Immunogenetics (1995) 42: 392–397

ORIGINAL PAPER

Miles P. Davenport · Ivan A. P. Ho Shon
Adrian V. S. Hill

An empirical method for the prediction of T-cell epitopes

Received: 23 May 1995/Revised: 13 July 1995

Abstract Identification of T-cell epitopes from foreign proteins is the current focus of much research. Methods using simple two or three position motifs have proved useful in epitope prediction for major histocompatibility complex (MHC) class I, but to date not for MHC class II molecules. We utilized data from pool sequence analysis of peptides eluted from two *HLA-DR13* alleles to construct a computer algorithm for predicting the probability that a given sequence will be naturally processed and presented on these alleles. We assessed the ability of this method to predict known self-peptides from these *DR-13* alleles, *DRB1*1301* and **1302*, as well as an immunodominant T-cell epitope. We also compared the predictions of this scoring procedure with the measured binding affinities of a panel of overlapping peptides from hepatitis B virus surface antigen. We concluded that this method may have wide application for the prediction of T-cell epitopes for both MHC class I and class II molecules.

Introduction

The recognition of pathogen-derived peptide in the context of host major histocompatibility complex (MHC) molecules is crucial to the induction and maintenance of an effective antigen-specific immune response. Recent work has shown that MHC molecules bind peptides of approximately 8–10 amino acids in length for class I, and 12–20 amino acids in length for class II. In addition, analysis of the allele specificity of peptide binding has shown that key 'anchor residues' are important in determining the ability of peptides to bind to MHC molecules (Falk et al. 1991; Rudensky et al. 1991). Peptide binding motifs based on a knowledge of these anchor positions have now been determined for MHC class I and class II alleles. These motifs have been used to accurately predict T-cell epitopes for MHC class I in *Listeria monocytogenes* (Pamer et al. 1991), *Plasmodium falciparum* (Hill et al. 1992), and other pathogens.

MHC class II restricted T-cell epitopes have traditionally been mapped using recombinant fragments of proteins and/or overlapping peptides in T-cell proliferation assays. Early attempts at predicting T-cell epitopes directly from protein sequences relied on theoretical concepts of peptide binding (Rothbard et al. 1988). The peptide binding motifs so far identified for MHC class II molecules are generally more degenerate than MHC class I motifs (Chicz et al. 1993), perhaps due to the relatively strong interactions between the MHC molecule and the backbone carbon atoms of the peptide (Brown et al. 1993). Thus these simple motifs have been of less use in the prediction of T-cell epitopes for MHC class II than they have been in the case of MHC class I.

More recently an empirical approach to epitope prediction has utilized peptide binding assays to compare the affinities of short peptides where each sequence position is sequentially mutated to all possible amino acids (Reay et al. 1994; Hammer et al. 1994b). This binding information is then used in a computer program to scan proteins for candidate epitopes. This method has the advantage that, unlike the simple motifs employed in MHC class I, it may provide information about small preferences for particular amino acids and also about the negative effects on binding of some residues. We describe here a novel method for T-cell epitope prediction which relies on the use of pool sequence data from eluted self-peptides and a computer program to iteratively scan proteins for candidate epitopes. We test the ability of this program to accurately predict known self-peptides and T-cell epitopes for two *HLA-DR13* alleles and also compare the predictions of this program with in vitro measurements of peptide binding.

M. P. Davenport (✉) A. V. S. Hill
Institute for Molecular Medicine, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK

