BIOINFORMATICS

Vol. 19 no. 17 2003, pages 2263–2270
DOI: 10.1093/bioinformatics/btg312

Towards the in silico *identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction*

I.A. Doytchinova and D.R. Flower*

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

Received on November 22, 2002; revised on February 26, 2003; accepted on May 20, 2003

ABSTRACT

Motivation: The immunogenicity of peptides depends on their ability to bind to MHC molecules. MHC binding affinity prediction methods can save significant amounts of experimental work. The class II MHC binding site is open at both ends, making epitope prediction difficult because of the multiple binding ability of long peptides.

Results: An iterative self-consistent partial least squares (PLS)-based additive method was applied to a set of 66 peptides no longer than 16 amino acids, binding to DRB1*0401. A regression equation containing the quantitative contributions of the amino acids at each of the nine positions was generated. Its predictability was tested using two external test sets which gave $r_{\text{pred}} = 0.593$ and $r_{\text{pred}} = 0.655$, respectively. Furthermore, it was benchmarked using 25 known T-cell epitopes restricted by DRB1*0401 and we compared our results with four other online predictive methods. The additive method showed the best result finding 24 of the 25 T-cell epitopes.

Availability: Peptides used in the study are available from http://www.jenner.ac.uk/JenPep. The PLS method is available commercially in the SYBYL molecular modelling software package. The final model for affinity prediction of peptides binding to DRB1*0401 molecule is available at http://www.jenner.ac.uk/MHCPred. Models developed for DRB1*0101 and DRB1*0701 also are available in MHCPred.

Contact: darren.flower@jenner.ac.uk

INTRODUCTION

The recognition of antigen peptides by T-cell receptors (TCR) is a central event in cellular immunity against pathogens. The immunogenicity of peptides is strongly influenced by their ability to bind to MHC molecules. Because of this, T-cell epitope predictive algorithms are, in practice, based on binding affinity prediction. A broad spectrum of predictive methods is now available (Flower *et al.*, 2002). Beginning with early motif searching (Rammensee *et al.*, 1995; D'Amaro *et al.*, 1995; Meister *et al.*, 1995) and different scoring schemes based on the hypothesis of independent binding of side chains (IBS-hypothesis) (Parker *et al.*, 1994), through artificial neural networks (ANN) (Honeyman *et al.*, 1998; Brusic *et al.*, 1998) to the free energy scoring function FRESNO (Rognan *et al.*, 1999). More recent methods include positional scanning— synthetic combinatorial libraries (PS-SCL) (Udaka *et al.*, 2000) and 3D-QSAR studies (Doytchinova and Flower, 2001, 2002a–c). Although most methods have been developed for MHC class I binding peptides, a set of scoring matrices for class II peptides is also available (Hammer *et al.*, 1994; Marshall *et al.*, 1995; Southwood *et al.*, 1998; Brusic *et al.*, 1998; Borrás-Cuesta *et al.*, 2000; Mallios, 2001). The incorporation of these predictive methods in the initial *in silico* step of epitope identification can save great amounts of subsequent experimental work and is, therefore, increasingly important in the process of T-cell epitope search.

Peptides that bind to MHC class II molecules are usually between 10 and 20 residues long, with sizes between 13 and 16 amino acids being the most frequently observed (Rudensky *et al.*, 1991; Hunt *et al.*, 1992; Chicz *et al.*, 1992, 1993). X-ray data from peptide/MHC class II (Dessen *et al.*, 1997) and TCR/peptide/MHC class II complexes (Hennecke and Wiley, 2002) indicate that nine amino acids are bound in an extended conformation deep in the binding groove of HLA-DR4. A dozen hydrogen bonds between MHC α-helices and peptide main chain carbonyl and amide groups are formed. There is one deep pocket that binds the side chain at peptide position 1 (P1) and there are four shallow pockets that bind side chains at positions P4, P6, P7 and P9. Side chains at positions P2, P3, P5 and P8 project prominently toward the T-cell. The peptide binding groove of class II molecules is open at both ends and this allows a given peptide to bind in many different ways. This multiple binding ability of peptides results in a lower accuracy for prediction methods compared with those for class I peptides (Brusic *et al.*, 1998).

Recently, an additive method for binding affinity prediction was developed (Doytchinova *et al.*, 2002). The method is based on the assumption that the binding affinity of a peptide

*To whom correspondence should be addressed.

could be presented as a sum of the contributions of the amino acids at each position and certain interactions between them. The method is universal and can be applied to any peptide–protein interaction. It has been applied to 12 different MHC class I molecules (Doytchinova *et al.*, 2002; Guan *et al.*, 2003; Doytchinova and Flower, 2003) and these models are included in a web site called MHCPred, which is accessible via the internet: http://www.jenner.ac.uk/MHCpred (Guan *et al.*, 2003). In the present study we have applied the method to a set of 82 peptides of 16 amino acids or less, which bind to the HLA-DRB1*0401 molecule. In order to solve the problem of multiple binding an iterative self-consistent (ISC) PLS-based algorithm was used to select the binding set. Eighty percent of the peptides formed the training set and 20% a test set. Another set of peptides, all longer than 16 amino acids, was used as a second test set. The scoring model has been included in the MHCPred web site.

