Available online at www.sciencedirect.com
SCIENCE DIRECT®
Methods 34 (2004) 444–453
METHODS
www.elsevier.com/locate/ymeth

Quantitative structure–activity relationships and the prediction of MHC supermotifs

Irini A Doytchinova, Pingping Guan, Darren R Flower*

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

Accepted 21 June 2004

Abstract

The underlying assumption in quantitative structure–activity relationship (QSAR) methodology is that related chemical structures exhibit related biological activities. We review here two QSAR methods in terms of their applicability for human MHC supermotif definition. Supermotifs are motifs that characterise binding to more than one allele. Supermotif definition is the initial in silico step of epitope-based vaccine design. The first QSAR method we review here—the additive method—is based on the assumption that the binding affinity of a peptide depends on contributions from both amino acids and the interactions between them. The second method is a 3D-QSAR method: comparative molecular similarity indices analysis (CoMSIA). Both methods were applied to 771 peptides binding to 9 HLA alleles. Five of the alleles (A*0201, A*0202, A*0203, A*0206 and A*6802) belong to the HLA-A2 superfamily and the other four (A*0301, A*1101, A*3101 and A*6801) to the HLA-A3 superfamily. For each superfamily, supermotifs defined by the two QSAR methods agree closely and are supported by many experimental data.

Keywords: Superfamilies; Supermotifs; HLA-A2; HLA-A3; QSAR; Additive method; CoMSIA; Peptides; MHC; Binding affinities

1. Introduction

The T cell, a specialised immune cell mediating cellular immunity, patrols the body searching for antigen-derived epitopes, peptide fragments from pathogenic proteins that are delivered to the cell surface by major histocompatibility complex (MHC)[1] molecules [1]. T cells recognise these complexes and kill infected cells. There are two classes of MHC molecules. Class I MHC molecules deliver peptides originating in the cytosol and are recognised by CD8+ T cells. Class II MHC molecules deliver peptides originating in the vesicular system and are recognised by CD4+ T cells.

The major part of the class I MHC molecule is formed by a transmembrane heavy chain of 44 kDa folded into 3 domains α1, α2, and α3 [2]. α1 and α2 form the peptide-binding domain, containing the peptide-binding groove and the site of interaction with T cell receptors [3,4]. Although not all nine amino acids interact strongly with the binding site, all of them contact it [5]. X-ray data indicate that the MHC peptide-binding site has a 30 Å long solvent accessible surface [6], within which six pockets (A to F) have been described. Certain pockets are non-polar and make hydrophobic contacts. Others contain polar atoms and could hydrogen bond to bound peptides. Six peptide residues fall into these pockets: they are defined as primary (positions 2 (P2) and 9 (P9)) and secondary (positions 1 (P1), 3 (P3), 6 (P6), and 7 P(7)) anchor positions. The remaining three amino acids (peptide positions 4 (P4), 5 (P5), and 8 (P8)) are solvent accessible and can interact with T cell receptors. They are able to affect MHC-binding affinity in several

* Corresponding author. Fax: +44-0-1635-577901.
E-mail address: darren.flower@jenner.ac.uk (D.R Flower).
[1] *Abbreviations used:* MHC, major histocompatibility complex; TCR, T cell receptor; QSAR, quantitative structure–activity receptor.

1046-2023/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.ymeth.2004.06.007

ways: through direct non-bonded interactions with the MHC, by causing conformational changes in anchor residues, and by altering dynamic properties of the whole peptide.

Sequence analysis has shown the peptide domains α1 and α2 to be polymorphic. Twenty residues are the most variable [7]. Most of these residues contact the peptide, giving MHCs a broad specificity and allowing them to bind a wide variety of peptides [8]. Sette et al. [9] grouped class I alleles into superfamilies based on the overlap between their binding motifs (supermotifs). Four superfamilies are known: HLA-A2 [10], HLA-A3 [11], HLA-B7 [12], and HLA-B44 [9]. Supermotif identification has direct practical implications in epitope-based vaccine development for the prevention of infectious diseases and cancer. Epitope identification is the initial step in the design of epitope-based vaccine and often begins with an in silico motif search.

We review here the application of quantitative structure–activity relationship (QSAR) methods to the definition of A2 and A3 supermotifs. Previously, we applied QSAR methods to peptides binding to HLA-A*0201 allele [13,14] and now apply them to peptides binding to the HLA-A2 and A3 supertypes and define revised A2- and A3-supermotifs [14–18]. The HLA-A2 family is the largest and most diverse allele family at the HLA-A locus, consisting of 55 alleles and is common in all ethnic groups [19–21]. Within the HLA-A2 family, the most frequent alleles are A*0201, A*0202, A*0203, A*0206, and A*6802. These alleles differ by 1–7 amino acids and these sequence differences alter the peptide-binding selectivity of the different A2 alleles. The HLA-A3 supertype covers 44% of the human population and includes 5 main alleles: A*0301, A*1101, A*3101, A*3301, and A*6801 [19]. The supermotif for this supertype is characterised by a hydroxyl containing (Ser or Thr) or hydrophobic (Leu, Ile, Val or Met) residue at P2 and a positively charged amino acid (Arg or Lys) at the C-terminus [19].

