Immunology and Cell Biology (2002) **80**, 270–279

Special Feature

Quantitative approaches to computational vaccinology

IRINI A DOYTCHINOVA and DARREN R FLOWER

Edward Jenner Institute for Vaccine Research, Compton, Berkshire, United Kingdom

Summary This article reviews the newly released JenPep database and two new powerful techniques for T-cell epitope prediction: (i) the additive method; and (ii) a 3D–Quantitative Structure Activity Relationships (3D–QSAR) method, based on Comparative Molecular Similarity Indices Analysis (CoMSIA). The JenPep database is a family of relational databases supporting the growing need of immunoinformaticians for quantitative data on peptide binding to major histocompatibility complexes and to the Transporters associated with Antigen Processing (TAP). It also contains an annotated list of T-cell epitopes. The database is available free via the Internet (http://www.jenner.ac.uk/JenPep). The additive prediction method is based on the assumption that the binding affinity of a peptide depends on the contributions from each amino acid as well as on the interactions between the adjacent and every second side-chain. In the 3D–QSAR approach, the influence of five physicochemical properties (steric bulk, electrostatic potential, local hydrophobicity, hydrogen-bond donor and hydrogen-bond acceptor abilities) on the affinity of peptides binding to MHC molecules were considered. Both methods were exemplified through their application to the well-studied problem of peptides binding to the human class I MHC molecule HLA-A*0201.

Key words: 3D–QSAR, additive method, binding affinity prediction, CoMSIA, HLA-A*0201, MHC.

Introduction

One of the principal goals of bioinformatic research within immunology is to develop computer-aided vaccine design, or computational vaccinology, as a practical science applied to the quest for new vaccines. The recognition of antigenic epitopes by the immune system, either small discrete T-cell epitopes or large conformational epitopes recognized by B cells and soluble antibodies, is the key molecular event at the heart of the immune response to pathogens. Within the development of rational vaccine design, the rapid and reliable identification of epitopes, particularly T-cell epitopes, is currently the focus of considerable endeavour.

Within the context of cellular immunology, the immunogenicity of peptides strongly depends on their ability to bind to MHC and to be recognized subsequently by TCR.[1] Traditionally, T-cell epitopes have been identified by examining T-cell responses to overlapping peptides generated from target antigens. This is adequate, if labour intensive, for the study of a single, small protein, but the experimental overhead becomes prohibitive for the study of genomes from large viruses, bacteria or parasites, which may contain thousands, if not tens of thousands, of gene products. The computational analysis of a pathogenic proteome can, through the prediction of peptide binding to MHC proteins, reduce significantly subsequent experimental work. Immunoinformatics, a newly emergent branch of bioinformatics, offers a range of techniques suitable for T-cell epitope searches and predictions. Experimental work suggests that only peptides that bind with high affinity to MHC molecules are recognized as T-cell epitopes.[2] In terms of a competition assay, the IC_{50} (the concentration required for 50% inhibition of a standard labelled peptide by the test peptide) must be less than 500 nmol/L. As only peptides that bind well to MHC can become T-cell epitopes, MHC-binding prediction is a necessary preliminary to the identification of such epitopes. In what follows, we use the terms 'MHC binding' and 'T-cell epitope identification' synonymously. A broad spectrum of predictive methods is currently available.[3] These began with the development of the early motif searching methods,[4,5] and include a variety of ever more sophisticated approaches: peptide-scoring schemes based on the hypothesis for independent binding of side-chains (IBS-hypothesis);[6,7] the artificial neural networks (ANN);[8] free energy scoring function (Fresno);[9] and positional scanning–synthetic combinatorial libraries (PS–SCL).[10,11]

In this article we review our contribution to the rapidly developing field of computational vaccinology, including a discussion of our newly released JenPep database and two powerful new techniques for T-cell epitope prediction. One is a 2D–Quantitative Structure Activity Relationships (2D–QSAR) approach, which we have called the 'additive' method,[12] and the other is a 3D–QSAR approach, based on Comparative Molecular Similarity Indices Analysis (CoMSIA).[13,14]

JenPep

The JenPep database is a family of relational databases supporting the growing needs of immunoinformaticians for quantitative data on peptide binding to MHC and to the TAP peptide transporter, as well as an annotated list of T-cell epitopes.[15] The database, and a hypertext markup language (HTML) interface for searching, is available free via the Internet (http://www.jenner.ac.uk/JenPep).

Correspondence: Dr Darren Flower, Edward Jenner Institute for Vaccine Research, High Street, Compton, Berkshire RG20 7NN, UK. Email: darren.flower@jenner.ac.uk

Received 17 December 2001; accepted 21 January 2002.

