ARTICLE

OBC
www.rsc.org/obc

New horizons in mouse immunoinformatics: reliable *in silico* prediction of mouse class I histocompatibility major complex peptide binding affinity†

Channa K. Hattotuwagama,* Pingping Guan, Irini A. Doytchinova and Darren R. Flower
Edward Jenner Institute for Vaccine Research, Compton, Berkshire, UK RG20 7NN.
E-mail: channa.hattotuwagama@jenner.ac.uk; Fax: +44 (0) 1635 577901 / 577908;
Tel: +44 (0) 1635 577954

Received 25th June 2004, Accepted 11th August 2004
First published as an Advance Article on the web 16th September 2004

Quantitative structure–activity relationship (QSAR) analysis is a main cornerstone of modern informatic disciplines. Predictive computational models, based on QSAR technology, of peptide-major histocompatibility complex (MHC) binding affinity have now become a vital component of modern day computational immunovaccinology. Historically, such approaches have been built around semi-qualitative, classification methods, but these are now giving way to quantitative regression methods. The additive method, an established immunoinformatics technique for the quantitative prediction of peptide–protein affinity, was used here to identify the sequence dependence of peptide binding specificity for three mouse class I MHC alleles: H2–D^b, H2–K^b and H2–K^k. As we show, in terms of reliability the resulting models represent a significant advance on existing methods. They can be used for the accurate prediction of T-cell epitopes and are freely available online (http://www.jenner.ac.uk/MHCPred).

Introduction

Quantitative structure–activity relationship (QSAR) analysis, as a predictive tool of wide applicability, is one of the main cornerstones of modern cheminformatics and increasingly, bioinformatics. Immunoinformatics, a newly emergent sub-discipline of bioinformatics which addresses informatic problems within immunology, uses QSAR technology to tackle the crucial issue of epitope prediction. As high throughput biology reveals the genomic sequences of pathogenic bacteria, viruses, and parasites, such prediction will become increasingly important in the post-genomic discovery of novel vaccines, reagents, and diagnostics.

In the unending war between host and pathogen, the adaptive immune system has been the primary vertebrate defence for 500 million years. At the heart of cellular adaptive immunity is a set of molecular recognition events: premier amongst them is the cell surface recognition of peptide-bound major histocompatibility complexes (MHC) by T-cells. The T-cell is a specialised type of immune cell that mediates cellular immunity. T-Cells contribute to immune defences in two main ways: regulating the complex workings of the immune system and, more directly, by eliminating infected or malignant cells. The short antigenic peptides recognised by T-cells are a form of epitope: in this case, markers of foreign or host proteins. The biological role of MHC proteins is thus to bind small peptides derived from both pathogen and host protein and to "present" these for inspection by T-cells. T-Cells recognise peptide-MHC (pMHC) complexes *via* a special form of receptor: the T-cell receptor (TCR). Class I MHC molecules present endogenously synthesised antigens, including host and viral proteins, inducing a cytotoxic T-cell response. Class II MHC molecules present exogenously derived proteins, *e.g.* bacterial products or viral capsid proteins. MHC class I and II are distinct at the level of sequence and structure. This is also reflected in the geometry of their peptide-binding grooves and their peptide selectivities. The binding site of class I MHCs accommodates 8–11 amino acid peptides while the open-ended class II sites allows binding of much longer peptides, some in excess of 20 amino acids. The cell biology and expression pattern of each class of MHC is tailored to meet its distinct role. MHC class I molecules bind peptides in the endoplasmic reticulum (ER), which are generated continuously in the cytoplasm through protein degradation, mainly by the proteasome. Peptides of ~8–18 amino acids are specifically transported across the ER membrane by a heterodimeric transporter, known as transporter associated with antigen processing or TAP, where they then bind to class I MHC molecules.

The ability to predict the recognition of epitopes accurately is a principal goal of modern *in silico* immunology. Within the human population there are a vast number of different variant genes, or alleles, coding for class I and class II MHC proteins. Each allele exhibits different peptide selectivity: peptides are bound which have particular sequence patterns and with an affinity dependent on those sequence patterns. Typically, human alleles bind nonameric peptide sequences. Peptide selectivities of class I MHCs are most often rationalised in terms of a characteristic motif with a preference for particular amino acids at two or more positions. Such motifs have enjoyed a wide popularity within immunology, as they are both easy to use and easy to understand. Motifs characterise a short peptide in terms of dominant anchor positions with a strong preference for certain amino acids. Sette and co-workers[1,2] defined the first allele-dependent sequence motifs using the mouse alleles I–E^d and I–A^d. There are fundamental problems with motifs, however, as they produce significant numbers of both false positives and false negatives, and are overly reliant on the choice of anchors. Subsequently, much more sophisticated methods have arisen.[3] These include many using artificial intelligence techniques, such as artificial neural networks,[4–7] hidden Markov models,[8,9] support vector machines,[10,11] and profiles.[12]

For understandable reasons—the desire to generate new vaccines and diagnostics, for example—much work has hitherto focussed on human alleles. The mouse—the primary experimental animal in immunology—has received some attention, but not as much as its pre-eminent position as an instrument of immunological investigation might warrant. The H2 genes are part of the mouse MHC and forms a multi-gene cluster containing three major gene classes: class I located in the H2–D, H2–K (as discussed here), Qa and H2–T18 regions and class II located in the H2–I region and class III in the H2–S region.[13] MHC class I gene products of the H2–D and H2–K regions are

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b409656h

Table 1 List of peptides used in this study of the H2–D^b mouse allele

No.	Epitope	Exp. logIC_{50}	Pred. logIC_{50}	Ref.	No.	Epitope	Exp. logIC_{50}	Pred. logIC_{50}	Ref.
1	AAAENAEAA	7.357	7.380	26	34	RSVINIVII	5.854	5.980	20
2[a]	***AEDTNVSLI***	***3.357***	***5.732***	24	35	SAIENLEYM	7.721	7.810	26
3	AENENMRTM	5.712	6.400	20	36	SEVSNVQRI	5.797	5.710	20
4	AMIENLEYM	7.620	7.990	26	37	SFYRNLLWL	6.542	6.690	22
5	ASNENIDTM	8.699	8.200	21	38	SGVENPGGY	4.881	4.980	25
6	ASNENMETM	7.750	7.960	20	39	SLLGNATAL	6.796	6.930	24
7	ASNENMRTM	8.155	7.480	20	40	SLLYNLDLM	8.097	7.850	20
8	CDFNNGITI	5.344	5.250	20	41	SMAENLEYM	7.222	7.080	26
9	CKGVNKEYL	7.409	7.130	20	42	SMIANLEYM	6.848	6.990	26
10	FAPGNYPAL	8.091	7.900	20	43	SMIEALEYM	6.796	6.950	26
11	FCGVNSDTV	6.799	6.740	22	44	SMIENAEYM	7.523	7.420	26
12	FQLCNSYDL	7.886	8.030	24	45	SMIENLAYM	6.780	6.950	26
13	FQPQNGQFI	8.067	8.210	20	46	SMIENLEAM	7.699	7.450	26
14	FRGPNVVTL	5.925	5.700	20	47	SMIENLEYA	7.538	7.470	26
15[a]	***GFKSNFNKI***	***3.357***	***6.303***	24	48	SMIENLEYM	7.871	7.570	26
16	IISHNFCNL	6.027	5.990	20	49	SSVIGVWYL	5.854	5.910	23
17[a]	***IKPSNSEDL***	***5.538***	***7.699***	20	50	SSVVGVWYL	6.268	6.390	23
18	ISANNDSEI	6.056	6.190	24	51	SSVVNVWYL	7.244	7.220	23
19	ISNGNSDCL	6.503	6.990	24	52	TAGANPMDL	4.658	4.840	24
20	ISVSNPGDL	6.658	6.250	24	**53**[a]	***TALANTIEV***	***8.444***	***5.747***	22
21	ITYKNSTWV	6.570	6.340	22	54	TGICNQNII	7.699	7.540	22
22	KAVYNFATC	6.484	6.440	27	**55**[a]	***TGKLNLENL***	***4.754***	***7.097***	24
23	KICQNFILL	5.606	5.730	24	56	VENPGGYCL	4.475	3.940	25
24	LIDYNKAAL	5.714	5.960	20	57	VKYPNLNDL	5.878	6.090	20
25	LLVFNYPGI	5.287	5.270	24	58	VLSFNLGDM	4.202	4.570	24
26	LTFTNDSII	5.835	5.780	27	59	VLSTNGDTL	6.370	6.590	24
27	LTFTNDSSI	5.824	5.760	20	60	WLVTNGSYL	6.911	7.100	20
28	NGLWNLDVI	8.000	8.080	20	61	YAIENAEAL	7.658	7.610	26
29	QAPTNRWML	8.252	8.530	24	62	YAIENAKAL	6.959	7.060	26
30	QGINNLDNL	7.824	7.890	20	63	YAIKNAEAL	7.678	7.610	26
31[a]	***QLPPNSLLI***	***3.533***	***6.193***	24	64	YASDNQAIL	6.319	6.300	24
32	RGVINIVII	5.692	5.590	20	65	YSQGNSGLM	6.051	5.930	24
33	RLIQNSLTI	6.967	6.610	22					

[a] Peptides highlighted in bold and italics indicate where peptide has been removed (outlier) during calculation.

found on most cells except in very early embryos and function in cytolytic immune responses. Allogenic differences at these loci induce vigorous graft rejection and strong primary *in vitro* cytotoxic responses.

