Protein Science (2000), *9*:1838–1846. Cambridge University Press. Printed in the USA.

Structure-based prediction of binding peptides to MHC class I molecules: Application to a broad range of MHC alleles

ORA SCHUELER-FURMAN,[1,3] YAEL ALTUVIA,[1,3] ALESSANDRO SETTE,[2] AND HANAH MARGALIT[1]

[1]Department of Molecular Genetics and Biotechnology, The Hebrew University, Hadassah Medical School, POB 12272, Jerusalem 91120, Israel
[2]Epimmune, Inc., 5820 Nancy Ridge Drive, Suite 100, San Diego, California 92121

(RECEIVED April 5, 2000; FINAL REVISION June 22, 2000; ACCEPTED June 22, 2000)

Abstract

Specific binding of antigenic peptides to major histocompatibility complex (MHC) class I molecules is a prerequisite for their recognition by cytotoxic T-cells. Prediction of MHC-binding peptides must therefore be incorporated in any predictive algorithm attempting to identify immunodominant T-cell epitopes, based on the amino acid sequence of the protein antigen. Development of predictive algorithms based on experimental binding data requires experimental testing of a very large number of peptides. A complementary approach relies on the structural conservation observed in crystallographically solved peptide-MHC complexes. By this approach, the peptide structure in the MHC groove is used as a template upon which peptide candidates are threaded, and their compatibility to bind is evaluated by statistical pairwise potentials. Our original algorithm based on this approach used the pairwise potential table of Miyazawa and Jernigan (Miyazawa S, Jernigan RL, 1996, *J Mol Biol 256*:623–644) and succeeded to correctly identify good binders only for MHC molecules with hydrophobic binding pockets, probably because of the high emphasis of hydrophobic interactions in this table. A recently developed pairwise potential table by Betancourt and Thirumalai (Betancourt MR, Thirumalai D, 1999, *Protein Sci 8*:361–369) that is based on the Miyazawa and Jernigan table describes the hydrophilic interactions more appropriately. In this paper, we demonstrate how the use of this table, together with a new definition of MHC contact residues by which only residues that contribute exclusively to sequence specific binding are included, allows the development of an improved algorithm that can be applied to a wide range of MHC class I alleles.

Keywords: knowledge-based prediction; MHC class I; pairwise potential; peptide-protein binding; predictive algorithm

Cytotoxic T-cells (CTLs) recognize short peptides of 8–10 residues, the processing products of a protein antigen, bound to major histocompatibility complex (MHC) class I molecules. The MHC-peptide complex is presented to the T-cell receptor on the cell surface of antigen presenting cells (APCs). Since CTLs play a major role in the immune system, the prediction of potential peptides that can elicit a T-cell response has important implications for rational design of peptide vaccines and cancer therapy. One of the major factors that determine which peptides, along a protein sequence, will be immunogenic is their ability to bind to MHC molecules. Furthermore, it was shown that high binding affinity correlates with immunogenicity (Chen et al., 1994; Sette et al., 1994b; Ressing et al., 1995). Hence, most of the antigenic peptides are expected to be included within the subgroup of high affinity binding peptides, and therefore prediction of MHC binding peptides should narrow down remarkably the number of candidates for T-cell epitopes.

Binding of peptides to MHC molecules is allele specific. Sequence requirements for binding were defined by investigating sets of peptides with the capability to bind to a given MHC molecule. These studies enabled the characterization of simple sequence motifs that have been used extensively to predict antigenic peptides along a protein sequence (Rammensee et al., 1995). An alternative to the motif based prediction is the weight matrix approach (Sette et al., 1989; Ruppert et al., 1993; Parker et al., 1994; Gulukota et al., 1997; Sturniolo et al., 1999). By this approach, a matrix with weights for each of the amino acid residues in every position along the peptide can be generated for a given MHC allele, based on experimental binding data of large ensembles of sequence variants. Peptide candidates can be assigned scores by summing up the position specific weights, based on their sequence and the appro-

Reprint requests to: Hanah Margalit, Department of Molecular Genetics and Biotechnology, The Hebrew University, Hadassah Medical School, PO Box 12272, Jerusalem 91120, Israel; e-mail: hanah@md2.huji.ac.il.
[3]These authors contributed equally to this work.

priate matrix. An advantage of using the matrix approach, in comparison to the motif-based approach, is that it covers a wider range of peptides with binding potential and that it gives a quantitative score to each peptide, enabling the ranking of the predicted peptides. A serious obstacle is that the generation of a matrix for each MHC allele requires the experimental testing of hundreds of peptides.

An alternative approach to the sequence-based prediction schemes relies on structural information (Altuvia et al., 1995, 1997). Here, no experimental testing of a large set of different peptides for each allele is required. It is based on known structures of peptide-MHC complexes and evaluates the compatibility of different peptides to fit into the binding groove of a distinct MHC molecule. This can be obtained by threading the peptides through their structural templates and obtaining a rough estimate for the binding energy of a peptide in the groove, based on the interactions between the peptide and the MHC residues, as defined in a solved crystal. A solved structure, or a structural model based on a similar MHC allele, provides the information for generating a list of MHC contact residues for each peptide position. The contributions of the pair interactions between each peptide residue and its neighbors in the MHC structure are summed up according to a knowledge-based pairwise potential matrix. The peptide score is obtained by summing up the "energy" values along all peptide positions, enabling the ranking of different peptides by these scores, and choosing the peptides with lowest scores ("binding energies") as best candidates for MHC ligands (Altuvia et al., 1995, 1997).

The application of the structural approach is based on three determinants: (1) The availability of structural templates for peptides bound to different MHC alleles; (2) the choice of a pairwise potential table; and (3) the criterion for determining MHC and peptide positions that are in contact.