The currently available version of JenPep (Version 1.0) is composed of three subdatabases: (i) a compilation of quantitative binding measures for peptides to class I and class II MHC; (ii) a compendium of dominant and subdominant T-cell epitopes; and (iii) a set of quantitative data for peptide binding to the TAP peptide transporter. The T-cell section contains 2300 T-cell epitopes, the MHC binding section contains 6000 peptides and the TAP section covers 400 peptides. JenPep contains binding data on a wide variety of different MHC alleles: for class I MHC molecules, JenPep has data for 68 different restriction alleles with more than 50 genotype variations. For class II MHC molecules there are over 40 restriction alleles with 52 genotype designations. Peptide lengths for class I MHC molecules are in the range of 7–16 residues and for class II MHC molecules are in the range of 9–35 residues. The database itself is a relational system, currently constructed using MS ACCESS and is searchable through a graphical user interface (GUI) built using active server pages (ASP). Together with the peptide sequence, JenPep includes various kinds of binding measures, MHC restriction and, where such data are known, the protein from which the peptide originates. Data on T-cell epitopes are currently limited to a list of binders. While there are many different ways to identify T-cell epitopes, including T-cell killing, proliferation assays such as thymidine uptake, and so on, the quantitative data produced by such assays, are not consistent enough to be used outside of particular experimental conditions. Since there is no real meaning to the idea of a partial T-cell epitope, we have relied on the immunological judgement of experimental immunologists to define what are, or are not, T-cell epitopes. For MHC binding we have used a number of alternative measures of binding affinity, which are currently in common currency. These include radiolabelled[16–18] and fluorescent[19–21] IC_{50} values, BL_{50}[22–24] (half maximal binding level calculated from a mean fluorescence intensity MFI of MHC-expressing RMA-S cells) and SC_{50}[25–27] (the concentration inducing half of the maximal up-regulating effect calculated in a peptide-binding stabilization assay), and half-lives.[6,28,29]

We are actively developing the database beyond its current limitations, and expect to release a much larger and more complete quantitative database in due course. Much useful data are still locked into written, hard-copy literature, presented as tabulated values or in a graphical form. It is an on-going challenge to find and extract these data into a machine-readable format. We also look forward to the day when immunologists submit their experimental binding data to an online archive, such as ours, much as molecular biologists must submit their data to a publicly curated sequence database. JenPep is the first database in immunology to concentrate on quantitative measurements, complementing existing systems. This compilation of binding data underlies our attempts to derive statistically sound QSAR tools for the accurate prediction of peptide binding to immunological molecules.

2D–QSAR method for binding affinity prediction

One of the predictive techniques developed in our group is based on the so-called additivity concept, whereby each substituent makes an additive and constant contribution to the biological activity regardless of variation in the rest of the molecule.[30] The IBS hypothesis, developed by Parker,[6,7,31] is the immunological analogue of this idea. We extended this concept by adding additional terms that account for near-neighbour side-chain interactions.[12] The binding affinity of a peptide will depend on contributions from each amino acid as well as their interactions with adjacent and every second side-chain: (1)

$$\text{binding affinity} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2},$$

where the *const* accounts, at least nominally, for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acids contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, and $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interaction.

Four hundred and twenty IC_{50} values for 340 nonamer peptides were used in the development of the additive method. The peptide sequences and their binding affinities to the HLA-A*0201 molecule were extracted from the JenPep database. Eighty IC_{50} values are higher than 500 nmol/L (low binders), 182 values are between 50 and 500 nmol/L (intermediate binders) and 158 are less than 50 nmol/L (high binders). More than one IC_{50} value was found for some of the peptides. The binding affinities (IC_{50} values) were originally obtained from a quantitative assay, based on the inhibition of binding of a radiolabelled standard peptide to detergent-solubilized MHC molecules.[16,17] As is common practice amongst QSAR practitioners, IC_{50} values were converted to p-units (negative decimal logarithm). Many amino acids are present at a certain position only once. However, by disregarding these single amino acids, one runs the risk of eliminating legitimate predictors. This problem will be minimized as the size of our database increases.

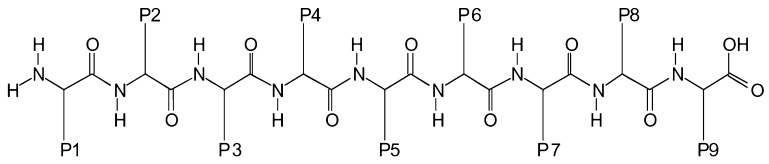
The development of the additive method is described in Fig. 1. A program was developed to transform the nine amino acid peptide sequence into a row of the table presented in Fig. 1. A term is equal to 1 when a certain amino acid at a certain position, or a certain interaction exists, and 0 when they are absent. Thus a matrix of 420 rows and 6180 columns was generated. One hundred and eighty columns account for the contributions of the amino acids (20 amino acids × 9 positions), 3200 for the adjacent side-chains, or 1–2 interactions (20 × 20 × 8), and 2800 for the 1–3 side-chain interactions (20 × 20 × 7). To reduce the number of columns, the program omits columns that contain only zeros. The final matrix consists of 420 rows and 2158 columns. To resolve this matrix requires prohibitive amounts of computer time, and so we divided the equation into three:

$$pIC_{50} = const + \sum_{i=1}^{9} P_i \quad (2)$$

$$pIC_{50} = const + \sum_{i=1}^{8} P_i P_{i+1} \quad (3)$$

$$pIC_{50} = const + \sum_{i=1}^{7} P_i P_{i+2} \quad (4)$$

Equation 3 gives the amino acid contributions, while equations 4 and 5 give the contributions for side-chain interactions. The contributions of the interactions below ±0.030 were omitted and the three equations were combined into one.



