Journal of Computer-Aided Molecular Design (2005) 19: 203–212
DOI 10.1007/s10822-005-3993-x

Towards the chemometric dissection of peptide – HLA-A*0201 binding affinity: comparison of local and global QSAR models

Irini A. Doytchinova, Valerie Walshe, Persephone Borrow & Darren R. Flower*
Edward Jenner Institute for Vaccine Research, RG20 7NN, Compton, Berkshire, UK

Received 22 November 2004; accepted in revised form 15 March 2005

Key words: GA, peptides, PLS, stepwise regression, *z*-scales

Summary

The affinities of 177 nonameric peptides binding to the HLA-A*0201 molecule were measured using a FACS-based MHC stabilisation assay and analysed using chemometrics. Their structures were described by global and local descriptors, QSAR models were derived by genetic algorithm, stepwise regression and PLS. The global molecular descriptors included molecular connectivity χ indices, κ shape indices, E-state indices, molecular properties like molecular weight and log *P*, and three-dimensional descriptors like polarizability, surface area and volume. The local descriptors were of two types. The first used a binary string to indicate the presence of each amino acid type at each position of the peptide. The second was also position-dependent but used five *z*-scales to describe the main physicochemical properties of the amino acids forming the peptides. The models were developed using a representative training set of 131 peptides and validated using an independent test set of 46 peptides. It was found that the global descriptors could not explain the variance in the training set nor predict the affinities of the test set accurately. Both types of local descriptors gave QSAR models with better explained variance and predictive ability. The results suggest that, in their interactions with the MHC molecule, the peptide acts as a complicated ensemble of multiple amino acids mutually potentiating each other.

Introduction

Foreign proteins originating from pathogenic organisms (parasites, bacteria, fungi, or viruses) are cleaved into short peptides of 8–11 amino acids in the cell and bind to major histocompatibility complex proteins (MHCs). Peptide–MHC complexes are presented on the cell surface, where they are recognised by T cells. T cells are specialised immune cells which mediate cellular immunity. Peptides bound to MHC and recognised by T cells are known as epitopes. Most epitopes bind well to MHC molecules [1], but only certain binders are recognised as T-cell epitopes. Peptide binding to MHC molecules is thus a prerequisite for T-cell recognition. T-cell epitopes are key components of immunotherapeutics and vaccines. Vaccine research is one of the main components of publicly funded health-care programs world wide, with the current world human vaccine market valued at $5 billion. With the incipient threat from bioterrorism, the AIDS pandemic, and the growth of antibiotic resistance, interest in vaccines has increased dramatically in the last decade: the number of companies focussing on vaccines having risen from around 10 companies in 1990 to over 100 today. Research into therapeutic vaccines, or pharmaccines, is now extremely active.

MHCs bind peptides derived through the proteolytic degradation of proteins, the details of such events are complex and remain poorly understood. Part of this complexity allows

*To whom correspondence should be addressed. Fax: +44-1635-577901/577908; e-mail: darren.flower@jenner.ac.uk

peptides with post-translational modifications, such as phosphorylation or glycosylation, to form pMHCs and be recognised subsequently by TCRs. Chemically modified peptides and peptidomimetic compounds also bind MHCs [2]. Many drug-like molecules also bind MHCs, exhibiting pathological effects through this mechanism [3]. Thus MHCs are very catholic in terms of the molecules they bind and are not restricted to peptides. Since the disadvantageous metabolic and physicochemical properties of long peptides have caused them to fall from favour as putative leads or drugs, it has been suggested recently that the formation of a MHC–drug–TCR complex may be a useful approach for the development of immunotherapeutic drug-like small molecule inhibitors of T cell mediated processes [3]. The resulting complex would resemble the sort of complex formed by FK506 with the FK binding protein that together act as an inhibitor of calcinerin [4].

There are two classes of MHC molecules: class I and class II. They are encoded by six separate genetic loci, three for each class. MHC genes are amongst the most polymorphic genes in higher vertebrates, indeed over 1500 human alleles have been characterised [5]. The allele HLA-A*0201 is one of the most frequent class I alleles in many populations [6]. It is expressed in 50% of Caucasians [7] and it has been demonstrated to play a critical role in antigen presentation of viral and tumour antigens [8–10]. Because of its high incidence, HLA-A*0201 allele has been widely and intensively studied, and its three-dimensional structure has been solved by X-ray crystallography [11, 12]. Peptides that bind to HLA-A*0201 have a size of 8–11 amino acids, and require a combination of anchor residues: two primary anchors, at position 2 and the C-terminus, plus several other positions (1, 3 and 7), the so-called secondary anchors [12–14].

