ARTICLE

OBC
www.rsc.org/obc

The HLA-A2-supermotif: a QSAR definition †

Irini Doytchinova and Darren Flower *
Edward Jenner Institute for Vaccine Research, Compton, Berkshire, UK RG20 7NN

Received 22nd January 2003, Accepted 12th June 2003
First published as an Advance Article on the web 23rd June 2003

Identification of epitopes capable of binding multiple HLA types will significantly rationalise the development of epitope-based vaccines. A quantitative method assessing the contribution of each amino acid at each position was applied to over 500 nonamer peptides binding to 5 MHC alleles — A*0201, A*0202, A*0203, A*0206 and A*6802 — which together define the HLA-A2-like supertype. FXIGXI (L)IFV was identified as a supermotif for the A2-supertype based on the contributions of the common preferred amino acids at each of the nine positions. The results indicate that HLA-A*6802 is an intermediate allele standing between A2 and A3 supertypes: at anchor position 2 it is closer to A3 and at anchor position 9 it is nearer to A2. Models are available free on-line at http://www.jenner.ac.uk/MHCPred and can be used for binding affinity prediction.

Introduction

The biological function of HLA molecules is to bind antigenic peptides (epitopes) and to present them to T cells.[1] The elimination of epitope-bearing cells by cytotoxic T lymphocytes (CTLs) plays a central role in the eradication of both infectious diseases and cancer by the immune system.[2,3] This process of antigenic peptide recognition underlies the development of epitope-based vaccines.[4–6] X-Ray data show that the peptide-binding site has a 30 Å long surface accessible to a solvent probe.[7] There are six pockets in the surface denoted A through F. Some of the pockets are non-polar and can form hydrophobic contacts, but others contain polar atoms that can make hydrogen bonds with the peptide side chains. HLA polymorphism tends to be concentrated in these hypervariable binding pockets suggesting a structural basis for allelic specificity in antigen binding. The stereoelectronic and hydrophobic complementarity between the side chain at position 2 of the peptide and pocket B of the MHC molecule, as well as the C-terminal and pocket F, are of primary importance for determining peptide affinity.[8] They are denoted as primary anchor residues. Within a potential epitope, the presence of primary anchors is necessary, but not sufficient, for high-affinity binding. Prominent roles for several other positions (1, 3, 6 and 7), so-called secondary anchor residues, have also been demonstrated.[9–11] Tangri *et al.*[12] found that substitutions at positions 3, 5 and 7 gave rise to heteroclitic peptides (peptides which are more antigenic than wild-type peptide). The side chains at positions 4 and 8 are solvent-exposed in the complex with the MHC molecule and therefore they can contact the TCR. These positions are called "flag" positions.[11]

The combination of two or more anchor residues is called a binding motif.[8] Beside the extreme polymorphism exhibited by the HLA molecules, it was found that some alleles recognise very similar motifs. They were grouped into HLA supertypes. Motif binding to the same supertype is called a supermotif.[5,13] Four different HLA supertypes and corresponding supermotifs have been defined based on two primary anchor positions (position 2 and C-terminal) of the binding peptides: for HLA-A2,[14] HLA-A3,[4] HLA-B7[15] and HLA-B44[16]-like alleles. These four supertypes cover 80–90% of the general population[13] and the development of a single peptide capable of binding to each supertype is a tempting goal. Definition of supermotifs will significantly rationalise the development of epitope-based vaccines.

Recently, more detailed A2-supermotifs based on four anchors for 9-mers and on six anchors for 10-mers were published.[17] In the present study we define the preferred and deleterious amino acids at each position applying the additive method to binding data for 5 alleles: A*0201, A*0202, A*0203, A*0206 and A*6802. The additive method is a two-dimensional quantitative structure–activity relationships (2D-QSAR) method, which we developed recently.[18] It is based on the Free–Wilson's concept[19] whereby each substituent makes an additive and constant contribution to the biological activity regardless of substituent variation in the rest of the molecule. Parker's hypothesis[20,21] that each amino acid side chain binds independently of the rest of the peptide (IBS hypothesis) is also derived from this concept. We extended Free–Wilson's additive concept with terms accounting for the possible interactions between amino acid side chains. Because of the twisted conformation of the binding peptide only the adjacent and every second side-chain interactions will contribute to the affinity. Thus, the binding affinity of a nonamer peptide could be represented by eqn. (1):

$$p\mathrm{IC}_{50} = \text{constant} + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2} \qquad (1)$$

where $p\mathrm{IC}_{50}$ is the binding affinity measured in a radiolabeled assay and represented as a negative logarithm, the constant accounts, at least nominally, for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acid contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, and $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interaction. Initially, we applied the additive method to HLA-A*0201 binding data.[18] Recently we extended its application to alleles belonging to the HLA-A3 superfamily.[22] In the present study the method deals with peptides binding to another four alleles from the A2-superfamily. The method is, however, universal and can be applied to any peptide protein interaction where quantitative binding data is known.