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

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……
420	……	……0……1……0	……0……1……0	……0……1……0

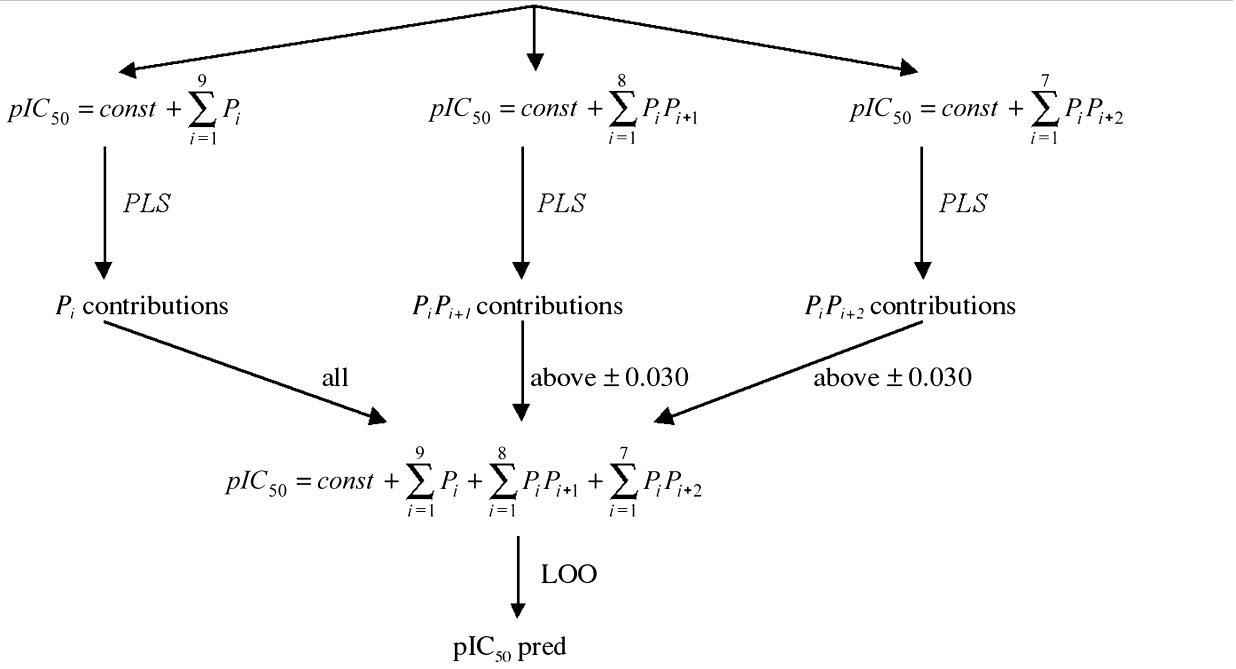


Figure 1 Protocol of the additive method.

As the columns are more numerous than the rows, the equations were solved using partial least square (PLS), as implemented in SYBYL 6.7.[32] pIC_{50} was added as a dependent variable. The predictive power was assessed by the cross-validated q^2 (as generated by 'leave-one-out' cross-validation [LOO-CV]), standard error of predictions (SEP) and residuals between the experimental and predicted by LOO-CV pIC_{50} values. According to the residuals, peptides could be classified into three categories: (i) very well-predicted peptides with |residuals| ≤ 0.5; (ii) well-predicted peptides with |residuals| between 0.5 and 1.0; and (iii) poorly predicted peptides with |residuals| > 1.0. A mean |residual| value and standard deviation for the set was also calculated. The non-cross-validated model was assessed by multiple linear regression (MuLR) parameters: explained variance (r^2), standard error of estimate (*SEE*) and *F* ratio.

The final equation derived by the additive method consists of 1815 terms including the constant. It contains the contributions of the amino acids and the contributions of the significant side-chain interactions. Its LOO-CV and MuRL parameters are given in Table 1. In the cases of multiple pIC_{50} values for one peptide, the $pIC_{50}pred$ were calculated omitting all available $pIC_{50}exp$ values. There were 172 very well-predicted peptides (50.5%), 128 well-predicted peptides (37.5%) and only 41 poorly predicted peptides (12.0%). The contributions of the amino acids at different positions are presented in Fig. 2. The contributions of the more important adjacent or 1–2 side-chain interactions are plotted in Fig. 3a, and the more important 1–3 side-chain interactions are in Fig. 3b.