Various knowledge-based pairwise potentials have been derived from known protein structures (reviewed in Jernigan & Bahar, 1996; Jones & Thornton, 1996; Skolnick et al., 1997). A basic approximation underlying these potentials is that the total "free energy" of a protein can be expressed as a sum of independent pairwise interactions. The frequency of amino acid pair residues that are in contact in protein structures is assumed to represent the interaction preference between the two amino acids involved. This interaction preference between two amino acids is expressed by its comparison with their affinity to a "reference state." Various matrices have been published, differing in their character essentially due to the distinct reference states used (Skolnick et al., 1997). In one of the widely used tables of Miyazawa and Jernigan (1985, 1996), the solvent has been used as the reference state, and the resulting table puts much emphasis on hydrophobic interactions. Recently, Betancourt and Thirumalai (1999) have modified the table of Miyazawa and Jernigan by changing the reference state from solvent to a defined, single solvent-like molecule: the amino acid threonine. The resulting matrix represents the physical character of the hydrophilic interactions more adequately. In our previous studies, we have used the pair potentials of Miyazawa and Jernigan to successfully identify binding peptides to MHC class I alleles of hydrophobic character (Altuvia et al., 1995, 1997). However, the algorithm failed when applied to other MHC class I alleles with hydrophilic pockets (Altuvia et al., 1997). Here, we demonstrate how the pairwise potential table of Betancourt and Thirumalai (1999) leads to a significant improvement in the performance of the algorithm, enabling its application to MHC alleles with binding pockets of various character.

The algorithm uses a list of MHC contact residues for each peptide position. Peptide-MHC binding involves general as well as sequence-specific interactions. The general interactions involve a conserved network of hydrogen bonds between MHC side-chain atoms and the peptide backbone. The sequence-specific interactions can be largely represented by contacts between MHC residues and the peptide side chains. For the algorithm, we need to capture side-chain–side-chain interactions that constitute the binding specificity. However, an MHC side chain may be in contact with a peptide side chain as a consequence of its interaction with the peptide backbone, and its explicit contribution to specificity cannot be ascertained. Here, we demonstrate that when the definition of MHC contact residues takes this consideration into account, the algorithm is significantly improved.

In summary, the use of a pairwise potential matrix that describes the hydrophilic interactions more appropriately, together with the new definition of MHC contact residues by which only residues that contribute exclusively to sequence specific binding are included, allows the development of an improved algorithm that can be applied to a wide range of MHC class I alleles.

Results

In an attempt to improve the performance of the threading algorithm, especially for MHC alleles with dominant hydrophilic interactions, four alternative versions of the algorithm were tested. They vary in at least one of the following parameters (see Materials and methods for definition): (1) Definition of MHC contact residues (*SS*, MHC *S*ide chain − peptide *S*ide chain interactions; or *SS-SB*, *SS* without MHC *S*ide chain − peptide *B*ackbone interactions). (2) Table of statistical pairwise potentials (*MJ*, table of *M*iyazawa and *J*ernigan (1996); or *BT*, table of *B*etancourt and *T*hirumalai (1999)).

Our basic strategy was to evaluate the performance of each of these versions in predicting the binding of peptides to different MHC class I alleles. The best version of the algorithm should perform reasonably well for a broad range of MHC alleles.

Our analyses were based on the currently available structural data of MHC class I–peptide complexes. Since the length of the peptide that constitutes the structural template determines the length of the threaded peptides, we have organized a data set of known binding peptides of that length for each MHC allele (see Materials and methods). A total of 14 different data sets were compiled, based on the combinations of MHC class I allele and peptide length. For each version of the algorithm, the performance was evaluated using the rank analysis, i.e., by ranking a known binding peptide among a list of overlapping peptides of same length derived from its source protein sequence. The version that performed best in the training set was applied to the test set. In addition, experimentally measured binding values were available to us for peptides in 7 out of the 14 different groups, and the best version of the algorithm was further tested on these data.

Choosing the optimal parameters for the algorithm

A good algorithm is expected to rank good MHC binders high within all overlapping same-length peptides spanning the corresponding source protein sequences. The peptides used for this analysis were either immunogenic peptides (IGPs), or naturally processed peptides (NPPs), or peptides from the cocrystals (CPs) (see Materials and methods). All these are *known* binding peptides.

This set of peptides (and their corresponding protein sequences) was divided into a training set and a test set. All CPs were included in the training set. The IGPs and NPPs were grouped according to their MHC restriction and length, and the peptides in each group were equally and randomly distributed between the training and the test sets. The final training and test sets included 123 and 106 peptides, respectively (Table 1). For each combination of parameters, the algorithm was applied to rank all the peptides in the training set.

The performance of the different algorithms on the training set was compared for each group separately (i.e., on the 14 different MHC allele–peptide length combinations), as well as for the whole group of sequences in the training set. The peptides were ranked according to their energy values. The first rank was assigned to the peptide predicted to be the best binder, i.e., to the peptide with lowest energy. Subsequently, the ranks were normalized and expressed in fractions. The ranks of all known binding peptides within their source protein were recorded. Three criteria were used to evaluate the prediction: (1) The number of known binding peptides ranked within the first 15% of all overlapping peptides of same length within their respective sequences (rank ≤ 0.15). (2) Average rank of all the known binding peptides. (3) "Robustness" of algorithm ("average success"). For each group, the fraction of known binding peptides predicted within the first 15% was calculated. The average of these values represents a measure that is less sensitive to the large differences in the size of the different groups and should therefore reflect the overall robustness of the algorithm.

Table 2A summarizes the performance of different versions of the algorithm on the training set. The best overall performance was obtained by the algorithm that used the *SS-SB* definition of MHC contact residues and the table *BT* of statistical pairwise contact potentials derived by Betancourt and Thirumalai (1999). It ranked 87 out of 123 peptides (71%) within the first 15%. The average rank for all the training set peptides was 0.13 and the robustness 0.7. Detailed examination of the table leads to similar conclusions. For 11 out of the 14 different groups, this combination of parameters resulted in the highest number of known binding peptides ranked within the first 15%. For HLA-B2705 the pairwise potential table *BT* performed better than table *MJ*; however, in this case, the MHC contact residues definition *SS* worked better. For HLA-A0201 (nonamers) and HLA-A6801 (decamers), table *MJ* seemed to work better. The choice of the *BT*/*SS-SB* parameter combination is also supported by the average rank criterion, although its advantage is less eminent when looking at each MHC allele separately.