2. Quantitative structure–activity relationship methods

The underlying assumption in QSAR methodology is that related chemical structures exhibit related biological activities [23–25]. To derive such relationships, chemical structure must be translated to a quantitative description, followed by mathematical modelling that can relate this structural description to the observed biological activity. The aims of QSAR modelling are to predict the activity of untested compounds, to gain an understanding of which chemical descriptors have a large influence on the activity, and how one can use this information to optimise chemical structure.

The extant QSAR literature has seen a great accumulation of different structural descriptors and mathematical methods since the early 1960s, when Hansch et al. [26] first used physicochemical properties and statistical methods in QSAR studies [27]. The biological effect is seldom dependent on just a single factor, and so multiple linear regression (MLR) has been used to investigate this multidimensional problem. MLR frequently exhibits a low predictivity for models including more variables than compounds. An alternative to MLR for "short and fat" matrices is partial least squares regression in latent variables, more commonly known as PLS, in combination with cross-validation. PLS handles data matrices with more variables than observations very well, and the data can be highly collinear and noisy. In such cases, conventional statistical methods, such as MLR, or artificial intelligence techniques produce formulae that fit the training data well but are unreliable in prediction. PLS forms new X variables (called principal components), as linear combinations of the old ones, and then uses them to predict biological activity [28].

Cross-validation (CV) is a powerful method for testing model predictivity. It has become a standard in PLS analysis and can be found in all available PLS software [28]. CV is performed by splitting the data into several groups, developing parallel models from the reduced data with some of the groups omitted, and then predicting the activity of excluded compounds. When the number of omitted groups is equal to the number of compounds, the procedure is called "leave-one-out" (LOO-CV). Predictivity is expressed by the cross-validated coefficient q^2 the standard error of prediction (SEP), and the mean absolute error (MAE):

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

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

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

where p is the number of omitted peptides, $p\mathrm{IC}_{50\,\mathrm{pred}}$ is the experimental value and $p\mathrm{IC}_{50\,\mathrm{pred}}$ is the cross-validated predicted. The explained variance r^2, standard error of estimate (SEE), and the F ratio assessed the non-cross-validated models.

We review two QSAR methods—one 2D and one 3D—in terms of their suitability for supermotif definition. Peptide sequences and corresponding binding affinities were extracted from JenPep database [29] (http://www.jenner.ac.uk/Jenpep). All peptides were nonamers and some bound more than one allele. The binding affinities (IC_{50}), which range over almost 5 log orders of magnitude (from 4.5 to 9.0), were originally assessed by a quantitative assay based on the inhibition of a radiolabelled standard peptide to solubilised MHC

molecules [30,31]. Negative logarithms of IC_{50} values were used as they relate to free energy of binding changes [32]. SYBYL6.7 was used for molecular modelling, PLS and CV [33].

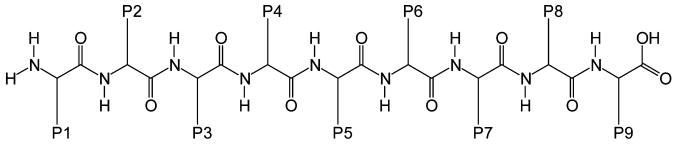
3. Additive method

The additive method is a 2D-QSAR method, based on Free–Wilson's concept [34] that each substituent makes an additive and constant contribution to biological activity regardless of variation in the rest of the molecule. Parker's hypothesis [35,36] that each amino acid side-chain binds independently of the rest of the peptide (IBS hypothesis) is the immunological counterpart to this concept. We extended Free–Wilson's concept to include terms that account for interactions between peptide side chains. Because the twisted conformation of binding peptide allows adjacent and every second side-chain to interact, we only include 1–2, and 1–3 interactions as contributions to affinity. Thus, the binding affinity of a nonamer peptide is represented by Eq. (1):

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

where $p\mathrm{IC}_{50}$ is the measured affinity, the const accounts for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino-acid contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of 1–2 side-chain interactions, and $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of 1–3 side-chain interactions. The data flow for the additive method is presented in Fig. 1. The nine residue peptide sequence is transformed into a row of 6180 terms. Amino-acid contributions account for 180 columns (20 aa × 9 positions), 1–2 interactions for 3200 columns (20 × 20 × 8) and 1–3 interactions for 2800 columns (20 × 20 × 7). A term is equal to 1 when a certain amino acid or a certain side-chain interaction exists, and 0 otherwise. Matrices with 6180 columns and rows equal in number to peptides in the set were generated. Columns containing only 0s were omitted. To deal with these huge, sparse matrices PLS was used.

Two types of models were created: one based solely on the amino acid contributions (amino acids model:



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

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

PLS

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

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

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

LOO-CV

pIC_{50} *pred*

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

Fig. 1. Data flow in the additive method.

AAM) and another based on both amino acid contributions and amino-acid interactions (amino acids and interactions models: AAIM) [16,18]. According to the q^2 values, the AAMs are more predictive than the AAIMs. This is because certain interactions occur only once. In cross-validation, they appear as missing terms in the equation used for affinity prediction. Prediction error is proportional to the number of missing terms. Missing terms in AAMs are less frequent and so their predictivity is higher. In contrast, r^2 was slightly lower for the single amino acid models than for the AAIM. The decrease in r^2 shows that the amino acid side-chain interactions are important for the explanation of variance and should be considered in the modelling of the binding process. The statistical parameters for these models are collected in Table 1.