Historically, two distinct mechanistic paradigms have, consciously or unconsciously, dominated thinking about ligand–receptor binding: one global, one local. In a global model, hydrophobic molecules are thought to depot into a hydrophobic binding site. In a local model, specific regioselective interactions, between receptor and ligand, drive affinity. Strictly, all receptor–ligand binding involves a balance between multiple interactions of a ligand with a weakly ionic aqueous solvent and multiple interactions with a binding site, each with both an enthalpic and an entropic component, and it is solvent interactions such as these that lead to solvation, desolvation, and hydrophobic effects. These opposing, if reconcilable, notions of the mechanisms underlying binding have led to the modelling of affinity in terms of either global descriptors of whole molecule bulk properties, such as hydrophobicity (often identified with the water–octanol partition coefficient or log *P*), or local, atomistic descriptions of site-specific interactions between ligand and receptor. Because peptides are polymeric in nature, receptor–peptide, and particularly MHC–peptide, binding provides an excellent system for comparison of local and global affinity models. Recently, we used the additive method, which considers peptide binding affinity to be the sum of independent contributions from each amino acid at each position [15], to design 10 peptides binding to HLA-A*0201 based on a chemometric study of 90 peptides [16]. The measured affinity of the newly designed peptides revealed that all had high binding affinities to HLA-A*0201 molecule, significantly greater than the highest previously reported, [1] suggesting that some co-operative synergism may be operating. Extending the set of in-house peptides studied, we have attempted to compare the relative contributions from global molecule properties versus those of two distinct site-specific methods: the additive and *z*-scale-based methods [17].

Methods

Peptides and binding affinities

One hundred and seventy-seven nonameric peptides we used in the present study. Their three-dimensional structures were built using SYBYL 6.9 software [18] starting from the X-ray structure of TLTSCNTSV [12]. The peptide backbone was kept in its X-ray conformation and full geometry optimization using the standard Tripos molecular mechanics force field was performed. Partial atomic charges were generated using single point calculations by AM1 semiempirical method as implemented in SYBYL. Finally, structures were imported into MDL QSAR software [19].

All peptides used in the present study were ordered from Mimotopes (Pensby, UK). Peptide binding to HLA-A2 was assessed using a FACS-based MHC stabilization assay [20] with

modifications as described elsewhere [21]. Briefly, T2 cells were incubated in 96-well flat-bottom plates at 2×10^5 cells per well in a 200 μl volume of AIM V medium (Life Technologies, Paisley, UK) with human β_2-microglobulin at a final concentration of 100 nM (Scipac, Sittingbourne, UK) with and without peptides at concentrations between 200 and 0.04 μM for 16 h at 37 °C. Cells were then washed and surface levels of HLA-A2 were assessed by staining with FITC-conjugated A2.1-specific mAb BB7.3 (BD Biosciences, Oxford, UK) or a FITC-conjugated isotype control Ab (BD Biosciences). Cells were fixed at 4 °C in 4% paraformaldehyde and analyzed on a FACSCalibur (BD Biosciences) using CellQuest software. Each peptide was tested in triplicate. These with an error higher than 0.5 μM were tested two more times. Results are expressed as fluorescence index (FI) values. These were calculated as the test mean fluorescence intensity (MFI) minus the no peptide isotype control MFI divided by the no peptide HLA-A2-stained control MFI minus the no peptide isotype control MFI. The half-maximal binding level (BL_{50}) which is the peptide concentration yielding the half-maximal FI of the reference peptide in each assay was calculated and presented as pBL_{50} ($-\log BL_{50}$). The HLA-A2 high binder FLPSDFFPSV (IC_{50} = 2.6 nM) [22] was used as the reference peptide.

Training and test sets

As the amino acids did not vary in a balanced way at each peptide position, the training and test sets were selected stepwise. First, all peptides with amino acids available three or less times at each position were included in the training set. For the rest of the peptides Tanimoto coefficients (TC) were calculated [23]. They ranged from 0 (no one common amino acid) to 0.8 (8 common amino acids). All peptides with TC less than 0.8 were also included in the training set. For pairs with TC = 0.8, the more active peptide was put into the training set and the less active into the test set. Thus, the training set consisted of 131 peptides and the test set comprised 46 peptides (Table 1). This selection procedure allowed the test set to be fully representative, i.e. not to include peptides with missing or poorly presented (less than three times) amino acids in the training set. The training set was used for models development and the test set for external validation.

Global molecular descriptors

One hundred and twenty-eight molecular descriptors, belonging to five different types, were computed using the software MDL QSAR version 2.2 [19]. The first type of descriptor includes the molecular connectivity χ indices, [24] which represent molecular structure by encoding significant topological features of whole molecule. There are five categories of structural information described by χ indices: degree of branching (low order χ indices), variable branching pattern (path χ indices), position and influence of heteroatoms (valence χ indices), patterns of adjacency (χ cluster and path/cluster indices) and degree of cyclicity (χ chain indices). The second group of descriptors – the κ shape indices – are a family of graph-based structure descriptors that represent shape [25]. The third group comprises the electrotopological state (E-state) indices, which are atom level molecular descriptors computed for each atom in the molecule [25]. They represent the electron density at each atom and the ability of those electrons to participate in intermolecular interactions. A variety of molecular properties – weight, log *P*, number of rings, number of hydrogen bond donors and acceptors, etc. – was defined as the fourth group. The last group consisted of three-dimensional molecular properties such as polarizability, surface area, volume, etc.

Additive model

The additive method for binding affinity prediction is described in detail elsewhere [15]. Briefly, the peptide sequence was represented as a set of 180 local descriptors (20 aa $\times$ 9 positions), which could take the values 1 (present) or 0 (absent) depending on whether a certain amino acid exists at a certain position. pBL_{50} values (*y* or dependent variable) were included as the first column. Columns containing only 0s were omitted. As there were 19 missing amino acids at different positions, the final number of variables was reduced to 161. In the derived additive model each AA at each position has a regression coefficient accounting for its contribution to the affinity. Thus, AA with positive coefficients make positive contributions to

Table 1. Peptides used in the study and their experimental binding affinities.