The selected combination of parameters (*SS-SB* and table *BT*) was subsequently tested on the independent test set (see Table 2B). The overall performance was similar, but somewhat reduced in the test set. The relative rank of immunogenic peptides was slightly worse: 0.15 compared to 0.13 in the training set. Sixty-four percent of the known binding peptides were ranked within the first 15% of all peptides of same length derived from the source proteins (compared to 71% in the training set), and the overall robustness decreased from 0.7 to 0.62. The performance in the different groups was similar for the test and training sets. An improvement could be seen for some groups, especially in HLA-B2705, while a drop in performance was observed for others, especially in HLA-A6801 (decamers) and H-2D^b (nonamers). The two evaluation measures did not always show the same tendency: in four cases the test set gave better performance with one of the measures and worse per-

Table 1. *MHC-peptide solved complexes and experimental binding data used in this study*

	Data for ranking analysis					Binding data[b]			
MHC allele[a]	*IGP*s	*NPP*s	*CP*s	Train	Test	Good binders	Nonbinders	All[c]	Structural data[d]
HLA-A0201 (9)	86	—[e]	4	46	44	62	202	518	1hhi[1], 1hhk[1], 1ao7[2], 1bd2[3], 1hhj[1], 1akj[4], 1hhg[1], 1b0g[5]
HLA-A0201 (10)	26	—[e]	2	15	13	27	100	265	1hhh[1]
HLA-A6801 (9)	—	2	1	2	1	21	35	130	a68*[6]
HLA-A6801 (10)	1	2	—	2	1	10	7	51	1tmc[7]
HLA-B2705 (9)	6	16	—	11	11	11	44	66	1hsa[8], b27*[8]
HLA-B3501 (8)	2	—	1	2	1				1a1n[9]
HLA-B3501 (9)	11	2	—	7	6				1a9b[10], 1a9e[10]
HLA-B5301 (9)	—	4	2	4	2				1a1m[11], 1a1o[11]
HLA-B0801 (8)	4	1	1	4	2				1agb[12], 1agc[12], 1agd[12], 1age[12], 1agf[12]
H-2D^b (9)	11	2	2	8	7	10	7	24	1hoc[13], 1ce6[14], 1bz9[5]
H-2D^d (10)	3	—	1	2	2				1bii[15], 1ddh[16]
H-2K^b (8)	15	4	3	12	10	22	12	63	2vaa[17], 1bqh[18], 1osz[19], 1vac[20], 2ckb[21]
H-2K^b (9)	2	—	1	2	1				2vab[17], 1vad[20]
H-2L^d (9)	9	—	2	6	5				1ld9[22], 1ldp[23]
Total	176	33	20	123	106				

[a] Length of peptide is given in parentheses.

[b] Binding was measured by competition experiments and expressed in relative binding values (see Materials and methods).

[c] Includes good, intermediate, weak and nonbinders.

[d] PDB code, or * coordinates kindly provided by the authors. Underlined names indicate that the coordinates are available for more than one complex. The references for the different structures are listed in the following: [1] Madden et al. (1993); [2] Garboczi et al. (1996); [3] Ding et al. (1998); [4] Gao et al. (1997); [5] Zhao et al. (1999); [6] Silver et al. (1992); [7] Collins et al. (1995); [8] Madden et al. (1992); [9] Smith et al. (1996b); [10] Menssen et al. (1999); [11] Smith et al. (1996a); [12] Reid et al. (1996); [13] Young et al. (1994); [14] Glithero et al. (1999); [15] Achour et al. (1998); [16] Corr et al. (1993); [17] Fremont et al. (1992); [18] Kern et al. (1998); [19] Ghendler et al. (1998); [20] Fremont et al. (1995); [21] Garcia et al. (1998); [22] Balendiran et al. (1997); [23] Speir et al. (1998).

[e] *NPP*s were not included for HLA-A0201, since this allele is already overrepresented.

Table 2. *Performance of algorithm in ranking a peptide within its protein sequence*

	(A) Training set								(B) Test set		
MHC allele[a]	Number of peptides	*MJ/SS*[b]		*MJ/SS-SB*		*BT/SS*		*BT/SS-SB*	Number of peptides	*BT/SS-SB*	
HLA-A0201 (9)	46	0.07[c]	*41*[d]	0.09	*40*	0.17	*28*	0.10 *35*	44	0.15	*27*
HLA-A0201 (10)	15	0.14	*9*	0.15	*9*	0.22	*9*	0.18 *9*	13	0.08	*11*
HLA-A6801 (9)	2	0.52	*0*	0.47	*0*	0.04	*2*	0.07 *2*	1	0.02	*1*
HLA-A6801 (10)	2	0.10	*2*	0.10	*2*	0.12	*1*	0.18 *1*	1	0.69	*0*
HLA-B2705 (9)	11	0.52	*0*	0.57	*0*	0.08	*8*	0.15 *6*	11	0.07	*10*
HLA-B3501 (8)	2	0.33	*0*	0.27	*0*	0.27	*1*	0.20 *1*	1	0.04	*1*
HLA-B3501 (9)	7	0.24	*2*	0.23	*2*	0.29	*3*	0.27 *3*	6	0.25	*2*
HLA-B5301 (9)	4	0.24	*2*	0.17	*2*	0.12	*2*	0.08 *3*	2	0.13	*1*
HLA-B0801 (8)	4	0.30	*0*	0.21	*0*	0.02	*4*	0.01 *4*	2	0.06	*2*
H-2D[b] (9)	8	0.19	*3*	0.15	*4*	0.14	*5*	0.07 *7*	7	0.27	*1*
H-2D[d] (10)	2	0.70	*0*	0.72	*0*	0.31	*1*	0.43 *1*	2	0.33	*0*
H-2K[b] (8)	12	0.12	*9*	0.09	*9*	0.09	*9*	0.08 *10*	10	0.09	*9*
H-2K[b] (9)	2	0.19	*1*	0.17	*1*	0.10	*2*	0.10 *2*	1	0.02	*1*
H-2L[d] (9)	6	0.23	*2*	0.30	*2*	0.28	*3*	0.26 *3*	5	0.19	*2*
Total	123	0.19	*71*	0.19	*71*	0.16	*78*	0.13 *87*	106	0.15	*68*
Robustness[e]		0.37		0.38		0.66		0.7		0.62	

[a]Length of peptide is given in parentheses.
[b]Versions of algorithm: Table *MJ* or Table *BT* in combination with MHC contact residue definition *SS* or *SS-SB* (see Materials and methods).
[c–e]Measures of performance: [c]average rank of peptides within all peptides of same length derived from their respective protein sequences; [d]number of peptides ranked within the first 15% (in italics); and [e]robustness, average percent of peptides ranked within the first 15% in different sets of MHC allele-peptide length.

formance with the other (see, for example, the performance for H-2L[d]). Obviously, for some of the groups the size of the training and test sets was not large enough.