I. A. P. Ho Shon
Concord Repatriation General Hospital, Hospital Road, Concord, NSW, Australia

Materials and methods

Peptide scoring matrix

Pool sequencing of self peptides was performed as previously described (Davenport et al. 1995). Briefly, homozygous B-cell lines HHKB (*DRB1*1301*; Gorski et al. 1986) and WT-47 (*DRB1*1302*; Tiercy et al. 1989) were grown and their pellets lysed in phosphate buffered saline with 2% nonidet-P40 (NP-40), 2 μg/ml leupeptin, 2 μg/ml pepstatin, and 5 mM ethylenediaminetetraacetic acid (EDTA). After centrifugation at 100000 × g for 90 min at 4 °C, lysates were loaded onto an L-243 affinity column. After washing, HLA-DR molecules were eluted with 0.05 M diethylamine, 150 mM NaCl, 0.1% NP-40 (pH = 11.5) and then concentrated on a CENTRICON-10 column (Amicon, Beverly, MA) followed by elution of peptides with 0.1% trifluoroacetic acid. Eluted peptides were separated by reverse phase HPLC and any large peptide peaks (representing largely invariant chain and MHC class I-derived peptides) removed before pool sequencing on a liquid protein sequencer (473A; Applied Biosystems, Foster City, CA; Falk et al. 1994). Sequencing was performed for 15 cycles. Cysteine was not quantitated. Data were not corrected for variations in the yield of individual amino acids arising from, for example, carry-over from the previous sequencing cycle. We performed pool sequences on three independent preparations from each cell line.

Data were analyzed as previously described (Kubo et al. 1994). Briefly, we expressed the concentration of each amino acid at each position as the percentage of the total amino acids detected at that position. We then averaged these values for the three data sets for each allele. We then took the average value for each amino acid over the length of the peptide by averaging the values at all positions except the first. Finally, each amino acid at each position was expressed as a deviation from the average value for that amino acid over the whole run (Fig. 1).

Computer analysis

We then developed a computer program in visual basic under Microsoft Excel 5 for windows, which iteratively scanned proteins one amino acid at a time and calculated a score for each amino acid at each position based on the scoring matrix derived from pool sequencing. The only alteration to this empirical data was to adjust the scores for all amino acids to equal 1 at the first position. This was performed because we believe that the strong bias in favor of small amino acids and against charged amino acids is an artefact of the methodology used, since it is also seen when pool sequences of random synthetic peptides are performed (M. Davenport, unpublished observations). Thus, for example, a peptide commencing Leu-Ile-Gly-Ala-...analyzed for *DRB1*1301* would score 1.0 for position 1, 0.73 for 2, 1.39 for position 3, 0.82 for position 4, and so on. The sum of the scores for all 15 amino acids is then divided by 15 to get an average score for the whole peptide. Thus each position of a protein is allocated a score, which we term PPP (the Probability of being Processed and Presented),

Fig. 1 Scoring matrix for peptides eluted from *DRB1*1301* and *DRB1*1302*. Three independent pool sequences of 15 cycles were performed for each allele after removal of dominant peptide peaks observed on reversed phase HPLC. The concentration of each amino acid at each position was first expressed as a proportion of the total amino acids in that cycle. These values for each amino acid at each position were averaged over the three runs for each allele. These results were then expressed as the deviation from the overall average concentration of the amino acid over all cycles except the first. For clarity, an increase of >1.2-fold above the average is indicated by the presence of a *letter* on the *left hand side* of the column. An increase of >1.5 is shown in *bold*, and of >2 is *underlined*. Similarly, where the concentration falls to <0.75 of the average a *letter* is shown on the *right*, and where it falls below 0.5 it is indicated in *bold*. Note that for computer calculations position 1 was set to 1 for all amino acids to avoid a systematic error due to an artefact of the sequencing methods used (see text)