Peptides	pBL_{50exp}	Peptides	pBL_{50exp}	Peptides	pBL_{50exp}
ALCRWGLLL	4.91	ILDPFPVTN	5.29	RLWPFYHNV	5.72
ALIHHNTHL	4.30	ILDPFPVTP	5.82	RLWPIYHNV	5.77
ALPYWNFAT	4.66	ILDPFPVTQ	5.28	RLWPLYPNV	5.57
CLTSTVQLV	4.93	ILDPFPVTS	4.78	SIISAVVGI	4.47
FLCKQYLNL	5.21	ILDPFPVTT	5.54	SLHVGTQCA	3.79
FLDQVPFSV	5.98	ILDPFPVTV	8.65	SLNFMGYVI	4.00
FLLSLGIHL	5.17	ILDPFPVTW	4.71	SLYADSPSV	5.24
FLLTRILTI	4.95	ILDPFPVTY	3.19	TLGIVCPIC	4.68
FLNPFYPNV	6.16	ILDPIPPTV	7.30	TLHEYMLDL	4.94
FLWPFYHNV	5.99	ILDQVPFSV	6.09	TTAEEAAGI	3.39
FLWPFYPNV	5.89	ILFPGPVTA	6.23	VCMTVDSLV	4.20
FLWPIYHDV	6.16	ILKEPVHGV	5.59	VLHSFTDAI	4.54
FLWPIYHNV	6.37	ILWPIYHNV	6.24	VLIQRNPQL	5.06
FLWPLYPNV	6.14	ILWQVPFSV	5.91	VLLDYQGML	4.52
FVTWHRYHL	4.21	IMDPFPVTV	7.21	VTWHRYHLL	4.38
GLLGWSPQA	5.13	IMDQVPFSV	5.71	WILRGTSFV	4.06
GLSRYVARL	4.78	INDPFPVTV	4.78	WLDQVPFSV	5.23
GLYSSTVPV	5.15	IPDPFPVTV	5.10	YAIDLPVSV	5.63
HLESLFTAV	3.79	IQDPFPVTV	6.05	YLAPGPVTA	5.74
HLLVGSSGL	3.91	ISDPFPVTV	5.50	YLAPGPVTV	6.00
HLYSHPIIL	5.41	ITAQVPFSV	4.43	YLCPGPVTA	6.18
IADPFPVTV	5.76	ITDPFPVTV	6.08	YLEPGPVTL	5.41
ICDPFPVTV	5.45	ITDQVPFSV	4.48	YLFDGPVTA	5.50
IDDPFPVTV	4.36	ITFQVPFSV	4.42	YLFNGPVTA	5.80
IFDPFPVTV	4.89	ITWQVPFSV	5.01	YLFNGPVTV	5.65
IGDPFPVTV	3.92	IVDPFPVTV	6.21	YLFPCPVTA	6.63
IHDPFPVTV	4.96	IWDPFPVTV	5.13	YLFPDPVTA	6.09
IIDPFPVTV	6.31	IYDPFPVTV	5.41	YLFPGPETA	5.81
IISCTCPTV	5.17	KIFGSLAFL	4.40	YLFPGPFTV	5.81
ILDDFPVTV	7.16	KLHLYSHPI	4.77	YLFPGPMTA	5.98
ILDDLPPTV	7.14	KLPQLCTEL	4.50	YLFPGPMTV	5.85
ILDPFPPEV	7.68	KTWGQYWQV	4.43	YLFPGPSTA	5.69
ILDPFPPTV	8.17	LLFGYPVYV	5.45	YLFPGPVQA	6.14
ILDPFPVTA	6.32	LLMGTLGIV	4.21	YLFPGPVTA	6.31
ILDPFPVTC	5.65	LLWFHISCL	4.13	YLFPGPVTG	5.22
ILDPFPVTD	2.94	LQTTIHDII	3.90	YLFPPPVTA	5.75
ILDPFPVTE	3.13	MLDLQPETT	4.36	YLFPPPVTV	6.19
ILDPFPVTF	5.674	MLGTHTMEV	5.37	YLNPGPVTA	5.53
ILDPFPVTG	6.662	NLQSLTNLL	3.96	YLSPGPVTA	5.44
ILDPFPVTH	3.604	NLSWLSLDV	4.75	YLWQYIPSV	5.17
ILDPFPVTI	6.688	NMVPFFPPV	5.60	YLYPGPVTA	5.77
ILDPFPVTK	4.590	PLLPIFFCL	5.32	YMNGTMSQV	4.67
ILDPFPVTL	7.033	RLLQETELV	4.83	YTDQVPFSV	4.80
ILDPFPVTM	6.126	RLMKQDFSV	4.97		
Test set n = 46					
ALMPLYACI	5.08	ILKPLYHNV	5.25	YLFDGPVTV	4.96
FLDDHFCTV	6.68	ILNPFYHNV	6.16	YLFPFPITV	6.68
FLFPGPVTA	6.18	ILNPFYPDV	6.11	YLFPGPFTA	5.65

Table 1. (Continued).