Testing the algorithm on experimental binding data

Distinction between good binders and nonbinders

A good predictive algorithm is expected to distinguish between binding and nonbinding peptides. Peptide sequences of good binders (relative binding value ≥ 0.1) and nonbinders (relative binding value ≤ 0.0001) were extracted from seven sets of peptides with experimental binding values (see Table 1). The differences between the predicted energy values of the good binders and nonbinders were evaluated by the Mann–Whitney test. As shown in Table 3, the algorithm could discriminate between good binders and nonbinders for six out of the seven sets. Notably, three of the MHC alleles, HLA-A6801, HLA-B2705, and H-2D[b], have at least one hydrophilic binding pocket. Table 3 provides also a comparison with the results based on the parameter combination used in our previous work (Table *MJ* in combination with MHC contact residue definition *SS*). As can be seen, the current algorithm succeeded much better than the previous one to predict binding of peptides to MHC alleles with dominant hydrophilic pockets, im-

Table 3. *Comparison between the predicted energy values of good binders and nonbinders*

	Current algorithm (*BT/SS-SB*)			Previous algorithm (*MJ/SS*)		
MHC allele	Good binders[a]	Nonbinders[a]	P-value[b]	Good binders	Nonbinders	P-value
HLA-A0201 (9)	-5.28 ± 0.68	-3.04 ± 0.95	$p < 0.0001$	-133 ± 6.5	-110 ± 8	$p < 0.0001$
HLA-A0201 (10)	-6.09 ± 0.68	-3.29 ± 1.18	$p < 0.0001$	-128 ± 4.8	-102 ± 7	$p < 0.0001$
H-2K[b] (8)	-3.32 ± 0.60	-1.78 ± 0.90	$p < 0.0005$	-110 ± 5.7	-94 ± 5.9	$p < 0.0005$
HLA-B2705 (9)	-1.54 ± 0.32	-0.94 ± 0.44	$p = 0.001$	-69 ± 4.2	-78 ± 7.8	—
HLA-A6801 (9)	-2.13 ± 0.75	-1.02 ± 0.74	$p = 0.0041$	-63 ± 4.2	-61 ± 10	—
H-2D[b] (9)	-3.77 ± 0.84	-3.02 ± 0.81	$p = 0.05$	-101 ± 7.2	-95 ± 6.6	$p = 0.05$
HLA-A6801 (10)	-3.65 ± 0.78	-2.88 ± 1.07	—	-116 ± 9.7	-106 ± 5	—

[a]Median energy $\pm$ average deviation in *RT* units, where *R* is the gas constant and *T* is the absolute temperature. Tables *MJ* and *BT* use a different reference state; therefore, the ranges of the values differ.
[b]*P*-value was obtained by Mann–Whitney test.

pairing only slightly the prediction of binding of peptides to MHC alleles with dominant hydrophobic amino acids.

Relative Operating Characteristic (ROC) curves

ROC is a method to measure the distinguishing power of a classification (e.g., Swets, 1988). Here, we used the ROC analysis to compare the accuracy of the algorithm used above for distinguishing binders from nonbinders.

A predicted peptide belongs to one of the following four categories: (1) *TP* (true positive), a binding peptide that is predicted to bind; (2) *FP* (false positive), a nonbinding peptide that is predicted to bind; (3) *TN* (true negative), a nonbinding peptide that is predicted not to bind; and (4) *FN* (false negative), a binding peptide that is predicted not to bind. The classification of a peptide to one of the four categories is dependent on a threshold value used to distinguish between binding and nonbinding peptides. Two parameters are used in the ROC analysis: (1) "*hits*",$TP/(FN + TP)$, representing the fraction of the binding peptides that were predicted to bind; and (2) "*false alarms*," $FP/(FP + TN)$, representing the fraction of the nonbinding peptides that were predicted to bind. In the ROC analysis, *hits* are plotted against *false alarms* for various threshold values. This curve shows the tradeoff between the *TP* and *FP* proportions as the threshold is changed. The fraction of the area lying below the curve is indicative of the distinguishing power of the algorithm.

Representative ROC curves are shown in Figure 1 for HLA-A0201 (nonamers), HLA-A6801 (nonamers), and HLA-B2705 (nonamers). The new version of our algorithm was more discriminative than the previous one for the alleles with hydrophilic pockets HLA-A6801 and especially for HLA-B2705, while it was only slightly impaired for HLA-A0201.

Correlation analysis

Given a set of peptides with known binding values, the correlation between the experimental binding values and the predicted energy values could be calculated. In practice, since the prediction provides pseudo-energy binding values and the experimental binding values are expressed in IC_{50} (where K, the binding constant is proportional to $1/IC_{50}$, and the free energy $\Delta G/RT = \ln(1/K)$), the correlation was calculated between the predicted values and the natural logarithm of IC_{50} values.

The algorithm was applied separately to each of the sets above (see Table 1), and the correlation between the predicted energy values and the known binding values was calculated using the Spearman rank correlation test. The results are summarized in Table 4. A significant correlation coefficient was obtained for six out of the seven sets, consistent with the results for distinction between binders and nonbinders. Clearly, the correlation coefficients for MHC alleles with hydrophilic binding pockets are much improved in comparison to the results obtained by the previous algorithm.

