

Analysis of Peptide–Protein Binding Using Amino Acid Descriptors: Prediction and Experimental Verification for Human Histocompatibility Complex HLA-A*0201

Pingping Guan,* Irini A. Doytchinova, Valerie A. Walshe, Persephone Borrow, and Darren R. Flower

Edward Jenner Institute for Vaccine Research, Compton, Berkshire RG20 7NN, U.K.

Received June 6, 2005

Amino acid descriptors are often used in quantitative structure–activity relationship (QSAR) analysis of proteins and peptides. In the present study, descriptors were used to characterize peptides binding to the human MHC allele HLA-A*0201. Two sets of amino acid descriptors were chosen: 93 descriptors taken from the amino acid descriptor database AAindex and the z descriptors defined by Wold and Sandberg. Variable selection techniques (SIMCA, genetic algorithm, and GOLPE) were applied to remove redundant descriptors. Our results indicate that QSAR models generated using five z descriptors had the highest predictivity and explained variance (q^2 between 0.6 and 0.7 and r^2 between 0.6 and 0.9). Further to the QSAR analysis, 15 peptides were synthesized and tested using a T2 stabilization assay. All peptides bound to HLA-A*0201 well, and four peptides were identified as high-affinity binders.

Introduction

The first study of amino acid descriptors was undertaken by Sneath,¹ who used physicochemical semiquantitative data to derive descriptors for the 20 naturally occurring amino acids. Since then, a number of studies have generated many other descriptors. Kidera et al. analyzed 188 amino acid indices and divided them into groups according to the properties they represent.² On the basis of Kidera's data and some new additions, Nakai et al. carried out another cluster analysis of 222 amino acid indices, dividing them into four major groups: α and β turn propensities, β propensity, hydrophobicity, and physicochemical properties.³ In the same year, Fauchere et al. chose a group of 15 physicochemical descriptors and applied them to 20 natural and 26 synthetic amino acids.⁴ To facilitate public access to these descriptors, Kawashima collected most published descriptors and established a database named AAindex.⁵

Quantitative structure–activity relationships (QSARs) relate the biological activity of a molecule to its structure. QSAR techniques can be 2D or 3D. The former uses physicochemical descriptors, while the latter also takes the spatial features of the molecules into account.^{4,6,7} Statistical methods, such as partial least-squares (PLS) analysis, are used in QSAR to produce models that relate changes in activity to molecular properties.^{8,9}

Amino acid descriptors are often used in peptide QSAR studies.^{10–13} They describe the physicochemical properties of the peptides quantitatively. Many of these properties are measured experimentally by methods such as TLC, HPLC, and spectroscopy.^{14–16} Other properties cannot be measured but can be calculated, such as molecular surface areas and atomic charges. The quality of the amino acid descriptors is an important factor in producing models with good predictivity.¹⁷

Another set of amino acid descriptors commonly used in peptide QSAR studies is the z descriptors, which were obtained by applying principal component analysis to groups of physicochemical variables.¹⁸ These descriptors were used to analyze the activity of bradykinin potentiating pentapeptides and analogue peptide sets.^{19,20} Three z descriptors mainly explain hydrophilicity, size, and electronegativity of the amino acids.²¹ Later, Sandberg et al. used the z descriptors to classify 89 synthetic elastase substrates and 29 neurotensin peptide analogues.²² Their work generated models with high predictivity and a high level of explained variance ($q^2 = 0.77$ and $r^2 = 0.83$ for elastase substrates and $q^2 = 0.78$ and $r^2 = 0.93$ for neurotensin analogues). Recent applications of the z descriptors include modeling the relationship between the functions of the peptides and their amino acid composition,²³ predicting the activity of β -lactam antibiotics,²⁴ and applying genetic algorithms to optimize models generated using z descriptors.²⁵

The present study uses amino acid descriptors and 2D-QSAR techniques to describe the binding of peptides to the human class I MHC allele HLA-A*0201. This allele was chosen because it is well-studied and because binding data were abundant. Two sets of descriptors were used: the AAindex descriptors and the z descriptors. The question was how to identify only those descriptors that were relevant to the problem from a large selection of potential variables. To address this, three variable selection techniques were applied to this problem: partial least squares as implemented in the soft independent modeling of class analogy (SIMCA), genetic algorithm (GA), and generating optimal linear partial least-squares estimations (GOLPE).

Methods

Peptides. Two-hundred-sixty-six A*0201 nonamer peptides were used as a training set in MHC binding analysis, all of which were taken from the AntiJen

* To whom correspondence should be addressed. Phone: +44 1603 450852. Fax: +44 1603 450045. E-mail: pingping.guan@bbsrc.ac.uk.

Table 1. The Three z Descriptors Developed by Wold and the Five z Descriptors Developed by Sandberg

	z3 descriptors			z5 descriptors				
	z1	z2	z3	z1	z2	z3	z4	z5
A	0.07	-1.73	0.09	0.24	-2.32	0.60	-0.14	1.30
C	0.71	-0.97	4.13	0.84	-1.67	3.71	0.18	-2.65
D	3.64	1.13	2.36	3.98	0.93	1.93	-2.46	0.75
E	3.08	0.39	-0.07	3.11	0.26	-0.11	-3.04	-0.25
F	-4.92	1.30	0.45	-4.22	1.94	1.06	0.54	-0.62
G	2.23	-5.36	0.30	2.05	-4.06	0.36	-0.82	-0.38
H	2.41	1.74	1.11	2.47	1.95	0.26	3.90	0.09
I	-4.44	-1.68	-1.03	-3.89	-1.73	1.71	-0.84	0.26
K	2.84	1.41	-3.14	2.29	0.89	-2.49	1.49	0.31
L	-4.19	-1.03	-0.98	-4.28	-1.30	-1.49	-0.72	0.84
M	-2.49	-0.27	-0.41	-2.85	-0.22	0.47	1.94	-0.98
N	3.22	1.45	0.84	3.05	1.62	1.04	-1.15	1.61
P	-1.22	0.88	2.23	-1.66	0.27	1.84	0.70	2.00
Q	2.18	0.53	-1.14	1.75	0.50	-1.44	-1.34	0.66
R	2.88	2.52	-3.44	3.52	2.50	-3.50	1.99	-0.17
S	1.96	-1.63	0.57	2.39	-1.07	1.15	-1.39	0.67
T	0.92	-2.09	-1.40	0.75	-2.18	-1.12	-1.46	-0.40
W	-4.75	3.65	0.85	-4.36	3.94	0.59	3.44	-1.59

database.^{26,27} Experimentally measured IC_{50} values (pIC_{50}) were used as the dependent variable.²⁸⁻³⁶