Although crystallographic analysis confirms the high overall similarity of human and mouse MHC structures, there are, nonetheless, clear differences in their peptide specificities: for example, experimental analysis of eluted mouse peptides indicates a preference for both nonameric (nine amino acid) and octameric (eight amino acid) peptides. The mouse class I peptide binding is formed by the α-1 and α-2 domains of the alpha chain. Eight anti-parallel β-strands form the floor of the cleft while its sides are formed by two α-helices. The cleft is closed at each end: bound peptides are anchored at each end and bow in the middle. Crystallography shows that peptide amino acid side chains are accommodated by "pockets" within the binding groove of class I MHC molecules. Primary and secondary anchor residues are buried in a number of complementary pockets, which are designated A–F.[14] T-Cell receptor (TCR) and TCR-pMHC structures show that the TCR alpha and beta chain variable regions form an immunoglobulin-type combining site with residues from complementary regions 1, 2, 3 making contact with α-1 and α-2 domains of the MHC, as well as with exposed amino acid side chains of the bound peptide.

We have recently developed an immunoinformatic technique for the prediction of peptide–MHC affinities, known as the Additive method, which is based on the Free-Wilson principle,[15] whereby the presence or absence of groups is correlated with biological activity. For a peptide, the binding affinity is thus represented as the sum of amino acid contributions at each position. We have extended the classical Free-Wilson model with terms, which account for interactions between amino acids side chains. Using literature data, we applied the additive method to peptides binding to several human class I,[16–18] and class II alleles.[19] In order to better understand the sequence dependence of peptide–MHC binding of the mouse MHC, we have now used our approach to explore the amino acid preferences of three mouse alleles: H2–D^b (nonamers), H2–K^b (octamers) and H2–K^k (octamers). This paper exemplifies the first use of the additive method for octameric, as well as nonameric, peptides, and, as we show, these models represent, in terms of reliability, significant improvements over existing methods.

Results

For the H2–D^b model, 65 nonameric peptides were used as the initial training set (Table 1).[20–27] six peptides with residual values ≥2.0 log units were omitted, reducing the dataset to 59 peptides. Based on the peptide sequences used here, the distribution of amino acids at the nine peptide positions is shown in Table 2. Previous analysis of H2–D^b binding indicates a strong preference for Asn at position 5 and Ile, Leu and Met residues at position 9. For the H2–D^b model, the LOO-CV parameters are $q^2 = 0.401$, SEP = 0.840 and NC = 4, while the non-cross validation parameters are $r^2 = 0.946$ and SEE = 0.252 (Table 3). The quantitative contributions of amino acids at each position for this model are shown in Fig. 1 (black bars). An extended motif, as defined by this model, is summarised in Table 4. The results show that the peptide sequences in the H2–D^b allele have the same anchor positions (Asn at position 5 and Met, Ile or Leu at the C terminus (position 9)) as seen previously.[20–23] This was expected because of the limited number of amino acids at these positions. Inspection of Fig. 1 allows us to determine the influence on binding affinity to H2–D^b of certain amino acids at each non-anchor positions. For example, hydrophobic residues such as Phe, Ile, Leu, Val and Pro were found at strong binding positions 1, 3 and 8. Amino acids Ser, Thr and Cys were the only residues with hydroxyl or sulfhydryl containing side chains found at strong binding positions (position 4). It is interesting to note that certain amino acids, such as positively charged (His and Lys), neutral (Asn) or small (Ala) residues, exhibited low occupancy of non-anchor position.

Table 2 Amino acid abundance for the H2–D^b, H2–K^b and H2–K^k alleles

	H2–D^b									H2–K^b								H2–K^k							
	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
A	6	12	3	3	1	6	4	9	2	5		1	1	1	3	8		2	4	1	1	2	1	1	1
R	3	1		1		1	2	1		6		1	4		4	7		1	1	1	1	1	1	1	1
N	1		6	3	61		3	3		3	3		9		1	6		2	1	1	1	2	127	1	19
D		1	1	1		3	7	6		1	3		1		4			2	1	1	1	1	1	2	1
C	2	1	1	2			1	2	1	1			1		4									1	
Q	3	2	1	3		2	2			2	2	4	5		7	3		2	1	1	1	1	1	1	1
E		3		17			16	2			1	2	3		3			1	130	1	1	1	2	2	1
G	1	6	4	5	3	5	4	2		2	5	1	2	1	2	9		1	2	1	2	130	3	1	1
H				1						3	1	2	2	1		2		2	1	2	2	1	1	1	1
I	6	3	15	3		3	3	5	14	6	8	13	4	1	3	5	5	1	1	3	1	1	1	2	113
L	4	8	5	1		16	3	4	28	8	4	4	7	2	8	6	45	2	2	2	2	1	2	130	2
K	2	3	2	2		2	1	1		2	1		3		3	4		1	1	2	2	1	3	1	1
M		9				3	1	1	16	6	2		1				5	1	1	1	1	1	2	1	1
F	6	2	3	2		4		1		3	1	4	4	30	1	1		130	1	4	1	2	2	1	1
P			6	4		3	3				4	1	2		7	3		1	1	1	3	1	1	2	1
S	17	11	4	4		8	4	1		7	14	6	5	1	6	4		1	1	128	2	3	1	1	1
T	4	3		5		1	2	10		1	4	4		2	4			1	2	1	127	1	1	1	1
W	5		3	3		3	2	14	1	2	2	15	1	21		1		1	1	1	1	1	1	1	1
Y	1			1			4	2			1		3	1		1		1	1	1	2	1	1	1	1
V	4		11	4		5	3	1	3	4	6	4	4	1	2	2	7	1	1	1	2	2	2	2	5

Table 3 QSAR statistics of the additive models for three class I alleles

Allele		H2–D^b	H2–K^b	H2–K^k
No. of peptides in original dataset		65	62	154
No. of peptides removed (outliers)		6	7	2
NC [a]		4	6	6
Cross-validation leave- one-out	q^{2b}	0.401	0.454	0.456
	SEP[c]	0.840	0.894	0.565
Non-cross validation	r^2	0.946	0.989	0.933
	SEE[d]	0.252	0.128	0.198

[a] Optimal number of components. [b] q^2 obtained after LOO-CV. [c] Standard error of prediction. [d] Standard estimate of error.