Results

Peptide database

The peptide sequences and their binding affinities were extracted from the JenPep database[23] (http://www.jenner.ac.uk/

† Electronic supplementary information (ESI) available: matrices for A*6802, A*0206, A*0203, A*0202 and A*0201. See http://www.rsc.org/suppdata/ob/b3/b300707c/

DOI: 10.1039/b300707c

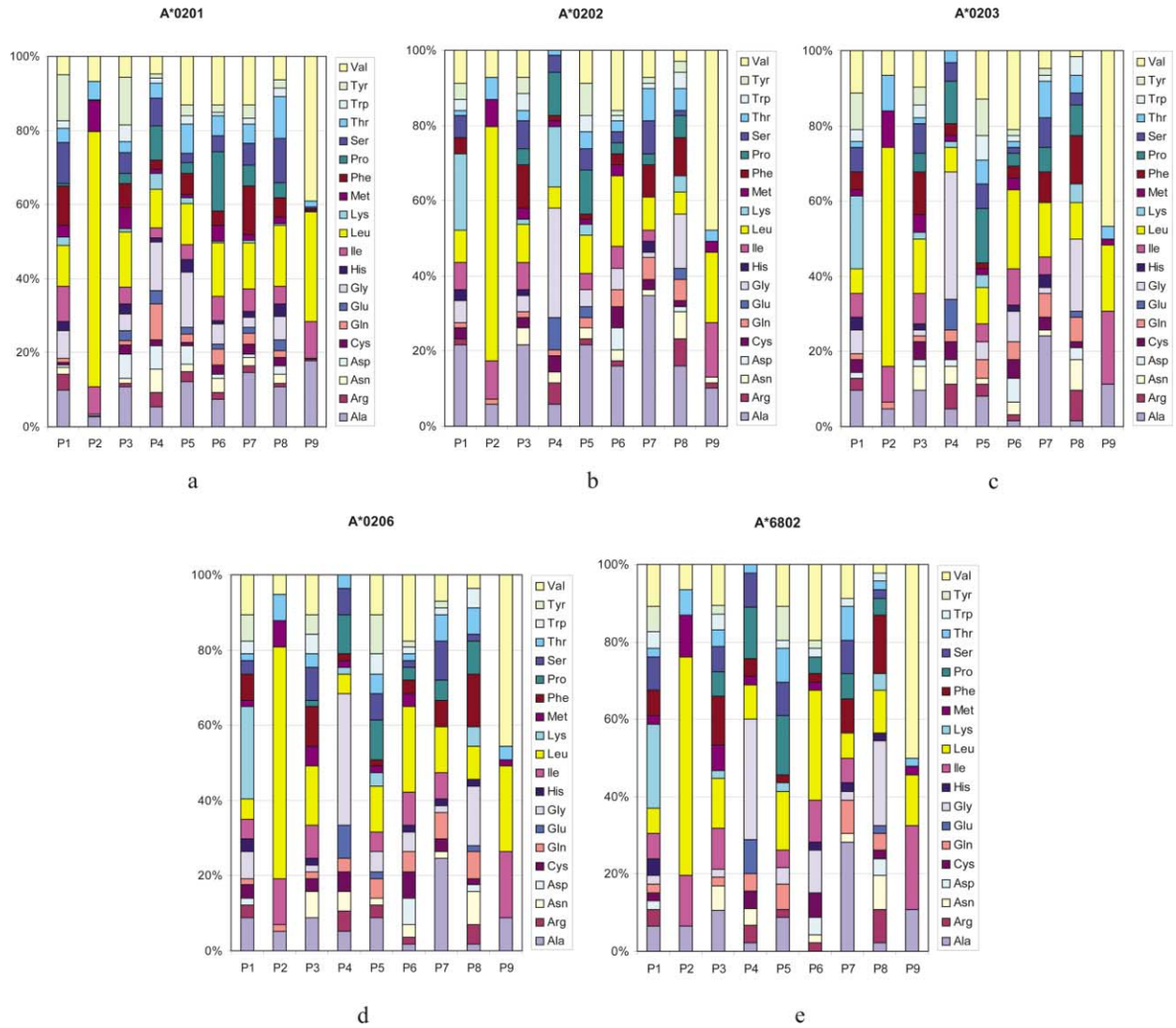


Fig. 1 Amino acids presented at each position in the peptide sets for (a) A*0201, (b) A*0202, (c) A*0203, (d) A*0206 and (e) A*6802 alleles.