SYSTEMS AND METHODS

Peptide database

The JenPep database (Blythe *et al.*, 2002), URL: http://www.jenner.ac.uk/JenPep, was used as a source for peptide sequences and their binding affinities to the MHC class II molecule HLA-DRB1*0401. The binding affinities (IC_{50}) were originally assessed by a quantitative assay based on the inhibition of binding of a radiollabelled standard peptide to detergent-solubilized MHC molecules (Ruppert *et al.*, 1993; Sette *et al.*, 1994). A set of 96 peptides was obtained. In order to make tractable the calculation of multiple subsequence binding, only peptides with 16 or less amino acids were chosen. They were 82 such peptides and these were divided into training and test sets. Sixteen peptides (20%) were randomly selected to cover the total affinity range and used as a test set for external validation. The other 66 peptides (80%) were used as a training set. The remaining 14 peptides longer than 16 amino acids were used as an additional test set.

Additive method, PLS method, 'leave-one-out' cross-validation

Each nonamer was transformed into a binary bit string of 180 bins (9 positions × 20 amino acids). A term is equal to 1 when a certain amino acid exists at a certain position, and 0 when it is absent. To simplify the matrix, only amino acid contributions were taken into account. 1–2 and 1–3 interactions were neglected. To reduce the multiple binding options only subsequences bearing anchor amino acids (Y, F, W, L, I, M and V) at position 1 were selected. The initial matrix consists of 185 rows and 181 columns (180 x variables $+ 1y$ variable). The matrix was solved by the partial least squares (PLS) method.

As a projection method PLS handles data matrices with more variables than observations very well, and the data can be both noisy and highly collinear. In this situation, conventional statistical methods like multiple regression produce a formula that fits the training data but is unreliable for prediction. PLS forms new x variables, named *principal components*, as linear combinations of the old ones, and then uses them as predictors of the biological activity (Wold, 1995). We used the PLS method as implemented in the QSAR module of SYBYL6.7 (Tripos Inc.). The IC_{50} values were presented as negative logarithms and were used as the dependent variable y. The scaling method was set to 'none'. The column filtering was switched off. The predictive ability of the models was assessed by 'leave-one-out' cross-validation and by external validation using a test set.

Cross-validation (CV) is a practical and reliable method for testing the predictive power of the models. It has become a standard in PLS analysis and is incorporated in all available PLS software (Wold, 1995). In principle, CV is performed by dividing the data into a number of groups, developing a number of parallel models from the reduced data with one of the groups omitted, and then predicting the biological activities of the excluded compounds. When the number of the groups omitted is equal to the number of the compounds in the set, the procedure is named 'leave-one-out' (LOO). The predictive power of the models was assessed by the cross-validated coefficient q^2 and the standard error of prediction (SEP):

$$q^2 = 1 - \frac{\text{PRESS}}{\text{SSQ}}$$

$$\text{SEP} = \sqrt{\frac{\text{PRESS}}{p-1}}$$

where PRESS is the predictive sum of squares $\left[\sum_{i=1}^{n}(\text{pIC}_{50} \times \text{exp} - \text{pIC}_{50}\text{pred})^2\right]$, SSQ—the sum of squares of $\text{pIC}_{50}\text{exp}$ corrected for the mean $\left[\sum_{i=1}^{n}(\text{pIC}_{50}\text{exp} - \text{pIC}_{50}\text{mean})^2\right]$, p is the number of the peptides omitted, $\text{pIC}_{50}\text{pred}$ is that predicted by the LOO-CV value. The optimal number of components (NC) found by LOO-CV was used in the non-cross-validated models, which were assessed by the explained variance r^2. The experimental versus predicted binding affinities of the test peptides were fitted by linear regression and a r_{pred} was determined.

ALGORITHM

Data flow in the iterative self-consistent (ISC) PLS-based additive method is shown in Figure 1. The training set of 66 long peptides is presented as a set of nonamers accompanied by the pIC_{50} values of the parent peptide. Only nonamers bearing anchor amino acids (Y, F, W, L, I, M, V) at position 1 were selected. The matrix is solved by PLS. LOO-CV is applied to extract the optimum number of components subsequently used to generate the non-cross-validated model. The last model is used to predict pIC_{50} values and a new set is extracted. The best predicted nonamers were selected for each