Amino acids with contributions greater than 0.2 were considered as preferred for a particular allele at the specific position and those with contributions lower than −0.2 were considered as deleterious. Residues identified as preferred for two or more A2/A3-alleles without being deleterious for others were considered as preferred. Residues identified as deleterious for two or more alleles were considered as deleterious in the common motif. The supermotifs defined by the additive method are given in Fig. 2

4. Comparative molecular similarity indices analysis

3D-QSAR methods provide a powerful combination of rigorous statistical analysis, an understandable molecular description, unambiguous graphical display of results [32,37], and can produce accurate quantitative predictions. The CoMSIA method, a key 3D-QSAR technique, correlates ligand similarities with binding affinity changes [38–40], using 3D fields describing steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor properties. Each field can be visualised in 3D maps, which show regions around the superimposed ligand series where the presence or absence of a particular physicochemical property will increase or decrease

Table 1
Statistics of the additive models

Model	n	q^2	NC	SEP	r^2	SEE	F	MAE
HLA-A2 superfamily								
A*0201								
AAM[a]	335	0.377	6	0.694	0.731	0.456	148.66	0.546
AAIM[b]	340	0.337	5	0.726	0.898	0.285	588.88	0.573
A*0202								
AAM	69	0.317	9	0.606	0.943	0.193	109.10	0.546
AAIM	68	0.283	2	0.621	0.748	0.368	96.65	0.511
A*0203								
AM	62	0.327	6	0.841	0.963	0.197	239.30	0.652
AAIM		<0.300						
A*0206								
AAM	57	0.475	6	0.576	0.989	0.085	728.52	0.443
AAIM		<0.300						
A*6802								
AAM	46	0.500	7	0.647	0.983	0.119	313.30	0.517
AAIM		<0.300						
HLA-A3 superfamily								
A*0301								
AAM	72	0.436	6	0.680	0.959	0.181	246.90	0.504
AAIM	70	0.305	4	0.699	0.972	0.136	557.37	0.527
A*1101								
AAM	62	0.458	2	0.572	0.829	0.321	143.00	0.507
AAIM	62	0.428	3	0.593	0.977	0.119	821.10	0.467
A*3101								
AAM	30	0.482	3	0.710	0.892	0.325	71.36	0.502
AAIM	31	0.453	6	0.727	0.990	0.098	399.96	0.602
A*6801								
AAM	38	0.531	4	0.594	0.959	0.175	194.85	0.418
AAIM	37	0.370	4	0.664	0.974	0.136	297.48	0.485

[a] AAM, amino acids model.
[b] AAIM, amino acids and interactions model.

A

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

B

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

C

Preferred	S M	I T	F	F R Q	-	S	F I	R L Y	R
Position	1	2	3	4	5	6	7	8	9
Deleterious	A L Q	N	L	S	G H S	-	-	K S E	Y

Fig. 2. Supermotifs defined by the additive method. (A) A2 supermotif, based on A*0201, A*0202, A*0203, A*0206, and A*6802 alleles. (B) A2 supermotif, based on A*0201, A*0202, A*0203, and A*0206 alleles; (C) A3 supermotif, based on A*0301, A*1101, A*3101, and A*6801 alleles.

affinity. Initially, we applied CoMSIA to peptides binding to the HLA-A*0201 allele and obtained a model with good predictivity [14]. We have recently extended our treatment to include HLA-A2 and HLA-A3 superfamilies [15,17]. By comparison of the favoured and disfavoured areas, in the five property maps, we identify a set of common features which allows detailed supermotifs to be defined. In contrast to other studies, our supermotifs cover all the nine positions of the binding nonamer peptides.

Molecular modelling and QSAR calculations were performed on a Silicon Graphics octane workstation using the SYBYL 6.7 molecular modelling software [33]. The X-ray structure of the nonameric viral peptide TLTSCNTSV [5] bound to HLA-A*0201 molecule was used as a starting conformation for all alleles as there are no X-ray peptide-binding data for other A2-and A3-supertype alleles. The structures were subjected to fully geometry optimisation using the standard Tripos molecular mechanics force field. The peptide backbone was fixed in the X-ray conformation and kept as an aggregate. The partial atomic charges were computed using the AM1 semiempirical method [41] available in the MOPAC program. MOPAC V6 was used as implemented in SYBYL. Single-point calculations were performed. Molecular alignment was based on the backbone atoms, as defined as an aggregate in the optimisation process.

CoMSIA studies were undertaken using the QSAR option of SYBYL. Five physicochemical properties (steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor) were evaluated, using a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, and hydrogen-bond donor and acceptor properties +1. The grid extended beyond the molecular dimensions by 2.0 Å in all directions. Different resolution steps were tested: from 1.0 to 4.0 Å in steps of 0.5 Å. Different column filterings σ_{min} (from 0.0 to 4.0 in steps of 0.5) and attenuation factors α (from 0.1 to 0.8 in steps of 0.1) were also analysed. The predictive power of the models was assessed by the cross-validated coefficient q^2, the standard error of prediction (SEP), and the mean absolute error (MAE).

