science 304:587–590

Yewdell JW, Reits E, Neefjes J (2003) Making sense of mass destruction: quantitating MHC class I antigen presentation. Nat Rev Immunol 3:952–961

Yusim K, Kesmir C, Gaschen B, Addo MM, Altfeld M, Brunak S, Chigaev A, Detours V, Korber BT (2002) Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV) proteins reveal imprints of immune evasion on HIV global variation. J Virol 76:8757–8768