Jenpep). The selected set of peptides binding to A*0201 consists of 335 nonamers. Unfortunately, for the rest of the alleles belonging to the A2-supertype there were far fewer data. All the peptides included in the study were nonamers. The set for A*0202 included 69 peptides, 62 peptides for A*0203, 57 peptides for A*0206, and 46 peptides for A*6801. Some of the peptides bind more than one allele. The binding affinities (IC_{50}) we used were originally assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules.[10,24] The negative logarithms of IC_{50} values were used as they are related to changes in the free energy of binding.[25] The magnitude of measured binding affinity ranges over almost 5 orders: from 4.5 to 9.0 in log units. The number of each type of amino acid at each position for the five alleles is given in Fig. 1. The number of missing amino acids ranges from 21 for A*0201 allele to 65 for A*6802. Most of them are at positions 2 and 9. Many amino acids are presented only once at a certain position. From this, one might presume their contributions and even more the contributions of their 1–2 and 1–3 interactions with other positions are spurious, achieving significance by chance. However, by disregarding these single amino acids one runs the risk of eliminating legitimate predictors. This problem will reduce greatly as the database of peptides grows.

Matrix construction

The data flow in the additive method is presented in Fig. 2. The nine amino acid peptide sequences were transformed into rows consisting of 6180 terms. One hundred and eighty columns account for the amino acids contributions (20 aa × 9 positions), 3200 for the adjacent side-chains, or 1–2 interactions (20 × 20 × 8) and 2800 for every second side-chain, or 1–3 interactions (20 × 20 × 7). A term is equal to 1 when a certain amino acid at a certain position or a certain interaction between two side-chains exists, and 0 when they are absent. Thus, matrices with 6180 columns and a number of rows equal to the number of peptides in the set were generated. To reduce the column number, columns containing only 0s were omitted. To deal with these matrices a robust multivariate statistical method, named partial least squares (PLS), was used.

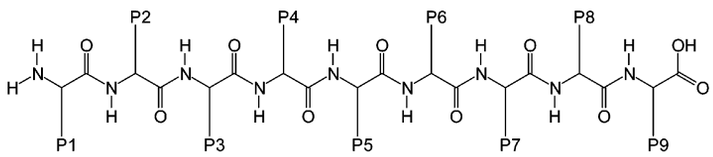
The predictive power of the models was assessed by the cross-validated coefficients q^2_{LOO} and q^2_{CV5}, the standard error of prediction (SEP), and the residuals between the experimental (pIC_{50exp}) and predicted by LOO-CV binding affinity (pIC_{50pred}), as described in the Computational details section. The non-cross-validated models were assessed by the MLR parameters: explained variance r^2, standard error of estimate (SEE), and the F ratio.

Algorithm implementation

Initially, we applied eqn. (1) to the binding data for A*0202, A*0203, A*0206 and A*6802 alleles. For the A*0201 allele we used the equation published in our previous paper.[18] The number of columns in the matrices generated for the former 4 alleles was 13–17 times higher than the numbers of rows (number of peptides in the set). For the A*0201 allele this difference was

Table 1 Additive models

Parameter	A*0201		A*0202		A*0203		A*0206		A*6802	
n	335		69		62		57		46	
q^2_{LOO}	0.377		0.317		0.327		0.475		0.500	
q^2_{CV5}	0.360		0.309		0.316		0.452		0.478	
NC	6		9		6		6		7	
SEP	0.694		0.606		0.841		0.576		0.647	
r^2	0.731		0.943		0.963		0.989		0.983	
SEE	0.456		0.193		0.197		0.085		0.119	
F ratio	148.661		109.101		239.300		728.521		313.298	
Res. ≤ |0.5|	188	56%	39	57%	29	47%	36	63%	24	52%
|0.5| < res. ≤ |1.0|	103	31%	20	29%	21	34%	19	33%	18	39%
Res. > |1.0|	44	13%	10	14%	12	19%	2	4%	4	9%
Mean |residual|	0.546		0.546		0.652		0.443		0.517	
Standard deviation	0.417		0.361		0.453		0.310		0.317	



$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2}$$

peptide	pIC_{50}	20AA x 9 180 columns	20 x 20 x 8 3200 columns	20 x 20 x 7 2800 columns
1	6.398	1……0…	1……0…	1……0…
…….	……	……1……0	……1……0	……1……0
…….	……	……1……	……1……	……1……
n	……	……0……1……0	……0……1……0	……0……1……0

PLS

pIC_{50} = 6.691 + (0.007) * 1A - (0.030) * 1R - (0.004) * 1C - (0.004) * 1Q + …… + (0.162) * 9V

- (0.019) * 1A2A + (0.013) * 1A2I - (0.002) * 1A2L + (0.029) * 1A2T - …… - (0.010) * 8V9V

- (0.023) * 1A3A - (0.011) * 1A3N + (0.020) * 1A3I - (0.003) * 1A3M + …… + (0.043) * 7V9V

LOO-CV

pIC_{50} *pred*

$res = pIC_{50}$ *exp* - pIC_{50} *pred*

Fig. 2 Data flow in the additive method.

only 5 times. The resolutions of the full model (amino acid contributions plus side chain–side chain interactions) by PLS gave low q^2 values for alleles A*0202, A*0203, A*0206 and A*6802. Analyzing the matrices a great number of columns containing only one 1 was found. Most of them were among the columns accounting for the adjacent or every second side-chain interactions. The presence of many unique interactions affects the predictivity of the models because in a "leave-one-out" cross-validation (LOO-CV) these interactions appear as missing values.