DRB1*1301

	position														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	**A 1.78**	A 1.24	1.11	0.82	0.81	1.11	0.91	0.94	0.92	0.99	1.03	A 1.23	1.10	1.14	1.10
D	0.83	0.65 D	0.59 D	0.89	0.87	0.92	1.13	1.01	1.23	1.16	D 1.22	0.69	1.14	D 1.23	D 1.47
E	0.84	1.06	1.18	1.10	0.88	0.87	0.81	0.84	0.76	0.68 E	0.72 E	0.76	0.83	0.96	1.09
F	1.05	0.72	0.89	F 1.31	F 1.22	1.18	F 1.48	F 1.31	0.95	0.95	0.84	0.95	F 1.20	1.06	0.92
G	**G 1.85**	**G 1.68**	G 1.39	1.01	0.83	0.91	0.70	0.75	0.78	0.79	0.82	0.89	0.88	0.91	0.94
H	**0.44 H**	**0.38 H**	0.74 H	0.74 H	0.53 H	0.61 H	0.79	0.95	0.99	H 2.04	**H 1.95**	H 1.30	H 1.26	0.73 H	0.99
I	1.12	0.73 I	0.98	I 2.06	I 2.02	**I 1.62**	I 1.46	0.88	0.92	0.79	0.81	0.97	0.83	0.75	0.67
K	1.14	0.82	K 1.31	**0.49 K**	**0.44 K**	0.55 K	0.57 K	1.12	**K 1.53**	**K 1.50**	K 1.23	1.11	0.85	1.16	0.97
L	1.15	0.54 L	1.10	L 1.47	L 1.30	L 1.48	L 1.49	L 1.41	0.73 L	0.76	0.80	0.80	0.73	0.64	0.67
M	0.83	**0.49 M**	0.53 M	0.96	0.84	1.03	1.14	1.03	1.15	0.80	M 1.38	M 1.34	M 1.38	M 1.42	1.19
N	**0.48 N**	0.53 N	0.52 N	0.77	0.86	0.83	0.99	0.82	1.07	1.10	1.17	0.93	N 1.21	N 1.55	**N 1.60**
P	**0.26 P**	P 3.05	**P 1.66**	1.00	0.77	0.70 P	0.74 P	0.99	1.01	0.77	0.87	0.79	0.72 P	0.77	0.86
Q	**0.31 Q**	0.56 Q	0.71 Q	0.62 Q	Q 1.20	1.05	0.83	0.96	0.99	1.03	Q 1.20	Q 1.22	1.16	1.09	1.11
R	**0.24 R**	**0.30 R**	**0.25 R**	**0.21 R**	**0.24 R**	**0.35 R**	0.63 R	0.76	R 1.34	**R 1.91**	R 1.42	R 1.45	**R 1.80**	1.10	0.90
S	S 2.96	**S 1.51**	1.17	0.89	**S 1.81**	0.67 S	0.64 S	0.82	0.74 S	0.70 S	0.78	0.61 S	0.72 S	0.67 S	0.82
T	**T 1.63**	0.94	0.86	0.84	T 1.27	0.85	0.65 T	0.85	0.80	0.84	0.90	0.77	0.84	0.82	0.90
V	0.97	0.87	1.19	V 1.35	**V 1.54**	**V 1.73**	**V 1.58**	V 1.33	1.19	0.85	0.79	0.86	0.52 V	0.92	0.85
W	0.96	1.08	1.08	0.86	0.81	0.85	0.85	0.97	1.13	0.87	0.85	0.83	0.90	1.08	1.16
Y	0.74 Y	0.54 Y	0.76	0.89	0.90	0.87	1.06	0.95	0.61 Y	0.90	1.04	**Y 1.68**	Y 1.46	Y 1.22	Y 1.33