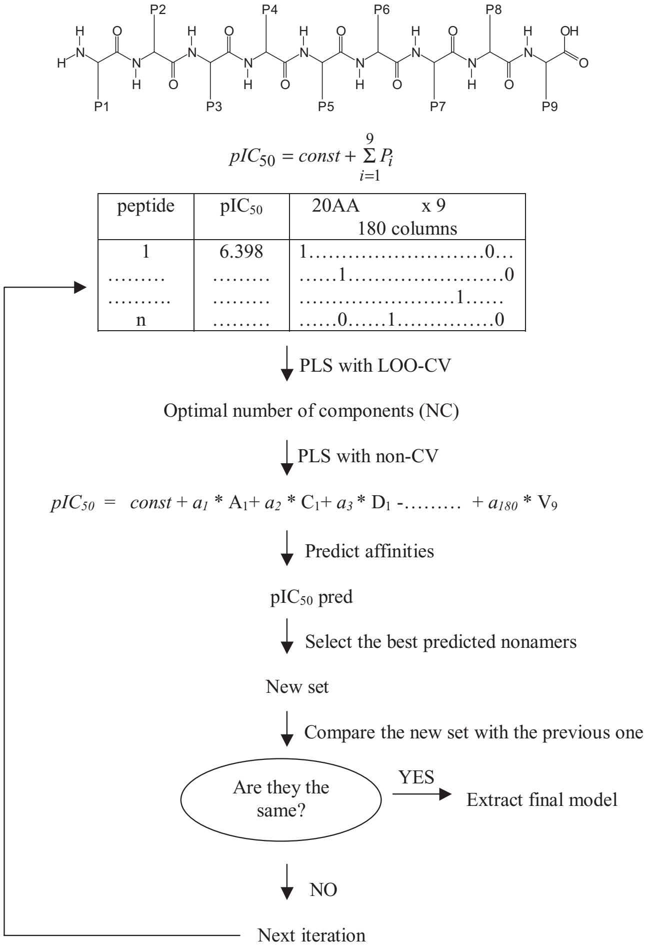

P2
P4
P6
P8
P1
P3
P5
P7
P9
$pIC_{50} = const + \sum_{i=1}^{9} P_i$
peptide
pIC₅₀
20AA x 9
180 columns
1
6.398
1………………………0…
……1……………………0
……………………1……
n
……0……1……………0
PLS with LOO-CV
Optimal number of components (NC)
PLS with non-CV
$pIC_{50} = const + a_1 * A_1 + a_2 * C_1 + a_3 * D_1 - \dots + a_{180} * V_9$
Predict affinities
pIC₅₀ pred
Select the best predicted nonamers
New set
Compare the new set with the previous one
Are they the same?
YES
Extract final model
NO
Next iteration

Fig. 1. Data flow in the ISC PLS-based additive method.

peptide, i.e. those with the lowest residual between the experimental and predicted pIC$_{50}$. The new set is compared with the previous one and if they are the same the final model is obtained. Otherwise, the selection procedure is repeated. The coefficients in the final non-cross-validated model represent the quantitative contributions of each amino acid at each position.

IMPLEMENTATION

The first model had poor predictivity: $q^2 = 0.152, \mathrm{NC} = 1, r^2 = 0.396, n = 185$. Self-consistency was achieved on the seventh iteration. The final model had excellent predictivity with $q^2 = 0.716$, NC $= 4$, $r^2 = 0.967$. The coefficients of the final model are shown in Table 1.

All class II prediction methods must overcome the problem of the multiple binding ability of the peptides. This arises both from the indeterminacy of the problem—we do not know a priori which subsequence is the dominant binder—and from the possible degeneracy of the binding process itself. Where a single dominant binding sequence is absent, the measured affinity is a canonical average of the binding of several subsequences. These phenomena arise from the binding groove of class II molecules being open at both ends. We may posit that, from a thermodynamic viewpoint, the actual nonameric binding subsequence should have the highest pIC$_{50}$ or lowest binding energy, among all the nonamers originating from the same long parent peptide. However, our analysis of the training set indicates that the predicted value closest to the experimental pIC$_{50}$, is seldom the highest predicted value. We tried three different selection rules to deal with this problem when applied to the test sets: mean, highest value (max) and a combination of both (combi). The last rule selects the mean pIC$_{50}$ when the difference between the highest and lowest predicted pIC$_{50}$ is less than one log unit. Otherwise, it selects the highest predicted value. The statistics are shown in Table 2. For both test sets the highest predictivity is given by the combination rule with $r_{\mathrm{pred}} = 0.593$ (test set I) and $r_{\mathrm{pred}} = 0.655$ (test set II). The graphs of best models for the test sets are shown in Figure 2. The performance of the combination rule is not surprising, because when an easily distinguished good binder is not available in the peptide sequence, the binding affinity is a degenerate average of affinities from several binding subsequences.

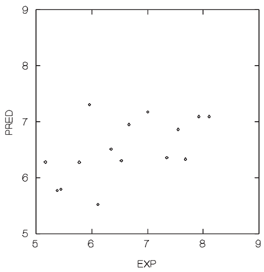
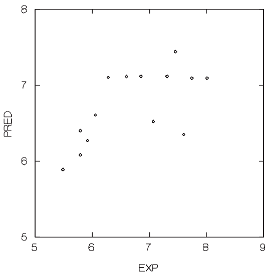
DISCUSSION