Amino Acid Descriptors. The descriptors used in the first part of the analysis were taken from the amino acid descriptor database AAindex,⁵ which can be accessed at <http://www.genome.ad.jp/dbget/aaindex.html>. The AAindex database is composed of two sections: AAindex 1 containing amino acid indices (437 descriptors at the time of study) and AAindex 2 containing amino acid mutation matrices (71 amino acid mutation matrices at the time of study). The descriptors used in the study were taken from AAindex 1. The z descriptors were also used in the analysis. The three z descriptors were originally defined in a peptide QSAR study by Wold and co-workers.¹⁸ Later, Sandberg reexamined these descriptors and added two other properties ($z4$ and $z5$) to explain further molecular properties for both natural and synthetic amino acids.²² The main difference between the two descriptor sets is that AAindex includes large numbers of molecular descriptors such as side chain volume, pK_a , isoelectric point, etc., while the z descriptors are a small group of descriptors focused on four molecular properties: steric bulk, electrostatic, hydrophobicity, and electronic effects. The z descriptors are listed in Table 1.

Partial Least Squares (PLS) Method. PLS is an effective technique for finding the relationship between the properties of a molecule and its structure. In mathematical terms, PLS relates a matrix \mathbf{Y} of dependent variables to a matrix \mathbf{X} of molecular structure descriptors.³⁷ PLS decomposes the matrix \mathbf{X} into several latent variables that correlate best with the activity of the compounds. The latent variables are used to predict the activity (\mathbf{Y}). The PLS method as implemented in Sybyl 6.9 was used. Both the column filtering and the scaling were turned off. The optimal number of components was found by running cross-validation using SAMPLS.³⁸

Cross-Validation (CV). Models produced by PLS are validated by leave-one-out cross-validation (LOO-CV) and cross-validation in five groups. In LOO-CV each peptide in the model is omitted once. The following parameters are generated by the calculation and are used to assess the predictive ability of the models: the

cross-validated coefficient q^2 and the standard error of prediction SEP:

$$q^2 = 1 - \frac{\sum_{i=1}^n (pIC_{50exp} - pIC_{50pred})^2}{\sum_{i=1}^n (pIC_{50exp} - pIC_{50mean})^2}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^n (pIC_{50exp} - pIC_{50pred})^2}{n-1}}$$

where n represents the number of the peptides included in the model (for LOO-CV, n equals the number of peptides) and where pIC_{50pred} and pIC_{50exp} are the values predicted by LOO-CV for the binding affinity and from the binding experiments, respectively.

After cross-validation, a non-cross-validated model is generated by PLS using the number of principal components (PC) derived from CV. Three values are obtained in the calculation: the variance explained by the model (r^2), the standard error of estimate (SEE), and the F ratio.

PLS Implemented in SIMCA. The soft independent modeling of class analogy (SIMCA) is a multivariate data analysis program that groups variables with similar properties and decreases descriptor redundancy.³⁹ SIMCA-P 8.0 was used to perform the calculations. SIMCA uses PLS to build QSAR models. The correlation between the individual variable and the data matrix was observed through the variable importance in projection (VIP). VIP is the sum of the variable influence over all model dimensions and is a measure of variable importance. Higher VIP values ($VIP > 0.7$) indicated good correlation between the variable and the data. To improve the model quality, variables with low VIP values were excluded from the model in a stepwise manner.

Genetic Algorithm (GA). The GA-PLS used in this study was implemented by Shun Jin Chou at the Laboratory for Molecular Modelling, School of Pharmacy, University of North Carolina at Chapel Hill (<http://mmlin1.pha.unc.edu/~jin/QSAR/analyze.html>). The equation

$$\text{predictivity} = 1 - \frac{(n-1)(1-q^2)}{n-c}$$

was used as the fitness function in the Web implemented GA-PLS calculation. q^2 is the predictivity of the model, n is the number of compounds in the data set, and c is the optimal number of components.

GOLPE. Generating optimal linear partial least-squares estimations (GOLPE)⁴⁰ improves the predictivity of the model by excluding unwanted variables. In this way, models generated by GOLPE have a higher level of predictivity than ones generated by PLS alone. Fractional factorial design (FFD) was applied to reduce redundancy in the present data set, which generates models with a random subset of variables and compares the impact of missing variables on the PLS model.

T2 Stabilization Assay. The binding of peptides to HLA alleles was measured using a quantitative T2 cell surface stabilization assay.^{41,42} Briefly, cells (2×10^5 cells/well) were incubated with 100 μ L of test and control peptides (0.04–200 μ M) in the presence of AIMV (Life Technologies, Paisley, U.K.) and 100 nM β_2 -microglobulin. The plates were stored at 37 °C with 5% CO₂ overnight. After incubation, the cells were washed and surface levels of HLA-A*0201 were assessed by staining with FITC conjugated mouse antihuman HLA-A2 monoclonal antibody or mIgG_{2b}-FITC isotype control antibody. The cells were fixed with 4% paraformaldehyde. The MHC-bound fluorescence level was measured by facscalibur analysis (FACS), and the results were analyzed with the program Cellquest.

The fluorescence level of the peptides bound to HLA-A2 molecules was converted to fluorescence index (FI) values using⁴³

$$FI = \frac{F_S - F_B}{F_{T2} - F_B} \times 100.00$$

where F_S is the mean fluorescence index (MFI) of the test peptides, F_B is the no-peptide isotype antibody-stained control MFI, and F_{T2} is the no-peptide HLA-A2 antibody-stained control MFI. The binding affinities of the test peptides to HLA-A*0201 were obtained by converting their FI values to BL₅₀ values, which is the peptide concentration yielding the half-maximal FI value.

Results

HLA-A*0201 Models with AAindex Descriptors.