In order to reduce the number of missing values, matrices comprising only amino acid (AA) columns were generated and solved by PLS. The statistics of the new models are shown in Table 1. q^2 for the new models ranges from 0.317 to 0.500. The explained variance r^2 is above 90% except for the A*0201 model. Its r^2 is 0.731, 17% lower than the r^2 value of the model including the side-chain interactions (0.898).[18] Obviously, these interactions account for 17% of the explained variance in the set. The increased number of peptides in the studied set requires additional terms responsible for the variance. Therefore, the IBS-hypothesis can be usefully applied to small sets of peptides but not to larger sets.

The contributions of the amino acids at each position to the affinity of the different alleles are presented in Fig. 3. Amino acids with contributions above 0.2 were defined as preferred and these ones with contributions below −0.2 as deleterious. Residues identified as preferred for three or more A2-supertype molecules, without being deleterious for any molecule, may be

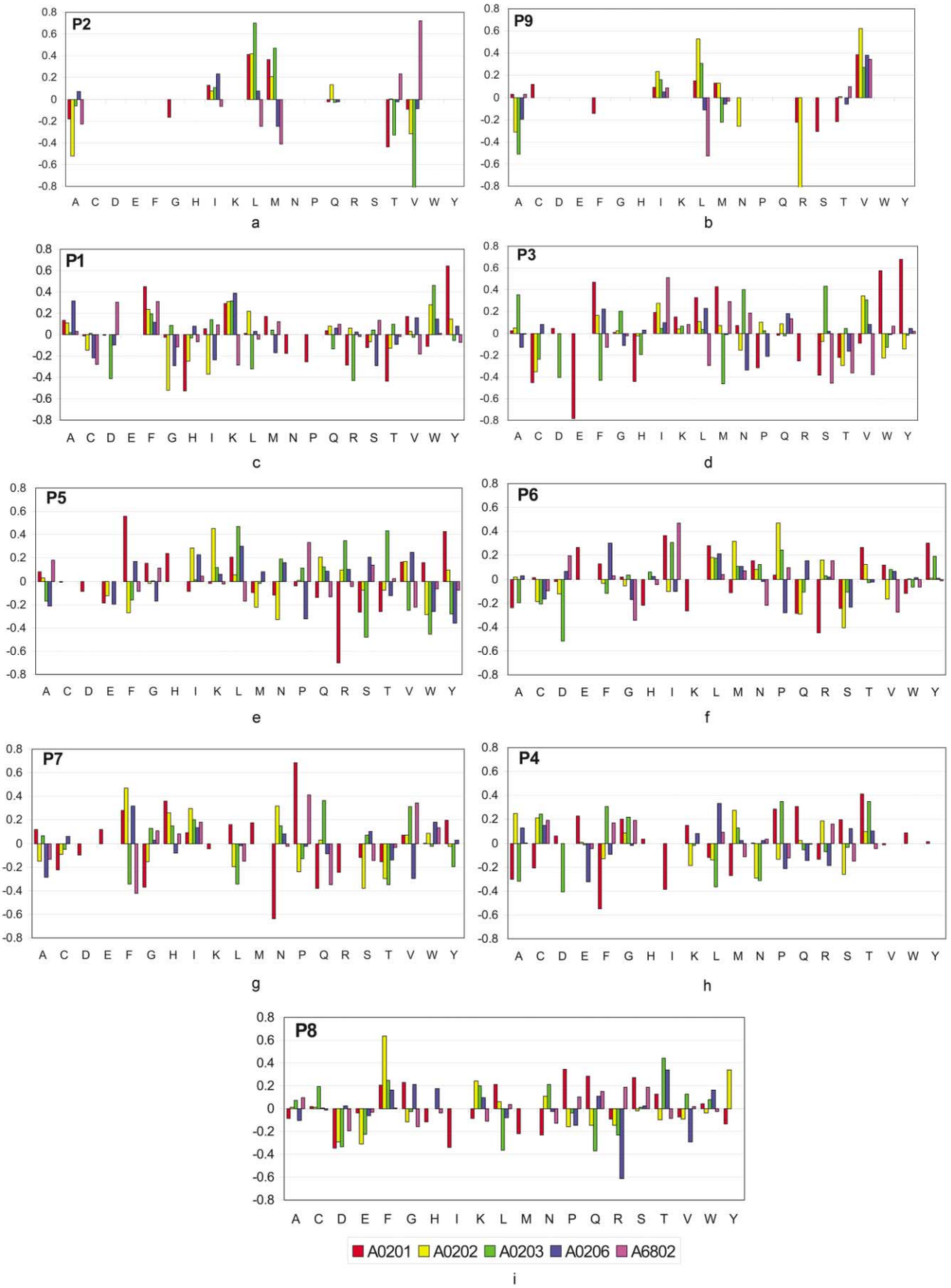


Fig. 3 Amino acid contributions (in $-\log IC_{50}$ units) to affinity to A2-supertype alleles at position 2 (a), 9 (b), 1 (c), 3 (d), 5 (e), 6 (f), 7 (g), 4 (h) and 8 (i). The constants in the regression models are 5.846 for A*0201, 6.169 for A*0202, 6.804 for A*0203, 6.568 for A*0206 and 6.290 for A*6802.