Discussion

Previously, we have shown that the threading approach can be successfully applied to the prediction of peptides that bind to MHC class I alleles of hydrophobic character (Altuvia et al., 1995, 1997). An adjustment of this algorithm allows us now to apply it to MHC class I alleles with a wide range of pocket types. This is demonstrated by an efficient estimation of binding energies for a large number of different peptides. The computed values are well correlated with the experimentally derived binding values (Table 4) and allow the distinction between binders and nonbinders (Table 3; Fig. 1). We use this approach to predict the binding peptide from its source protein sequence (Table 2). The successful application to alleles that include hydrophilic pockets is accompanied by only a slight reduction in performance for the alleles that contain pockets with hydrophobic preferences that were successfully predicted by the original algorithm.

As stated before, the successful application of threading is mainly dependent on three features: (1) The list of MHC-peptide contacts; (2) the pairwise potential table; and (3) the set of template structures. In the following, we discuss how these influence the performance of our algorithm.

List of contacts

For the evaluation of the binding ability of a peptide to an MHC molecule, one derives for each peptide position the contacting MHC residues. This is expected to capture the sequence-specific aspect: If a sequence fits optimally to a certain fold, there will be many favorable residue–residue interactions involving side chains, since these represent the sequence specific character. Analysis of solved structures of peptides bound to MHC class I molecules has revealed a conserved peptide backbone structure, connected to MHC side chains by a net of conserved hydrogen bonds. Thus,

Table 4. *Correlation between predicted energy values and experimental binding values*

MHC allele	Number of peptides	Current algorithm (*BT/SS-SB*)		Previous algorithm (*MJ/SS*)	
		Correlation coefficient	*P*-value[a]	Correlation coefficient	*P*-value
HLA-A0201 (9)	518	−0.57	$p < 0.0005$	−0.68	$p < 0.0005$
HLA-A0201 (10)	265	−0.61	$p < 0.0005$	−0.67	$p < 0.0005$
H-2K^b (8)	63	−0.46	$p < 0.0005$	−0.50	$p < 0.0005$
HLA-B2705 (9)	66	−0.39	$p < 0.001$	0.27	—
HLA-A6801 (9)	130	−0.2	$p < 0.025$	−0.02	—
H-2D^b (9)	24	−0.47	$p < 0.025$	−0.41	$p < 0.025$
HLA-A6801 (10)	51	−0.07	—	−0.06	—

[a] *P*-value was obtained by a Spearman rank correlation test.

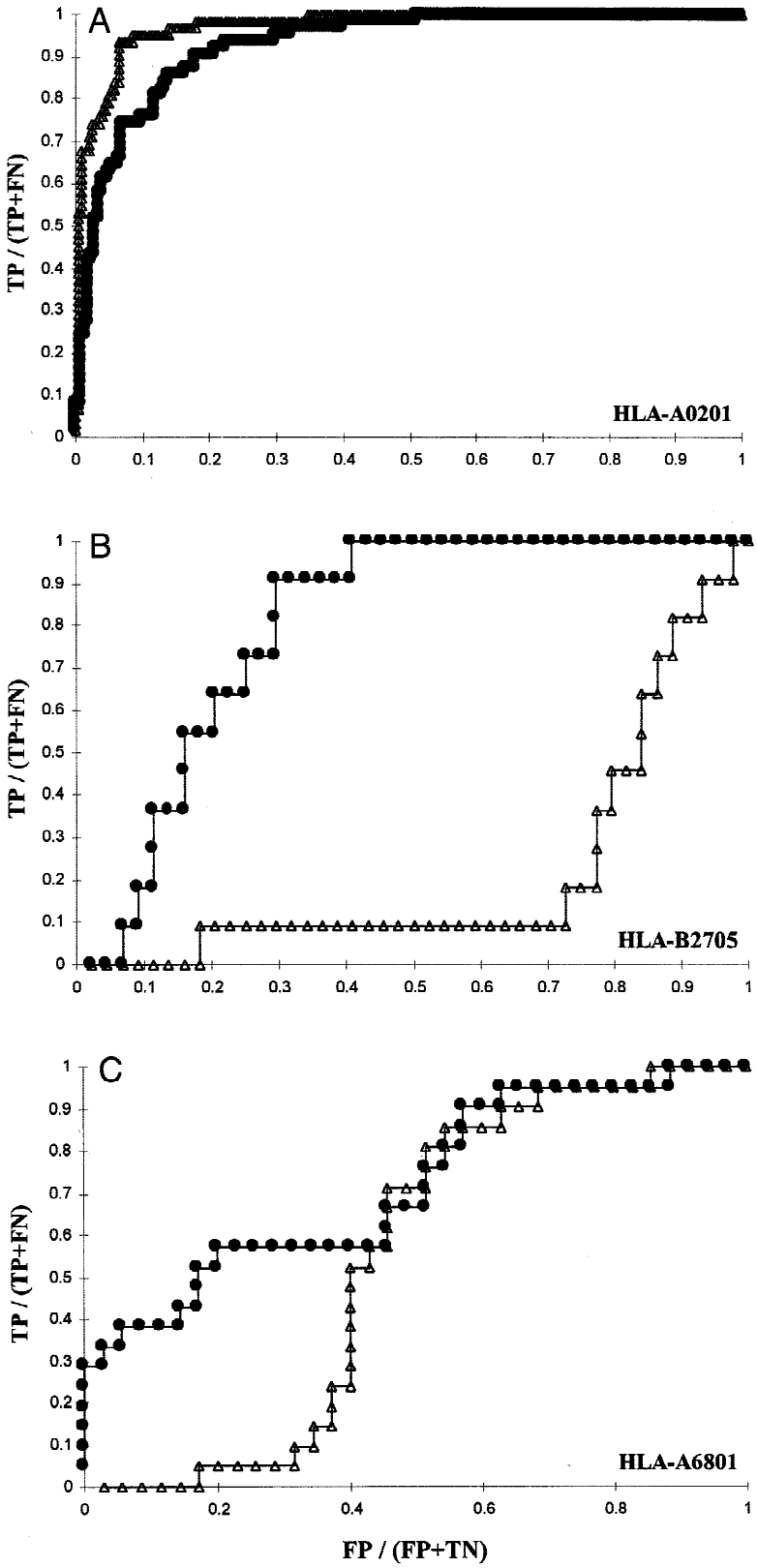

Fig. 1. ROC curves assess the accuracy of the algorithm in distinguishing between binders and nonbinders. The algorithm was applied with two different parameter combinations: White triangles and black circles represent the results of the original parameter combination (i.e., *MJ/SS*) and the new parameter combination (i.e., *BT/SS-SB*), respectively. The curves for three different MHC alleles are shown: (**A**) HLA-A0201 (nonamers), (**B**) HLA-B2705 (nonamers), and (**C**) HLA-A6801 (nonamers). The current version of the algorithm is clearly better for HLA-B2705 and HLA-A6801, while the performance for HLA-A0201 is only slightly affected. The absolute number of good binders and nonbinders is given in Table 1. See text for more details.