Many of the descriptors in the AAindex database described whole protein properties such as helix and β -sheet conformations. Because the present study was focused on short peptides, only descriptors that described individual amino acid properties were collected from the database. A total of 93 descriptors were selected, covering four major aspects: hydrophobicity, steric bulk, flexibility, and electrostatic forces (see Supporting Information).

QSAR models were generated for the HLA-A*0201 data set using PLS implemented in SIMCA. The training set includes 266 nonamer peptides, and pIC₅₀ values range from 4.3 to 9. The 93 descriptors were applied to each position of the nonamer peptide, generating a total of 837 (93 \times 9) columns in the matrix. Initially, the q^2 value was low for both LOO-CV ($q^2 = 0.259$) and CV in seven groups ($q^2 = 0.268$). The VIP value of each variable was calculated, and q^2 was improved by excluding variables with VIP values lower than 0.7. A list of q^2 values after variable exclusion is given in Table 2. There was a 4% improvement in q^2 after excluding nearly half of the variables (model 2, $q^2 = 0.305$); there were no significant changes in q^2 in the subsequent models. There was another slight increase in q^2 when two-thirds of the variables were excluded (model 8, $q^2 = 0.332$), after which q^2 started to decrease (models 9 and 10). In contrast, r^2 was the highest when all variables were included and was decreased as the number of variables dropped (model 2, $r^2 = 0.358$). The r^2 values were generally low for all the models. The best q^2 value was 0.332 from model 8, which indicated that

Table 2. Changes in q^2 of A*0201 Models after Variable Selection in SIMCA

model	variable no.	q^2	no. of components	r^2
1	837	0.268	1	0.458
2	440	0.305	1	0.358
3	337	0.317	1	0.361
4	276	0.323	1	0.363
5	260	0.324	1	0.364
6	237	0.327	1	0.365
7	229	0.329	1	0.368
8	223	0.332	1	0.370
9	216	0.329	1	0.366
10	206	0.324	1	0.361

Table 3. Results of the Three z Descriptors Models Calculated by Three Methods: GOLPE, GA, and SIMCA (SIMCA Does Not Report SEP or SEE Values)

	QSAR Models Using $z1$ – $z3$ Descriptors				
	q^2	SEP	NC	r^2	SEE
GOLPE	0.424	0.517	4	0.510	0.477
GA	0.396	0.534	3	0.528	0.472
SIMCA	0.292		2	0.383	

the model had only moderate predictivity. Similar results were obtained when applying GOLPE and GA to this descriptor set ($q^2 = 0.298$, $r^2 = 0.445$ for GOLPE and $q^2 = 0.260$ and $r^2 = 0.410$ for GA). The low q^2 and r^2 values obtained suggested that the models generated using AAindex descriptors were not predictive and were not appropriate for the analysis of peptide–MHC interactions.

A*0201 Models with the z Descriptors. The z descriptors are a class of properties obtained by PCA analysis. The $z1$ scale is the hydrophobicity scale where negative values indicate hydrophobicity and positive values indicate hydrophilicity. The $z2$ scale describes steric properties. A negative value in $z2$ corresponds to small amino acids with low molecular weights and small surface areas, while a positive value corresponds to large, bulky amino acids with large surface areas. The $z3$ scale describes electronic properties. Amino acids with negative $z3$ values are electronegative, and those with positive $z3$ values are electropositive. Compared with the AAindex descriptors, the z descriptors are fewer in number but represent large numbers of redundant, degenerate descriptors. The three z descriptors were used first. A total of 27 (3 \times 9) descriptors were applied to the training set. QSAR models were built using SIMCA, GOLPE, and GA. Results of the QSAR models are listed in Table 3. q^2 for the SIMCA model was the lowest among the three ($q^2 = 0.29$). The predictivities of the GA and the GOLPE models were 0.396 and 0.424, respectively. The relative coefficients of the variables in each model were plotted in Figure 1. The coefficients reflected the contributions of each variable at each position. Positive coefficients from all three models indicate that a property is favored, and negative coefficients from the three models indicate that a property is disfavored.

Considering the individual properties, hydrophobic amino acids were favored at P2, P3, P6, and P7 (with positive $z1$ coefficients) and disfavored at P4 and P8 (with negative $z1$ coefficients). Large amino acids were favored at P2, P3, P4, and P6 (with positive $z2$ coefficients) and disfavored at P5, P7, and P8 (negative $z2$ coefficients). Electronegative residues were preferred at

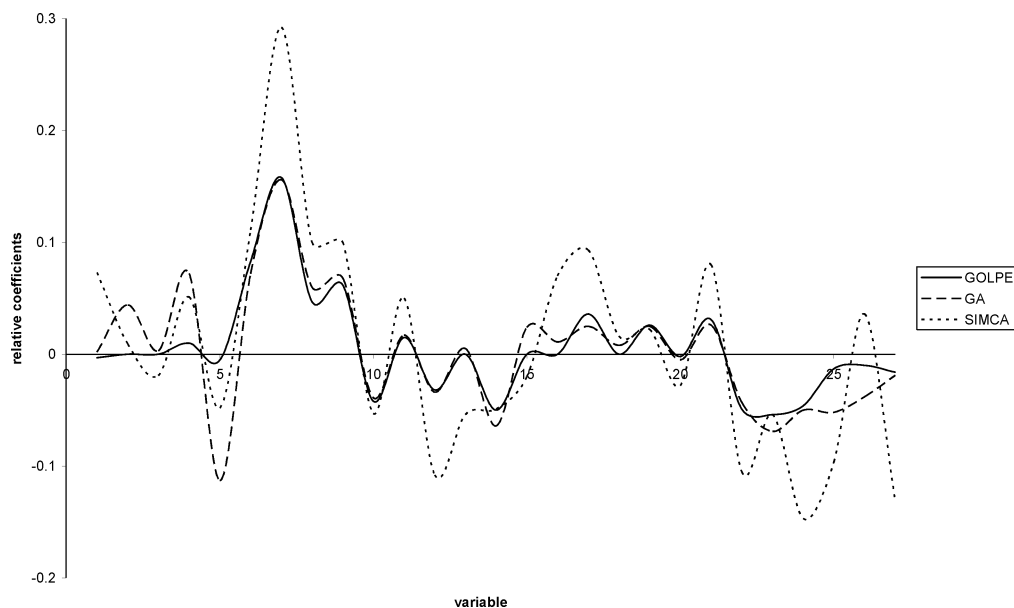


Figure 1. Relative coefficients of the QSAR models built by GOLPE, GA, and SIMCA using the three z descriptors.