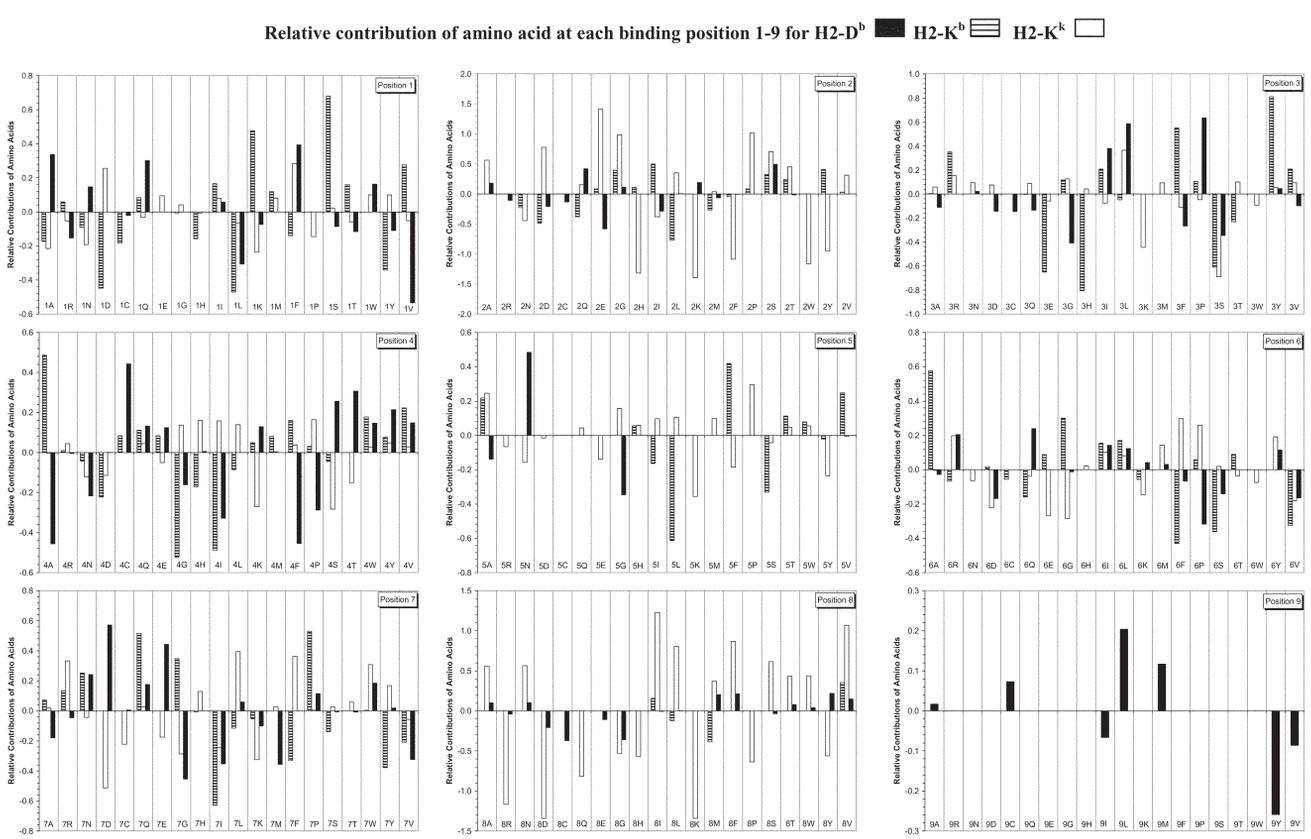

Fig. 1 Relative contributions of position-wise amino acids at each binding positions 1–9 for the H2–D^b, H2–K^b and H2–K^k alleles. The contribution made by different individual amino acids at each position of the 9mer H2–D^b, H2–K^b and H2–K^k binding peptide. The contribution is equivalent to a position-wise amino acid regression coefficient obtained by PLS regression (as described in the text).

Table 4 Non-anchor residues related with strong and weak binding for amino acids only

		P1	P2	P3	P4	P5	P6	P7	P8	P9
Favoured binding	**H2–D^b**	A, Q, F	Q, S	I, L, P	C, T	N	*Q*	D, E	*Y*	*L*
	H2–K^b	K, S	G, I, S, Y	R, F, Y	A	F	A, G	Q, G, P	V	
	H2–K^k	*F*	A, D, E, G, L, P, S, T, V	L	*P*	*P*	*F*	R, L, F, W	A, N, I, L, M, F, S, T, W, V	
	H2–K^k	*A*	N, H, I, K, F, W, Y	K, S	*S*	K	*G*	D, K	R, D, Q, G, H, K, P, Y	
Disfavoured binding	**H2–K^b**	D, L, Y	D, Q, L	E, H, S	G, I	L, S	F, S, V	I, F, Y	M	
	H2–D^b	L, V	E	G, S	A, I, F	G	P	G, I, M, V	C, G	*Y*

[a] A cut-off value of >±0.3 is applied to favoured and disfavoured binding amino acids as shown in Fig. 1. [b] Where no amino acid residue lies outside the cut-off limit (>±0.3), the next best residue is chosen (as shown in italics).

For the H2–K^b model, 62 octameric peptides were used for the initial training set (Table 5).[22,24,27–30] 7 peptides with residual values ≥2.0 log units were omitted, reducing the dataset to 55 peptides. Based on the peptide sequences used here, the distribution of amino acids at the eight peptide positions is shown in Table 2. LOO-CV parameters of the model are $q^2 = 0.454$, SEP = 0.894 and NC = 6, while the non-cross validated parameters are $r^2 = 0.989$ and SEE = 0.128 (Table 3). The amino acids contributions at each position according to this model are shown in Fig. 1 (stripy bars). H2–K^b binding peptides are usually octamers with major MHC anchor binding positions at positions 5 and 8.[20–23] An extended motif, as defined by this model, is summarised in Table 4. From these results it is clear that the highest positive contributions at the anchor positions belong to Phe at position 5 and Val at the C-terminus (position 8). Both residues are hydrophobic. It is clear that hydrophobic amino acids are also concentrated at non-anchor positions: Ile, Gly, Tyr (position 2); Tyr, Phe (position 3) and Ala (position 4 and 6). Two polar amino acid residues (Ser and Lys) occupy the strong binding position 1, which interacts with a network of hydrogen bonding side chains directly involved in binding the N-terminus of the peptide.

For the H2–K^k model, 154 octameric peptides were used as the initial training set (Table 6).[31–33] two peptides with residual values ≥2.0 log units were omitted, reducing the dataset to 152 peptides. The distribution of amino acids at each of the eight peptide positions is shown in Table 2. The LOO-CV parameters for this model are $q^2 = 0.456$, SEP = 0.565 and NC = 6, while the non-cross validated parameters are $r^2 = 0.933$ and SEE = 0.198 (Table 3). The quantitative contributions at each position are shown in Fig. 1 (white bars). The H2–K^k binding peptides were octameric but unlike the H2–K^b allele, its anchor positions are at P2 and P8.[34–37] An extended motif, as defined by this model, is summarised in Table 4. For H2–K^k, the major amino acid residues at the anchor positions were Ala, Asp, Glu, Gly, Leu, Pro, Ser, Thr and Val (position 2) and Ala, Asn, Ile, Leu, Met, Phe, Ser, Thr, Trp and Val (position 8). There is a resemblance to the H2–K^b model in that many small polar or charged amino acids are associated negatively with pocket positions 1, 2, 3, 4, 6 and 7. Similarly, there is an abundance of strong hydrophobic interactions in the same pockets. Lys seems to be a common weak binding amino acid in all binding pockets. It is somewhat amphipathic: most of the side chain is long and hydrophobic, whereas its terminal functional group is positively charged, which is why most of the side chain is buried and only the charged part is exposed.

As can be seen from Tables 1, 5 and 6, which show both experimental and predicted affinities, most of the outliers are found at the ends of the IC_{50} value distribution, where fewer observations are available. For example, for the mouse H2–D^b system, five out of 11 observations at the lower end of the distribution are outliers. This may indicate chance effects, or a lack of model reliability, or deviations from linearity in this region, or the existence of separate binding modes at different affinities, or, indeed, several other possibilities. While we are wary of dismissing outliers out of hand, as they sometimes provide valuable insights, currently the quality and quantity of data available precludes us from distinguishing between these alternative explanations.

We further exemplify the strength of our models through use of our online prediction algorithm MHCPred,[38,39] which implements the additive method, comparing it with other internet-enabled prediction methods: RANKEP,[40] BIMAS,[41] and SYFPEITHI.[42] A data set of 20 new class I mouse H2–K^b and H2–D^b epitopes, not used to train these models, were collected from the literature.[43–54] The corresponding protein sequences, from which the epitopes were identified, were retrieved from either SWISS-PROT,[55] or Genbank,[56] and used as the input to the prediction algorithm. Algorithms used by the servers' vary.[3] SYFPEITHI[42] uses peptide binding motifs for both class I and II MHC alleles available in the literature