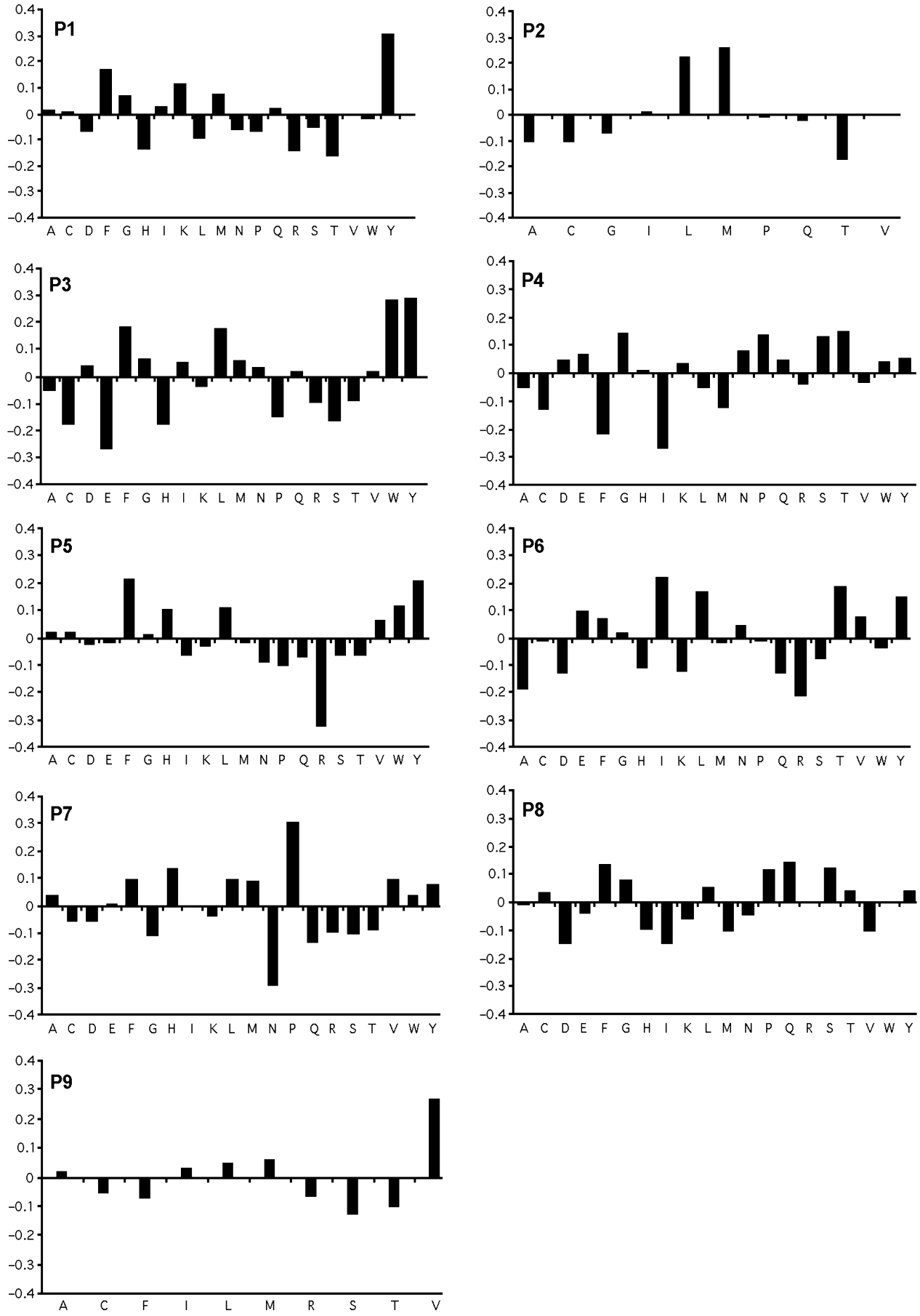


Figure 2 Amino acid contributions to HLA-A*0201 binding affinity according to the additive method.

Table 1 Statistics of the additive and Comparative Molecular Similarity Indices Analysis (CoMSIA) models for HLA-A*0201 binding

Parameter	Additive method	CoMSIA
n	340	266
q^2_{LOO}	0.337	0.683
q^2_{CV5}*	–	0.656
q^2_{LHO}†	–	0.558
NC	5	7
SEP_{LOO}	0.726	0.443
r^2	0.898	0.891
$r^2_{bootstrap}$‡	–	0.924
SEE	0.285	0.260
F-ratio	588.883	265.082
Fractions		
Steric	–	0.145
Electrostatic	–	0.320
Hydrophobic	–	0.210
Hydrogen-bond donor	–	0.161
Hydrogen-bond acceptor	–	0.164
Residuals		
Residual ≤ |0.5|	172 (63.2%)	168 (50.5%)
|0.5| < residual ≤ |1.0|	128 (29.3%)	78 (37.5%)
Residual > |1.0|	40 (7.5%)	20 (12.0%)
Mean |residual|	0.573	0.489
Standard deviation	0.442	0.462

*Mean value of 20 runs; †mean value of 50 runs; ‡mean value of 20 runs; LOO, 'leave-one-out' cross-validation; SEP, standard error of predictions; SEE, standard error of estimate.