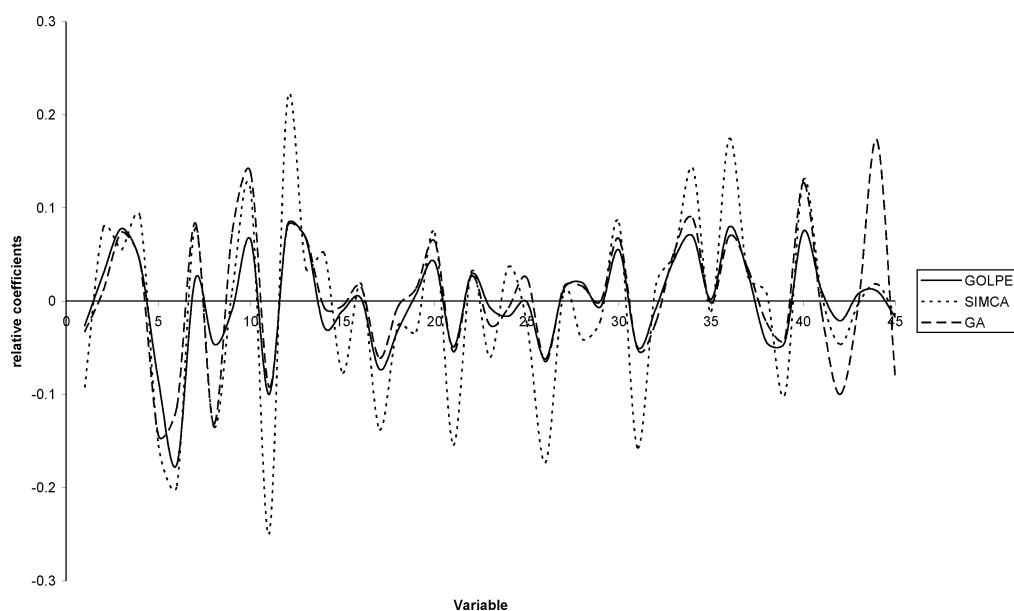


Figure 2. Relative coefficients of the QSAR models built by GOLPE, GA, and SIMCA using the five z descriptors.

Table 4. Results of the Five z Descriptors Models Calculated by Three Methods: GOLPE, GA, and SIMCA (SIMCA Does Not Include SEP or SEE Values in the Result)

	QSAR Models Using z_1 – z_5 Descriptors				
	q^2	SEP	NC	r^2	SEE
GOLPE	0.619	0.452	4	0.684	0.412
GA	0.606	0.464	4	0.732	0.383
SIMCA	0.702		2	0.897	

P4 and P8 (with negative z_3 coefficients), and electro-positive ones were preferred at P2, P3, P6, and P7 (positive z_3 coefficients).

A second set of models was created using the five z descriptors. A total of 45 (5×9) descriptors were applied to the training set. Results of the QSAR models are listed in Table 4. Relative coefficients of each position are shown in Figure 2. q^2 values of the five z descriptor models ranged from 0.6 to 0.7, which indicated good predictivity. The SIMCA model had the highest q^2 of 0.702 with the first two components, and its r^2 value

was 0.897, which was also the highest. The q^2 values of the GOLPE and the GA model were slightly lower: 0.619 and 0.606, respectively. Compared with other descriptor models, the five z descriptors of the QSAR model gave the best results. The relative coefficients of the z_5 descriptors revealed that hydrophilic amino acids are favored at P4 and P8 (with positive z_1 coefficients) but disfavored at P1, P2, P3, P5, P6, and P7 (with negative z_1 values) (Figure 2). Large bulky amino acids were preferred at P1, P2, P3, P5, P6, and P8 (with positive z_2 coefficients), while small amino acids were preferred at P4 and P9 (with negative z_2 coefficients). More polar amino acids are likely to appear at P1 and P4 (with positive z_3 and z_4 coefficients) but not at P2, P3, P5, and P7 (with negative z_3 coefficients).

The QSAR models generated by the three z and five z descriptors gave similar results. Both the three and the five z descriptors model showed that P2 favored bulky, nonpolar amino acids. P9 preferred small amino

Table 5. A*0201 Test Peptides and Their Experimental Binding Affinities

peptide	exptl pBL ₅₀	peptide	exptl pBL ₅₀
FLWPYHNV	4.94	YLFDPVTA	6.58
YLFPGPMTA	5.43	YLFPGPVTG	5.5
YLFDPGPVTA	4.5	YLFPGPMTV	6.09
YLFPGPFTA	4.72	YLFDPGPVTV	5.38
YLFPPPVTG	5.25	YLFPGPFTV	5.98
YLCPPGVTA	5.84	YLFPPPVTV	6.34
YLFPGVVTA	5.66	VLFNGPVTV	6.06
YLFPCPVTA	5.81		

acids in the five *z* descriptors model. No consensus was found at P9 in the three *z* descriptors model. Secondary anchor positions P3 and P7 favored nonpolar amino acids in both models. Bulky hydrophobic amino acids were identified at P1 in the five *z* descriptors model. P4 and P8 accept hydrophilic amino acids. P6 favored bulky amino acids in both models.

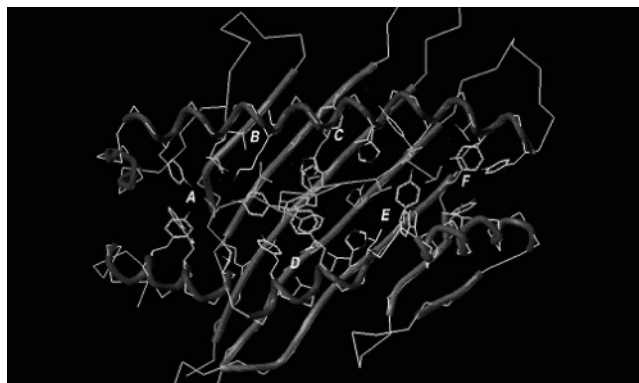
Peptide–MHC Binding Experiment. After defining an HLA-A*0201 binding model, 15 high-affinity peptides were designed and tested, using a T2 stabilization assay, with peptide binding affinities measured using BL₅₀ values. Because the original affinity of the HLA-A*0201 training set taken from Antigen was present as IC₅₀ values, previously published binding data on H-2Db restricted peptides were collected to compare the IC₅₀ and BL₅₀ measurements.^{44–46} These data, together with an A2 peptide binding study,⁴⁷ showed that there was a good linear relationship between IC₅₀ values and BL₅₀ values despite the different techniques used.