Table 5 List of peptides used in the study of the H2–K^b mouse allele

No.	Epitope	Exp. logIC_{50}	Pred. logIC_{50}	Ref.	No.	Epitope	Exp. logIC_{50}	Pred. logIC_{50}	Ref.
1	RGYVYQGL	8.137	8.040	28	***32***[a]	***MWYWGPSL***	***5.125***	***7.581***	30
2	SIINFEKL	8.138	8.180	28	33	VLLDYQGM	5.477	5.620	30
3	APGNYPAL	6.558	6.480	22	34	YSILSPFL	5.954	5.890	30
4	FSVIFDRL	6.971	6.870	22	35	ANEGYDAL	4.924	4.880	24
5	IGRFYIQM	7.770	7.840	22	36	DDEEYVIL	3.907	3.910	24
6	KSSFYRNL	7.066	6.900	22	37	GTYHFTKL	7.745	7.710	24
7	KVVRFDKL	7.310	7.490	22	38	HDQLFSLL	5.639	5.600	24
8	LSYSAGAL	7.523	7.600	22	39	HPTLFKVL	6.208	6.150	24
9[a]	***MGLIYNRM***	***8.337***	***6.213***	22	40	HPYLYRLL	6.712	6.830	24
10	MITQFESL	7.398	7.780	22	41	ISFAFCQL	8.886	8.800	24
11	MMIWHSNL	6.564	6.480	22	42	LIFNYPGV	7.398	7.250	24
12	MNIQFTAV	7.602	7.650	22	43	LIYNYPGV	8.387	8.040	24
13	MNYYWTLL	7.284	7.220	22	44	LMSGFRQM	5.162	5.120	24
14	RFYRTCKL	7.377	7.220	22	***45***[a]	***LQQRYSRL***	***9.222***	***5.795***	24
15	RGYVFQGL	8.509	8.480	22	46	LVYNYPGV	7.638	7.580	24
16	RSYLIRAL	7.174	7.340	22	47	NHPVFSPL	7.252	7.320	24
17	RTFSFQNI	8.013	7.980	22	***48***[a]	***NTVVFDAL***	***3.810***	***6.968***	24
18	SSIEFARL	8.770	8.820	22	49	QESCYGRL	6.463	6.440	24
19	SSISFCGV	8.678	8.740	22	***50***[a]	***QPQNYLRL***	***4.287***	***9.493***	24
20	SSLPFQNI	8.056	8.170	22	51	SIILFLPL	9.000	8.810	24
21	VYIEVLHL	7.699	7.650	22	***52***[a]	***SKLQYKII***	***3.810***	***6.955***	***24***
22	VYINTALL	7.886	7.810	22	53	VDYNFTIV	7.444	7.300	24
23	AIIKFAAL	8.046	8.030	29	54	ALISFLLL	6.030	6.080	27
24	RGYKYQGL	7.854	7.870	29	55	GVYQFKSV	8.000	8.030	27
25	ASARFSWL	6.523	6.620	30	56	ISHNFCNL	6.431	6.650	27
26	CLIFLLVL	5.222	5.150	30	57	IVTMFEAL	7.174	7.010	27
27	FIIFLFIL	5.301	5.440	30	58	LVSIFLHL	5.553	5.430	27
28	FVQWFVGL	6.824	6.900	30	59	NSHHYISM	5.507	5.380	27
29[a]	***IIFLFILL***	***5.125***	***7.853***	30	60	SQTSYQYL	5.729	5.850	27
30	ILSPFLPL	6.329	6.320	30	61	TSYQYLII	7.469	7.600	27
31	LSSIFSRI	5.477	5.520	30	62	YTVKYPNL	6.770	6.830	27

[a] Peptides highlighted in bold and italics indicate where peptide has been removed (outlier) during calculation.

and scores test peptide sequences accordingly. BIMAS[41] and RANKPEP[40] are based on quantitative matrices. BIMAS estimates the binding affinities of peptides by their half-time dissociation rates with class I MHC proteins. RANKPEP uses Position Specific Scoring Matrices (PSSM) in the prediction. The PSSM is produced by an 'ungapped' block alignment of known MHC proteins and identifies sequence similarities among peptides binding to specific both class I and II MHC proteins. With 90% correct predictions, MHCPred was the most reliable algorithm method in the test, followed by SYFPEITHI, which had 65%. BIMAS (35% correct) and RANKPEP (10% correct) performed poorly. Compared to validation methods favoured by computer scientists, such as ROC analysis, the submission of whole protein sequences, as used in this assessment, mirrors how these algorithms would be used in practice, providing a useful and unbiased assessment of a algorithm's ability of to predict T-cell epitope prediction in a "real life" situation. Our results are thus a powerful vindication and validation of the predictive power of the additive method and the utility of both this method, and MHCPred, in predicting mouse epitopes.

Discussion

Herein we report the development of quantitative, systematic models, based on literature IC_{50} values, for the mouse class I alleles: H2–D^b (nonamers), H2–K^b and H2–K^k (both octamers). The results are in good agreement with previous studies of the preferred primary anchor positions: 5 and 9 (nonamers), 2/5 and 8 (octamers—H2–K^k and H2–K^b, respectively). All three models also agree with previous analyses of the preferred residue type at the anchor positions. For H2–D^b: Asn at position 5 and Leu at position 9; for H2–K^b: Phe at position 5 and Val at position 8; and for H2–K^k: Glu, Pro, Gly (best three favoured residues) at position 2 and Ile, Val, Phe (best three favoured residues) at position 8. The nonameric and octameric alleles show both similarities and differences in amino acids preferred at various binding positions (Table 4). Preferences for primary anchors show certain similarities: all models exhibit some preference for small amino acids (H2–D^b (Asn), H2–K^b (Val) and H2–K^k (Pro, Ala)), while C-terminal amino acids are strongly hydrophobic: H2–D^b (Leu), H2–K^b (Val) and H2–K^k (Ile, Val). The most noticeable difference between the nonameric and octameric alleles is at position 5, where H2–D^b exhibits a preference for polar Asn, while H2–K^b shows a preference for Phe (aromatic hydrophobic residue) and H2–K^k for Pro (small amino acid residue).

As well as refining and confirming our understanding of sequence dependence at anchor positions, our results throw new light on all other positions within the peptide. Although this study supports the importance of both primary and secondary anchor residues, it is clear that other positions also play a key role in peptide-binding.[20] Table 4 shows a summary of non-anchor residues associated with both favoured and disfavoured binding to all three alleles. Looking at Table 4, for weak binding peptides, hydrophobic residues are present at position 1 (Phe) and position 3 (Leu, Ile, Tyr, Phe) in abundance, and there is a probable electrostatic repulsion of both negatively charged polar side chains (Asp and Glu) and positively charged polar side chains (Lys, Arg and His).

Although the additive method is a quantitative, rather than a qualitative, prediction method, to explore our results further, we have compared the favoured binding anchor positions, as derived by the additive method, to existing literature anchor motifs, as collated in SYFPEITHI.[42] Table 7 shows the favoured amino acid residues identified by our method (as shown in italics, with residues showing a cut-off value of >+0.3 from Fig. 1) compared with the anchor residues from SYFPEITHI (as shown in bold) at positions P2, P3, P5, P8 and P9. The table indicates our preferences are in accord with those from SYFPEITHI. For example, the H2–D^b allele shows Asn and Leu at the anchor positions P5 and P9, respectively; the H2–K^b allele has Tyr, Phe and Val at positions P3, P5 and P8, respectively; for the H2–K^k allele, additive method and SYFPEITHI motifs share Glu (P2) and Ile and Val (P8). Generally, SYFPEITHI motifs are a subset of our refined, improved, and updated extended-motifs.