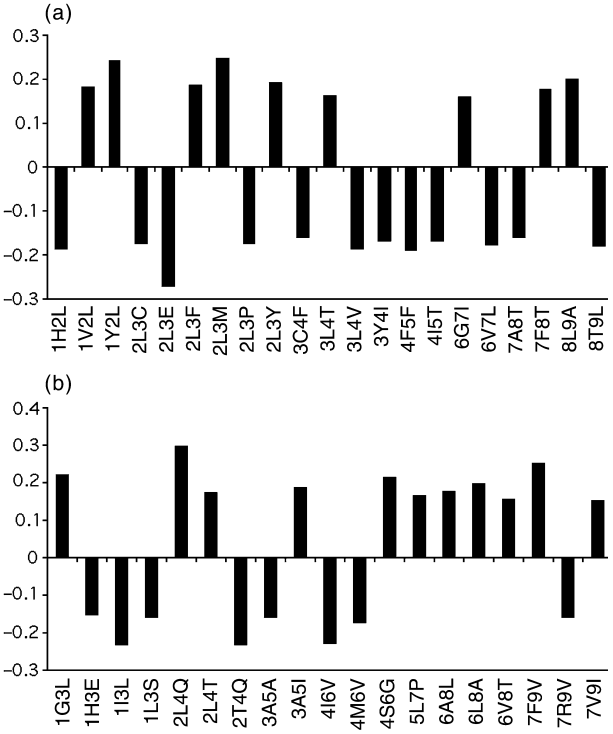


Figure 3 Contributions of some of the more important (a) adjacent; and (b) every second side-chain interactions according to the additive method. The presented contributions are above ±0.15.

3D–QSAR method for binding affinity prediction

One of the most reliable methods for investigating the structure-activity trends within sets of biological molecules is 3D–QSAR.[33,34] The explanatory power of 3D–QSAR methods is considerable, manifest not only in their ability to accurately predict binding affinities, but also in their capacity to display advantageous and disadvantageous 3-D interaction potential mapped onto the structures of molecules being investigated. We have applied the 3D–QSAR method (CoMSIA)[35–38] to gain an understanding of the relationship between physicochemical properties (steric bulk, electrostatic potential, local hydrophobicity, hydrogen-bond donor and hydrogen-bond acceptor abilities) and the affinities of peptides that bind to the MHC molecule HLA-A*0201.[13,14]

Two hundred and sixty-six nonamer peptides are included in the CoMSIA study. Their IC_{50} values were collected from the JenPep database and converted to p-units. All molecular modelling and QSAR calculations were performed on a Silicon Graphics octane workstation using the SYBYL 6.7 molecular modelling software.[32] The X-ray structure of the nonameric viral peptide TLTSCNTSV[39] was used as a starting conformation. The structures of the remaining peptides were built to this conformation. The partial atomic charges used in CoMSIA were computed using the AM1 semiempirical method,[40] available in MOPAC. A sybyl programming language (SPL) script for automatic building, optimization and AM1 calculation of the peptides was created within SYBYL. The program uses a text file containing the peptide sequences and a protein databank (pdb) file of the starting conformation.

Five types of similarity index (steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor) were calculated, using a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, hydrogen-bond donor and acceptor properties +1.[35–38] Since only the combination of all fields provided a complete insight, only an all-fields model was analysed further. The same parameters as for the additive method were used to assess the predictive power of the final model: q^2, *SEP* and residuals (Table 1). Three types of cross-validation were performed: (i) LOO-CV; (ii) CV in five groups; and (iii) CV in two groups. The non-cross-validated model was assessed by r^2, *SEE* and *F*-ratio. A bootstrap analysis[41] was performed in 20 runs and the mean r^2 is given as $r^2_{bootstrap}$. The non-cross-validated model was used to display the coefficient contour maps.

The initial CV model had low q^2 *and* r^2 values. This result was not surprising, given the great diversity of peptides collected from a variety of sources. One hundred and fifty-one were very well-predicted peptides (with |residuals| ≤ 0.5), 83 were well-predicted peptides (with |residuals| between 0.5 and 1.0), and 32 peptides were poorly predicted (with |residuals| > 1.0). The mean |residual| was 0.553. The model was improved by excluding a limited number of poorly predicted peptides in a stepwise manner, beginning with the peptide with the highest residual. The final CV model had significantly higher parameter values: $q^2 = 0.683$ at seven components and $r^2 = 0.891$. This model was used to predict the binding affinities of the excluded peptides. The predictions were better not only for the group of very well-predicted peptides, but also for the group of poorly predicted peptides. The mean |*residual*| value for this model was 0.489.