The presence of the anchor residues greatly influenced the binding affinity in the A*0201 analysis; therefore, all designed peptides possess the anchor residues of A*0201 (Table 5). Leu is present at P2 in all the peptides, and most peptides have Val or Ala at P9 (eight peptides have Ala at P9 and seven have Val). For secondary anchor positions, aromatic amino acids are favored both in previous studies and in the *z* descriptors analysis. Most of the peptides have Trp or Phe at P3, apart from three peptides that have either aliphatic or polar amino acids. The other secondary anchor position P7 is occupied by aliphatic or aromatic amino acids.

Peptides with pBL₅₀ values above 4 were considered as intermediate to good binders, and those with pBL₅₀ values above 6 were considered to be very good binders. In this experiment, all peptides had BL₅₀ values above 4. Four peptides had values above 6. The experimental results were in agreement with the findings from the 2D descriptor analysis. The presence of aliphatic amino acids Leu or Met at P2 and Gly, Val, Ala, or Tyr at P9 was important in peptide binding. Also, the presence of Phe at P3 and Val at P7 increased the binding affinity of the peptide. Tyr was well accepted at P1, as was Pro at P4. The presence of negatively charged residues Lys or Arg greatly reduces the ability of peptides to bind to A*0201, although the peptide has the preferred anchor residues, indicating that MHC–peptide binding is much more complex than the simple motif requirement.

Discussion

The crystal structure of HLA-A*0201 (PDB accession number 1HHI) indicated that the antigen binding site is located in sequence domain $\alpha 1/\alpha 2$.⁴⁸ Six binding

**Figure 3.** Structure of the HLA binding site. The protein is crystallized by Madden et al.⁴⁸**Table 6.** Residues That Form the Peptide Binding Pockets of the HLA-A*0201 Molecule

pocket	residue									
A	5	7	59	63	66	99	159	163	167	171
	M	Y	Y	E	K	Y	Y	T	W	Y
B	7	9	24	34	45	63	66	67	70	99
	Y	F	A	V	M	E	K	V	H	Y
C	9	22	70	73	74	97	99	114	116	
	F	F	H	T	H	R	Y	H	Y	
D	99	114	155	156	159	160				
	Y	H	Q	L	Y	L				
E	97	114	116	147	152	155	156			
	R	H	Y	W	V	Q	L			
F	73	77	80	81	84	95	116	118	123	124
	T	D	T	L	Y	V	Y	Y	Y	I

pockets (A–F) are present in the binding site to accommodate the side chains of the antigenic peptide (Figure 3). The specificity of these pockets largely determines the specificity of the A*0201 allele.

Pocket A is located at the end of the binding groove and accommodates the side chain of P1. The surface of pocket A is dominated by five tyrosine residues: Tyr7, Tyr59, Tyr99, Tyr159, and Tyr171 (see Table 6), among which Tyr7, Tyr59, and Tyr159 are conserved among A2 alleles. The bottom of the pocket is occupied by Tyr7. The composition of the surface suggests a preference for aromatic residues in pocket A. The hydroxyl group on the side chain of tyrosine is a potential hydrogen bond acceptor and could interact with amino acids with hydrogen-bond donor side chains. The results of the present study confirmed that aromatic residues are favored at P1.

The side chain of P2 interacts with pocket B.⁴⁹ Residues lining the pocket are bulky and hydrophobic (Phe9, Met45, and Val67), which reduced the volume inside the pocket. Nonpolar residues Ala24 and Val34 are located at the bottom of the pocket. Results of the *z* descriptors model show that medium size hydrophobic residues Leu and Ile are preferred at P2. The side chains of these two amino acids can extend to the bottom of pocket B. This is confirmed in the binding experiment, in which all high binders of A*0201 possessed Leu at P2.

Aromatic residues such as Tyr and Trp are favored at P3 and P7. The side chain of the residue extends into pocket D and interacts with the hydrophobic residues inside the pocket: Leu156, Tyr99, and Tyr159. Pocket

E accepts the side chain of the amino acid at P7, which may interact with the aromatic residues Trp133 at the bottom of the pocket. Hydrophobic residues such as Val and Met are also accepted at P7. In the peptide binding experiment, the three high binders have Phe at P3 and Val or Met at P7 (YLFDPVTA, YLFGPMTV, VLFNG-PVTV).

The side chain of P6 interacts with pocket C, which is shallow with polar residues lining the inside (His70, Thr73, His74, and Arg97). The bottom of the pocket is defined by aromatic residue Phe9. In the present study, P6 accommodates a variety of amino acids; aromatic residues (Tyr, Trp and Phe) and medium-sized hydrophobic residues (Leu, Ile, and Pro) are all accepted. Tyr and Trp are potential hydrogen bond donors that could interact with the polar residues inside the pocket, while Phe can reach the bottom of the pocket and stabilize binding.

P9 is known to be an important anchor residue in HLA-A*0201. The side chain of P9 extends into pocket F, which is relatively deep with side chains of Leu81, Tyr123, and Tyr116 at the bottom. In the present study, P9 favors medium size, nonpolar residues such as Leu, Ile, and Met. Two small nonpolar amino acids Ala and Val are also well accepted, as demonstrated in the peptide binding experiment. Thr is the only polar residue favored at this position, which may form hydrogen bonds with Thr and Tyr residues in pocket F (Thr80, Tyr84, Thr143).

Side chains of P4, P5, and P8 are orientated toward the outside of the groove and interact with the T cell receptor. The amino acids at these positions are more diverse. P4 prefers small amino acids, and P5 prefers hydrophobic residues. In the peptide binding experiment, peptides with Pro at P4 and Gly at P5 are well accepted. P8 favors more hydrophilic residues.