Table 6 List of peptides used in the study of the H2–K^k mouse allele

No.	Epitope	Exp. $logIC_{50}$	Pred. $logIC_{50}$	Ref.	No.	Epitope	Exp. $logIC_{50}$	Pred. $logIC_{50}$	Ref.
1	AESKSVII	6.648	6.410	31	78	FESTGNLY	6.010	6.200	32
2	NEKSFKDI	6.910	6.290	31	79	FESTGNMI	7.612	7.620	32
3	QTFVVGCI	6.796	6.460	31	80	FESTGNNI	7.521	7.740	32
4	AESTGNLI	7.624	7.490	32	81	FESTGNPI	7.410	7.600	32
5	DESTGNLI	7.712	7.960	32	82	FESTGNQI	7.612	7.620	32
6	EESTGNLI	7.732	7.760	32	83	FESTGNRI	8.004	7.920	32
7	FASTGNLI	7.429	7.560	32	84	FESTGNSI	7.612	7.620	32
8	FDSTGNLI	7.814	7.350	32	85	FESTGNTI	7.652	7.650	32
9	FEATGNLN	8.178	8.070	32	86	FESTGNVI	7.421	7.530	32
10	FEDTGNLN	8.199	7.930	32	87	FESTGNWI	7.974	7.900	32
11	FEETGNLN	8.028	8.130	32	88	FESTGNYI	7.793	7.760	32
12	FEFTGNLN	8.000	7.900	32	89	FESTGPLI	8.302	8.310	32
13	FEGTGNLN	8.265	8.140	32	90	FESTGQLI	7.920	8.010	32
14	FEHTGNLN	7.982	8.050	32	91	FESTGRLI	8.222	8.240	32
15	FEITGNLN	8.197	8.090	32	92	FESTGSLI	7.992	8.080	32
16	FEKTGNLN	7.904	7.570	32	93	FESTGTLI	7.922	8.010	32
17	FELTGNLN	8.343	8.380	32	94	FESTGVLI	8.023	7.870	32
18	FEMTGNLN	8.222	8.110	32	95	FESTGWLI	7.872	7.970	32
19	FENTGNLN	8.224	8.100	32	96	FESTGYLI	8.215	8.250	32
20	FEPTGNLN	8.043	8.300	32	97	FESTHNLI	7.836	7.890	32
21	FEQTGNLN	8.217	8.100	32	98	FESTINLI	7.887	8.020	32
22	FERTGNLN	8.300	8.170	32	99	FESTKNLI	7.304	7.470	32
23	FESAGNLI	8.031	7.950	32	100	FESTLNLI	7.898	8.070	32
24	FESDGNLI	7.890	7.640	32	101	FESTMNLI	7.888	7.930	32
25	FESEGNLI	7.972	8.090	32	102	FESTNNLI	7.748	7.320	32
26	FESFGNLI	8.085	8.170	32	103	FESTPNLI	8.141	8.130	32
27	FESGGNLI	7.985	8.270	32	104	FESTQNLI	7.819	7.870	32
28	FESHGNLI	8.248	8.300	32	105	FESTRNLI	7.679	7.830	32
29	FESIGNLI	8.239	8.290	32	106	FESTSNLI	7.821	7.880	32
30	FESKGNLI	7.978	8.190	32	107	FESTTNLI	7.821	7.790	32
31	FESLGNLI	8.403	8.280	32	108	FESTVNLI	7.912	7.830	32
32	FESMGNLI	8.040	8.140	32	109	FESTWNLI	7.832	7.880	32
33	FESNGNLI	7.880	8.010	32	110	FESTYNLI	7.460	7.600	32
34	FESPGNLI	8.042	7.950	32	111	FESVGNLI	8.230	8.170	32
35	FESQGNLI	8.094	8.180	32	112	FESWGNLI	7.989	7.930	32
36	FESRGNLI	8.095	8.190	32	113	FESYGNLI	8.099	8.180	32
37	FESSGNLI	8.046	7.990	32	114	FETTGNLN	8.232	8.110	32
38	FESTANLI	7.994	8.170	32	115	FEVTGNLN	8.223	8.110	32
39	FESTDNLI	7.743	7.800	32	116	FEWTGNLN	8.225	8.110	32
40	FESTENLI	7.583	7.690	32	117	FEYTGNLN	8.176	8.070	32
41	FESTFNLI	7.895	7.940	32	118	FFSTGNLI	5.421	5.490	32
42	FESTGALI	7.964	8.050	32	119	FGSTGNLI	7.846	7.560	32
43	FESTGDLI	7.683	7.870	32	120	FHSTGNLI	5.122	5.260	32
44	FESTGELI	7.593	7.780	32	121	FISTGNLI	6.329	6.200	32
45	FESTGFLI	8.267	8.350	32	122	FKSTGNLI	5.026	5.180	32
46	FESTGGLI	7.946	7.760	32	123	FLSTGNLI	7.088	6.930	32
47	FESTGHLI	7.997	7.920	32	124	FMSTGNLI	6.863	6.610	32
48	FESTGILI	8.098	8.150	32	125	FNSTGNLI	6.244	6.130	32
49	FESTGKLI	7.927	7.900	32	126	FPSTGNLI	8.113	7.590	32
50	FESTGLLI	8.079	8.130	32	127	FQSTGNLI	7.013	6.730	32
51	FESTGMLI	7.979	8.050	32	***128***[a]	***FRSTGNLI***	***4.192***	***6.758***	32
52	FESTGNAI	7.602	7.610	32	129	FSSTGNLI	7.718	7.280	32
53	FESTGNDI	7.290	7.190	32	130	FTSTGNLI	7.547	7.030	32
54	FESTGNEI	7.541	7.650	32	131	FVSTGNLI	7.216	6.890	32
55	FESTGNFI	8.044	7.970	32	132	FWSTGNLI	5.325	5.420	32
56	FESTGNGI	7.209	7.300	32	133	FYSTGNLI	5.592	5.620	32
57	FESTGNHI	7.742	7.720	32	134	GESTGNLI	7.665	7.740	32
58	FESTGNII	7.551	7.650	32	135	HESTGNLI	7.607	7.610	32
59	FESTGNKI	7.159	7.260	32	136	IESTGNLI	7.715	7.960	32
60	FESTGNLA	7.455	7.600	32	137	KESTGNLI	7.308	7.470	32
61	FESTGNLD	5.010	5.420	32	138	LESTGNLI	7.716	7.640	32
62[a]	***FESTGNLE***	***4.707***	***6.563***	32	139	MESTGNLI	7.716	7.780	32
63	FESTGNLF	7.848	7.630	32	140	NESTGNLI	7.736	7.510	32
64	FESTGNLG	6.051	6.230	32	141	PESTGNLI	7.426	7.140	32
65	FESTGNLH	6.000	6.190	32	142	QESTGNLI	7.727	7.800	32
66	FESTGNLI	8.046	7.860	32	143	RESTGNLI	7.544	7.420	32
67	FESTGNLK	5.010	5.420	32	144	SESTGNLI	7.641	7.560	32
68	FESTGNLL	7.737	7.720	32	145	TESTGNLI	7.535	7.650	32
69	FESTGNLM	7.212	7.130	32	146	VESTGNLI	7.545	7.350	32
70	FESTGNLN	7.000	7.320	32	147	WESTGNLI	7.740	7.820	32
71	FESTGNLP	5.919	6.120	32	148	YESTGNLI	7.740	7.800	32
72	FESTGNLQ	5.687	5.940	32	149	DGLGGKLV	7.959	8.630	33
73	FESTGNLR	5.232	5.590	32	150	FAFPGELL	7.022	7.410	33
74	FESTGNLS	7.525	7.370	32	151	FAFWAFVV	7.523	7.550	33
75	FESTGNLT	7.293	7.190	32	152	FLHPSMPV	7.149	7.430	33
76	FESTGNLV	7.626	7.830	32	153	HAIHGLLV	7.319	7.760	33
77	FESTGNLW	7.293	7.080	32	154	LEILNGEI	7.921	7.430	33

[a] Peptides highlighted in bold and italics indicate where peptide has been removed (outlier) during calculation.

Table 7 Comparison of favoured binding positions between additive method and SYFPEITHI database

	P2		P3		P5		P8		P9	
	Additive method	SYFPEITHI	Additive method	SYFPEITHI	Additive method	SYFPEITHI	Additive method	SYFPEITHI	Additive method	SYFPEITHI
H2–D^b					*N*	**N**			*L*	**L, M, I**
H2–K^b			*Y*, *R*, *F*	**Y**	*F*	**F, Y**	*V*	**V, L, M, I**		
H2–K^k	*E*, *A*, *D*, *G*, *L*, *P*, *S*, *T*, *V*	**E**					*I*, *V*, *A*, *N*, *L*, *M*, *F*, *S*, *T*, *W*	**I, V**		

Amino acid residues in bold represent well-tolerated anchors. Amino acid residues in italics represent favoured residues from additive method

Each class I mouse MHC allele binds a mixture of structurally diverse peptides, typically 8–10 amino acids in length, with each allele exhibiting defined peptide specificity. From our work,[16–19,57–59] previous peptide binding experiments, and X-ray crystallographic studies of human class I MHC molecules, it is clear that the molecule binds short peptides, most of which are nonamers.[60] Topologically position 1 corresponds to pocket A of the cleft of the peptide-binding site on HLA-A*0201.[14] Anchor residues at position 2 and at the C-terminus (position 9) are seen to be of primary importance for binding, where pocket B interacts with the side chain of the residue at position 2. The structure of pocket A is mainly polar residues and consists of a network of hydrogen bonding residues. A hydrophobic ridge cuts through the binding cleft forcing the peptide to arch between position 5 and the carboxyl-terminal residue (position 9) which are anchored into the D and F pockets in the floor of the cleft.[61] Equivalent data for mice shows clear differences and significant similarities. The crystal structure of several mouse class I molecules has revealed that the peptide binding cleft is also closed at both ends, that the length of the cleft is similar for all class I molecules,[62–66] and that the carboxyl-terminal peptide position is an anchor residue deeply buried in the F pocket. Analysis of the structure and binding results of the H2–K^b and H2–K^k octameric complex reveals that there is a strong preference for an aromatic and hydrophobic residues Tyr and Phe (H2–K^b) and Leu (H2–K^k) at positions 3 and 5 and for a strong hydrophobic residue Val (H2–K^b) and Ile, Val and Phe (H2–K^k) at position 8, which is in accordance to the studies of Falk.[67] It is found that in H2–K^b the B pocket is large enough to accommodate a bulky Ile residue at position 2, which is in accordance with the crystal structure of the antigenic peptide from the ovalbumin complex OVA-8 (SIINFEKL). In H2–K^b and H2–K^k alleles, the results showed that Tyr, Phe and Leu are all favoured in position 3,[61] which is situated in part of pocket D and would significantly deepen the depth and volume of the D pocket and is complementary to the pocket. The anchor carboxyl-terminal (position 8) prefers hydrophobic residues, which fall into pocket F.