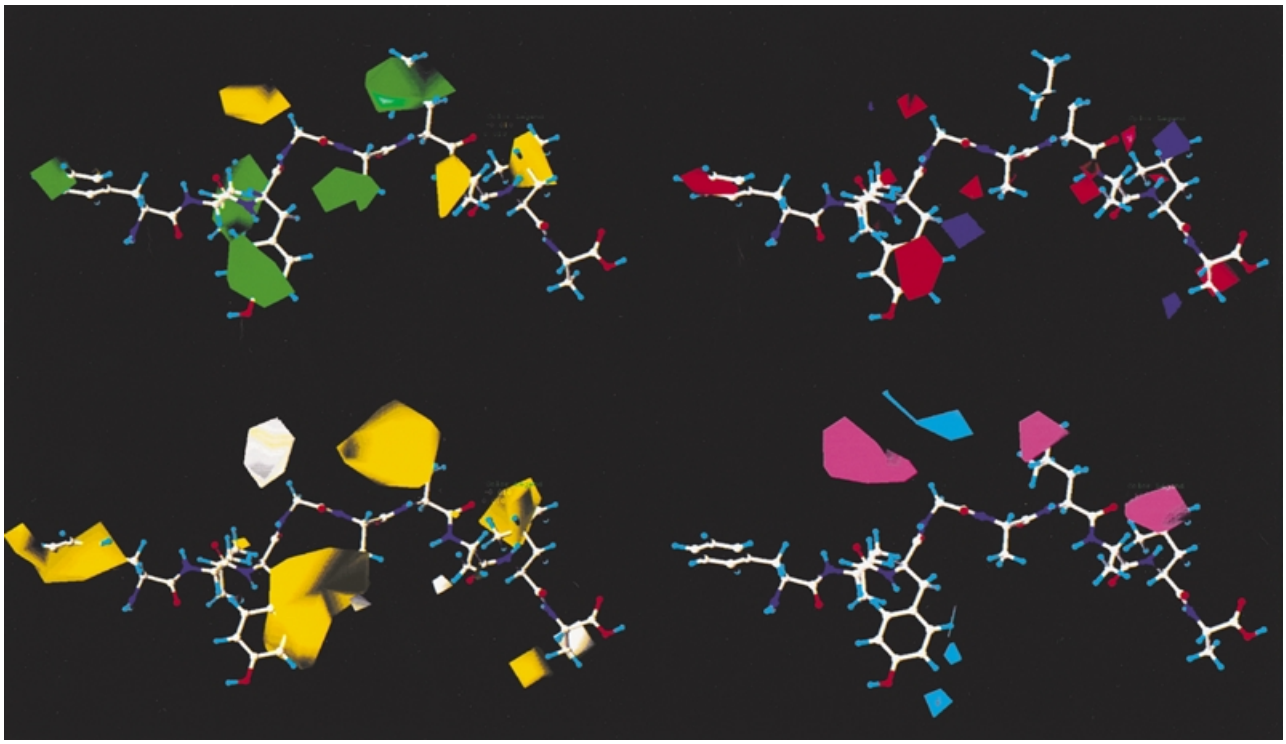


Figure 4 CoMFA StDev*Coeff maps contoured by actual values. Peptide FLYGALALA is shown inside the field. Upper left: steric map. Green (level +0.01) and yellow (level –0.01) polyhedra indicate regions where more steric bulk or less steric bulk, respectively, will enhance the affinity. Upper right: electrostatic map. Red (level +0.03) and blue (level –0.03) polyhedra indicate regions where negative potential or positive potential, respectively, will enhance the affinity. Lower left: hydrophobic map. Yellow (level +0.01) and white (level –0.01) polyhedra indicate regions where hydrophobic or hydrophilic groups, respectively, will enhance the affinity. Lower right: H-bond abilities. Cyan (level +0.01) and magenta (level +0.01) polyhedra indicate regions where hydrogen bond acceptor or donor groups, respectively, on the receptor will enhance the binding.

The statistical stability of the final CoMSIA model was tested by CV in two and five groups. The mean q^2 value of 20 runs for a CV in five groups was 0.656, which is very close to the LOO-CV value. The 'leave-half-out' CV (CV in two groups) gave a lower value for q^2 (the mean of 50 runs is 0.558), but it is still close to the other two q^2 values. From the fractions of the fields, the electrostatic and hydrophobic fields have the greatest influence, followed by the hydrogen-bond-formation fields and the steric field.

The visualization of the results from the CoMSIA analysis has been performed using the 'StDev*Coeff' mapping option contoured by actual values, and the peptide FLYGALALA (pIC_{50} = 8.620, one of the very well-predicted high binders) is shown inside the different fields in Fig. 4. The peptide is positioned with the N-terminus to the left.[42] Positions within the peptide are defined as P1 to P9. The contours of the CoMSIA steric map are shown in green (more bulk is favoured) and yellow (less bulk is favoured). The electrostatic map has red (negative potential is favoured) and blue (positive potential is favoured) coloured contours, and CoMSIA hydrophobic fields are coloured yellow (hydrophobic amino acids enhance affinity) and white (hydrophilic groups enhance affinity). The hydrogen-bond field contours show regions where hydrogen-bond acceptors (cyan) on the receptor (hydrogen-bond donors on the ligand) and hydrogen-bond donors (magenta) on the receptor (hydrogen-bond acceptors on the ligand) enhance the binding. (The Tripos implementation of CoMSIA uses a nomenclature opposite to that used in Ref. 37, in accordance with modifications made by the original authors).

As is evident from Fig. 4, steric bulk is well tolerated at P1, and there are also significant favourable areas at P2, P3, P5 and P6. Disfavoured areas exist at P4, P7 and P8. In the electrostatic map, negative potentials are favoured at most positions, while disfavoured regions lie between P3 and P5, and at P8. Areas of favourable local hydrophobicity exist at P1, P3, P5, P6, P8 and P9. Favoured hydrophilic groups are located at P4 and P9. Hydrogen-bond donor fields on the ligand are favoured around P3 and at P4; hydrogen-bond acceptor fields on the ligand are favoured at P4, P6 and P8.