Peptides	pBL_{50exp}	Peptides	pBL_{50exp}	Peptides	pBL_{50exp}
FLFPLPPEV	6.53	ILWPLFHEV	6.03	YLFPGPVWA	5.59
FLKPFYHNV	5.73	ILWPLYPNV	6.06	YLFPGTVTA	6.16
FLNPIYHDV	6.16	ILYQVPFSV	5.06	YLFPGVVTA	6.17
FTDQVPFSV	4.76	ITSQVPFSV	4.06	YLFQGPVTA	5.21
GILTVILGV	4.57	LLAQFTSAI	4.51	YLKPGPVTA	5.26
GLGQVPLIV	4.76	LMAVVLASL	3.99	YLMPGPVTA	5.27
GTLGIVCPI	5.23	RLNPFYHDV	4.24	YLWDHFIEV	6.36
ILDDFPPTV	7.08	RLNPLYPNV	5.37	YLWPGPVTV	5.70
ILDPFPITV	8.14	RLWPFYPNV	5.24	YLWQYIFSV	4.94
ILDPFPPPV	7.44	RLWPIYHDV	5.55	YMLDLQPET	5.28
ILDPLPPTV	7.15	SLDDYNHLV	5.27	YVITTQHWL	4.39
ILFPFPVEV	6.80	SVYDFFVWL	5.12		
ILFPFVHSV	6.58	VMGTLVALV	5.03		

affinity (they increase binding) and those with negative coefficients decrease affinity because of their negative contributions. Using the regression coefficients, the binding affinity of a peptide could be calculated easily as a sum of the contributions of the AA at each position in the peptide and the *const* term.

z-Scales

The *z*-scales, defined by Hellberg and collaborators [26], summarise the principal chemical and physical properties exhibited by amino acids. These scales were derived by PCA of a data matrix consisting of a multitude of physicochemical variables, such as molecular weight, pK_a's, ^{13}C NMR-shifts, etc. The first principle component reflects the hydrophobicity of amino acids, the second their size, and the third their electronic properties. The scores of these components are defined as z_1-, z_2- and z_3-scales, respectively. More recently, Sandberg and co-authors [17] extended the three *z*-scales to five, adding two additional *z*-scales, z_4 and z_5. By arranging the *z*-scales according to the amino acid sequence, it is possible to numerically quantify the structural variations within a series of related peptides. This parameterization has been shown to be useful in the selection of training sets of representative peptides by means of multivariate design [27]. In the present study, five *z*-scales were used to describe the peptide amino acid sequence. The *X*-matrix of the training set contains 45 local descriptors (nine positions times 5 *z*-scales) for 131 peptides.

Variable selection

A genetic algorithm (GA) [28] and stepwise regression, as implemented in the MDL QSAR package, were used as variable selection procedures in the present study. GA allows one to select a subset of the most significant predictors using two evolutionary operations: random mutation and genetic recombination (crossover). The algorithm was calibrated with regard to the size of initial population, choice of parents, types of crossover and mutation, and fitness function. The regression equations were generated on the basis of the selected variables by ordinary multiple linear regression (MLR). The stepwise regression was used in a forward mode with default value for F-to-enter (4.00) and F-to-remove (3.99). Final models were assessed by explained variance (r^2), standard error of estimate (SEE) and *F*-ratio.

PLS

The PLS linear regression, as implemented in SIMCA-P 8.0 [29], was used in the study. PLS can handle matrices, which have more variables than observations and also deal with noisy and highly collinear data. In this situation, conventional statistical methods, such as multiple regression, produce over-fitted models, i.e. models that fit the

training data well but are unreliable in prediction. PLS forms new variables, named principal components (PC), as linear combinations of the initial variables and then uses them as predictors of the dependent variable. The optimal number of PC was derived after leave-one-out cross-validation (LOO-CV). The models were assessed by $r^2$and q^2 (cross-validated r^2). Only models with positive q^2 values were considered further for external validation.

External validation

The test set was used for external validation. The predictive ability of the derived models was assessed by a correlation coefficient between the predicted and experimental pBL_{50} values (r_{pred}) and the explained variance (r^2_{pred}). Absolute average error (AAE) was also defined as the modulus of the averaged residual between predicted and experimental pBL_{50}s.

Results

QSAR model with global descriptors

One hundred and twenty-eight global descriptors were calculated for 131 training set peptides. As there were many inter-correlated descriptors, a pre-selection was applied: from pairs with $r \geq 0.7$, the descriptor best correlated with pBL_{50} was selected. Two variable selection procedures – GA and stepwise regression – were then used to select the most predictive variables. The best performance (highest r^2) for the GA was achieved with an initial population of size 32, tournament selection, uniform crossover, one-point mutation and Friedman's lack-of-fit scoring function with value 3. Identical models were derived by both procedures:

$$pBL_{50} = -0.441 * SHBint3\_acnt - 3.302 * MaxQp + 18.88 * MaxHp + 9.641$$

The statistics of the model is given in Table 2 and the predicted after LOO-CV versus experimental pBL_{50} values for the training and the test sets are plotted in Figures 1 and 2, respectively. The descriptor SHBint3_acnt, which is the count of internal hydrogen bonds between atoms separated by three skeletal bonds, is inversely correlated with affinity. Peptides bearing a lot of polar amino acids like Asp, Asn, Glu, Gln, Ser and Thr have high values of SHBint3_acnt and low pBL_{50}s. According to the intercorrelation matrix (data not shown) descriptor SHBint3_acnt correlates with the number of certain chemical groups in the molecules (numHBd, hydrogen bond donors; SssNH_acnt, count of all NH = groups; SsssN_acnt, count of all N atoms), several E-state indices (SHHBd, sum of atom-type hydrogen E-state indices for hydrogen bond donors; SssCH2, sum of all CH2 = E-state values in molecule; SsssN, sum of all – N = E-state values in molecule), and two χ indices (xvp9, valence 9th order path χ index; xvch5, simple 5th order chain χ index); MaxQp corresponds with the largest positive charge in the molecule and is also inversely correlated with

Table 2. Statistics of the QSAR models.

Descriptors	Model	R^2	Training set n = 131			q^{2b}	Test set n = 46		
			SEE[a]	F[a]	PC[b]		r_{pred}	r^2_{pred}	AAE
Global	GA + MLR	0.43	0.75	31.76			0.65	0.42	0.52
	Stepwise regression	0.43	0.75	31.76			0.65	0.42	0.52
	PLS	0.64			4	Negative	na[c]	na	na
Additive	GA + MLR	0.97	0.25	28.72			0.49	0.24	1.05
	Stepwise regression	0.85	0.43	21.93			0.45	0.20	1.12
	PLS	0.85			3	0.54	0.80	0.64	0.45
z-scales	GA + MLR	0.67	0.59	24.15			0.71	0.50	0.51
	Stepwise regression	0.67	0.59	24.15			0.71	0.50	0.51
	PLS	0.35			4	Negative	Na	na	na

[a]Calculated only for MLR.
[b]Calculated only for PLS.
[c]Not applicable, because of negative q^2 values.

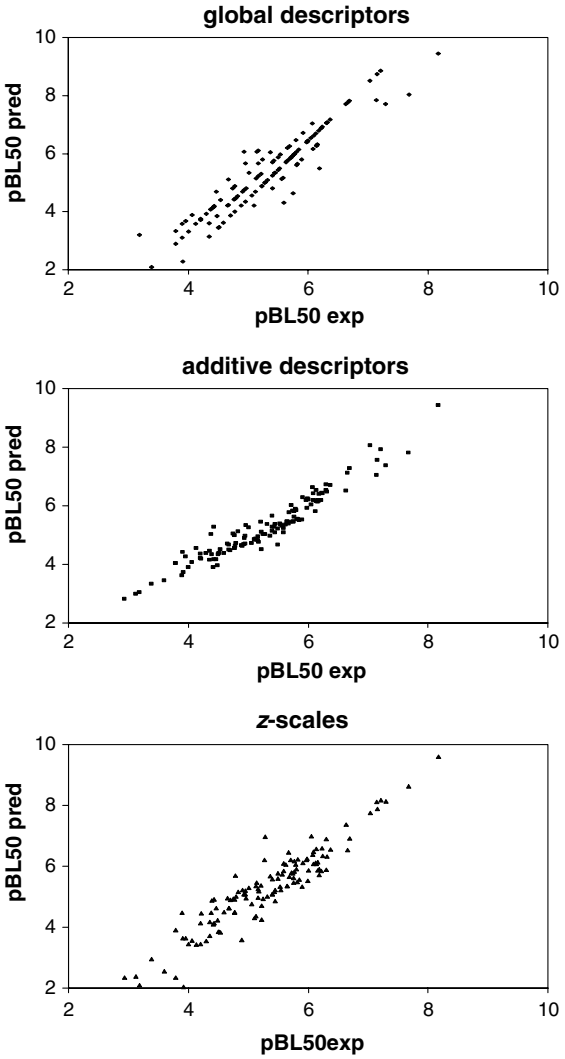

Figure 1. Predicted by LOO-CV versus experimental pBL_{50} values for the training set. Upper: QSAR model with global descriptors; middle: additive model; lower: QSAR models with *z*-scales. The statistics of the models are given in Table 2.