While traditional two-anchor motifs can generate reasonable binding predictions,[68] such motifs are clearly only a partial explanation of peptide–MHC affinity. Motifs are a deterministic method, giving yes or no answers, and have a significant error rate, missing many potential binders: peptides without dominant anchors can, and do, retain significant binding affinity. The sequences of binding peptides are very biased in terms of their amino acid composition.[4] This is particularly true of anchor positions, which often favour hydrophobic sequences, and arises from pre-selection resulting in self-reinforcement. Motifs are often used to reduce the experimental workload within epitope discovery: sparse sequence patterns are matched and the corresponding subset of peptides tested, with an enormous resulting reduction in sequence diversity. More sophisticated methods, such as ours, complement the motif approach, as they allow better identification of binders that do not fit the tight restrictions on allele anchors or whose non-anchors abolish binding. However, all efforts to generate reliable prediction methods are ultimately confounded by the data itself, as discussed below. Our methods probe the nature of binding and delineate the underlying structural trends upon which affinity is built, but only within our data set; they are less successful in extrapolating beyond it, thus reducing the universality of the resulting models. It is only through a synergistic interaction between experimental data gathering and *in silico* analysis—designing, testing, and analysing new peptides in an iterative manner—that these limitations can be overcome.[4]

However, we must temper our confidence and enthusiasm with caution, watchfulness, and prudence. The peptide sets we use are larger than is typical for QSAR studies in the literature, at least for affinity, rather than ADME/tox, prediction. The peptides are larger in themselves, and their physical properties more extreme, being multiply charged, zwitterionic, and/or exhibiting huge ranges of lipophilicity. The sequences and properties of the peptides are also heavily biased. This results in part from processes of pre-selection that result in self-reinforcement. As discussed above, simple motifs are often used to reduce the experimental burden of epitope identification: very sparse sequence patterns are use to match peptides, which are then tested, with an enormous concomitant reduction in peptide diversity. Moreover, affinity data is of an intrinsically inferior quality: multiple measurements of the same peptide may vary by several orders of magnitude, some values are clearly wrong, a mix of different standard peptides are used in radioligand competition assays, experiments are conducted at different temperatures and over different concentration ranges. We are also performing a "meta-analysis": almost certainly forcing many distinct binding modes into a single QSAR model. We are thus obliged to filter, albeit in a not altogether subtle manner, our data in order to attempt to remove outliers, which result from such inadequacies in the data. In an ideal world we would look at a variety of "internally rich" data, such as isothermal titration calorimetry, but to do this for all disease-related or frequent allele would be prohibitively time consuming and expensive, and to pursue this is beyond the scope of current methodology.

In order to obtain efficient immune responses with subunit vaccines, efficient adjuvant and delivery systems are required. However, ethical issues regarding the potential toxicity of human vaccines necessitates the use of experimental animals, such as mice, in order explore the nature of immunogenicity, *i.e.* T-cell responses, rather than simple MHC binding. The development of MHC affinity prediction algorithms for mice allows us to properly explore issues of predicting and manipulating immunogenicity, together with the opportunity to then test and validate such predictions experimentally. We will extend our efforts in this direction. Nonetheless, we will incorporate our present models into our web server for MHC-binding prediction: MHCPred, available at the URL: http://www.jenner.ac.uk/MHCPred.[38,39]

The results of the present study have opened up new horizons in mouse immunoinformatics, overhauling present understanding of the structural strategy by which class I mouse molecules are able to bind peptides. As high throughput

genomics reveals the sequences of pathogenic bacteria, viruses, and parasites, such an understanding will become increasingly important, aiding significantly the future discovery vaccines post-genomic discovery of reagents, diagnostics, and peptide and subunit vaccines.

Methodology

Additive method models were generated for three mouse class I alleles: H2–D^b (nonamers), H2–K^b (octamers) and H2–K^k (octamers). For each allele, two models using the Additive method were developed: the first contained just amino acid contributions (the amino acids only model) and the second contained both amino acid contributions and side chain-side chain interactions (the amino acids and interactions model). As these two models were roughly equivalent in terms of statistical quality, we applied the principle of Occam's razor and sought the simplest explanation, choosing the amino acids only model, which will be discussed below. Models were derived using partial least squares (PLS) and validated using leave-one-out cross-validation (LOO-CV); each model being assessed using the cross-validated coefficient (q^2_{LOO}), the standard error of prediction (SEP) and the residual between experimental ($IC_{50(exp)}$) and predicted ($IC_{50(pred)}$) binding affinity. Residuals were classified into three groups: very well predicted peptides with |residuals| $\leq$1.0 log unit, well predicted peptides with |residuals| between 1.0 and 2.0 log units and poorly predicted peptides with |residuals| $\geq$2.0 log units. To achieve a more self-consistent model, a small number of poorly predicted peptides with |residuals| $\geq$2.0 log units were excluded iteratively until the highest residual fell below 2.0 log units. According to present QSAR practice, predictions within 1.0 log unit are considered good.[69–71] This would result in mean residuals of around 0.5 log units. In ideal cases, QSAR methods allows for extrapolation in their predictions of up to 0.3 log units.[72] However, in our work, the experimental measurements we are trying to predict are much less accurate then those obtained for the smaller datasets typical in pharmaceutical applications. The experimental, or biological, error in these measurements is, in terms of logs, much greater which is why in our case we use peptides that have a residual cut-off value of no more than 2.0 log units. The optimal number of components (NC) leading to the highest q^2_{LOO} and the lowest SEP were used to derive the non-cross validated model. The non-cross validated models were assessed by the explained variance (r^2) and standard error of estimate (SEE). The QSAR statistics of the additive models for the three class I alleles are summarised in Tables 1, 5 and 6.

Peptide database and binding affinities

Peptides used in the study and their binding affinities were obtained from the JenPep database.[73,74] The database is freely available at the URL: http://www.jenner.ac.uk/JenPep. The peptide sequences of both nonamers and octamers were investigated in this study. The H2–D^b allele set included 65 nonamers, the H2–K^b allele set 62 octamers and the H2–K^k allele set 154 octamers. The binding affinity (IC_{50}) was used to quantify the interaction of the peptide and the MHC molecule. In this study the IC_{50} values were measured by a competition assay based on the inhibition of binding of a radiolabelled standard peptide to a detergent-solubilised MHC molecule.[75]

Additive method

The IC_{50} values were converted to $\log(1/IC_{50})$, $-\log(IC_{50})$, or (pIC_{50}) and used as a dependent variable in the QSAR regression. The classical Free-Wilson model was extended to allow for interactions between amino acid side chains. This means that the binding affinity of a nonamer is represented by eqn. (1):

$$pIC_{50}=\text{const.}+\sum_{i=1}^{9}P_i+\sum_{i=1}^{8}P_iP_{i+1}+\sum_{i=1}^{7}P_iP_{i+2}+\sum_{i=1}^{6}P_iP_{i+3}+\sum_{i=1}^{5}P_iP_{i+4}+\sum_{i=1}^{4}P_iP_{i+5}+\sum_{i=1}^{3}P_iP_{i+6}+\sum_{i=1}^{2}P_iP_{i+7}+P_1P_9 \quad (1)$$

where const. is the peptide backbone contribution,

$$\sum_{i=1}^{9}P_i$$

is the sum of amino acid contributions at each position,

$$\sum_{i=1}^{8}P_iP_{i+1}$$

is the sum of adjacent peptide side-chain interactions,

$$\sum_{i=1}^{7}P_iP_{i+2}$$

is the sum of every second side-chain interactions,

$$\sum_{i=1}^{6}P_iP_{i+3}$$

is the sum of every third side-chain interaction and so on. The binding affinity will depend primarily on the contributions of amino acid side-chains at each position and their interactions between the adjacent and every second side-chain, *e.g.* both positions (1)–(2) and (1)–(3) interactions are possible between the side chains, thus resulting in eqn. (2) (*amino acids and interactions* models):

$$pIC_{50}=\text{const.}+\sum_{i=1}^{9}P_i+\sum_{i=1}^{8}P_iP_{i+1}+\sum_{i=1}^{7}P_iP_{i+2} \quad (2)$$