Using single amino acid substituted analogs of the HA 307–319 peptide, Sette *et al.* (1993) defined an HLA-DRB1*0401-specific motif. This motif requires an aromatic or aliphatic anchor residue in position 1 (Y, W, F, L, I, V, M), and another anchor residue in position 6, defined as either a hydroxyl (S or T) or hydrophobic (L, V, I, or M) residue. In addition, no positive charges (K or R) are allowed in position 4 or 7 and no charges, either positive or negative (K, R, D or E) are allowed in position 9. Using an Y1-anchored peptide library Hammer *et al.* (1994) developed the first scoring scheme for prediction of affinity to HLA DRB1*0401. Marshall *et al.* (1995) developed a method based on the relative contributions of the 20 naturally occurring amino acids at the central 11 positions of a 13-residue monosubstituted polyalanine peptide. Southwood *et al.* (1998) applied the polynominal method (Gulukota *et al.*, 1997) to derive a scoring matrix. A genetic algorithm (GA) was successfully applied by Brusic *et al.* (1998) to discriminate between binders and non-binders. Comparing different algorithms Borrás-Cuesta *et al.* (2000) deduced a general motif for the prediction of binding to HLA-DR molecules. Mallios (2001) introduced an iterative stepwise discriminant analysis (SDA) meta-algorithm to classify peptides initially into binders and non-binders and later into non-binders, intermediate and high binders (Mallios, 2001). The ISC algorithm proposed here uses a related iterative procedure to select the best predicted binders, but it incorporates the PLS method, a robust multivariate statistical technique, for model generation.

Table 1. Additive model for binding affinity prediction to DRB1*0401 (the constant equals to 6.169)

	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6	Position 7	Position 8	Position 9
A	0	−0.130	−0.013	0.253	0.296	−0.223	−0.247	0.120	0.051
C	0	0	0	0	0	0	0	0	−0.128
D	0	−0.137	−0.246	−0.127	−0.093	−0.125	0.080	0.001	−0.013
E	0	−0.032	−0.008	−0.140	−0.271	−0.133	−0.032	0	0
F	0.136	−0.165	0.006	0.113	0	0	−0.227	0.103	−0.084
G	0	0.045	−0.063	0.013	0.038	−0.081	0.068	−0.025	−0.071
H	0	0.028	−0.128	0.039	0	−0.005	0	−0.093	0.114
I	−0.361	−0.051	−0.055	0.046	−0.029	−0.013	0	−0.173	−0.018
K	0	0.066	0.125	0	−0.056	−0.142	0.151	0.036	−0.005
L	−0.482	−0.068	0.330	−0.359	−0.206	0.194	0.178	0.303	−0.223
M	−0.005	0	0	0.193	0	−0.012	0	0.047	0
N	0	0.055	0.081	−0.085	0.284	0	−0.016	0	0.247
P	0	−0.072	0.050	0	0.213	−0.103	0.229	0.061	0.125
Q	0	−0.104	0.293	0.157	−0.014	0.035	−0.100	−0.032	0.061
R	0	0.193	−0.138	0	−0.048	−0.189	−0.174	−0.214	−0.064
S	0	0.071	−0.140	0.001	−0.284	0.313	0.042	−0.050	0.020
T	0	0	0	−0.072	0.127	0.207	−0.096	0	−0.016
V	0.297	0.238	−0.044	−0.031	−0.070	0.277	0.225	−0.154	0.004
W	0.272	−0.159	−0.012	0	0.114	0	0	0	0
Y	0.145	0.221	−0.037	0	0	0	−0.082	0.069	0

Table 2. Statistics of the test sets

Parameter	Test set I	Test set II
n	16	14
r_{pred} (mean)	0.547	0.480
r_{pred} (max)	0.459	0.596
r_{pred} (combi)	0.593	0.655
		

We compared the scores for the amino acids at each position for each of the five scoring matrices: Hammer's (Hammer *et al*., 1994) (code H1994), Marshall's (Marshall *et al*., 1995) (code M1995), Southwood's (Southwood *et al*., 1998) (code S1998), Brusic's (Brusic *et al*., 1998) (code B1998), Borrás-Cuesta's (Borrás-Cuesta *et al*., 2000) (code BC2000) and ours (D2003). No overall correlation was found between them (Table 3). Only the BC2000 matrix shows some correlation with H1994 and B1998 scoring functions ($r > 0.5$). The coefficients derived by the additive method (D2003) do not correlate well with the other scores except for BC2000. The correlation analysis for each position (data not shown) indicated that only for position 1 there is a correlation ($R_{\text{mean}} = 0.70$). The correlation coefficients for the remaining positions ranged from 0.49 for position 6 to 0.23 for position 5. Nevertheless, a consensus exists regarding the amino acid preferences at the anchor positions 1 and 6. Although Leu and Ile are considered as anchors at position 1 most of the scoring matrices (H1994, M1995, B1998, S1998) indicate that they make a negative contribution to the affinity. In the present study, Phe, Tyr, Trp and Val were found to contribute significantly to the affinity and Leu and Ile to contribute deleteriously. Met was found to make no significant contribution. Ser, Thr and Val are favoured amino acids at position 6 according to our model (D2003) and others (H1994, M1995, S1998). For position 4, Met is a preferred amino acid (H1994, B1998, S1998, BC2000), but our study shows that Ala and Gln are also well accepted here. No agreement exists for the favoured amino acid at position 7. Preferences are given to Met (H1994, S1998, B1998, BC2000), Asn (B1998), Cys (B1998), His (S1998), Val (H1994) and Tyr (M1995). According to our study Pro and Val were found to be favoured at this position. Asp and Glu at position 9 are deleterious for binding (H1994, M1995, S1998, B1998). In our training set, selected by the ISC algorithm, there was only one peptide bearing Asp at position 9 and no peptide with Glu at this position. A great variety of preferred amino acids was identified at the

Table 3. Correlation coefficients (R) between different predictive methods (n is the number of the common amino acids)