Previously, A*0201 motifs have been defined by quantitative binding assays⁵⁰ and by the grouping of eluted peptides and epitopes.⁵¹ Leu, Ile, Val, Ala, and Met at P2 and P9 have been identified as the most preferred amino acid from previous studies,⁵² which were in agreement with the present study. For nonanchor residues, we identified aromatic residues at P1, small hydrophilic residues at P4, and hydrophobic amino acids at P5, while the binding study by Drijfhout suggested that Lys, Tyr, and Thr are at P1 and that P4 and P5 accept both polar and nonpolar residues.⁵³ As previous studies indicate,⁵⁴ these positions are involved with TCR interaction and the amino acids that occupy these positions may vary greatly between epitopes from different organisms. Furthermore, the differences at the positions may also be due to the different peptide data set used. Results of the peptide binding experiments showed that peptides with the favored residues identified from the 2D-QSAR study bound to HLA-A*0201 with high affinity.

In the present study, the interactions between A*0201 and peptides were analyzed by two sets of descriptors: the 93 amino acid descriptors taken from AAindex and the z descriptors. Initially, QSAR models generated by the AAindex descriptors had low predictivity, which was improved after applying variable selection. However, as more variables were removed, the explained variance (r^2) was reduced. In comparison, models generated using

the three and five z descriptors had relatively higher predictivity. The explained variance of the z descriptor models was similar to that of models using 93 descriptors, although the latter had a significantly higher number of variables. The z descriptors were originally obtained by applying PCA and PLS to large numbers of variables. Thus, the great advantage of the z descriptors is that they represent a large number of similar descriptors and therefore reduce redundancy in the model, which often affects model predictivity. Our results suggest that in peptide QSAR analysis, the quality of models is not proportional to the number of descriptors used. Including many descriptors in a model is likely to increase redundancy and reduce model predictivity. Choosing a small but more representative set of descriptors leads to high-quality models.

Overall, the present study characterized peptide–A*0201 interactions using two sets of amino acid descriptors, and high-affinity peptides were designed on the basis of the results of this study. Previously, CoMSIA and the additive method, which also use PLS regression, were used to analyze peptide–MHC interactions.^{55–60} Compared with these analyses, the z descriptor models had similar levels of predictivity (q^2 between 0.5 and 0.7) and could be a promising tool for studying peptide–MHC interactions. The methods used in our analysis can be applied to study other MHC alleles, and the results can be used to synthesize peptide analogues that will allow us to model peptide–MHC or peptide–TCR interactions and design novel epitopes and immunomodulators.

This method is not limited to peptide–MHC interactions and can be applied to any QSAR study of protein–peptide binding events. The use of z scales also has two other important advantages. First, the use of scales as descriptors, rather than indicator variables, to represent amino acid identity allows us to deal with missing residues in the training set in a straightforward way. Prediction of MHC binding, for example, is often limited by omissions in the training data, where certain positions in a peptide, usually called anchors, exhibit a greatly reduced set of amino acids compared to much more complete distributions at other positions. This results in many potentially active peptides being essentially unpredictable. The use of scales, which contain numerical values for all 20 amino acids, neatly circumvents this problem. Second, z scales can be extended to include synthetic or rarely encountered natural amino acids, such as citrulline, other than the natural 20 amino acids. In the context of MHC, many such posttranslational modifications (PTMs) are known to be recognized within T cell epitopes.^{61,62} Moreover, other PTMs can also generate immunogenic epitopes, such as phosphorylation⁶³ and glycosylation.⁶⁴

The z scale based method described here, which utilizes variable selection, is an effective and efficient approach to the QSAR modeling of peptide–protein interactions of any type. In the context of peptide–MHC binding, we have demonstrated, and verified experimentally, that this method is as effective as other QSAR techniques based on an indicator variable representation of peptide sequences. The flexible and extensible nature of this approach is indicative of its future potential, and we are hopeful that it will contribute

significantly to the future development of vaccine design and discovery.

Acknowledgment. We thank Andrew Worth for his invaluable technical assistance. The Edward Jenner Institute for Vaccine Research thanks its sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the U.K. Department of Health.