considered as preferred for the A2-supermotif.[17] Residues identified as deleterious for three or more molecules can be considered as deleterious in the common motif.[17] The common A2-supermotif is shown in Fig. 4a.

Primary anchor positions

Position 2 (P2) and C-terminal (P9) are considered, at least nominally, as primary anchor positions. The most striking

Table 2 Polymorphic binding pockets on HLA-A2 supertype. A*1110 allele is given for comparison

	Pocket A										Pocket B										
	5	7	59	63	66	99	159	163	167	171	7	9	24	25	34	45	63	66	67	70	99
A*0201	M	Y	Y	E	K	Y	Y	T	W	Y	Y	F	A	V	V	M	E	K	V	H	Y
A*0202	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
A*0203	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
A*0206	—	—	—	—	—	—	—	—	—	—	—	Y	—	—	—	—	—	—	—	—	—
A*6802	—	—	—	N	N	—	—	—	—	—	—	Y	—	—	—	—	N	N	—	Q	—
A*1110	—	—	—	N	N	—	—	—	—	—	—	Y	—	—	—	—	N	N	—	Q	—

	Pocket C									Pocket D						Pocket E							
	9	22	70	73	74	97	99	114	116	99	114	155	156	159	160	97	114	116	133	147	152	155	156
A*0201	F	F	H	T	H	R	Y	H	Y	Y	H	Q	L	Y	L	R	H	Y	W	W	V	Q	L
A*0202	—	—	—	—	—	—	—	—	—	—	—	—	W	—	—	—	—	—	—	—	—	—	W
A*0203	—	—	—	—	—	—	—	—	—	—	—	—	W	—	—	—	—	—	—	—	E	—	W
A*0206	Y	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
A*6802	Y	—	Q	—	D	—	—	—	—	—	—	—	W	—	—	—	—	—	—	—	—	—	W
A*1110	Y	—	Q	—	D	I	—	R	D	—	R	—	Q	—	—	I	R	D	—	—	A	—	Q

	Pocket F												
	73	77	80	81	84	95	116	118	123	124	143	146	147
A*0201	T	D	T	L	Y	V	Y	Y	Y	I	T	K	W
A*0202	—	—	—	—	—	L	—	—	—	—	—	—	—
A*0203	—	—	—	—	—	—	—	—	—	—	—	—	—
A*0206	—	—	—	—	—	—	—	—	—	—	—	—	—
A*6802	—	—	—	—	—	I	—	—	—	—	—	—	—
A*1110	—	—	—	—	—	I	D	—	—	—	—	—	—

a)

Preferred	F		I	G		I L	I	F	V
	1	2	3	4	5	6	7	8	9
Deleterious			T		W	S		D	A

b)

Preferred	F K	L	I V L	G T	I L	I L Y	H I	F K T	V L
	1	2	3	4	5	6	7	8	9
Deleterious		V T	C H T	A N	S W Y	Q S	L T	D E R	A T

Fig. 4 A2-supermotif: a) based on A*0201, A*0202, A*0203, A*0206 and A*6802 alleles; b) based on A*0201, A*0202, A*0203 and A*0206 alleles.

differences in the amino acid preferences in a common A2-supermotif are at P2. Hydrophobic aliphatic residues such as Leu, Met and Val have well known preferences for this position.[8–11,20] However, the results in the present study indicate that Leu and Met are preferred amino acids only for A*0201, A*0202 and A*0203. Leu is deleterious for A*6802 and Met is deleterious for A*0206 and A*6802 (Fig. 3a). Val and Thr are preferred for A*6802. Sidney *et al.*[17] also reveal similar differences in P2 specificities although not so strong as in the present study. Comparing the residues forming the pocket B in the different alleles[26] four differences are evident (Table 2). Three of them (Glu63, Lys66 and His70) are disposed at the pocket rim and one (Phe9) at the inner wall. The Phe9 ⟶ Tyr9 substitution makes the pocket shallower and long side chains, such as Leu and Met, are no longer accommodated here. Molecular modeling studies hypothesize a possible conformational shift of the aromatic ring of Tyr9 into the cavity.[27] This conformational change would narrow the size of the B pocket and weaken the entirely hydrophobic state of this pocket. The preferred Val and Thr for A*6802 allele, being deleterious or negative for the other A2-sypertype molecules, denote another point of discrepancy between A*6802 and the remaining A2 alleles. Comparison of the residues forming pocket B shows identity between A*6802 and A*1110, A*2502, A*2613, A*6604, A*6601, A*6602, A*3403, A*3404, A*3402 alleles. None of them except for A*6802 was classified as an A2-like allele. A*1110 is shown in Table 2 for comparison.