The number of components (NC) with the highest q^2 and the lowest SEP defined the optimum NC used for non-validated modes. These were used to display coefficient contour maps. When each field makes a statistical contribution to the predicted binding affinity, the model that combines all fields provides the fullest insight. In

Table 2
Statistics of CoMSIA models

Model	n	q^2	NC	SEP	r^2	SEE	F	MAE
HLA-A2 superfamily								
A*0201	236	0.683	7	0.443	0.891	0.260	265.08	0.340
A*0202	63	0.534	8	0.509	0.935	0.190	97.20	0.393
A*0203	60	0.621	6	0.595	0.966	0.179	247.30	0.434
A*0206	54	0.523	12	0.505	0.991	0.071	363.76	0.443
A*6802	45	0.385	4	0.652	0.944	0.197	168.15	0.519
HLA-A3 superfamily								
A*0301	69	0.486	6	0.629	0.959	0.177	241.82	0.585
A*1101	59	0.496	8	0.588	0.972	0.141	167.67	0.443
A*3101	30	0.700	4	0.551	0.921	0.282	73.18	0.179
A*6801	39	0.430	5	0.674	0.950	0.119	126.22	0.516

the present study all fields were significant, so only an all fields model was considered. The statistics of the models are presented in Table 2.

Areas identified as favoured for two or more A2/A3-supertype molecules, without being disfavoured for any molecule, may be considered as preferred for the supermotif. Areas identified as disfavoured for two or more molecules can be considered as deleterious in the common motif. The supermotifs defined by CoMSIA are shown in Fig. 3.

5. HLA-A2 supermotif

There are two A2-supermotifs already defined in the literature. The oldest is Sette's "L_2V_9" [42] based on preferred aliphatic residues at primary anchors P2 and P9. P1, P3, P6, and P7 are considered as secondary anchors [30,6]. Recently, the "L_2V_9" supermotif was extended to include commonly preferred aromatic residues at secondary anchor P1 and an aversion to charged residues (Arg, Lys, Asp, and Glu) at P3 [43]. It was also noted that preferences at P2 for the A*6802 allele and the set of A*02 alleles were different: Val and Thr are preferred for A*6802, Leu and Met for A*0201 and A*0203, and Gln for A*0202 and A*0206 [44].

The definitions given by the additive and CoMSIA methods are shown in Figs. 2 and 3, respectively [15,18]. According to the additive method, Phe is the only preferred amino acid for P1 in the A2 supermotif. Lys is preferred for all alleles except for A*6802, where it is deleterious. CoMSIA indicates that hydrophobic aromatic amino acids at P1 are preferred for the A2-supermotif.

The most striking difference in amino acid preferences defined by the additive method is at P2. Preferences at this position for hydrophobic aliphatic residues (Leu, Met, and Val) are well known [5,6,30,35,45–49]. However, our results indicate that Leu and Met are only preferred for A*0201, A*0202, and A*0203. Leu is deleterious for A*6802 and Met is deleterious for A*0206 and A*6802. Val and Thr are preferred for A*6802. CoMSIA shows that hydrophobic bulky side chains at P2 are preferred for the A2-supermotif. This preference was defined on the basis of the preferences for A*0201, A*0202, and A*0203 alleles, P2 for A*0206 and A*6802 remains silent—neither favoured nor disfavoured areas exist here.

Ile is the only preferred amino acid at P3 and Thr is the only deleterious one. Leu and Val are preferred for A*02 alleles but are deleterious for A*6802. Non-hydrogen-bond forming amino acids are preferred at this position according to CoMSIA. P4 is solvent-exposed and may form contacts with the TCR [6]. Gly is preferred here. Thr is preferred for all A2 alleles except for A*6802. CoMSIA maps indicate that small aliphatic H-bond donors are well accepted at P4.

Leu is preferred at P5 for all A2-supertype molecules except for A*6802 where it makes a negative contribution. Ile is preferred for A*0202 and A*0206 with insignificant contributions for the other alleles. Trp is deleterious for three of the five alleles. The only CoMSIA requirement for the residue at P5 is for it to be aliphatic. Ile and Leu are preferred at P6 and Ser is deleterious. Hydrophobic and bulky substituents are required according to CoMSIA.

For affinity to A2-supertype molecules Ile is preferred at P7. CoMSIA requires small and aliphatic substituents here. Phe is a preferred residue at P8 and Asp is deleterious for four of the five A2 molecules. CoMSIA shows preferences for small, aliphatic, hydrophilic, and hydrogen-bond forming amino acids here.

There is better agreement at P9 between the preferences of different alleles. Val is favoured here, while Ala is deleterious. The uncharged side-chain of Tyr116 occupies the end of the pocket F causing the binding site to be complementary to small hydrophobic side chains [5,8]. CoMSIA indicates that hydrophobic and bulky substituents increase affinity at P9.