Supporting Information Available: A table including 93 descriptors used in the study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Sneath, P. H. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **1966**, *12* (2), 157–195.
- Kidera, A.; Konishi, Y.; Poka, M.; Ooi, T.; Scheraga, H. A. Statistical analysis of the physical properties of the 10 naturally occurring amino acids. *J. Protein Chem.* **1985**, *4*, 23–55.
- Nakai, K.; Kidera, A.; Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **1988**, *2* (2), 93–100.
- Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32* (4), 269–278.
- Kawashima, S.; Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **2000**, *28* (1), 374.
- Collantes, E. R.; Dunn, W. J., 3rd. Amino acid side chain descriptors for quantitative structure–activity relationship studies of peptide analogues. *J. Med. Chem.* **1995**, *38* (14), 2705–2713.
- Norinder, U. Theoretical amino acid descriptors. Application to bradykinin potentiating peptides. *Peptides* **1991**, *12* (6), 1223–1227.
- Felipe-Sotelo, M.; Andrade, J. M.; Carlosena, A.; Prada, D. Partial least squares multivariate regression as an alternative to handle interferences of Fe on the determination of trace Cr in water by electrothermal atomic absorption spectrometry. *Anal. Chem.* **2003**, *75* (19), 5254–61.
- Hasegawa, K.; Funatsu, K. Partial least squares modeling and genetic algorithm optimization in quantitative structure–activity relationships. *SAR QSAR Environ Res.* **2000**, *11* (3–4), 189–209.
- Cui, M.; Huang, X.; Luo, X.; Briggs, J. M.; Ji, R.; Chen, K.; Shen, J.; Jiang, H. Molecular docking and 3D-QSAR studies on gag peptide analogue inhibitors interacting with human cyclophilin A. *J. Med. Chem.* **2002**, *45* (24), 5249–5259.
- Gupta, M. K.; Mishra, P.; Prathipati, P.; Saxena, A. K. 2D-QSAR in hydroxamic acid derivatives as peptide deformylase inhibitors and antibacterial agents. *Bioorg. Med. Chem.* **2002**, *10* (12), 3713–3716.
- Eriksson, L.; Jonsson, J.; Hellberg, S.; Lindgren, F.; Skagerberg, B.; Sjostrom, M.; Wold, S. Peptide QSAR on substance P analogues, enkephalins and bradykinins containing L- and D-amino acids. *Acta Chem. Scand.* **1990**, *44* (1), 50–55.
- Nadasdi, L.; Medzihradsky, K. A study of the applicability of QSAR calculation for peptide hormones. *Biochem. Biophys. Res. Commun.* **1981**, *99* (2), 451–457.
- Sun, H. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 748–757.
- Zhao, Y.; Jona, J.; Chow, D. T.; Rong, H.; Semin, D.; Xia, X.; Zanon, R.; Spancake, C.; Maliski, E. High-throughput logP measurement using parallel liquid chromatography/ultraviolet/mass spectrometry and sample-pooling. *Rapid Commun. Mass Spectrom.* **2002**, *16* (16), 1548–1555.
- Xing, L.; Glen, R. C. Novel methods for the prediction of logP, pK(a), and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 796–805.
- Hunt, P. A. QSAR using 2D descriptors and Tripos' SIMCA. *J. Comput.-Aided Mol. Des.* **1999**, *13* (5), 453–467.
- Hellberg, S.; Sjostrom, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30* (7), 1126–1135.
- Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjostrom, M.; Skagerberg, B.; Wold, S.; Andrews, P. Minimum analogue peptide sets (MAPS) for quantitative structure–activity relationships. *Int. J. Pept. Protein Res.* **1991**, *37* (5), 414–424.
- Hellberg, S.; Sjostrom, M.; Wold, S. The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure–activity relationship. *Acta Chem. Scand. B* **1986**, *40* (2), 135–140.
- Eriksson, L.; Jonsson, J.; Sjostrom, M.; Wold, S. Multivariate parametrization of coded and non-coded amino acids by thin-layer chromatography. *Prog. Clin. Biol. Res.* **1989**, *291*, 131–134.
- Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjostrom, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
- Siebert, K. J. Quantitative structure–activity relationship modeling of peptide and protein behavior as a function of amino acid composition. *J. Agric. Food. Chem.* **2001**, *49* (2), 851–858.
- Wanchana, S.; Yamashita, F.; Hara, H.; Fujiwara, S.; Akamatsu, M.; Hashida, M. Two- and three-dimensional QSAR of carrier-mediated transport of beta-lactam antibiotics in Caco-2 cells. *J. Pharm. Sci.* **2004**, *93* (12), 3057–3065.
- Bayram, E.; Santago, P., 2nd; Harris, R.; Xiao, Y. D.; Clauset, A. J.; Schmitt, J. D. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *J. Comput.-Aided Mol. Des.* **2004**, *18* (7–9), 483–493.
- McSparron, H.; Blythe, M. J.; Zygouri, C.; Doytchinova, I. A.; Flower, D. R. JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1276–1287.
- Toseland, C. P.; Taylor, D. J.; McSparron, H.; Hemsley, S. L.; Blythe, M. J.; Paine, K.; Doytchinova, I. A.; Guan, P.; Hattotuwagama, C. K.; Flower, D. R. AntiPep: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data. *Immunome Res.* **2005**, *1*:4.
- Tsai, Y.; Southwood, S.; Sidney, J.; Sakaguchi, K.; Kawakami, Y.; Appella, E.; Sette, A.; Celis, E. Identification of subdominant CTL epitopes of the GP100 melanoma-associated tumor antigen by primary in vitro immunization with peptide-pulsed dendritic cells. *J. Immunol.* **1997**, *158* (4), 1796–1802.
- Rongcun, Y.; Salazar-Onfray, F.; Charo, J.; Malmberg, K. J.; Evrin, K.; Maes, H.; Kono, K.; Hising, C.; Petersson, M.; Larsson, O.; Lan, L.; Appella, E.; Sette, A.; Celis, E.; Kiessling, R. Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas. *J. Immunol.* **1999**, *163* (2), 1037–1044.
- Rivoltini, L.; Kawakami, Y.; Sakaguchi, K.; Southwood, S.; Sette, A.; Robbins, P. F.; Marincola, F. M.; Salgaller, M. L.; Yannelli, J. R.; Appella, E.; et al. Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1. *J. Immunol.* **1995**, *154* (5), 2257–2265.
- Parkhurst, M. R.; Fitzgerald, E. B.; Southwood, S.; Sette, A.; Rosenberg, S. A.; Kawakami, Y. Identification of a shared HLA-A*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2). *Cancer Res.* **1998**, *58* (21), 4895–4901.
- Kast, W. M.; Brandt, R. M.; Sidney, J.; Drijfhout, J. W.; Kubo, R. T.; Grey, H. M.; Melief, C. J.; Sette, A. Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J. Immunol.* **1994**, *152* (8), 3904–3912.
- Sette, A.; Vitiello, A.; Reherman, B.; Fowler, P.; Nayarsina, R.; Kast, W. M.; Melief, C. J.; Oseroff, C.; Yuan, L.; Ruppert, J.; et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **1994**, *153* (12), 5586–5592.
- del Guercio, M. F.; Sidney, J.; Hermanson, G.; Perez, C.; Grey, H. M.; Kubo, R. T.; Sette, A. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol.* **1995**, *154* (2), 685–693.
- Vitiello, A.; Sette, A.; Yuan, L.; Farness, P.; Southwood, S.; Sidney, J.; Chesnut, R. W.; Grey, H. M.; Livingston, B. Comparison of cytotoxic T lymphocyte responses induced by peptide or DNA immunization: implications on immunogenicity and immunodominance. *Eur. J. Immunol.* **1997**, *27* (3), 671–678.
- Parkhurst, M. R.; Salgaller, M. L.; Southwood, S.; Robbins, P. F.; Sette, A.; Rosenberg, S. A.; Kawakami, Y. Improved induction of melanoma-reactive CTL with peptides from the melanoma antigen gp100 modified at HLA-A*0201-binding residues. *J. Immunol.* **1996**, *157* (6), 2539–2548.
- Wold, S. PLS for Multivariate Linear Modelling. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195–218.
- Bush, B. L.; Nachbar, R. B., Jr. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7* (5), 587–619.
- Eriksson, L.; Johansson, E. Multivariate design and modeling in QSAR. *Chemom. Intell. Lab. Syst.* **1996**, *34* (1), 1–19.
- Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37* (16), 2589–2601.