	H1994	M1995	S1998	B1998	BC2000	D2003
	(Hammer	(Marshall	(Southwood	(Brusic *et al.*,	(Borrás-Cuesta	This study
	et al., 1994)	*et al.*, 1995)	*et al.*, 1998)	1998)	*et al.*, 2000)	
H1994	$n = 159$					
(Hammer *et al.*, 1994)	$R = 1.000$					
M1995	$n = 159$	$n = 167$				
(Marshall *et al.*, 1995)	$R = 0.138$	$R = 1.000$				
S1998	$n = 148$	$n = 156$	$n = 156$			
(Southwood *et al.*, 1998)	$R = 0.154$	$R = 0.057$	$R = 1.000$			
B1998	$n = 159$	$n = 167$	$n = 156$	$n = 56$		
(Brusic *et al.*, 1998)	$R = 0.274$	$R = 0.072$	$R = 0.170$	$R = 1.000$		
BC2000	$n = 36$	$n = 38$	$n = 35$	$n = 38$	$n = 38$	
(Borrás-Cuesta *et al.*, 2000)	$R = 0.590$	$R = 0.454$	$R = 0.282$	$R = 0.546$	$R = 1.000$	
D2003	$n = 130$	$n = 131$	$n = 124$	$n = 131$	$n = 28$	$n = 131$
This study	$R = 0.258$	$R = 0.056$	$R = 0.200$	$R = 0.098$	$R = 0.450$	$R = 1.000$

remaining positions (2, 3, 5 and 8). This is not surprising as the side chains of the amino acids at these positions are oriented toward the T-cell and have less influence on binding to the MHC molecule.

To benchmark our method, 25 known T-cell epitopes binding to DRB1*0401 were collected from JenPep and evaluated by different predictive methods available online: SYFPEITHI (Rammensee *et al.*, 1999), MHC-Thread (Brooks, 1999, http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/), RANKPEP (Reche *et al.*, 2002), ProPred (Singh and Raghava, 2001). The additive method, as implemented in MHCPred (Guan *et al.*, 2003), was used to make predictions. The scores are shown in Table 4. Comparison of these methods proved problematic because of the different scoring functions. Ideally, we would wish to compare the enrichment in predicted binders versus a random selection, where whole proteins had been analysed for T-cell epiotopes using overlapping peptides. Unfortunately, fully controlled experiments such as this are costly and are seldom performed. Instead, we focussed on prediction of known class II restricted epitopes. As SYFPEITHI and MHC-Thread deal with peptides longer than 15 or 13 amino acids, respectively, scores for some epitopes could not be calculated. For these two methods the whole proteins were evaluated and the top 50% binders were considered as epitopes. SYFPEITHI found 10 from 13 epitopes and MHC-Tread identified 10 out of 19. RANKPEP detects 20 of the 25 epitopes at the default binding threshold of 4.85. Using a score above 0 as a *de facto* threshold, 19 of the evaluated by ProPred peptides had positive scores and the remaining were negative. The predictions made by the additive method showed that 24 of the 25 epitopes have IC$_{50}$ values below 500 nM and only one peptide has IC$_{50}$ value slightly higher than 500 nM (577 nM). Affinity below 500 nM is widely accepted as a threshold for potential T-cell epitopes.

In conclusion, it was shown that the additive method, as modified in this paper, is a reliable quantitative method for binding affinity prediction for peptides binding to the MHC class II molecule DRB1*0401. As this method is universal, it could be applied to any peptide–protein interaction where the overall sequence length is unrestricted but binding is localized to a fixed, but unknown part, of the peptide. Rather than simply ranking or qualitatively scoring peptides, the ISC-PLS additive method produces a quantitative prediction of a real measured affinity. It is easy and fast to use and interpretation is facile. The model derived in this paper is implemented in an updated version of MHCPred.

Models for MHC class II molecules DRB1*0101 and DRB1*0701 were also developed. The DRB1*0101 model achieved self-consistency at the 13th iteration and had the following statistics: $n = 90$, $q^2 = 0.808$, $\mathrm{NC} = 8$, $r^2 = 0.994$. For the DRB1*0701 model the self-consistency was achieved on the 11th iteration and its statistics were $n = 84$, $q^2 = 0.649$, $\mathrm{NC} = 7$, $r^2 = 0.999$. Both models are included in MHCPred. As data becomes available, other class II models will be forthcoming.

In the general case, effective and protective vaccines may be required to act through stimulation of both the humoral and cellular immune systems. Likewise, T-cell mediated immunity may function through either or both class I or class II mediated mechanisms. In order to make computational vaccinology a pragmatic and useable reality, we must be able to predict all aspects of the immune response. Our extension of the additive method to deal with class II restricted MHC presentation is a pivotal step toward that goal.