Our studies cast doubt on whether the A*6802 allele should be part of the A2-supertype. Sequence comparison shows one or two differences in the residues forming the 6 pockets of A*0201, A*0202, A*0203, and A*0206

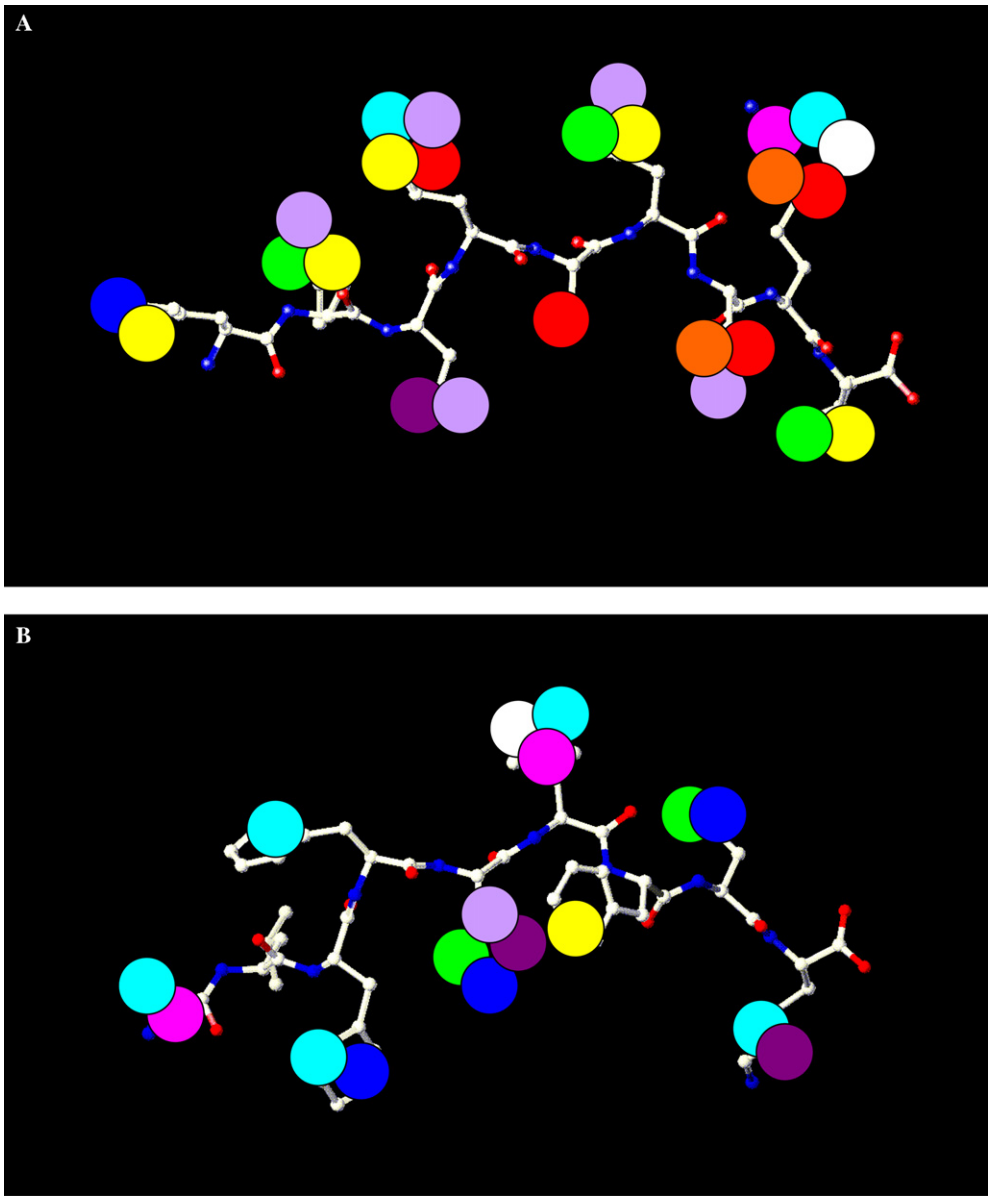


Fig. 3. Supermotifs defined by CoMSIA. (A) A2 supermotif, based on A*0201, A*0202, A*0203, A*0206, and A*6802 alleles. Peptide FLLLADARV is shown inside the fields, N terminal is positioned on the left side. (B) A3 supermotif, based on A*0301, A*1101, A*3101, and A*6801 alleles. Peptide ALFFIIFNK is shown inside the fields, N terminal is positioned on the left side. Legend: steric bulk favoured; steric bulk disfavoured; negative electrostatic potential favoured; Negative electrostatic potential disfavoured; hydrophobicity favoured; hydrophobicity disfavoured; H-bond donor group favoured; H-bond donor group disfavoured; H-bond acceptor group favoured; and H-bond acceptor group favoured.

molecules. Differences between A*6802 and A*02 molecules number seven: of which five concern pockets A, B, and C. These are so substantial that they alter the amino acid preferences at the primary anchor P2 and the secondary anchors P1 and P6. Preferences for Val and Thr at P2 bring the A*6802 allele close to the A3-supertype [19], but the A3 supermotif requires positively charged residues at the C-terminus [50]. This is not true for A*6802, suggesting that it is intermediate between A2 and A3 supertypes: at P2 it is closer to A3 and at P9 it is nearer to A2. Excluding A*6802 allele, the range of preferred and deleterious amino acids expands. (Fig. 2). The expansion concerns all positions but especially P2, with one to three new additions to each position's preferred and deleterious amino acids.