Peptide structure analysis

It has long been known that all nine side-chains of the bound peptides contact the HLA-A*0201 molecule and influence the energetics of binding.[39] The antigen-binding groove has a 30Å long surface accessible to a solvent probe. There are six pockets in the surface denoted A F.[43] Some of them are non-polar and can form hydrophobic contacts, while others contain polar atoms and can make hydrogen bonds with the side-chains.

As statistical approaches, the additive method and CoMSIA seek to correlate relative differences in discriminating

molecular descriptor values to a dependent property (e.g. the binding affinity). In that respect, CoMSIA is a method able to map similarities or dissimilarities between molecules. The additive method is able to quantify the contributions made to the binding affinity by each amino acid, at each position, and by the interactions between them. Comparing, in detail, the results of the additive method and CoMSIA, we have found a remarkable degree of congruence, and, where experimental evidence is available to support our results, we have provided them.

Hydrophobic steric bulk with negative potential is well tolerated at P1. The most suitable amino acids for this position seem to be Phe and Tyr. In our first CoMSIA study, many areas of hydrogen-bond donor groups were found near the N-terminus.[13] These areas are absent in the present map (Fig. 4, lower right). This is because there are no changes in the hydrogen atom positions near the N-terminal due to the automatic building of peptides. According to the additive method, Tyr is the favourite amino acid for P1. Phe and Lys also make positive contributions. Arg, His and Thr are not preferred at P1. The remaining amino acids make negligibly small contributions. Topologically, P1 corresponds to pocket A.[44] The surface of this pocket is predominantly polar: five Tyr hydroxyl groups (Tyr7, Tyr59, Tyr99, Tyr159 and Tyr171), a carboxyl group (Glu63), and an ε-amino group (Lys66). Tyr7, Tyr59 and Tyr171 form a network of hydrogen bonds that interact directly with the peptide N-terminus. Tyr159 hydrogen bonds to the carbonyl oxygen of the first peptide amino acid residue (P1).[45] Independently of our studies, it has recently been reported that the substitution of Ile at P1 with Phe or Tyr in the HIV reverse transcriptase (RT) peptide (309–317) (ILKEPVHGV) increased threefold the cell surface half-life of complexes.[46,47] A π–π stacking interaction between Trp167 and the aromatic P1 residues was proposed to account for this change.[46] Moreover, Tourdot *et al*. report that the P1Y substitution in 10 non-immunogenic low-affinity peptides exhibited a 2.3- to 55-fold higher binding affinity and/or stabilized the HLA-A2.1 for at least 2 h more than the corresponding native peptides.[47]

The steric map at P2 indicates that long side-chains such as Leu, Ile and Met are well tolerated here (Fig. 4, upper left). The additive method distinguishes two favourite amino acids for this position (Met and Leu). The contribution of Met is surprisingly higher than that of Leu. Ala, Cys, Gly and Thr have negative contributions. The contributions of the other amino acids are negligibly small. This is in good agreement with many experimental data.[6,48–50] The side-chain at P2 falls into pocket B of the peptide-binding site on HLA-A*0201. This pocket has a polar rim and hydrophobic inner walls made up of Val67, Phe9 and Met45.[44]

Hydrophobic volume with negative potential is preferred at P3. The side-chains of the amino acids at this position fall into pocket D. Pocket D has been defined as a 'loose' pocket,[43] and it belongs to the so-called secondary binding pockets. It is a hydrophobic cavity located between the aromatic rings of Tyr99 and Tyr159, also including residues 155, 156 and 160.[51] This pocket prefers large hydrophobic residues like Phe and Trp.[52] The hydrogen-bonding ability map indicates that amino acids able to form hydrogen bonds will also be well accepted here. Tyr and Trp have the greatest positive contributions for this position (Fig. 2), but Leu and Phe are also well accepted. Glu is deleterious here for the affinity. Cys, His, Pro and Ser contribute negatively.

Short hydrophilic amino acids able to form hydrogen bonds are well tolerated at P4. Ser or Thr would be well tolerated here. Kirksey *et al*. suggest the formation of a hydrogen bond between Tyr at P1 and Glu at P4 bridged by a water molecule.[47] This should make the bound peptide more rigid and easily recognized by T-cell receptors. The side-chain at P4 is called 'flag' residue because it is solvent-exposed in the complex with the MHC molecule;[43] therefore, it can contact the TCR. According to the additive method, there is no favourite amino acid at P4. Ile and Phe are deleterious, Cys and Met make significant negative contributions, and Gly, Pro, Ser and Thr are well accepted here.

The maps indicate that amino acids with hydrophobic, branched or aromatic side-chains ending with small hydrophilic groups are well tolerated at P5. Figure 2 shows that favourite amino acids for P5 are Phe and Tyr. His, Leu and Trp also contribute positively, while Arg should be avoided at this position.

Amino acids with long hydrophobic side-chains are preferred at P6. Hydrogen-bond ability is an additional priority. The bar chart for P6 shows that Ile, Leu, Thr and Tyr are well accepted here. Ala, Arg, Asp, Gln, His and Lys contribute negatively. This side-chain falls into pocket C.[44] This pocket is predominantly polar, made up of Thr73, His70, His74 and Arg97. This explains the acceptance of the hydrophilic Thr and Tyr, but it cannot explain the preference for the hydrophobic Ile and Leu.