Table 4. Comparison of MHC class II predictions

T-cell epitope	Source	Reference	SYFPEITHI[a]	MHC-Tread[b]	RANKPEP[c]	ProPred[d]	MHCPred[e]
QNLLKAEKGNKAAAQR	Histone H1-like protein HC1	Gaston *et al*. (1996)	20[f]/26[g]	764[f]/2366[g]	4.930[h]	0.7[i]	108[j]
LLESIQQNLLKAEKGN	Histone H1-like protein HC1	Gaston *et al*. (1996)	8/26	1820/2366	9.263	−1.8	48
EYLNKIQNSLSTEWSPCSVT	Circumsporozoite protein	Calvo-Calle *et al*. (1997)	18/26	2524/4347	2.828	2.4	90
AGFKGEQGPKGEP	Collagen alpha 1(II) chain	Fugger *et al*. (1996)	—	367/3435	10.854	−0.4	196
FFRMVISNPAATHQDIDFLI	Glutamate decarboxylase, 65 kDa isoform	Endl *et al*. (1997)	18/28	2177/4066	8.067	4.3	104
LPRLIAFTSEHSHF	Glutamate decarboxylase, 65 kDa isoform	Endl *et al*. (1997)	—	2277/4066	0.27	1.1	129
MNILLQYVVKSFD	Glutamate decarboxylase, 65 kDa isoform	Wicker *et al*. (1996)	—	2062/4066	4.996	3.48	164
IAFTSEHSHFSLK	Glutamate decarboxylase, 65 kDa isoform	Wicker *et al*. (1996)	—	1626/4066	5.660	3.4	346
PKYVKQNTLKLATGMRNVP	Hemagglutinin [Fragment]	Carmichael *et al*. (1996)	14/28	2478/4922	34.032	4.5	27
GYKVLVLNPSVAAT	Genome polyprotein	Diepolder *et al*. (1997)	—	1284/—	5.518	4.08	81
KHKVYACEVTHQGLSS	Ig kappa chain C region	Kovats *et al*. (1997)	22/26	2253/3871	21.661	2.4	148
KVQWKVDNALQSGNS	Ig kappa chain C region	Kovats *et al*. (1997)	22/26	1594/3871	12.485	4.4	89
KVDNALQSGNS	Ig kappa chain C region	Dong *et al*. (2000)	—	—	−3.493	−4.9	175
QPLALEGSLQK	Insulin	Congia *et al*. (1998)	—	—	6.683	2.9	577
YVIEGTSKQ	Integrin alpha-L	Gross *et al*. (1998)	—	—	20.971	7.3	182
EFVVEFDLPGIKA	18 kDa antigen	McNicholl *et al*. (1995)	—	1582/3066	20.469	2.7	80
LSRFSWGAEGQRPGFGYGG	Myelin basic protein	Muraro *et al*. (1997)	22/28	2979/3214	17.707	−1.7	308
WNRQLYPEWTEAQRLD	Melanocyte protein Pmel 17	Li *et al*. (1998)	26/28	1851/3581	9.752	4.3	290
AKYDAFVTALTE	Major pollen allergen Pha a 5.3	de Lalla *et al*. (1999)	—	—	13.606	−1.5	228
AFNDEIKASTGG	Pollen allergen Phl p 5a	de Lalla *et al*. (1999)	—	—	−3.066	−2.6	317
VIVMLTPLVEDGVKQC	Protein-tyrosine phosphatase-like N	Honeyman *et al*. (1998)	20/28	1004/4643	2.982	2.3	45
AKFYRDPTAFGSG	Proteoglycan link protein	Hammer *et al*. (1995)	—	1438/4762	16.850	3.9	342
QYIKANSKFIGITEL	Tetanus toxin	Reece *et al*. (1993)	6/28	1011/4513	9.334	1.5	34
QNILLSNAPLGPQFP	Tyrosinase	Topalian *et al*. (1996)	8/28	1532/4066	12.504	0.5	60
DYSYLQDSDPDSFQD	Tyrosinase	Topalian *et al*. (1996)	22/28	1395/4066	10.128	1.3	205

Known T-cell epitopes with affinity to HLA-DRB1*0401 are evaluated using different epitope prediction programs available free online.

[a] http://syfpeithi.bmi-heidelberg.com/

[b] http://www.csd.abdn.ac.uk/∼gjlk/MHC-Thread/

[c] http://www.mifoundation.org/Tools/rankpep.html

[d] http://www.imtech.res.in/ raghava/propred/

[e] http://www.jenner.ac.uk/MHCPred

[f] The highest score of a 9mer included in the T cell epitope.

[g] The highest score of a 9mer included in the whole protein.

[h] Binding threshold:4.85.

[i] The highest score achievable by any peptide is 8.6.

[j] IC_{50} value in nM.

REFERENCES

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

Borrás-Cuesta,F., Golvano,J.-J., García-Granero,M., Sarobe,P., Riezu-Boj,J.-I., Huarte,E. and Lasarte,J.-J. (2000) Specific and general HLA-DR binding motifs: comparison of algorithms. *Hum. Immunol.*, **61**, 266–278.

Brooks,T. (1999) *MHC–Thread*. University of Aberdeen.

Brusic,V., Rudy,G., Honeyman,M., Hammer,J. and Harrison,L. (1998) Prediction of MHC class II-binding peptides and artificial neural network. *Bioinformatics*, **14**, 121–130.

Calvo-Calle,J.M., Hammer, J., Sinigaglia,F., Clavijo,P., Moya-Castro,Z.R. and Nardin,H. (1997) Binding of malaria T cell epitopes to DR and DQ molecules *in vitro* correlates with immunogenicity *in vivo*. *J. Immunol.*, **159**, 1362–1373.