At the C-terminal there is a good agreement between the preferences of different alleles. Val is the favored amino acid at this position, Ala is deleterious (Fig. 3b). Pocket F appears to be the most conserved pocket in the HLA binding cleft.[11] The side chain of Tyr116 occupies the end of the pocket F and is uncharged, so that the binding site is complementary to small hydrophobic side chains.[7,9]

Secondary anchor positions

Positions 1 (P1), 3 (P3), 5 (P5), 6 (P6) and 7 (P7) are secondary anchor positions.[10,11] Phe is the only one preferred amino acid for P1 in the common motif (Fig. 3c). Lys is preferred for all alleles except for A*6802. For the last allele Lys is apparently deleterious. The main differences in the amino acid sequences forming this pocket are residues 63 and 66 (Table 2). Glu63 and Lys66 are substituted for Asn63 and Asn66 in A*6802 allele.[26] Obviously, the negatively charged Glu63 favored the presence of positively charged Lys at P1, while the neutral Asn63 is not electrostatically complementary to Lys.

Ile is the only one preferred amino acid at P3 and Thr is the common deleterious one (Fig. 3d). Leu and Val are preferred for A*02 alleles but is deleterious for A*6802. P3 side chains of bound peptides fall into pocket D which is a hydrophobic cavity.[28] There is only one difference in the sequences forming this pocket (Table 2). Leu156 in A*0201 and A*0206 is substituted for Trp in A*0202, A*0203 and A*6802 making a bulky ridge across the center of the cleft.

Leu was found to be a preferred residue at P5 for affinity to A2-supertype molecules except for A*6802 where it is negative but not deleterious (Fig. 3e). Trp is deleterious for three of the five MHC molecules.

Ile and Leu are preferred at P6 and Ser is deleterious (Fig. 3f). The side chain of P6 falls into pocket C. The most dramatic difference between A*6802 and A*02 alleles concerns this pocket. A deep negatively charged pocket at A*6802 is formed by the substitution of Asp for His at position 74 and Gln for His at position 70 (Table 2). This pocket seems suited to bind polar atoms, especially a positively charged side-chains or N-terminus (Lys).[29] Unfortunately, we could not find any published peptide, with Lys at P6, tested for affinity to A*6802.

For affinity to A2-supertype molecules Ile is preferred at P7 (Fig. 3g). The side chain at P7 falls into pocket E (Table 2). Two-thirds of the surface area in this pocket is hydrophobic, but Arg97 provides a large polar patch on one side of the pocket.[7] Pocket E can accommodate a variety of complementary peptide side chains, but an incompatible side chain need not prevent complex formation.[11]

"Flag" positions

Positions 4 and 8 are solvent-exposed and may form contacts with the TCR.[11] Gly is preferred here (Fig. 3h). Thr is preferred or positive for the A2-supertype alleles except for A*6802. Phe is a preferred common residue at P8 and Asp is deleterious for four of the five A2 molecules (Fig. 3i).

Discussion

Amino acid contributions to the affinity of peptides binding to five HLA alleles — A*0201, A*0202, A*0203, A*0206 and A*6802 — were analysed quantitatively using PLS. Amino acids, identified as preferred for three or more HLA molecules without being deleterious for any of the rest, were considered as preferred for the A2-supermotif (Fig. 4a). Amino acids, identified as deleterious for three or more molecules, were considered as deleterious for the supermotif.

Certain discrepancies between A*6802 and A*02 molecules concerning the amino acid preferences at P1–P9 were seen in the present study. These discrepancies throw doubt on whether the A*6802 allele belongs to the A2-supertype. The sequence comparison showed that there are only one or two differences in the residues forming the 6 pockets of A*0201, A*0202, A*0203 and A*0206 molecules (Table 2). The number of these differences between A*6802 and A*02 molecules is seven residues. Five of them concern pockets A, B and C and are so substantial that they alter the amino acid preferences at the primary anchor P2 and the secondary anchors P1 and P6. The preferred Val and Thr for P2 brings the A*6802 allele closer to the A3-supertype [30] rather than to the A2-one. But the A3 supermotif requires positively charged residues, such as Arg and Lys, at the C-terminus,[30] which is not true in the case of the A*6802 allele. Obviously, A*6802 is an intermediate allele standing between A2 and A3 supertypes: in anchor position 2 it is closer to A3 and in anchor position 9 it is nearer to A2.