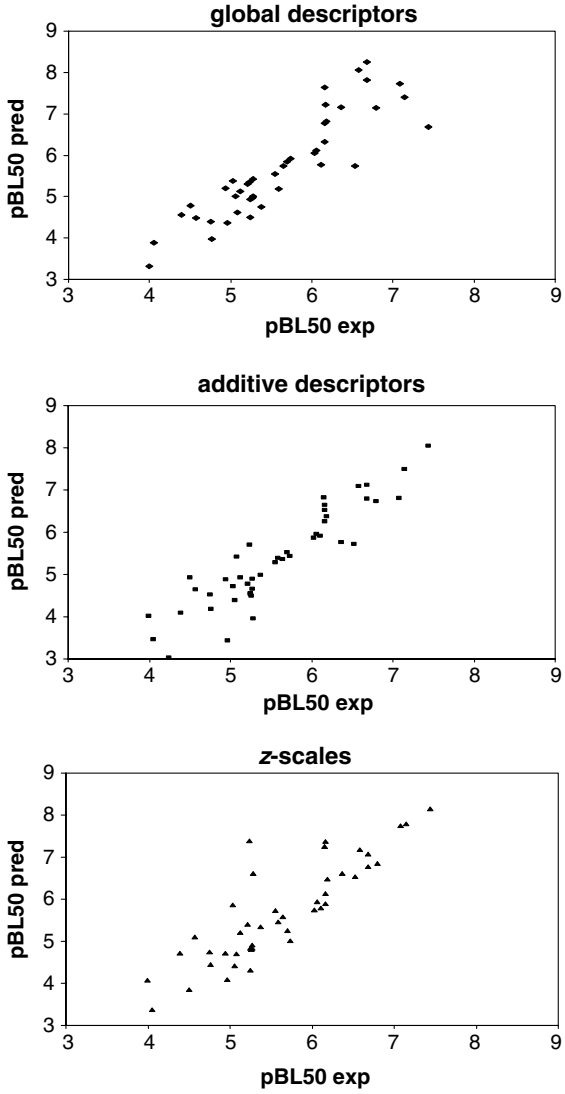

Figure 2. Predicted versus experimental pBL_{50} values for the test set. Upper: QSAR model with global descriptors; middle: additive model; lower: QSAR models with *z*-scales. The statistics of the models are given in Table 2.

affinity. Peptides with Phe, Trp and Tyr at p9 have the highest positive charges at $9C_\alpha$ atom. Descriptor MaxQp correlates with the number of Carbon atoms in the molecule (SddC_acnt) and the sum of their E-state values (SddC). MaxHp is the largest positive charge on a hydrogen atom and it does not correlate with other descriptors. The hydrogen atoms with the highest positive charge belong to Asp and Glu carboxy groups. This is in a good agreement with the additive model (see below), where Asp and Glu were found to be preferred at p4 and p8, respectively. As is evident from r^2 and r^2_{pred}, these descriptors are not enough either to explain the variance in the training set, or to predict test set affinities accurately.

As the QSAR model with global descriptors, solved by PLS, gave negative q^2 value, it was not considered further for external validation.

Additive model

The matrix of 162 columns (1 dependent variable pBL_{50} + 20 aa × 9 positions – 19 missing aa) and 131 rows was solved by PLS. Three principal

components explain 85% of the variance in the training set and predict the affinity of the test set with $r^2_{pred} = 0.64$ (Table 2, Figures 1 and 2). Additionally, both variable selection procedures (GA and stepwise regression) were applied here to reduce the number of variables, but the predictivities of the new models were poorer although their explained variances improved (Table 2).

The additive regression model is shown in Table 3. One key ergonomic advantage of this model is its easy interpretation. The quantitative contributions at each position indicate, directly, favoured and disfavoured amino acids. As is evident from the regression model (Table 2), preferred amino acids at position 1 (p1) are Ile and Phe, at position 2 (p2): Leu and Met, at position 3 (p3): Asp and Tyr, at position 4 (p4): Asp and Pro, at position 5 (p5): Ile, at position 6 (p6): Pro and Val, at position 7 (p7): Pro, at position 8 (p8): Thr and at position 9 (p9): Val. Many of these preferences are well known: Ile and Phe for p1, Leu and Met for p2, Val for p9 [30]. Some of the high positive contributions should be treated very carefully because they concern poorly represented amino acids in the training set (less than 3 times). In the present model, these are Cys and Pro at p1, Val at p2, Lys at p3, Cys at p5, Ile at p7, Tyr at p8, and Met at p9.

QSAR model with z-scales

z-Scales describe the main physicochemical properties of the amino acids, such as hydrophobicity, size and electronic properties. Both variable selection procedures led to identical models:

$$\begin{aligned} pBL_{50} = & 0.257 * 4zz5 - 0.523 * 9zz2 \\ & - 0.221 * 2zz1 - 0.105 * 1zz1 \\ & - 0.148 * 9zz1 + 0.221 * 6zz3 + 0.104 \\ & * 4zz3 - 0.166 * 2zz2 + 0.076 * 5zz2 \\ & + 0.225 * 9zz4 + 2.335 \end{aligned}$$

The statistics of the model is given in Table 2 and the predicted after LOO-CV versus experimental pBL_{50} values for the training and the test sets are plotted in Figures 1 and 2, respectively. The first number in the descriptor name corresponds to the position in the peptide sequence and the last to the *z*-property. In terms of affinity, this model indicates the importance of almost all positions in the peptide, with p9 making the most significant

Table 3. Additive QSAR model for peptides binding to HLA-A*0201 molecule.