DRB1*1302

	position														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	**A 1.73**	A 1.40	1.17	0.71 A	0.73 A	1.08	1.01	0.85	1.08	1.01	1.18	A 1.35	1.02	1.15	1.12
D	D 1.36	1.12	1.04	0.85	0.82	0.95	0.98	0.93	0.84	0.75	0.70 D	0.54 D	0.59 D	0.70 D	1.06
E	0.81	1.03	1.14	1.07	0.99	0.94	0.78	0.81	0.73 E	0.70 E	0.73 E	0.68 E	0.92	1.17	1.16
F	1.05	0.80	F 1.49	F 2.28	**F 1.97**	F 1.26	F 1.29	F 1.28	0.96	0.76	0.71 F	0.60 F	0.73 F	0.69 F	0.63 F
G	G 1.40	G 1.30	1.07	0.90	0.82	0.71 G	0.63 G	0.72 G	0.87	1.08	1.04	0.88	1.18	1.15	1.15
H	0.51 H	**0.41 H**	0.66 H	0.56 H	0.50 H	0.66 H	0.97	H 1.45	0.84	0.76	0.62 H	0.65 H	0.73 H	0.71 H	**0.49 H**
I	0.98	0.97	1.13	**I 1.89**	I 2.06	I 1.40	I 1.26	0.89	0.95	0.89	0.76	I 1.49	I 1.23	0.78	0.83
K	1.18	0.65 K	0.80	0.60 K	0.72 K	1.04	1.02	1.19	K 1.22	K 1.26	K 1.29	1.18	K 1.27	K 1.32	1.13
L	1.35	0.68 L	0.79	L 1.34	L 1.23	**L 1.74**	L 1.36	1.14	0.88	1.00	1.03	1.07	0.58 L	0.84	0.91
M	0.95	**0.44 M**	0.52 M	0.77	0.82	M 1.32	M 1.48	M 1.25	0.92	0.81	M 1.65	M 1.33	M 1.42	M 1.30	1.09
N	**0.31 N**	**0.46 N**	**0.43 N**	**0.35 N**	0.58 N	0.78	**N 1.82**	**N 1.56**	N 1.30	N 1.28	1.16	0.89	0.83	0.88	0.87
P	**0.27 P**	P 2.93	**P 1.63**	1.01	0.80	0.69 P	0.83	0.76	1.03	0.87	0.78	0.71 P	0.73 P	0.70 P	0.90
Q	**0.21 Q**	0.99	0.73 Q	0.62 Q	0.83	0.96	0.56 Q	0.95	Q 1.25	1.19	Q 1.46	0.77	1.08	1.11	0.91
R	**0.39 R**	**0.19 R**	**0.24 R**	**0.25 R**	**0.32 R**	**0.49 R**	0.75	0.82	R 1.45	R 1.44	R 1.46	**R 1.64**	**R 1.57**	R 1.35	R 1.23
S	S 2.60	**S 1.73**	**S 1.59**	1.04	0.93	1.11	0.99	1.01	0.97	0.61 S	0.82	0.71 S	0.79	0.79	0.85
T	T 1.26	1.00	0.98	0.88	0.91	0.80	0.62 T	0.63 T	0.82	0.89	0.86	0.76	0.69 T	0.73 T	0.98
V	1.16	0.82	V 1.34	1.01	1.14	1.18	1.11	V 1.38	1.16	1.17	0.95	1.03	1.11	V 1.22	1.13
W	0.95	0.68 W	0.64 W	0.65 W	0.96	0.73	0.65 W	0.89	0.91	1.07	1.08	0.86	0.94	0.87	1.01
Y	0.59 Y	0.73 Y	1.11	**Y 1.74**	**Y 1.70**	1.02	1.06	1.01	0.73 Y	0.92	0.95	1.18	Y 1.23	1.05	0.95

Eluted peptide/*T-cell epitope*	Restriction	score	rank	centile
Invariant chain(97-120)	*DRB1*1301*	1.227	1/232	99.6
Beta-2-microglobulin(4-20)	*DRB1*1301*	1.402	1/99	99.0
HLA-B*0701(104-123)	*DRB1*1301*	1.254	2/338	99.4
HLA-DQB1*0603(21-36)	*DRB1*1301*	1.172	11/244	95.5
Invariant chain(97-120)	*DRB1*1302*	1.257	1/232	99.6
Invariant chain(66-80)	*DRB1*1302*	1.211	2/232	99.1
Transferin receptor(215-232)	*DRB1*1302*	1.236	4/760	99.5
Apolipoprotein-B-100(43-58)*	*DRB1*1302*	1.149	164/4560	96.4
Apolipoprotein-B-100(3342-61)*	*DRB1*1302*	1.232	25/4560	99.5
Cathepsin-S(21-42)	*DRB1*1302*	1.109	16/331	95.2
HLA-C*0501	*DRB1*1302*	1.269	1/342	99.7
Flu Haemaglutinin(309-320)	*DRB1*1302*	1.190	3/329	99.1

Fig. 2 Ability of algorithm to predict self-peptides and T-cell epitopes. The program scanned the entire sequence of the proteins from which the peptides were derived and then the score for the observed peptide binding to the appropriate restriction element was compared with that of other peptides within the protein. The ranking and the centile band of the observed peptide are listed. * Apoliprotein peptides are believed to be derived from cattle sources, due to similarity to human apolipoprotein sequence. The human sequence was used in this analysis, since the cattle sequence is unknown. The protein sequence was modified to conform with the observed sequence from the eluted peptides

representing the predicted probability that a 15′ mer peptide commencing at that position may be processed and presented on that MHC.