If the interaction terms are neglected, eqn. (2) is reduced to eqn. (3) (*amino acids only* model—as chosen in this study):

$$pIC_{50}=\text{const.}+\sum_{i=1}^{9}P_i \quad (3)$$

Partial least squares, cross-validation and leave-one-out cross-validation

Partial least squares (PLS),[76] is an extension of multiple linear regression (MLR) that describes and/or predicts differences in one or more dependent variables from differences in descriptor values. The PLS method was implemented within the QSAR module using SYBYL 6.9.[77] The experimental IC_{50} values (pIC_{50}) were used as the dependent variable in the study. Both the column filtering and scaling were turned off and the optimal number of components (NC) was obtained by cross-validation (CV),[78] using SAMPLS.[79] CV is an approach that benchmarks the predictivity of models and is preformed by dividing the total data set into a number of groups, developing several parallel models from the reduced data, and then predicting the biological activities of the excluded peptides. When the number of excluded groups is equal to the number of compounds in the set, the technique is called leave-one-out cross-validation method (LOO-CV). The predictive power of the model is validated using the following terms: cross-validated coefficient (q^2), and the standard error of prediction (SEP) are defined in eqns. (4) and (5).

$$q^2=1.0-\frac{\sum_{i=1}(pIC_{50_{(exp)}}-pIC_{50_{(pred)}})^2}{\sum_{i=1}(pIC_{50_{(exp)}}-pIC_{50_{(mean)}})^2} \text{ Or simplified to } q^2=1.0-\frac{PRESS}{SSQ} \quad (4)$$

where $pIC_{50(pred)}$ is a predicted value, $pIC_{50(exp)}$ is an actual or experimental value, $pIC_{50(mean)}$ is the best estimate of the mean of all values that might be predicted. The summations are over the same set of pIC_{50} values. PRESS is the PRedictive Error Sum of Squares and SSQ is the sum of squares of $pIC_{50\ (exp)}$ corrected for the mean.

$$SEP=\sqrt{\frac{PRESS}{p-1}} \quad (5)$$

where p is the number of the peptides omitted from the data set, also known as residuals (outliers).

The non-cross validated models were assessed using standard MLR validation terms, explained by variance r^2, standard error of estimate (SEE) are defined in eqns. (6) and (7).

$$r^2=\frac{PRESS}{SSQ} \quad (6)$$

$$SEE=\sqrt{\frac{PRESS}{n-c-1}} \quad (7)$$

where n is the number of peptides and c is the number of components. In the present case, a component in PLS is an independent trend relating measured biological activity to the underlying pattern of amino acids within a set of peptide sequences. Increasing the number of components improves the fit between target and explanatory properties; the optimal number of components corresponds to the best q^2. Both SEP and SEE are standard errors of prediction and assess the distribution of errors between the observed and predicted values in the regression models.

Server comparison

MHCPred[38,39] was compared with three other internet-enabled prediction algorithms: RANKPEP,[40] BIMAS,[41] and SYFPEITHI,[42] in order to examine and find T-cell epitopes in protein sequences. To avoid replicating data from existing databases, only epitopes that have been published within the last two years were used.[43–54] The cut-off points for evaluation were different for each algorithm; if the epitope is above the cut-off, then the algorithm was scored as predicting the epitope. For RANKPEP and BIMAS default thresholds were used, which were 2 and 3% of generated peptide, respectively. Most algorithms listed all the generated peptides and their predicted binding affinities, but in real life situations, people are more interested in the first five or ten peptides as they are more likely to be the epitopes. SYFPEITHI does not give a suggested cut-off point, therefore in the second test the cut-off was set to top 30 peptides for both MHCPred and SYFPEITHI. For BIMAS, a peptide-MHC dissociation half-life of 5 minute was used.

Abbreviations

PLS	Partial least squares
LOO-CV	Leave-one-out cross-validation
SEP	Standard error of prediction
NC	Number of components
SEE	Standard error of estimate

References

1 S. Buus, A. Sette, S. M. Colon, C. Miles and H. M. Grey, *Science*, 1987, **235**, 1353.
2 A. Sette, S. Buus, S. Colon, J. A. Smith, C. Miles and H. M. Grey, *Nature*, 1987, **328**, 395.
3 D. R. Flower, I. A. Doytchinova, K. Paine, P. Taylor, D. Lamponi, C. Zygouri, P. Guan, H. McSparron and H. Kirkbride, *Computational vaccine design. Drug Design: Cutting Edge Approaches*, Royal Society of Chemistry, Cambridge, 2002. p. 136.
4 I. A. Doytchinova, V. A. Walshe, N. A. Jones, S. E. Gloster, P. Borrow and D. R. Flower, *J. Immunol.*, 2004, **172**, 7495.
5 H. P. Adams and J. A. Koziol, *J. Immunol. Methods*, 1995, **185**, 181.
6 M. C. Honeyman, V. Brusic, N. L. Stone and L. C. Harrison, *Nat. Biotechnol.*, 1998, **16**, 966.
7 V. Brusic, G. Rudy, M. Honeyman, J. Hammer and L. Harrison, *Bioinformatics*, 1998, **14**, 121.
8 H. Mamitsuka, *Proteins*, 1998, **33**, 460.
9 V. Brusic, N. Petrovsky, G. Zhang and V. B. Bajic, *Immunol. Cell Biol.*, 2002, **80**, 280.
10 P. Donnes and A. Elofsson, *BMC Bioinformatics*, 2002, **3**, 25.
11 Y. Zhao, C. Pinilla, D. Valmori, R. Martin and R. Simon, *Bioinformatics*, 2003, **19**, 1978.
12 P. A. Reche, J. P. Glutting and E. L. Reinherz, *Hum. Immunol.*, 2002, **63**, 701.
13 J. I. Bell, D. W. Denny, Jr. and H. O. McDevitt, *Immunol. Rev.*, 1985, **84**, 51.
14 M. A. Saper, P. J. Bjorkman and D. C. Wiley, *J. Mol. Biol.*, 1991, **219**, 277.
15 H. Kubinyi and O. H. Kehrhahn, *J. Med. Chem.*, 1976, **19**, 578.
16 I. A. Doytchinova, M. J. Blythe and D. R. Flower, *J. Proteome Res.*, 2002, **1**, 263.
17 P. Guan, I. A. Doytchinova and D. R. Flower, *Protein Eng.*, 2003, **16**, 11.
18 C. K. Hattotuwagama, P. Guan, I. A. Doytchinova, C. Zygouri and D. R. Flower, *J. Mol. Graphics*, 2003, **22**, 195.
19 I. A. Doytchinova and D. R. Flower, *Bioinformatics*, 2003, **19**, 1.
20 D. Hudrisier, H. Mazarguil, F. Laval, M. B. A. Oldstone and J. E. Gairin, *J. Biol. Chem.*, 1996, **271**, 17829.
21 G. E. Price, R. Ou, H. Jiang, L. Huang and D. Moskophidis, *J. Exp. Med.*, 2000, **191**, 1853.
22 A. Vitiello, L. Yuan, R. W. Chesnut, J. Sidney, S. Southwood, P. Farness, M. R. Jackson, P. A. Peterson and A. Sette, *J. Immunol.*, 1996, **157**, 5555.
23 D. A. Ostrov, M. M. Roden, W. Shi, E. Palmieri, G. J. Christianson, L. Mendoza, G. Villaflor, D. Tilley, N. Shastri, H. Grey, S. C. Almo, D. Roopenian and S. G. Nathenson, *J. Immunol.*, 2002, **168**, 283.
24 B. Wizel, B. C. Starcher, B. Samten, Z. Chroneos, P. F. Barnes, J. Dzuris, Y. Higashimoto, E. Appella and A. Sette, *J. Immunol.*, 2002, **169**, 2524.
25 J. E. Gairin, H. Mazarguil, D. Hudrisier and M. B. A. Oldstone, *J. Virol.*, 1995, **69**, 2297.
26 D. Hudrisier, H. Mazarguil, M. B. A. Oldstone and J. E. Gairin, *Mol. Immunol.*, 1995, **32**, 895.
27 R. G. Van der Most, K. Murali-Krishna, J. L. Whitton, C. Oseroff, J. Alexander, S. Southwood, S. Sidney, R. W. Chesnut, A. Sette and R. Ahmed, *Virology*, 1998, **240**, 158.
28 M. G. Rudolph, J. A. Speir, A. Brunmark, N. Mattsson, M. R. Jackson, P. A. Peterson, L. Teyton and I. A. Wilson, *Immunity*, 2001, **14**, 231.
29 A. Franco, T. Yokoyama, D. Huynh, C. Thomson, S. G. Nathenson and H. M. Grey, *J. Immunol.*, 1999, **162**, 3388.
30 A. Sette, C. Oseroff, J. Sidney, J. Alexander, R. W. Chesnut, K. Kakimi, L. G. Guidotti and F. V. Chisari, *J. Immunol.*, 2001, **166**, 1389.
31 H. V. Nielsen, S. L. Lauemoller, L. Christiansen, S. Buus, A. Fomsgaard and E. Petersen, *Infect. Immun.*, 1999, **67**, 6358.
32 A. Stryhn, P. S. Anderson, L. O. Pederson, A. Svejgaard, A. Holm, C. J. Thorpe, L. Fugger, S. Buus and J. Engberg, *Proc. Nat. Acad. Sci. USA*, 1996, **93**, 10338.
33 S. L. Lauemoller, A. Holm, J. Hilden, S. Brunak, M. H. Nissen, A. Stryhn, L. O. Pederson and S. Buus, *Tissue Antigens*, 2001, **57**, 405.
34 J. Cossins, K. G. Gould, M. Smith, P. Driscoll and G. G. Brownlee, *Virology*, 1993, **193**, 289.
35 M. Norda, K. Falk, O. Rotzschke, S. Stevanovic, G. Jung and H. G. Rammensee, *J. Immunother*, 1993, **14**, 144.
36 K. G. Gould, H. Scotney and G. G. Brownlee, *Virology*, 1991, **65**, 5401.
37 G. G. Burrows, K. Ariail, B. Celnik, J. E. Gambee, B. F. Bebo, Jr., H. Offner and A. A. Vandenbark, *J. Neurosci. Res.*, 1996, **45**, 803.
38 P. Guan, I. A. Doytchinova, C. Zygouri and D. R. Flower, *Appl. Bioinformatics*, 2003, **2**, 63.
39 P. Guan, I. A. Doytchinova, C. Zygouri and D. R. Flower, *Nuc. Acids Res.*, 2003, **31**, 3621.
40 P. A. Reche, J. P. Glutting and E. L. Reinherz, *Hum. Immunol.*, 2002, **63**, 701.
41 K. C. Parker, M. A. Bednarek and J. E. Coligan, *J. Immunol.*, 1994, **152**, 163.
42 H.-G. Rammesee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor and S. Stevanovic, *Immunogenetics*, 1999, **50**, 213.
43 A. H. Choi, M. M. McNeal, M. Basu, J. A. Bean, J. L. VanCott, J. D. Clements and R. L. Ward, *Vaccine*, 2003, **21**, 761.
44 A. Diaz-Quinonez, N. Martin-Orozco, A. Isibasi and V. Ortiz-Navarrete, *Infect. Immun.*, 2004, **72**, 3059.