6. HLA-A3 supermotif

P2 and P9 are generally accepted to be primary anchors for the A3 superfamily [51–54]. The peptide side-chain at P2 falls into pocket B and the C-terminus is

buried in pocket F [8,55,56]. Peptides usually have a positively charged residue Arg or Lys at P9 and a variety of hydrophobic residues at P2. A peptide-binding motif for the HLA-A3 superfamily has been defined previously by Sidney et al. [50] and Rammensee et al. [57]. The supermotif defined in our studies, while in good agreement with previous supermotifs, is more extensive, covering all the nine positions that contact the MHC molecule [16,17].

P1 is a secondary anchor position. According to the additive method Ser and Met are preferred here. CoMSIA shows that the common favoured property for P1 was hydrogen-bond donor ability.

Although there was wide variation in preferences at P2, Ile and Thr were found to be preferred for two of the alleles without being deleterious for the other two. In the CoMSIA study, it was found that the steric bulk was favoured at P2 for A*3101 and A*0301 but disfavoured in A*6801 and A*1101. A great variety in the electron density, hydrophobicity and hydrogen bond acceptance maps existed at this position, which was consistent with the broad spectrum of amino acids accommodated here. This is explained by the polymorphism of residues forming pocket B. Phe9 in A*0301 is substituted to Tyr9 in A*6801 and A*1101, and to Thr9 in A*3101 [22]. The hydroxyl group of Tyr9 points towards the inside of the pocket and prevents larger amino acids from reaching the bottom of the pocket [58]. Because of this, larger residues like Leu are deleterious for A*6801 and A*1101 but are preferred for A*0301. The change from Glu63 to Asn63 in A*6801 and A*1101 also changes the conformation of the pocket and stops large amino acids from binding [56]. A previous study of pocket B revealed Val67 was reoriented in A*6801 and affected amino acid selection [59].

P3 prefers the hydrophobic residue Phe. Sidney and co-workers [50] found that peptides with aromatic residue, like Tyr, Phe, and Trp, had a 31 fold increase in binding affinity to A*0301. The electron density at P3 is preferred for three of the alleles. Phe, Arg, and Tyr are favoured at P4. Electron density and hydrogen-bond ability are important here.

No amino acid is favoured at P5. Ser, Gly, and His are disfavoured here. This position as well as P6 is not particularly important in determining the affinity of peptide binding but may participate in T cell recognition. CoMSIA indicates that bulky side chains with high electron density are preferred at P5. Hydrophilic amino acids capable of forming hydrogen bonds are well accommodated at P6, which is in good agreement with the preferred Ser defined by the additive method.

P7 is another secondary anchor position [57]. Hydrophobic residues are preferred here. Phe and Ile are strongly preferred by A*0301 and A*1101. Peptide-binding studies showed either P3 or P7, together with residues at P2 and P9, induced stable binding of the peptide [50]. Arg, Tyr, and Leu were slightly favoured at P8, while Ser, Lys, and Glu were deleterious. CoMSIA suggests that steric bulk is disfavoured here but negative electrostatic potential is well accepted.

Positively charged amino acid Arg is the common preferred amino acid at P9. A*6801 and A*3101 preferred Arg, A*1101 favoured the smaller residue Lys, while A*0301 accepted both. Tyr was deleterious at P9, possibly because its aromatic ring was too large for the pocket. CoMSIA shows that the most important property here is hydrogen-bond donor ability. It is favoured by A*6801 and A*3101, and was disfavoured by A*1101.

7. Concluding remarks

The development of poly-epitope vaccines is likely to prove of great utilitarian value in the treatment and prevention of autoimmune and infectious disease, allergy, and cancer. Informatics techniques have much to offer in this regard. We have pioneered the explicit application of QSAR techniques to problems in immunobiology [13–18,29,44,60–63]. These methods are particularly appropriate for the quantitative assessment of peptide MHC interaction.

We extended our additive method to examine peptides binding to MHC class II molecules [62] and shall widen its application to other MHC alleles in the future. We are also using in-house experimental cell surface stabilisation assays to test out the predictivity of our modelling approach [64] and to this end we are designing and testing synthetic super-binding peptides as well as developing comprehensive models for poorly characterised alleles using experimental design. Furthermore, we are also applying QSAR techniques to the iterative optimisation of heteroclitic peptides as potential cancer vaccines [64]. Heteroclitic, or altered, peptide ligands are, generally, mutated peptides whose substitutions lead to increased MHC-binding affinity, and which, often, also exhibit an enhanced T cell response. To make our methods publicly accessible, all the models derived by the additive method have been incorporated into a freely available MHC-binding prediction program: MHCPred (URL:http://www.jenner.ac.uk/MHCPred) [63].