Peptide binding assays

Peptide binding assays were performed as previously described (Davenport et al. 1995) using overlapping 15′ mer peptides from the hepatitis B virus surface antigen. Briefly, 0.1 μg of MHC class II was incubated with 0.02 μg biotinylated invariant chain (97–117) peptide and various concentrations of inhibitor peptide overnight at 37 °C at pH = 5. These solutions were then neutralized with 1 M tris (pH = 7.5) before transfer to wells precoated with L-243 antibody. Binding of biotinylated peptide was detected with ExtrAvidin-horseradish peroxidase conjugate (Sigma, St Louis, MO) and biotinylated avidin-specific monoclonal antibody (Sigma) and developed with o-phenylenediamine. The concentration of cold inhibitor required to produce 50% inhibition of the binding of the labeled invariant chain peptide (IC-50) was calculated, and the relative binding affinity of each peptide expressed as the reciprocal of the IC-50.

Results

Computer calculations of the PPP for self-peptides eluted from *DRB1*1301* and *DRB1*1302* were performed to determine whether this method was able to correctly predict that these peptides would be likely to be processed from the endogenous source proteins. Figure 2 shows that most self-peptides were in the top 1% of peptides in the source protein (i. e., >99′th centile). The peptide which performed worst was the Cathepsin-S-derived peptide, which scored only in the top 5% of peptides. It is interesting to speculate whether peptides derived from proteases involved in processing may be processed slightly differently due to their relative resistance to the degradative machinery in the class II processing pathway.

The success of these results must be tempered by consideration of the fact that the sequencing was performed on a pool which may have contained these peptides. However, we consider this is unlikely to play a major role in biasing the predictions because: 1) any large peaks on HPLC were removed prior to sequencing. The invariant chain and class I peptides formed part of the large peaks consistently removed and yet these still score highly; 2) analysis by mass spectrometry of multiple individual peptide peak from HPLC showed that there was a great abundance of peptides present even in apparently homogeneous HPLC peaks, making it unlikely that a few peptides contributed greatly to the pool; 3) a large number of different peptides were analyzed (seven in the case of *DRB1*1302*) which show no obvious sequence similarities except where they conform to a simple binding motif (Davenport et al. 1995). Therefore it is hard to see how each of these could sufficiently bias the pool enough to affect results.

Nonetheless we considered it important to assess the ability of this program to predict epitopes from pathogen-derived proteins. We did this by two independent means. First we looked at the ability of the program to predict the presentation of a known epitope. Few well-defined DR13-restricted T-cell epitopes are currently known for which the molecular HLA type is known. The influenza haemagglutinin protein peptide HA306–319 is the dominant epitope for the *DRB1*1302*-restricted response to this protein (C. Gelder, personal communication), and scored highly on the program (see Figure 2).

The other method we chose to analyze the efficacy of the program was to compare the binding of synthetic peptides with their calculated PPP (Fig. 3). In this case the PPP shown is the maximum value from the [N-terminal minus three] position to the [C-terminal minus 8] position. This was to take into account the fact that we were measuring only binding here, and that: 1) the first three amino acids of eluted peptides rarely if ever contain key anchor residues for MHC binding; 2) the score for the last amino acid of a peptide represents the score for the subsequent 15 amino acids. Therefore we considered that the [C-teminus minus 8] position is the last one which would allow sufficient length of the peptide to interact with the MHC molecule efficiently. Comparison of the observed binding and predicted PPP shows significant correlation (Fig. 3), indicating that the program is of value in predicting the binding efficiency of peptides independent of processing.