several MHC side chains contribute to the peptide binding by stabilizing its backbone. To catch only features that are specific to the peptide sequence, these MHC residues should not be included in the list. For example, the backbone of nonamers that bind to MHC allele HLA-A2 shows a conserved hydrogen bonding network in all solved structures. This network involves hydrogen bonds between the C-terminal peptide position and the following residues of the MHC chain: D77, T143, Y84, and K146 (except in *1ao7*). In addition to these residues, other residues contact the backbone in all structures, although not necessarily through a hydrogen bond: T80 and in some cases W147. Exclusion of these residues from the contact list leads to a better description of the pocket character of the C-terminal anchor position. For structure *1hhi* for example, the original list of contact residues (D77, L81, Y123, T143, and W147) is reduced to three residues of hydrophobic character (L81, Y123, and W147). The new list of MHC contact residues results in a significant improvement of our algorithm.

Pairwise potential

The Miyazawa and Jernigan table has been very useful in applications of threading to fold recognition (Miyazawa & Jernigan, 1996). There, a protein sequence is threaded through the backbone template of structures in a structural library and the best sequence-structure fit is searched for. The pairwise potential table is used to estimate the energies of the sequence threaded upon the different structural templates. The success of tables such as the Miyazawa and Jernigan table stems from the nature of the problem. The protein structure is mainly stabilized by hydrophobic interactions in the buried core, and these interactions determine essentially which sequence will fit to the protein fold. Pairwise potential tables with favorable hydrophobic interactions clearly succeed in this task. In fact, successful fold recognition is obtained even with a simple HP model (Hydrophobic-Polar model) that defines the HH interactions as favorable, while HP and PP contacts are equivalent to solvent contacts (e.g., Huang et al., 1995; Thomas & Dill, 1996). The application of threading to the peptide-MHC binding confronts a different situation. Although most peptides are relatively buried within the binding groove of the MHC molecule, one cannot assume that hydrophobic interactions are the only ones that will tell binding from nonbinding peptides apart. Since MHC alleles contain pockets of various characters, we need a table that will express nonhydrophobic interactions more adequately. In principle, an alternative table that may overcome these difficulties is a table derived from interactions in the interfaces of complexes. Indeed, Keskin et al. (1998) have recently applied the same approach of Miyazawa and Jernigan (1996) to derive a pairwise potential table for intermolecular interactions. However, the resulting table is very similar to the original *MJ* table for intramolecular interactions (Keskin et al., 1998), and therefore did not improve significantly the predictions for MHC alleles with hydrophilic pockets (data not shown). Betancourt and Thirumalai have modified the original *MJ* table by changing the reference state from solvent to threonine (Betancourt & Thirumalai, 1999). The new table does not rely on the estimation of contacts involving the solvent, as it is based on a well-defined reference state, composed of a single, solvent-like molecule. The amino acid threonine was chosen since as reference state it reproduces best the hydrophobicities of the different amino acid types. Now hydrophilic interactions such as charge–charge interactions are adequately presented. Indeed, this table results in significantly improved predictions for MHC alleles of nonhydrophobic specificity, while the prediction for hydrophobic molecules is not significantly affected.

Structural template

A structural template is essential to apply the threading approach to the prediction of binding peptides. The number of individual templates, as well as their quality, influences the performance of the prediction. As the number of solved structures increases, more MHC alleles can be evaluated. If several structures are available for a combination of peptide length–MHC allele, the algorithm is expected to be more robust. As described below for HLA-A6801, the abrupt cutoff used to define MHC contact residues can result in significant differences of the neighbor lists. Such effects can be reduced when more structural templates are available.

We would expect a similar performance of our algorithm for data sets of the same MHC allele in combination with peptides of different lengths. This is true for alleles HLA-A0201, HLA-B3501, and H-2K^b, but not for HLA-A6801. The performance for the latter allele, in combination with decamers, is unsatisfactory in all the tests that were presented here. The rank of the decamer peptide in the test set is 0.69, no significant correlation between the predicted and the experimental binding values, and no statistically significant distinction between good binders and nonbinders were obtained. In this case, the structural template may not be general enough. The structures of HLA-A6801 complexed with a nonamer and a decamer are largely conserved; however, the list of contact residues for equivalent peptide positions varies. For example, the C-terminal anchor position shares common neighbors I95 and D116, but in addition contains D74 for nonamers and L81 and W147 for decamers. According to the motif based on known peptide sequences that bind to HLA-A6801, this anchor position shows a preference for a positively charged amino acid (K or R). Obviously, the list of neighbors of the nonamer will capture this better than the list of the decamer, since it contains an additional negatively charged amino acid (D), instead of the two hydrophobic amino acids (L and W).