Motif definition is a requirement for the initial in silico step of epitope identification [21]. The more precisely a motif is defined the greater the accuracy of epitope prediction. In this sense, a quantitative prediction is more helpful than a qualitative one. Defining a supermotif also allows identification of promiscuous epitopes that can bind several alleles. Identification of such promiscuity is a powerful extension to established epitope-based vaccine design. The good agreement between the supermotifs defined by different QSAR methods shows that these methods are reliable tools in epitope-based

vaccine research. The application of other QSAR methods and techniques for vaccine research is work in progress in our laboratory. It is our expectation that QSAR will influence as strongly the search for new vaccines as it has the design and discovery of new drugs.

Acknowledgments

We thank Debra Taylor, Helen McSparron, Christianna Zygori, and Martin Blythe for their help with Jenpep. Valerie Walshe for her work on peptide-binding assays. We thank Dr Vladimir Brusic, Dr Persephone Borrow, and Prof Peter Beverley for help and illuminating discussions.

References

[1] C.A. JanewayJr., P. Travers, M. Walport, J.D. Capra (Eds.), Immunobiology, Elsevier Science, London, 1999, pp. 115–162.
[2] A.M. Krensky, C. Clayberger, Int. Rev. Immunol. 13 (1996) 173–185.
[3] M. Takiguchi, Nippon Rinsho. Japan. J. Clin. Med. 52 (1994) 2817–2823.
[4] E.Y. Jones, Curr. Opin. Immunol. 9 (1997) 75–79.
[5] D.R. Madden, D.N. Garboczi, D.C. Willey, Cell 75 (1993) 693–708.
[6] D.R. Madden, Annu. Rev. Immunol. 13 (1995) 587–622.
[7] P. Parham, C.E. Lomen, D.A. Lawlor, J.P. Ways, N. Holmes, H.L. Coppin, R.D. Salter, A.M. Wan, P.D. Ennis, Proc. Natl. Acad. Sci. USA 85 (1988) 4005–4009.
[8] M.A. Saper, P.J. Bjorkman, D.C. Wiley, J. Mol. Biol. 219 (1991) 277–319.
[9] A. Sette, J. Sidney, Curr. Opin. Immunol. 10 (1998) 478–482.
[10] M.A. Altfeld, B. Livingston, N. Reshamwala, P.T. Nguyen, M.M. Addo, A. Shea, M. Newman, J. Fikes, J. Sidney, F. Wentworth, et al., J. Virol. 75 (2001) 1301–1311.
[11] I. Kawashima, V. Tsai, S. Southwood, K. Takesako, A. Sette, E. Celis, Cancer Res. 59 (1999) 431–435.
[12] A.J. Coyle, J.C. Gutierrez-Ramos, Nature 363 (2001) 203–209.
[13] I.A. Doytchinova, D.R. Flower, J. Med. Chem. 44 (2001) 3572–3581.
[14] I.A. Doytchinova, D.R. Flower, Proteins 48 (2002) 505–518.
[15] I.A. Doytchinova, D.R. Flower, J. Comput. Aid. Mol. Des. 16 (2002) 535–544.
[16] P. Guan, I.A. Doytchinova, D.R. Flower, Protein Eng. 16 (2003) 11–18.
[17] P. Guan, I.A. Doytchinova, D.R. Flower, Bioorgan. Med. Chem. 11 (2003) 2307–2311.
[18] I.A. Doytchinova, D.R. Flower, Appl. Bioinform. 1 (2003) 167–176.
[19] J. Sidney, H.M. Grey, R.T. Kubo, A. Sette, Immunol. Today 17 (1996) 261–266.
[20] J.M. Ellis, V. Henson, R. Slack, J. Ng, R.J. Hartzman, C.K. Hurley, Hum. Immunol. 61 (2000) 334–340.
[21] A. Sette, B. Livingston, D. McKinney, E. Appella, J. Fikes, J. Sidney, M. Newman, R. Chesnut, Biologicals 29 (2001) 271–276.
[22] C. Schönbach, J.L.Y. Koh, X. Sheng, L. Wong, V. Brusic, Nucleic Acids Res. 28 (2000) 222–224.
[23] H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 1–13.
[24] P.M. Andersson, M. Sjöström, S. Wold, T. Lundstedt, J. Chemometrics 14 (2000) 629–642.
[25] H. Kubinyi, in: R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman (Eds.), Methods and Principles in Medicinal Chemistry, vol. 1, VCH, Weinheim, 1993.
[26] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, Nature 194 (1962) 178–180.
[27] H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 113–308.
[28] S. Wold, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195–218.
[29] M.J. Blythe, I.A. Doytchinova, D.R. Flower, Bioinformatics 18 (2002) 434–439.
[30] J. Ruppert, J. Sidney, E. Celis, R.T. Kubo, H.M. Grey, A. Sette, Cell 74 (1993) 929–937.
[31] A. Sette, J. Sidney, M.-F. del Guercio, S. Southwood, J. Ruppert, C. Dalberg, H.M. Grey, R.T. Kubo, Mol. Immunol. 31 (1994) 813–822.
[32] T.I. Oprea, C.L. Waller, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 11, Wiley-VCH, New York, 1997, pp. 127–182.
[33] SYBYL 6.7. Tripos Inc, 1699 Hanley Road, St. Louis, MO 63144, 2002.
[34] S.M. Free Jr., J.W. Wilson, J. Med. Chem. 7 (1964) 395–399.
[35] K.C. Parker, M.A. Bednarek, J.E. Coligan, J. Immunol. 152 (1994) 163–175.
[36] K.C. Parker, M. Shields, M. DiBrino, A. Brooks, J.E. Coligan, Immunol. Res. 14 (1995) 34–57.
[37] G. Greco, E. Novellino, Y.C. Martin, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 11, Wiley-VCH, New York, 1997, pp. 183–240.
[38] G. Klebe, U. Abraham, T. Mietzner, J. Med. Chem. 37 (1994) 4130–4146.
[39] G. Klebe, U. Abraham, J. Comput.-Aid. Mol. Des. 13 (1999) 1–10.
[40] M. Böhm, J. Stürzebecher, G. Klebe, J. Med. Chem. 42 (1999) 458–477.
[41] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902–3909.
[42] M.-F. Del Guercio, J. Sidney, G. Hermanson, C. Perez, H.M. Grey, R.T. Kubo, A. Sette, J. Immunol. 154 (1995) 685–693.
[43] J. Sidney, S. Southwood, D.L. Mann, M.A. Fernandez-Vina, M.J. Newman, A. Sette, Hum. Immunol. 62 (2001) 1200–1216.
[44] I.A. Doytchinova, D.R. Flower, Immunol. Cell Biol. 80 (2002) 270–279.
[45] K. Falk, O. Rötzschke, S. Stefanovic, G. Jung, H.-G. Rammensee, Nature 351 (1991) 290–296.
[46] T.J. Kirksey, R.R. Pogue-Caley, J.A. Frelinger, E.J. Collins, J. Biol. Chem. 274 (1999) 37259–37264.
[47] S. Tourdot, A. Scardino, E. Saloustrou, D.A. Gross, S. Pascolo, P. Cordopatis, F.A. Lemonnier, K.A. Kosmatopoulos, Eur. J. Immunol. 30 (2000) 3411–3421.
[48] R.T. Kubo, A. Sette, H.M. Grey, E. Appella, K. Sakaguchi, N.-Z. Zhu, D. Arnott, N. Sherman, J. Shabanowitz, H. Michel, W.M. Bodnar, T.A. Davis, D.F. Hunt, J. Immunol. 152 (1994) 3913–3924.
[49] K.C. Parker, M.A. Bednarek, L.K. Hull, U. Utz, B. Cunningham, H.J. Zweerink, W.E. Biddison, J.E. Coligan, J. Immunol. 149 (1992) 3580–3587.
[50] J. Sidney, H.M. Grey, S. Southwood, E. Celis, P.A. Wentworth, M.-F. del Guercio, R.T. Kubo, R.W. Chesnut, A. Sette, Hum. Immunol. 45 (1996) 79–93.
[51] T.P.J. Garrett, M.A. Saper, P.J. Bjorkman, J.L. Strominger, D.C. Wiley, Nature 342 (1989) 692–696.
[52] M. Matsamura, D.H. Fremont, P. Peterson, I.A. Wilson, Science 257 (1992) 927–934.
[53] K. Falk, O. Rtzschke, Seminar Immunol. 5 (1993) 81–94.
[54] R. Gavioli, M.G. Kurilla, P.O. De Campos-Lima, L.E. Wallace, R. Dolcetti, R.J. Murray, A.B. Rickinson, M.G. Masucci, J. Virol. 67 (1994) 1572–1578.