Carmichael,A., Jin,X. and Sissons,P. (1996) Analysis of the human env-specific cytotoxic T-lymphocyte (CTL) response in natural human immunodeficiency virus type 1 infection: low prevalence of broadly cross-reactive env-specific CTL. *J. Virol.*, **70**, 8468–8476.

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

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

Congia,M., Patel,S., Cope,A.P., De Virgiliis,S. and Sonderstrup,G. (1998) T cell epitopes of insulin defined in HLA-DR4 transgenic mice are derived from preproinsulin and proinsulin. *Proc. Natl Acad. Sci. USA*, **95**, 3833–3838.

D'Amaro,J., Houbiers,J.G.A., Drijfhout,J.W., Brandt,R.M.P., Schipper,R., Bavinck,J.N.B., Melief,C.J.M. and Kast,W.M. (1995) A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum. Immunol.*, **43**, 13–18.

Dessen,A., Lawrence,C.M., Cupo,S., Zaller,D.M. and Wiley,D.C. (1997) X-ray crystal structure of HLA-DR4 (DRA*0101, DRB1*0401) complexed with a peptide from human collagen II. *Immunity*, **7**, 473–481.

Diepolder,H.M., Gerlach,J.-T., Zachoval,R., Hoffmann,R.M., Jung,M.-C., Wierenga,E.A., Scholz,S., Santantonio,T., Houghton,M., Southwood,S. *et al*. (1997) Immunodominant CD4+ T-cell epitope within nonstructural protein 3 in acute hepatitis C virus infection. *J. Virol.*, **71**, 6011–6019.

Dong,X., An,B., Kierstead,L.S., Storkus,W.J., Amoscato,A.A. and Salter,D. (2000) Modification of the amino terminus of a class II epitope confers resistance to degradation by CD13 on dendritic cells and enhances presentation to T cells. *J. Immunol.*, **164**, 129–135.

Doytchinova,I. and Flower,D.R. (2001) Towards the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity to class I MHC molecule HLA-A*0201. *J. Med. Chem.*, **44**, 3572–3581.

Doytchinova,I.A. and Flower,D.R. (2002a) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure–activity relationship study. *PROTEINS*, **48**, 505–518.

Doytchinova,I.A. and Flower,D.R. (2002b) Quantitative approaches to computational vaccinology. *Immunol. Cell Biol.*, **80**, 270–279.

Doytchinova,I.A. and Flower,D.R. (2002c) A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput.-Aid. Mol. Des.*, **16**, 535–544.

Doytchinova,I.A., Blythe,M.J. and Flower,D.R. (2002) Additive method for the prediction of protein–peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.*, **1**, 263–272.

Doytchinova,I.A. and Flower,D.R. (2003) The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem.*, **1**, in press.

Endl,J., Otto,H., Jung,G., Dreisbusch,B., Donie,F., Stahl,P., Elbracht,R., Schmitz,G., Meinl,E., Hummel,M. *et al*. (1997) Identification of naturally processed T cell epitopes from glutamic acid decarboxylase presented in the contex of HLA-DR alleles by T lymphocytes of recent onset IDDM patients. *J. Clin. Invest.*, **99**, 2405–2415.

Fugger,L., Rothbard,J.B. and Sonderstrup-McDevitt,G. (1996) Specificity of an HLA-DRB1*0401-restricted T cell response to type II collagen. *Eur. J. Immunol.*, **26**, 928–933.

Flower,D.R, Doytchinova,I.A., Paine,K., Taylor,P., Blythe,M.J., Lamponi,D., Zygouri,C., Guan,P., McSparron,H. and Kirkbride,H. (2002) Computational vaccine design. In Flower,D.R. (ed.) *Drug Design: Cutting Edge Approaches*. Royal Society of Chemistry, Cambridge, pp. 136–180.

Gaston,J.S., Deane,K.H., Jecock,R.M. and Pearce,J.H. (1996) Identification of 2 Chlamydia trachomatis antigens recognized by synovial fluid T cells from patients with Chlamydia induced reactive arthritis. *J. Rheumatol.*, **23**, 130–136.

Gross,D.M., Forsthuber,T., Tary-Lehmann,M., Etling,C. Ito,K., Nagy,Z.A., Field,J.A., Steere,A.C. and Huber,B.T. (1998) Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. *Science*, **281**, 703–706.

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

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

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

Hammer,J., Gallazzi,F., Bono,E., Karr,R.W., Guenot,J., Valsasnini,P., Nagy,Z.A. and Sinigaglia,F. (1995) Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association. *J. Exp. Med.*, **181**, 1847–1855.

Hennecke,J. and Wiley,D.C. (2002) Structure of a complex of the human α/β T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J. Exp. Med.*, **195**, 571–581.

Honeyman,M.C., Brusic,V., Stone,N.L. and Harrison,L.C. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotech.*, **16**, 966–969.

Honeyman,M.C., Stone,N.L. and Harrison,L.C. (1998) T-cell epitopes in type 1 diabetes autoantigen tyrosine phosphatase IA-2: potential for mimicry with rotavirus and other environmental agents. *Mol. Med.*, **4**, 231–239.