Excluding A*6802 allele, the redefinition of the preferred and deleterious amino acids expands the A2-supermotif (Fig. 4b). Residues identified as preferred for two or more A*02 molecules, without being deleterious for any molecule, are considered as preferred. Residues identified as deleterious for two or more molecules are considered as deleterious in the common motif. The expansion concerns all positions and especially the anchor P2. One to three new amino acids are added to each position's preferred and deleterious amino acids.

Finally, we searched our database (http://www.jenner.ac.uk/Jenpep) using different combinations of the supermotif from Fig. 4a. We found only one peptide containing the maximum of five preferred amino acids at the proper positions. It was the decamer FLIFFDLFLV which is a good binder to A*02 alleles [IC_{50}(A*0201) = 12 nM, IC_{50}(A*0202) = 10 nM, IC_{50}(A*0203) = 5.9 nM, IC_{50}(A*0206) = 11 nM] and an intermediate binder to A*6802 [IC_{50}(A*6802) = 333 nM].[31] This peptide did not take part in our training set because it is a decamer. It contains only 5 of the preferred 9 amino acids and suggests the hopeful prospect that the best binder is still not found. Our preliminary experimental validation in this regard gives promising results.

The redefinition of the HLA-A2 supermotif presented in this paper both expands and strengthens the set of preferred and deleterious amino acids at each position of the binding peptide. However, the models developed for each allele also give quantitative predictions for the affinity to each allele. This can prove useful, particularly in the search for promiscuous heteroclitic peptides, where use of our models can accurately predict increases in peptide binding to a range of MHC alleles. Internet access to these models is available at http://www.jenner.ac.uk/MHCPred.[32] Expansion of the method to other alleles is progressing. Leveraged by the development of methods such as these, we would hope that computational immunovaccinology will have a similar effect on the fight against global disease, as mediated through the development of new vaccines, as similar informatics strategies have had on the discovery of novel small-molecule therapeutics.

Computational details

Multiple linear regression by partial least squares

Partial least squares (PLS) belongs to so called projection methods, which can handle data matrices with more variables than observations very well, and the data can be both highly collinear and noisy. In this situation, conventional statistical methods, such as multiple regression, or artificial intelligence techniques, such as artificial neural networks, tend to produce a formula that fits the training data well but is very unreliable for prediction. PLS forms new x variables, called principal components, as linear combinations of the old ones, and then uses them as predictors of biological activity.[33]

We used the PLS method implemented in the QSAR module of SYBYL 6.7.[34] pIC_{50} was put as a dependent variable. The scaling method was set to "none". The column filtering was switched off. The optimal number of components (NC) was

found by cross-validation using SAMPLS.[35] The non-cross-validated models were assessed by the MLR parameters as explained variance r^2, standard error of estimate (SEE), and F ratio. A cross-validation using the "leave-one-out" procedure assessed the predictive power of the models.

Cross-validation using the "leave-one-out" procedure

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.[33] In principle, CV is performed by splitting the data into a number of groups, developing a series of parallel models from the reduced data with one of the groups omitted, and then predicting the activities of the excluded compounds. When the number of 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 coefficients q^2_{LOO} (LOO-CV) and q^2_{CV5} (CV in 5 groups), the standard error of prediction (SEP), and the residuals between the experimental ($p\mathrm{IC}_{\mathbf{50exp}}$) and predicted binding affinity ($p\mathrm{IC}_{\mathbf{50pred}}$):

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (p\mathrm{IC}_{50\,\mathrm{exp}} - p\mathrm{IC}_{50\,\mathrm{pred}})^2}{\sum_{i=1}^{n} (p\mathrm{IC}_{50\,\mathrm{exp}} - p\mathrm{IC}_{50\,\mathrm{mean}})^2}$$

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n} (p\mathrm{IC}_{50\,\mathrm{exp}} - p\mathrm{IC}_{50\,\mathrm{pred}})^2}{p-1}}$$

$$\mathrm{residual} = p\mathrm{IC}_{50\,\mathrm{exp}} - p\mathrm{IC}_{50\,\mathrm{pred}}$$

where p is the number of peptides omitted, $p\mathrm{IC}_{\mathbf{50pred}}$ is that predicted by the CV-LOO value. The residuals between the experimental and predicted $p\mathrm{IC}_{50}$ values were classified into 3 categories: below |0.5|, from |0.5| to |1.0| and above |1.0|. A mean |residual| and its standard deviation were extracted as well.