[55] D.R. Madden, J.C. Gorga, J.L. Strominger, D.C. Wiley, Nature 353 (1991) 321–325.
[56] G. Vasmatzis, C. Zhang, J.L. Cornette, C. DeLisi, Mol. Immunity 33 (1996) 1231–1239.
[57] H.G. Rammensee, T. Friede, S. Stevanovic, Immunogenetics 41 (1995) 178–228.
[58] T. Sudo, N. Kamikawaji, A. Kimura, Y. Date, C.J. Savoie, H. Nakashima, E. Furuichi, S. Kuhara, T. Sasazuki, J. Immunol. 155 (1995) 4749–4756.
[59] H.C. Guo, D.R. Madden, M.L. Silver, T.S. Jardetzky, J.C. Gorga, J.L. Strominger, D.C. Wiley, Proc. Natl. Acad. Sci. USA 90 (1993) 8053–8057.
[60] I.A. Doytchinova, M.J. Blythe, D.R. Flower, J. Proteome Res. 1 (2002) 263–272.
[61] D.R. Flower, I.A. Doytchinova, K. Paine, P. Taylor, M.J. Blythe, D. Lamponi, C. Zygouri, P. Guan, H. McSparron, H. Kirkbride, in: D.R. Flower (Ed.), Drug Design: Cutting Edge Approaches, RSC publications, Cambridge, 2002, pp. 136–180.
[62] I.A. Doytchinova, D.R. Flower, Bioinformatics 19 (2003) 2263–2270.
[63] P. Guan, I.A. Doytchinova, C. Zygouri, D.R. Flower, Appl. Bioinform. 2 (2003) 63–66.
[64] I.A. Doytchinova, V.A. Walshe, N.A. Jones, S.E. Gloster, P. Borrow, D.R. Flower, J. Immunol. 172 (2004) 7495–7502.