The present algorithm can be applied to predict binding peptides to a wide range of MHC alleles. In practice, it is equivalent and therefore complementary to a weight matrix. For a given MHC allele, a defined set of contact residues is used, and the summation of the pairwise potential values for each amino acid in every peptide position establishes a weight matrix. Thus, the running time of the algorithm is short and it may evaluate large ensembles of peptides very quickly, enabling computational screening of peptide libraries or screening of all proteins in a genome. Still, if binding data are available the weight matrix approach based on these data (e.g., Sette et al., 1989; Ruppert et al., 1993; Parker et al., 1994; Kondo et al., 1995, 1997; Sidney et al., 1996a, 1996b; Gulukota et al., 1997; Southwood et al., 1998; Sturniolo et al., 1999) is preferable. Weight matrices that are derived from binding data are based on input of the same nature as the prediction output and probably therefore lead to better ranking of the known binding peptides. However, for many MHC alleles no (or not enough) binding values are available for the generation of such matrices. Since the present algorithm does not rely on binding data, it can be used for those alleles when a structural template is available. Even without available crystallographically solved complexes, the highly conserved structure of different MHC-peptide complexes suggests that the generation of structural models for other alleles and their use as templates for the threading approach should be possible.

Currently, a web version of the algorithm is available for all MHC alleles analyzed in this paper (http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/).

Materials and methods

The threading procedure

The main features of the threading procedure remain as described previously (see Altuvia et al., 1995, 1997). In brief, a peptide sequence is threaded onto the template, based on a crystal structure. The interaction energy of a peptide residue is calculated by summing up the pairwise energies according to a list of peptide–MHC contact residues. The score for the whole peptide is the sum of the interaction energies over all peptide positions.

The templates

The algorithm was applied to 10 MHC alleles. These are represented in 37 solved MHC-peptide cocrystals that contain 48 distinct complexes (listed in Table 1). The coordinate files of the MHC-peptide complexes were used to derive the lists of contact residues.

Definition of contact residues

Peptide and MHC residues were considered as contacting each other if at least one pair of their atoms was within 4 Å distance. For every peptide position, a list of contact MHC residues was created, based on the crystal structures. Intrapeptide interactions as well as intra-MHC interactions were not considered, since the former were shown to be of minor influence for a given backbone template (e.g., Schueler-Furman et al., 1998), and the latter are assumed to be constant. Two different lists were tested: (1) *SS*—Includes MHC residues which contact the peptide side chain via their side chain. (2) *SS-SB*—A subset of *SS* that includes MHC residues that contact the peptide residue via its side chain solely and do not have additional contacts via its backbone. *SS* and *SS-SB* lists were derived for all peptide positions in each available template. Lists were merged when several copies of an MHC-peptide pair appeared in the same crystal, or when the same MHC-peptide pairs appeared in different crystals. Merging was also applied to some cases with very similar peptide sequences (for example: *1agb*, *1agc*, *1agd*, *1age*, and *1agf*). The merged lists contained only MHC contact residues that were defined for at least 50% of the relevant templates.

Binding data

Two different types of binding data are available.

Qualitative data

Peptides that were detected in a complex with MHC molecules or are known to elicit an immune response. Only peptides for which the source protein sequence is known were included. The data were obtained from three different sources: (1) Immunogenic peptides (*IGP*)—peptides that elicit immune response by activating T-cells; derived from the database of Rammensee et al. (1995). (2) Naturally processed peptides (*NPP*)—peptides processed and presented by APCs; derived from the database of Rammensee et al. (1995). Since the former set is biased in favor of HLA-A2 peptides, those were not included in the *NPP* set. (3) Peptides from the MHC-peptide crystals (*CP*). For each of these peptides, its corresponding protein sequence was extracted. In total, the source sequences of 176 *IGP*s, 33 *NPP*s, and 20 *CP*s were extracted from three Protein Data Banks: SWISSPROT, PIR, and GENPEPT (see Table 1).

Quantitative data: Binding/nonbinding peptides (BNP)

These were synthetic peptides for which experimental binding values were available to us. Binding was measured in competition experiments with a standard peptide, and evaluated as the ratio between the IC_{50} of the standard peptide and that of the test peptide (hereinafter, relative binding) (Sette et al., 1994a). IC_{50} is the concentration required for 50% inhibition of the standard peptide. The peptides were defined as good/intermediate/weak binders and nonbinders according to their IC_{50} ratios (good binders, relative binding ≥ 0.1; intermediate binders, $0.1 >$ relative binding ≥ 0.01; weak binders, $0.01 >$ relative binding > 0.0001; nonbinders, relative binding ≤ 0.0001).

Acknowledgments

This study was supported by grants from the US-Israel Bi-national Science Foundation granted to H.M. and A.S., the Israel Cancer Research Fund granted to H.M., and in part by NIH-NIAID grant NOI-AI-95362 to A.S. O.S.-F. is supported by The Abisch-Fraenkel Foundation and The Clore Foundation.

References

Achour A, Persson K, Harris RA, Sundback J, Sentman CL, Lindqvist Y, Schneider G, Karre K. 1998. The crystal structure of H-2D^d MHC class I complexed with the HIV-1-derived peptide P18-I10 at 2.4 Å resolution: Implications for T cell and NK cell recognition. *Immunity 9*:199–208.

Altuvia Y, Schueler O, Margalit H. 1995. Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol 249*:244–250.

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.

Balendiran GK, Solheim JC, Young AC, Hansen TH, Nathenson SG, Sacchettini JC. 1997. The three-dimensional structure of an H-2L^d-peptide complex explains the unique interaction of L^d with β_2 microglobulin and peptide. *Proc Natl Acad Sci USA 94*:6880–6885.

Betancourt MR, Thirumalai D. 1999. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci 8*:361–369.

Chen Y, Sidney J, Southwood S, Cox AL, Sakaguchi K, Henderson RA, Appella E, Hunt DF, Sette A, Engelhard VH. 1994. Naturally processed peptides longer than nine amino acid residues bind to the class I MHC molecule HLA-A2.1 with high affinity and in different conformations. *J Immunol 152*:2874–2881.

Collins EJ, Garboczi DN, Karpusas MN, Wiley DC. 1995. The three-dimensional structure of a class I major histocompatibility complex molecule missing the α_3 domain of the heavy chain. *Proc Natl Acad Sci USA 92*:1218–1221.

Corr M, Boyd LF, Padlan EA, Margulies DH. 1993. H-2D^d exploits a four residue peptide binding motif. *J Exp Med 178*:1877–1892.

Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC. 1998. Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. *Immunity 8*:403–411.

Fremont DH, Matsumura M, Stura EA, Peterson PA, Wilson IA. 1992. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b. *Science 257*:919–927.