	p1	p2	p3	p4	p5	P6	p7	p8	p9
Ala	−0.011	0.070	−0.153	−0.212	0.047	−0.430	−0.096	−0.250	0.171
Arg	−0.173			−0.135	−0.113	−0.275		0.030	
Asn	−0.282	−0.963	−0.148	−0.045		−0.023	−0.469	−0.028	−0.436
Asp		−1.411	0.152	0.591	0.064	−0.064	−0.141	0.067	−2.925
Cys	0.249	−0.171	0.145	−0.121	0.820	−0.279		−0.317	−0.130
Gln		0.083	−0.469	−0.095	−0.341	−0.294	−0.282	−0.040	−0.446
Glu			−0.520	0.024	−0.145		−0.181	0.736	−2.726
Gly	−0.017	−1.877	−0.028	−0.158	−0.029	0.040	−0.330	−0.173	0.218
His	−0.300	−0.776	−0.070	−0.320	0.002	−0.472	0.053	−0.106	−2.219
Ile	0.156	0.109	−0.007	−0.290	0.342	−0.274	0.413	−0.124	0.114
Leu	−0.255	0.322	−0.114	−0.373	−0.077	−0.123	−0.039	−0.099	0.031
Lys	−0.150		0.218	−0.091					−1.174
Met	−0.286	0.525	−0.165		−0.171	−0.075	0.106	−0.294	0.454
Phe	0.138	−0.847	0.086	−0.273	0.042	−0.265	−0.010	−0.212	−0.025
Pro	0.242	−0.619	−0.168	0.147	0.145	0.216	0.569	0.021	0.134
Ser	−0.293	−0.202	−0.206	−0.021	0.066	−0.228	−0.337	−0.085	−0.974
Thr	−0.148	−0.448	−0.168	−0.105	−0.169	−0.184	−0.557	0.105	−0.252
Trp	−0.271	−0.594	−0.078	−0.315	0.031		−0.349		−1.052
Tyr	0.036	−0.299	0.108	0.082	−0.029	−0.064	−0.225	0.206	−2.661
Val	−0.130	0.258	−0.014	−0.666	−0.092	0.139	−0.052	−0.171	0.316

The constant is 4.893. The statistics of the model is given in Table 2.

contribution with three z-properties taking part in the model. P9 favours amino acids with negative z_1 and z_2 and positive z_4. Only Met satisfies these three requirements simultaneously, which corresponds well to the result from the additive model, but contrasts with the well known preference for Val at this position [30]. At p2, amino acids with negative z_1 and z_2 values such as Ile, Leu and Met are preferred. Positive z_3 and z_5 at p4 suggest Asp and Pro as favoured residues. Although p4 is considered as an anchor for the T-cell receptor (TCR) [12–14], its importance for the MHC binding could be connected with a conformational influence on the neighbouring anchor positions. Negative z_1 values at p1 broaden the range of favoured substituents to all hydrophobic amino acids. A similar broad range of small amino acids at p5 is suggested by the positive value of z_2. At p6, amino acids with positive z_3 are favoured. The PLS model with z descriptors gave low explained variance r^2 and negative q^2 value (Table 2).

Discussion

The relationship between structure and affinity for a set of 131 peptides binding to HLA-A*0201 molecule was examined by three different sets of QSAR descriptors – one of global and two of local ones. Three methods were applied – GA followed by MLR, stepwise forward regression and PLS. The QSAR model based on global descriptors treats each peptide as an ensemble of amino acids acting co-operatively. The additive model dissects out the contribution of each amino acid at each position quantitatively and is easily interpreted. However, it does not allow extrapolations outside the set of amino acids used in the training set at each position. This disadvantage could be overcome by the z-scaled-based model. The z-scales-based model assesses the contributions of the physicochemical properties of the amino acids at each position, indicating common properties. Present results indicated that the global properties selected by two variable selection procedures – GA and stepwise regression – are not sufficient to explain the variance in the training set, describing only 43%. This is reflected in the relatively poor predictivity of the model ($r_{\text{pred}} = 0.65$). The additive model explains 85% of the variance and has high predictivity ($r_{\text{pred}} = 0.80$). The z-scales-based model explains 67% of the variance and exhibits a modest predictivity ($r_{\text{pred}} = 0.71$). A series of models containing combinations of global and local descriptors, with and without variable selection, were also created (data not shown). These showed improvements in explained variance r^2 but the models produced were overfitted with reduced predictivity: $r^2 = 0.90$ and $r_{\text{pred}} = 0.53$ for the best model.

Previous studies on newly designed peptides using the additive method showed [16] that the experimental pBL_{50}s were higher than the predicted ones, suggesting a possible synergism between the amino acids in the peptide. One possibility is that this synergism results from global properties of the whole peptide, but it is clear from the present results that bulk properties are not able to adequately explain the observed co-operatively. Although the amino acids in a peptide act as a good team, their own separate contributions to the affinity are of primary importance for the whole peptide binding affinity to MHC molecule. Although the peptide binding affinity is more than a simple sum of the independent contributions of the amino acids at each position, it is inappropriate to consider the peptide acting as a single, undifferentiated molecule. Position dependent interactions are important. However, the combination of global and local models, though overfitted, suggests that bulk properties cannot be overlooked either. Properties, such as overall hydrophobicity, probably are important, but the current limitations in our parameterisation of the local models do not afford us the opportunity to properly partition contributions to affinity between bulk and amino acid-based descriptors. Clearly, the peptide acts as a complicated ensemble of multiple amino acids each mutually potentiating each other in their interaction with the MHC molecule.

In conclusion, the comparison between QSAR models, including global or local descriptors, made in this study, which explain the binding affinity of peptides binding to HLA-A*0201 molecule, showed that the local-descriptors-based models have better explanatory and predictive ability. These local models are more useful, than those based on global properties, in the continuing search for good binders to MHC molecules.