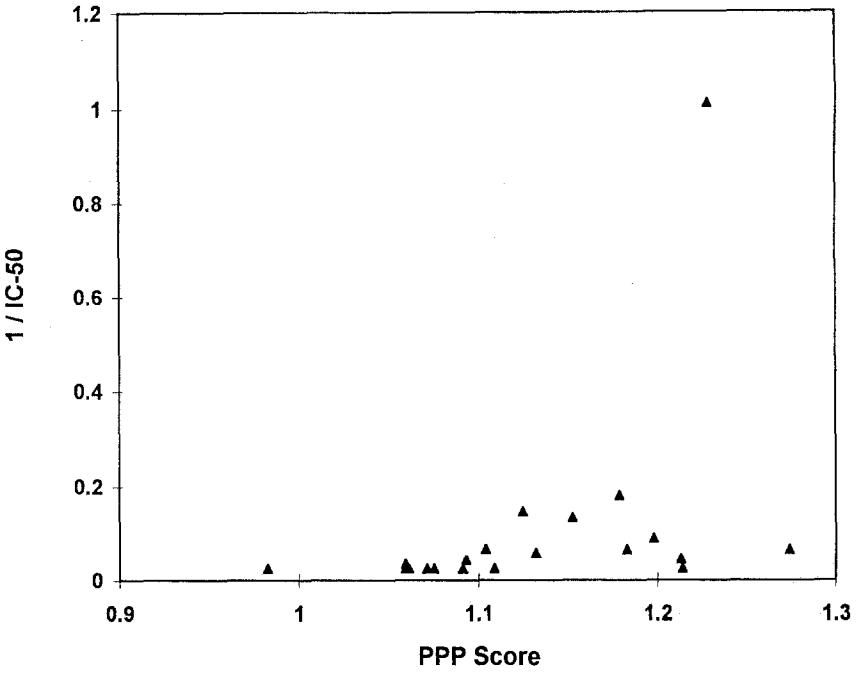

Fig. 3 Comparison of predicted scores with observed binding. The concentration of synthetic peptide required to inhibit 50% of the binding of a biotinylated peptide (IC-50) to *HLA-DRB1*1302* in vitro was measured. The reciprocal of this (1/IC-50) allows a comparison of the relative affinity of the peptides. Here the maximum PPP score for each peptide is plotted against the observed 1/IC-50 and shows a correlation between these two factors ($r = 0.388$, $P < 0.05$ (one-tailed p-value))

Discussion

The identification of allele-specific motifs for peptide binding to MHC molecules by analysis of individual and pool sequences of eluted peptides has allowed the prediction of MHC class I-restricted epitopes from pathogen-derived proteins. More recent evidence has suggested that binding motifs for MHC class I may be more complex than these simple motifs would suggest (Chen et al. 1994). In particular, the importance placed on 'anchor' residues is reliant on observations of an enrichment for one or a few related amino acids at certain positions. More subtle trends such as small preferences for amino acids or dominant negative effects of some amino acids at some positions may be missed by these methods. Motifs for MHC class II molecules have tended to be somewhat less specific than for MHC class I. In part this arises from the variable N-terminal extensions of the naturally processed peptides, which has hampered alignment studies of individual peptides and reduces the usefulness of pool sequencing (Chicz et al. 1992; Falk et al. 1994). The ability of class II molecules to exchange peptides has permitted the extensive use of binding assays (O'Sullivan et al. 1991) and even phage display libraries (Hammer et al. 1992). However, the use of multiple methods for motif determination has complicated interpretation of class II motifs. HLA-DR1 has been analyzed using pool sequencing (Falk et al. 1994) and individual peptide sequencing (Chicz et al. 1992) of eluted peptides, peptide binding (O'Sullivan et al. 1991), and phage display (Hammer et al. 1992), and the results of these have not always been consistent.

Two groups have recently described a method for analyzing the relative contributions of each amino acid at each position of peptides binding to MHC class II (Reay et al. 1994; Hammer et al. 1994b). Using peptides substituted with each amino acid at each position they were able to construct a hierarchy of peptide binding for each position. A computer algorithm was then used to score regions of a protein for their predicted binding ability. However, the use of peptide binding assays to derive the data used for epitope prediction in these systems raises some potential problems. On a purely practical level, large numbers of synthetic peptides and relatively large amounts of purified MHC are required for the binding assays. Moreover, these studies measure the ability of a peptide to bind in vitro. They cannot directly address whether this peptide will be presented in vivo for the following reasons: 1) this method can only address the question of binding and not processing, despite evidence that endogenous peptides seem to have a strong preference for some amino acids at positions thought to be outside the binding groove (most notably proline at position 2; Falk et al. 1994); 2) binding is performed in vitro in artificial solutions in which the MHC-peptide interactions may differ from those in vivo. In particular, pH may have unpredictable and inconsistent effects on the binding of different peptides (Sette et al. 1992); 3) only equilibrium binding is measured in these competition assays, therefore the ability of peptides to form long-lived stable complexes with MHC and remain on the cell surface for extended periods at neutral pH is not able to be assessed; 4) only 9 amino acid positions were analyzed, despite evidence that the MHC molecule may interact with more residues than this (Stern et al. 1994).