Fremont DH, Stura EA, Matsumura M, Peterson PA, Wilson IA. 1995. Crystal structure of an H-2K^b-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. *Proc Natl Acad Sci USA 92*:2479–2483.

Gao GF, Tormo J, Gerth UC, Wyer JR, McMichael AJ, Stuart DI, Bell JI, Jones EY, Jakobsen BK. 1997. Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2. *Nature 387*:630–634.

Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. 1996. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature 384*:134–141.

Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, Teyton L, Wilson IA. 1998. Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science 279*:1166–1172.

Ghendler Y, Teng MK, Liu JH, Witte T, Liu J, Kim KS, Kern P, Chang HC, Wang JH, Reinherz EL. 1998. Differential thymic selection outcomes stimulated by focal structural alteration in peptide/major histocompatibility complex ligands. *Proc Natl Acad Sci USA 95*:10061–10066.

Glithero A, Tormo J, Haurum JS, Arsequell G, Valencia G, Edwards J, Springer S, Townsend A, Pao YL, Wormald M, et al. 1999. Crystal structures of two H-2D^b/glycopeptide complexes suggest a molecular basis for CTL cross-reactivity. *Immunity 10*:63–74.

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.

Huang ES, Subbiah S, Levitt M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J Mol Biol 252*:709–720.

Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr Opin Struct Biol 6*:210–216.

Kern PS, Teng MK, Smolyar A, Liu JH, Liu J, Hussey RE, Spoerl R, Chang HC, Reinherz EL, Wang JH. 1998. Structural basis of CD8 coreceptor function revealed by crystallographic analysis of a murine CD8alpha(alpha) ectodomain fragment in complex with H-2K^b. *Immunity 9*:519–530.

Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. 1998. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci 7*:2578–2586.

Kondo A, Sidney J, Southwood S, del Guercio M-F, Appella E, Sakamoto H, Celis E, Grey HM, Chesnut RW, Kubo RT, Sette A. 1995. Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class~ I molecules. *J Immunol 155*:4307–4312.

Kondo A, Sidney J, Southwood S, del Guercio MF, Appella E, Sakamoto H, Grey HM, Celis E, Chesnut RW, Kubo RT, Sette A. 1997. Two distinct HLA-A*0101-specific submotifs illustrate alternative peptide binding modes. *Immunogenetics 45*:249–258.

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.

Madden DR, Gorga JC, Strominger JL, Wiley DC. 1992. The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell 70*:1035–1048.

Menssen R, Orth P, Ziegler A, Saenger W. 1999. Decamer-like conformation of a nona-peptide bound to HLA-B*3501 due to non-standard positioning of the C terminus. *J Mol Biol 285*:645–653.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol 256*:623–644.

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.

Rammensee HG, Friede T, Stevanoviic S. 1995. MHC ligands and peptide motifs: First listing. *Immunogenetics 41*:178–228.

Reid SW, McAdam S, Smith KJ, Klenerman P, O'Callaghan CA, Harlos K, Jakobsen BK, McMichael AJ, Bell JI, Stuart DI, Jones EY. 1996. Antagonist HIV-1 Gag peptides induce structural changes in HLA B8. *J Exp Med 184*:2279–2286.

Ressing ME, Sette A, Brandt RM, Ruppert J, Wentworth PA, Hartman M, Oseroff C, Grey HM, Melief CJ, Kast WM. 1995. Human CTL epitopes encoded by human papillomavirus type 16 E6 and E7 identified through in vivo and in vitro immunogenicity studies of HLA-A*0201-binding peptides. *J Immunol 154*:5934–5943.

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

Schueler-Furman O, Elber R, Margalit H. 1998. Knowledge-based structure prediction of MHC class I bound peptides: A study of 23 complexes. *Fold Des 3*:549–564.

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.

Sette A, Sidney J, del Guercio MF, Southwood S, Ruppert J, Dahlberg C, Grey HM, Kubo RT. 1994a. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol Immunol 31*:813–822.

Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief CJM, Oseroff C, Yuan L, Ruppert J, et al. 1994b. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol 153*:5586–5592.

Sidney J, Grey HM, Southwood S, Celis E, Wentworth PA, del Guercio M-F, Kubo RT, Chesnut RW, Sette A. 1996a. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Hum Immunol 45*:79–93.

Sidney J, Southwood S, del Guercio M-F, Grey HM, Chesnut RW, Kubo RT, Sette A. 1996b. Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. *J Immunol 157*:3480–3490.

Silver ML, Guo HC, Strominger JL, Wiley DC. 1992. Atomic structure of a human MHC molecule presenting an influenza virus peptide. *Nature 360*:367–369.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci 6*:676–688.

Smith KJ, Reid SW, Harlos K, McMichael AJ, Stuart DI, Bell JI, Jones EY. 1996a. Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity 4*:215–228.

Smith KJ, Reid SW, Stuart DI, McMichael AJ, Jones EY, Bell JI. 1996b. An altered position of the α_2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. *Immunity 4*:203–213.

Southwood S, Sidney J, Kondo A, del Guercio M-F, Appella E, Hoffman S, Kubo RT, Chesnut RW, Grey HM, Sette A. 1998. Several common HLA-DR types share largely overlapping peptide binding repertoires. *J Immunol 160*:3363–3373.

Speir JA, Garcia KC, Brunmark A, Degano M, Peterson PA, Teyton L, Wilson IA. 1998. Structural basis of 2C TCR allorecognition of H-2L^d peptide complexes. *Immunity 8*:553–562.

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol 17*:555–561.

Swets JA. 1988. Measuring the accuracy of diagnostic systems. *Science 240*:1285–1293.

Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol 257*:457–469.

Young AC, Zhang W, Sacchettini JC, Nathenson SG. 1994. The three-dimensional structure of H-2D^b at 2.4 Å resolution: Implications for antigen-determinant selection. *Cell 76*:39–50.

Zhao R, Loftus DJ, Appella E, Collins EJ. 1999. Structural evidence of T cell xeno-reactivity in the absence of molecular mimicry. *J Exp Med 189*:359–370.