Acknowledgements

This work was supported by GlaxoSmithKline, Medical Research Council, Biotechnology and Biological Sciences Research Council, and UK Department of Health.

References

1. Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W.M., Melief, C.J., Oseroff, C., Yuan, L., Ruppert, J., Sidney, J., Delguercio, M.F., Southwood, S., Kubo, R.T., Chestnut, R.W., Grey, H.M. and Chisari, F.V., J. Immunol., 153 (1994) 5586.
2. Krebs, S. and Rognan, D., Pharm. Acta Helv., 73 (1998) 173.
3. Pichler, W.J., Toxicology, 181–182 (2002) 49.
4. Griffith, J.P., Kim, J.L., Kim, E.E., Sintchak, M.D., Thomson, J.A., Fitzgibbon, M.J., Fleming, M.A., Caron, P.R., Hsiao, K. and Navia, M.A., Cell, 82 (1995) 507.
5. Marsh, S.G.E., Bodmer, J.G., Albert, E.D., Bodmer, W.F., Bontrop, R.E., Dupont, S., Erlich, H.A., Hansen, J.A., Mach, B., Mayr, W.R., Parham, P., Petersdorf, E.W., Sasazuki, T., Schreuder, G.M.T., Strominger, J.L. Svejgaard, A. and Terasaki, P.I., Eur. J. Immunogen., 28 (2001) 377.
6. Bodmer, J., Ciba Found. Symp., 197 (1996) 233.
7. Peoples, G.E., Goedegebuure, P.S., Smith, R., Linehan, D.C., Yoshino, I. and Eberlein, T.Y., Proc. Natl. Acad. Sci. USA, 92 (1995) 432.
8. McMichael, A.J., Parham, P., Brodsky, F.M. and Pilch, J.R., J. Exp. Med., 152(Suppl. 2) (1980) 195.
9. Rongcun, Y., Salazar-Onfray, F., Charo, J. Malmberg, K.-J., Evrin, K., Maes, H., Hising, C., Petersson, M., Larsson, O., Lan, L., Appella, E., Sette, A., Celis, E. and Kiessling, R., J. Immunol., 163 (1999) 1037.
10. Rivoltini, L., Kawakami, Y., Sakaguchi, K., Southwood, S., Sette, A., Robbins, P.F., Marincola, F.M., Salgaller, M.L., Yannelli, J.R., Appella, E. and Rosenberg, S.A., J. Immunol., 154 (1995) 2257.
11. Saper, M.A., Bjorkman, P.J. and Wiley, D.D., J. Mol. Biol., 219 (1991) 277.
12. Madden, D.R., Garboczi, D.N. and Wiley, D.C., Cell, 75 (1993) 693.
13. Ruppert, J., Sidney, J., Celis, E., Kubo, R.T., Grey, H.M. and Sette, A., Cell, 74 (1993) 929.
14. Madden, D.R., Annu. Rev. Immunol., 13 (1995) 587.
15. Doytchinova, I.A., Blythe, M.J. and Flower, D.R.J., Proteome Res., 1 (2002) 263.
16. Doytchinova, I.A., Walshe, V.A., Jones, N.A., Gloster, S.E., Borrow, P. and Flower, D.R., J. Immunol., 172 (2004) 7495.
17. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S., J. Med. Chem., 41 (1998) 2481.
18. SYBYL 6.9. Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
19. MDL QSAR 2.2. 14600 Catalina St. San Leandro CA 94577.
20. Stuber, G., Modrow, S., Hoglund, P., Franksson, L., Elvin, J., Wolf, H., Karre, K. and Klein, G., Eur. J. Immunol., 22 (1992) 2697.
21. Lopes, A.R., Jaye, A., Dorrell, L., Sabally, S., Alabi, A., Jones, N.A., Flower, D.R., Groot, A.De, Newton, P., Lascar, R.M., Williams, I., Whittle, H., Bertoletti, A., Borrow, P. and Maini, M.K., J. Immunol., 171 (2003) 307.
22. Chen, Y., Sidney, J., Southwood, S., Cox, A.L., Sakaguchi, K., Henderson, R.A., Appella, E., Hunt, D.F., Sette, A. and Engelhard, V.H., J. Immunol., 152 (1994) 2874.
23. Daylight .Theory Manual at http://www.daylight.com/release/f_manuals.html.
24. Hall, L.H. and Kier, L.B., In Devillers, J. (Ed.), Methods for QSAR Modeling, Gordon and Breach, London, 1999, pp.
25. Hall, L.H. and Kier, L.B., In Devillers, J. (Ed.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, London, 1999, pp.
26. Hellberg, S., Sjöström, M. and Wold, S., Acta Chem. Scand., B40 (1986) 135.
27. Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S. and Andrews, P., Int. J. Peptide Protein Res., 37 (1991) 414.
28. Leardi, R., Boggia, R. and Terrile, M., J. Chemometrics, 6 (1992) 267.
29. SIMCA-P 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.
30. Falk, K., Rötzschke, O., Stefanovic, S., Jung, G. and Rammensee, H.-G., Nature,, 351 (1991) 290.