The method we used utilizes information concerning what is naturally bound to the MHC class II molecule. Thus it reflects what has been naturally processed, bound to MHC class II in the appropriate compartment, and exists as a stable complex on the cell surface. This method is purely empirical and does not rely on theoretical concepts of what should bind or the conditions under which binding may occur. Thus, for example, the presence of proline at

position 2 receives a score of approximately 3 in each of our matrices, reflecting the natural abundance of proline at this position without requiring any assumptions as to why this occurs. In addition, whereas others have insisted on the necessity for a fixed spacing of putative MHC anchor residues (Hammer et al. 1994a), our method does not rely on this feature.

The *HLA-DR13* alleles analyzed have been associated with protection from chronic hepatitis B virus infection (Thursz et al. 1995) and from cervical cancer associated with human papilloma-virus infection (Apple et al. 1994). In addition, *DRB1*1302* but not *DRB1*1301* has been associated with protection from severe malaria in studies in West Africa (Hill et al. 1991). We are currently using this program to predict candidate epitopes from known hepatitis B virus and *Plasmodium falciparum* proteins. We believe this method should be easily generalizable for use with other MHC class II alleles and will be useful in identifying T-cell epitopes for these alleles without requiring the synthesis of large numbers of synthetic peptides. For example, if the top 1% of peptides were synthesized this would lead to the identification of peptides which are naturally processed and presented in approximately 75% (9/12) of the cases we observed. This methodology could also easily be applied to MHC class I using similar information from pool sequencing. As in the case for class II, this information would reflect not only MHC binding, but also any possible constraints on processing and transport of peptides for MHC class I. The key to our approach is that it does not rely on any theoretical concepts regarding which residues are important for binding or what is required for processing. Using this method we believe an empirical analysis of data from eluted peptides should lead to more rapid and efficient identification of T-cell epitopes in antigenic proteins.

Acknowledgments We are grateful to B. Askonas and C. Gelder for sharing unpublished results, to M. Thursz and H.C. Thomas for generously providing the synthetic peptides from the hepatitis B virus surface antigen, and to A.C. Willis for pool sequencing. M.P.D. was supported by the Lionel Murphy Foundation and the Wellcome Trust. A.V.S. H. is a Wellcome Trust Senior Clinical Fellow. Copies of the computer program used are available by contacting I. HoShon at the address given or by email; MX%"ivan_ho_shon@sydpcug.org.au".

References

Apple, R. J., Erlich, H. A., Klitz, W., Manos, M. M., Becker, T. M., and Wheeler, C. M. HLA DR-DQ associations with cervical carcinoma show papillomavirus-type specificity. *Nat Genet 6:* 157–162, 1994

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., and Wiley, D. C. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature 364:* 33–39, 1993

Chen, W., Khilko, S., Fecondo, J., Margulies, D. H., and McCluskey, J. Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues. *J Exp Med 180:* 1471–1483, 1994

Chicz, R. M., Urban, R. G., Lane, W. S., Gorga, J. C., Stern, L. J., Vignali, D. A. A., and Strominger, J. L. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature 358:* 764–768, 1992

Chicz, R. M., Urban, R. G., Gorga, J. C., Vignali, D. A. A., Lane, W. S., and Strominger, J. L. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J Exp Med 178:* 27–47, 1993

Davenport, M. P., Quinn, C. L., Chicz, R. M., Green, B. N., Willis, A. C., Lane, W. S., Bell, J. I., and Hill, A. V. S. Naturally processed peptides from two disease-resistance associated HLA-DR13 alleles show related sequence motifs and the effects of the dimorphism at position 86 of the HLA-DRβ chain. *Proc Natl Acad Sci USA 92:* 6567–6571, 1995