Hunt,D.F., Michel,H., Dickinson,T.A., Shabanowitz,J., Cox,A.L., Sakaguchi,K. and Appella,E. (1992) Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science*, **256**, 1817–1820.

Kovats,S., Whiteley,P.E., Concannon,P., Rudensky,A.Y. and Blum,J.S. (1997) Presentation of abundant endogenous class II DR-restricted antigens by DM-negative B cell lines. *Eur. J. Immunol.*, **27**, 1014–1021.

Mallios,R.R. (2001) Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics*, **17**, 942–948.

Marshall,K.W., Wilson,K.J., Liang,J., Woods,A., Zaller,D. and Rothbard,J.B. (1995) Prediction of peptide affinity to HLA DRB1*0401. *J. Immunol.*, **154**, 5927–5933.

McNicholl,J.M., Whitworth,W.C., Oftung,F., Fu,X., Shinnick,T., Jensen,P.E., Simon,M., Wohlhueter,R.M. and Karr,R.W. (1995) Structural requirements of peptide and MHC for DR(alpha, beta 1*0401)-restricted T cell antigen recognition. *J. Immunol.*, **155**, 1951–1963.

Meister,G.E., Roberts,C.G.P., Berzofsky,J.A. and De Groot,A.S. (1995) Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine*, **13**, 581–591.

de Lalla,C., Sturniolo,T., Abbruzzese,T., Abbruzzese,L., Hammer,J., Sidoli,A., Sinigaglia,F. and Panina-Bordignon,P. (1999) Identification of novel T cell epitopes in *Lol p5a* by computational prediction. *J. Immunol.*, **163**, 1725–1729.

Li,K., Adibzadeh,M., Halder,T., Kalbacher,T., Heinzel,S., Muller,C., Zeuthen,J. and Pawelec,G. (1998) Tumour-specific MHC-class-II-restricted responses after *in vitro* sensitization to synthetic peptides corresponding to gp100 and Annexin II eluted from melanoma cells. *Cancer Immunol. Immunother.*, **47**, 32–38.

Muraro,P.A., Vergelli,M., Kalbus,M., Banks,D.E., Nagle,J.W., Tranquill,L.R., Nepom,G.T., Biddison,W.E., McFarland,H.F. and Martin,R. (1997) Immunodominance of low-affinity major histocompatibility complex-binding myelin basic protein epitope (residues 111–129) in HLA-DR4 (B1*0401) subjects is associated with a restricted T cell receptor repertoire. *J. Clin. Invest.*, **100**, 339–349.

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

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

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

Reche,P., Glutting,J. and Reinherz,E. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.*, **63**, 701–709.

Reece,J.C., Geysen,H.M. and Rodda,S.J. (1993) Mapping the major human T helper epitopes of tetanus toxin. *J. Immunol.*, **151**, 6175–6184.

Rognan,D., Lauemøller,S.L., Holm,A., Buus,S. and Tschinke,V. (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.*, **42**, 4650–4658.

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

Ruppert,J., Sidney,J., Celis,E., Kubo,R.T., Grey,H.M. and Sette,A. (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell*, **74**, 929–937.

Sette,A., Sidney,J., Oseroff,C., del Guercio,M.-F., Southwood,S., Arrhenius,T., Powell,M.F., Colón,S.M., Gaeta,F.C.A. and Grey,H.M. (1993) HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. *J. Immunol.*, **151**, 3163–3170.

Sette,A., Sidney,J., del Guercio,M.-F., Southwood,S., Ruppert,J., Dalberg,C., Grey,H.M. and Kubo,R.T. (1994) Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.*, **31**, 813–822.

Singh,H. and Raghava,G.P. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, **17**, 1236–1237.

Southwood,S., Sidney,J., Kondo,A., del Guercio,M.-F., Appella,E., Hoffman,S., Kubo,R.T., Chesnut,R.W., Grey,H.M. and Sette,A. (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol.*, **160**, 3363–3373.

SYBYL 6.7. Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.

Topalian,S.L., Gonzales,M.I., Parkhurst,M., Li,Y.F., Southwood,S., Sette,A., Rosenberg,S.A. and Robbins,P.F. (1996) Melanoma-specific CD4+ T cells recognize nonmutated HLA-DR-restricted tyrosinase epitopes. *J. Exp. Med.*, **183**, 1965–1971.

Udaka,K., Wiesmüller,K.-H., Kienle,S., Jung,G., Tamamura,H., Yamagishi,H., Okumura,K., Walden.P., Suto,T. and Kawasaki,T. (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*, **51**, 816–828.

Wicker,L.S., Chen,S.-L., Nepom,G.T., Elliott,J.F., Freed,D.C., Bansal,A., Zheng,S., Herman,A., Lernmark,A., Zaller,D.M. *et al.* (1996) Naturally processed T cell epitopes from human glutamic acid decarboxylase identified using mice transgenic for the type 1 diabetes-associated human MHC class II allele, DRB1*0401. *J.Clin. Invest.*, **98**, 2597–2603.

Wold,S. (1995) PLS for multivariate linear modeling. In van de Waterbeemd, H. (ed.), *Chemometric Methods in Molecular Design*. VCH, Weinheim, pp. 195–218.