References

1 R. N. Germain, *Cell*, 1994, **76**, 287.
2 P. D. Greenberg, *Adv. Immunol.*, 1991, **49**, 281.
3 D. Kagi and H. Hengartner, *Curr. Opin. Immunol.*, 1996, **8**, 472.
4 S. C. Threlkeld, P. A. Wentworth, S. A. Kalams, B. M. Wilkers, D. J. Ruhl, E. Keogh, J. Sidney, S. Southwood, B. C. Walker and A. Sette, *J. Immunol.*, 1997, **159**, 1648.
5 A. Sette and J. Sidney, *Curr. Opin. Immunol.*, 1998, **10**, 478.
6 B. D. Livingston, C. Crimi, J. Fikes, R. W. Chesnut, J. Sidney and A. Sette, *Hum. Immunol.*, 1999, **60**, 1013.
7 M. A. Saper, P. J. Bjorkman and D. C. Wiley, *J. Mol. Biol.*, 1991, **219**, 277.
8 K. Falk, O. Rötzschke, S. Stefanovic, G. Jung and H.-G. Rammensee, *Nature*, 1991, **351**, 290.
9 D. R. Madden, D. N. Garboczi and D. C. Wiley, *Cell*, 1993, **75**, 693.
10 J. Ruppert, J. Sidney, E. Celis, R. T. Kubo, H. M. Grey and A. Sette, *Cell*, 1993, **74**, 929.
11 D. R. Madden, *Annu. Rev. Immunol.*, 1995, **13**, 587.
12 S. Tangri, G. Y. Ishioka, X. Huang, J. Sidney, S. Southwood, J. Fikes and A. Sette, *J. Exp. Med.*, 2001, **194**, 833.
13 J. Sidney, H. M. Grey, R. T. Kubo and A. Sette, *Immunol. Today*, 1996, **17**, 261.
14 M.-F. del Guercio, J. Sidney, G. Hermanson, C. Perez, H. M. Grey, R. T. Kubo and A. Sette, *J. Immunol.*, 1995, **154**, 685.
15 J. Sidney, M. F. del Guercio, S. Southwood, V. H. Engelhard, E. Appella, H. G. Rammensee, K. Falk, O. Rotzschke, M. Takiguchi and R. T. Kubo, *J. Immunol.*, 1995, **154**, 247.
16 H.-G. Rammensee, T. Friede and S. Stevanovic, *Immunogenetics*, 1995, **41**, 178.
17 J. Sidney, S. Southwood, D. L. Mann, M. A. Fernandez-Vina, M. J. Newman and A. Sette, *Hum. Immunol.*, 2001, **62**, 1200.
18 I. A. Doytchinova, M. J. Blythe and D. R. Flower, *J. Proteome Res.*, 2002, **1**, 263.
19 S. M. Free and J. W. Wilson, *J. Med. Chem.*, 1964, **7**, 395.
20 K. C. Parker, M. A. Bednarek and J. E. Coligan, *J. Immunol.*, 1994, **152**, 163.
21 K. C. Parker, M. Shields, M. DiBrino, A. Brooks and J. E. Coligan, *Immunol. Res.*, 1995, **14**, 34.
22 P. Guan, I. A. Doytchinova and D. R. Flower, *Protein Eng.*, 2003, **16**, 11.
23 M. J. Blythe, I. A. Doytchinova and D. R. Flower, *Bioinformatics*, 2002, **18**, 434.
24 A. Sette, J. Sidney, M.-F. del Guercio, S. Southwood, J. Ruppert, C. Dalberg, H. M. Grey and R. T. Kubo, *Mol. Immunol.*, 1994, **31**, 813.
25 T. I. Oprea and C. L. Waller, in *Reviews in Computational Chemistry*, eds. K. B. Lipkowitz and D. B. Boyd, Wiley-VCH, New York, 1997, vol. 11, p. 127.
26 C. Schönbach, J. L. Y. Koh, X. Sheng, L. Wong and V. Brusic, *Nucleic Acids Res.*, 2000, **28**, 222.
27 T. Sudo, N. Kamikawaji, A. Kimura, Y. Date, C. J. Savoie, H. Nakashima, E. Furuichi, S. Kuhara and T. Sasazuki, *J. Immunol.*, 1995, **155**, 4749.
28 P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger and D. C. Wiley, *Nature*, 1987, **329**, 506.
29 T. P. J. Garrett, M. A. Saper, P. J. Bjorkman, J. L. Strominger and D. C. Wiley, *Nature*, 1989, **342**, 692.
30 J. Sidney, H. M. Grey, S. Southwood, E. Celis, P. A. Wentworth, M.-F. del Guercio, R. T. Kubo, R. W. Chesnut and A. Sette, *Hum. Immunol.*, 1996, **45**, 79.
31 D. L. Doolan, S. L. Hoffman, S. Southwood, P. A. Wentworth, J. Sidney, R. W. Chesnut, E. Keogh, E. Appella, T. B. Nutman, A. A. Lal, D. M. Gordon, A. Oloo and A. Sette, *Immunity*, 1997, **7**, 97.
32 P. Guan, I. A. Doytchinova, C. Zygouri and D. R. Flower, *Nucl. Acids Res.*, 2003, **31**, in press.
33 S. Wold, in *Chemometric Methods in Molecular Design*, ed. H. van de Waterbeemd, VCH, Weinheim, 1995, p. 195.
34 SYBYL 6.7, 2002 Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
35 B. L. Bush and R. B. Nachbar Jr., *J. Comput-Aid. Mol. Des.*, 1993, **7**, 587.