Falk, K., Rötzschke, O., Stevanović, S., Jung, G., and Rammensee, H.-G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature 351:* 290–296, 1991

Falk, K., Rötzschke, O., Stevanović, S., Jung, G., and Rammensee, H.-G. Pool sequencing of natural HLA-DR, DQ, and DP ligands reveals detailed peptide motifs, constraints of processing, and general rules. *Immunogenetics 39:* 230–242, 1994

Gorski, J. and Mach, B. Polymorphism of human Ia antigens: gene conversion between two DR loci results in new HLA-D/DR specificity. *Nature 322:* 67–70, 1986

Hammer, J., Takacs, B., and Sinigaglia, F. Identification of a motif for HLA-DR1 binding peptides using M13 display libraries. *J Exp Med 176:* 1007–1013, 1992

Hammer, J., Belunis, C., Bolin, D., Papadopoulos, J., Walsky, R., Higelin, J., Danho, W., Sinigaglia, F., and Nagy, Z. A. High-affinity binding of short peptides to major histocompatibility complex class II molecules by short anchor combinations. *Proc Natl Acad Sci USA 91:* 4456–4460, 1994a

Hammer, J., Bono, E., Gallazzi, F., Belunis, C., Nagy, Z., and Sinigaglia, F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med 180:* 2353–2358, 1994b

Hill, A. V. S., Allsopp, C. E. M., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennet, S., Brewster, D., McMichael, A.J., and Greenwood, B. M. Common West African HLA antigens are associated with protection from severe malaria. *Nature 352:* 595–600, 1991

Hill, A. V. S., Elvin, J., Willis, A. C., Aidoo, M., Allsopp, C. E. M., Gotch, F. M., Gao, X. M., Takiguchi, M., Greenwood, B. M., Townsend, A. R. M., McMichael, A. J., and Whittle, H. C. Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature 360:* 434–439, 1992

Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, N.-Z., Arnott, D., Sherman, N., Shabanowitz, J., Michel, H., Bodnar, W. M., Davis, T. A., and Hunt, D. F. Definition of specific peptide motifs for four major HLA-A alleles. *J Immunol 152:* 3913–1924, 1994

O'Sullivan, D., Arrhenius, T., Sidney, J., DelGuercio, M.-F., Albertson, M., Wall, M., Oseroff, C., Southwood, S., Colon, S. M., Gaeta, F. C. A., and Sette, A. On the interaction of promiscuous antigenic peptides with different DR alleles. *J Immunol 147:* 2663–2669, 1991

Pamer, E. G., Harty, J. T., and Bevan, M. J. Precise prediction of a dominant class I MHC-restricted epitope of *Listeria monocytogenes. Nature 353:* 852–855, 1991

Reay, P. A., Kantor, R. M., and Davis, M. M. Use of global amino acid replacements to define the requirements for MHC binding and T cell recognition of moth cytochrome c (93–103). *J Immunol 152:* 3946–3957, 1994

Rothbard, J. B. and Taylor, W. R. A sequence pattern common to T cell epitopes. *EMBO 7:* 93–100, 1988

Rudensky, A. Y., Preston-Hurlburt, P., Hong, S.-C., Barlow, A., and Janeway, C. A. Sequence analysis of peptides bound to MHC class II molecules. *Nature 353:* 622–627, 1991

Sette, A., Southwood, S., O'Sullivan, D., Gaeta, F. C. A., Sidney, J., and Grey, H. M. Effect of pH on MHC class II-peptide interactions. *J Immunol 148:* 844–851, 1992

Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L., and Wiley, D. C. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature 368:* 215–221, 1994

Thursz, M. R., Kwiatkowski, D., Allsopp, C. E. M., Greenwood, B. M., Thomas, H. C., and Hill, A. V. S. Association between an HLA class II allele and clearance of hepatitis B virus in The Gambia. *N Engl J Med 332:* 1065–1069, 1995

Tiercy, J.-M., Gorski, J., Betuel, H., Friedel, A. C., Gebuhrer, L., Jeanner, M., and Mach, B. DNA typing of DRw6 subtypes: correlation with DRB1 and DRB3 allelic sequences by hybridisation with oligonucleotide probes. *Hum Immunol 24:* 1–14, 1989