45 S. D'Souza, V. Rosseels, M. Romano, A. Tanghe, O. Denis, F. Jurion, N. Castiglione, A. Vanonckelen, K. Palfliet and K. Huygen, *Infect. Immun.*, 2003, **71**, 483.
46 R. Greenwood, B. Wang, K. Midkiff, G. C. White 2nd, H. F. Lin and J. A. Frelinger, *J. Thromb. Haemost.*, 2003, **1**, 95.
47 G. E. Hancock, P. W. Tebbey, C. A. Scheuer, K. S. Pryharski, K. M. Heers and N. A. LaPierre, *J. Med. Virol.*, 2003, **70**, 301.
48 K. Honjo, X. Xu and R. P. Bucy, *Transplantation*, 2000, **70**, 1516.
49 M. Lyman, H. Lee, B. S. Kang, H. K. Kang and B. S. Kim, *J. Virol.*, 2002, **76**, 3125.
50 K. Mizumachi and J. Kurisaki, *Biosci. Biotechnol. Biochem.*, 2003, **67**, 712.
51 A. Saren, S. Pascolo, S. Stevanovic, T. Dumrese, M. Puolakkainen, M. Sarvas, H. G. Rammensee and J. M. Vuola, *Infect. Immun.*, 2002, **70**, 3336.
52 K. Schulze, E. Medina, G. S. Chhatwal and C. A. Guzman, *Infect. Immun.*, 2003, **71**, 7197.
53 D. Sun, Y. Zhang, B. Wei, S. C. Peiper, H. Shao and H. J. Kaplan, *Int. Immunol.*, 2003, **15**, 261.
54 B. Wu, L. V. Elst, V. Carlier, M. G. Jacquemin and J. M. Saint-Remy, *J. Immunol.*, 2002, **169**, 2430.
55 E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch, *Nucleic Acids Res.*, 2003, **31**, 3784.
56 D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler, *Nucleic Acids Res.*, 2004, **32,** Database issue: D23.
57 I. A. Doytchinova and D. R. Flower, *J. Comput.-Aid. Mol. Des.*, 2002, **16**, 535.
58 I. A. Doytchinova and D. R. Flower, *Proteins: Struct., Funct. Genet.*, 2002, **48**, 505.
59 P. Guan, I. A. Doytchinova and D. R. Flower, *Bioorg. Med. Chem.*, 2003, **11**, 2307.
60 P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger and D. C. Wiley, *Nature*, 1987, **329**, 506.
61 D. H. Fremont, M. Matsumura, E. A. Stura, P. A. Peterson and I. A. Wilson, *Science*, 1992, **257**, 919.
62 D. H. Fremont, E. A. Stura, M. Matsumara, P. A. Peterson and I. A. Wilson, *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 2479.
63 W. Zhang, A. C. Young, M. Imarai, S. G. Nathenson and J. L. Sacchettini, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 8403.
64 A. C. Young, W. Zhang, J. C. Sacchettini and S. G. Nathenson, *Cell*, 1994, **76**, 39.
65 K. J. Smith, S. W. Reid, D. I. Stuart, A. J. McMichael, E. Y. Jones and J. I. Bell, *Immunity*, 1996, **4**, 203.
66 K. J. Smith, S. W. Reid, A. J. Harlos, A. J. McMichael, D. I. Stuart, J. I. Bell and E. Y. Jones, *Immunity*, 1996, **4**, 215.
67 K. Falk, O. Rotzschke, S. Stevanovic, G. Jung and H. G. Rammensee, *Nature*, 1991, **351**, 290.
68 J. D'Amaro, J. G. Houbiers, J. W. Drijfhout, R. M. Brandt, R. Schipper, J. N. Bavinck, C. J. Melief and W. M. Kast, *Hum. Immunol.*, 1995, **43**, 13.
69 G. Klebe, U. Abrahman and T. Mietzner, *J. Med. Chem.*, 1994, **37**, 4130.
70 S. Sicsic, I. Serraz, J. Andrieux, B. Bremont, M. Mathe-Allainmat, A. Poncet, S. Shen and M. Langlois, *J. Med. Chem.*, 1997, **40**, 739–748.
71 P. Durcot, M. Legraverend and D. S. Grierson, *J. Med. Chem.*, 2000, **43**, 4098.
72 *Ligand-Based Design Manual, Sybyl 6.6*, Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
73 M. Blythe, I. A. Doytchiniva and D. R. Flower, *Bioinformatics*, 2002, **18**, 434.
74 H. McSparron, M. J. Blythe, C. Zygouri, I. A. Doytchinova and D. R. Flower, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1276.
75 J. Sidney, H. M. Grey, S. Southwood, E. Celis, P. A. Wentworth, M. F. del Guercol, R. T. Kubo, R. W. Chestnut and A. Sette, *Hum. Immunol.*, 1996b, **45**, 79.
76 D. Young. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, Wiley Inter-Science, New York, 2001, p. 243.
77 *Sybyl 6.9*, Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
78 S. Wold. *PLS for Multivariate Linear Modelling in Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995. p. 195.
79 B. L. Bush and R. B. Nachbar, Jr., *J. Comput.-Aid. Mol. Des.*, 1993, **7**, 587.