- (41) MyIntyre, C.; Rees, R.; Platts, K.; Cooke, C.; Smith, M.; Mulcahy, K.; Murray, A. Identification of peptide epitopes of MAGE-1, -2, -3 that demonstrate HLA-A3-specific binding. *Cancer Immunol. Immunother.* **1996**, *42*, 246–250.
- (42) Salter, R.; Cresswell, P. Impaired assembly and transport of HLA-A and -B antigens in a mutant TxB cell hybrid. *EMBO J.* **1986**, *5*, 943–949.
- (43) Yoon, H.; Chung, M. K.; Min, S. S.; Lee, H. G.; Yoo, W. D.; Chung, K. T.; Jung, N. P.; Park, S. N. Synthetic peptides of human papillomavirus type 18 E6 harboring HLA-A2.1 motif can induce peptide-specific cytotoxic T-cells from peripheral blood mononuclear cells of healthy donors. *Virus Res.* **1998**, *54* (1), 23–29.
- (44) Gairin, J. E.; Mazarguil, H.; Hudrisier, D.; Oldstone, M. B. Optimal lymphocytic choriomeningitis virus sequences restricted by H-2Db major histocompatibility complex class I molecules and presented to cytotoxic T lymphocytes. *J. Virol.* **1995**, *69* (4), 2297–2305.
- (45) Hudrisier, D.; Mazarguil, H.; Laval, F.; Oldstone, M. B. A.; Gairin, J. E. Binding of viral antigens to major histocompatibility complex class I H-2Db molecules is controlled by dominant negative elements at peptide non-anchor residues. *J. Biol. Chem.* **1996**, *271*, 17829–17836.
- (46) Hudrisier, D.; Mazarguil, H.; Oldstone, M. B.; Gairin, J. E. Relative implication of peptide residues in binding to major histocompatibility complex class I H-2Db: application to the design of high-affinity, allele-specific peptides. *Mol. Immunol.* **1995**, *32* (12), 895–907.
- (47) Doytchinova, I. A.; Walshe, V. A.; Jones, N. A.; Gloster, S. E.; Borrow, P.; Flower, D. R. Coupling in silico and in vitro analysis of peptide–mhc binding: A bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunol.* **2004**, *172*, 7495–7502.
- (48) Madden, D. R.; Garboczi, D. N.; Wiley, D. C. The antigenic identity of peptide–MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **1993**, *75*, 693–708.
- (49) Young, A. C.; Zhang, W.; Sacchettini, J. C.; Nathenson, S. G. The three-dimensional structure of H-2Db at 2.4 Å resolution: implications for antigen-determinant selection. *Cell* **1994**, *76* (1), 39–50.
- (50) Ruppert, J.; Sidney, J.; Celis, E.; Kubo, R. T.; Grey, H. M.; Sette, A. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **1993**, *74* (5), 929–937.
- (51) Rammensee, H. G.; Friede, T.; Stevanović, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **1995**, *41* (4), 178–228.
- (52) Falk, K.; Rotzschke, O. Consensus motifs and peptide ligands of MHC class I molecules. *Semin. Immunol.* **1993**, *5* (2), 81–94.
- (53) Drijfhout, J. W.; Brandt, R. M.; D'Amato, J.; Kast, W. M.; Melief, C. J. Detailed motifs for peptide binding to HLA-A*0201 derived from large random sets of peptides using a cellular binding assay. *Hum. Immunol.* **1995**, *43* (1), 1–12.
- (54) Falk, K.; Rotzschke, O.; Stevanović, S.; Jung, G.; Rammensee, H. G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **1991**, *351* (6324), 290–296.
- (55) Doytchinova, I.; Flower, D. The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem.* **2003**, *1* (15), 2648–2654.
- (56) Doytchinova, I. A.; Blythe, M. J.; Flower, D. R. Additive method for the prediction of protein–peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.* **2002**, *1* (3), 263–272.
- (57) Doytchinova, I. A.; Flower, D. R. A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput.-Aided Mol. Des.* **2002**, *16* (8–9), 535–544.
- (58) Guan, P.; Doytchinova, I. A.; Flower, D. R. A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. *Bioorg. Med. Chem.* **2003**, *11* (10), 2307–2311.
- (59) Guan, P.; Doytchinova, I. A.; Flower, D. R. HLA-A3 supermotif defined by quantitative structure–activity relationship analysis. *Protein Eng.* **2003**, *16* (1), 11–18.
- (60) Hattotuwa, C. K.; Guan, P.; Doytchinova, I. A.; Flower, D. R. New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity. *Org. Biomol. Chem.* **2004**, *2* (22), 3274–3283.
- (61) McAdam, S. N.; Fleckenstein, B.; Rasmussen, I. B.; Schmid, D. G.; Sandlie, I.; Bogen, B.; Viner, N. J.; Sollid, L. M. T cell recognition of the dominant I-A(k)-restricted hen egg lysozyme epitope: critical role for asparagine deamidation. *J. Exp. Med.* **2001**, *193* (11), 1239–1246.
- (62) Hill, J. A.; Southwood, S.; Sette, A.; Jevnikar, A. M.; Bell, D. A.; Cairns, E. Cutting edge: the conversion of arginine to citrulline allows for a high-affinity peptide interaction with the rheumatoid arthritis-associated HLA-DRB1*0401 MHC class II molecule. *J. Immunol.* **2003**, *171* (2), 538–541.
- (63) Andersen, M. H.; Bonfill, J. E.; Neisig, A.; Arsequell, G.; Sondergaard, I.; Valencia, G.; Neefjes, J.; Zeuthen, J.; Elliott, T.; Haurum, J. S. Phosphorylated peptides can be transported by TAP molecules, presented by class I MHC molecules, and recognized by phosphopeptide-specific CTL. *J. Immunol.* **1999**, *163* (7), 3812–3818.
- (64) Holm, L.; Kjellen, P.; Holmdahl, R.; Kihlberg, J. Identification of the minimal glycopeptide core recognized by T cells in a model for rheumatoid arthritis. *Bioorg. Med. Chem.* **2005**, *13* (2), 473–482.

JM0505258