Short side-chains are favoured sterically at P7. Pro is the favourite amino acid for this position according to the additive method. His makes a good contribution as well. Asn is deleterious here, Arg, Gln, Gly and Ser and Thr are not preferred. The side-chain at P7 falls into pocket E. Two-thirds of the surface area in this pocket is hydrophobic, but Arg97 provides a large polar patch on one side of the pocket.[44] Pocket E can accommodate a variety of complementary peptide side-chains, but an incompatible side-chain need not prevent complex formation. This pocket has been called a 'part-time' pocket,[43] and it belongs to the class of secondary binding pockets.

The side-chain at P8 should be short, with a hydrophobic core and an end capable of forming hydrogen bonds. No favourite amino acids could be distinguished for this position by the additive method, although Gln, Phe, Pro and Ser are well accepted here. The presence of Asp, Ile, His, Met or Val is not desired. Trp147 hydrogen bonds to the P8 carbonyl oxygen.[43] P8 is a 'flag' position as P4 is.

Amino acids with hydrophobic, short side-chains are required for P9. Val is the favourite amino acid here judged by the information in Fig. 2. The side-chain of Tyr116 occupies the end of pocket F and is uncharged, so that the binding site is complementary to small hydrophobic side-chains.[39,44] Interestingly, a small hydrophilic area carrying negative potential appears near P9, which is due to the Thr introduced here by the intermediate binder MLQDMAILT and the high binder YMLDLQPET. However, according to the additive method, Ser and Thr should be avoided. The Tyr116 side-chain hydroxyl group forms a hydrogen bond to Asp77 on the α_1 helix, stabilizing it in this orientation.[43] Tyr84, Thr143, and positively charged Lys146 bind to the carboxyl group of the C-terminal.[43]

Peptides bound to the HLA-A*0201 molecule assume extended but twisted conformations.[39] As a result, the adjacent side-chains protrude in largely opposite directions and, in practice, interactions between them are unlikely to exist. The interaction between the adjacent side-chains may be considered as a change in the backbone conformation caused by a certain amino acid at a certain position producing change in the conformation of the adjacent amino acid side-chain. However, the twisted conformation makes possible the interactions between every second amino acid side-chain. These interactions might have a steric, electrostatic, hydrophobic or hydrogen-bonding character. A conformational change is also possible here. Unfortunately, the additive method is unable to give any explanation regarding the nature of forces involved in such interactions, but it can assess quantitatively the significance for the affinity.

Among the adjacent side-chain interactions, the favourite ones are 1Y2L, 2L3M and 8L9A (Fig. 3a). The high contributions of the last two combinations are very unexpected as 3M, 8L and 9A are not among the favourite or highly positively contributing amino acids. The only reasonable explanation is a conformational change favouring the binding. The only delirious combination is 2L3E; Glu is disfavoured at P3.

The combinations 1G3L, 2L4Q, 4S6G and 7F9V have the highest contributions in binding affinity to HLA-A*0201 among the 1–3 side-chain interactions (Fig. 3b). 2L and 9V are favourite amino acids, 3L and 4S make significant positive contributions, and 1G, 4Q, 6G and 7F make negligibly small positive contributions. It is possible that conformational changes and steric interactions are responsible, rather than electrostatic, hydrophobic interactions, or intramolecular hydrogen-bond formation. Extremely disfavoured combinations are 1I3L, 2T4Q and 4I6V. The first of them seems counterintuitive because both 1I and 3L make positive contributions in the affinity. Furthermore, 3L in combination with 1G makes a positive contribution with a high value. Obviously, the steric bulk of 1I causes an inappropriate change in the conformation of the 3L side-chain. The intolerance of 2T4Q and 4I6V is probably due to the high negative contributions of 2T and 4I.

Discussion

We have described the development of quantitative approaches to the prediction of MHC binding built on our database of quantitative binding measures. Despite the principal differences between the additive method and CoMSIA, very good agreement was found between the results generated by both techniques. The combination of these two methods gives very useful results. CoMSIA can extrapolate, that is, predict the binding affinity of a peptide with an amino acid not presented in the initial training set, but it cannot assess the contribution of each amino acid at each position and the interactions between them. The opposite is true for the additive method: it can not extrapolate, but it can give a quantitative assessment of individual amino acid contributions at any position in the peptide. The two methods have been applied to sets of nonamers binding to the MHC class I molecule HLA-A*0201. An expansion to apply these to other alleles is in progress.

In developing these methods, we have encountered problems that are only rarely associated with QSAR analyses of small molecules. These include the size of the peptide molecules being studied; the sheer number of molecules being investigated, perhaps 10-fold greater than a small molecule study; and the great diversity of physicochemical properties associated with each position being examined. We have avoided issues of molecular alignment by assuming a constant backbone structure. It is clear from X-ray analyses that there are only small differences in backbone conformation for nonamer peptides.[44,45,51] Given the number of peptides under study, allowances for conformational flexibility in the backbone is not tractable. As 92.5% (CoMSIA model) of the peptides are either well- or very well-predicted, variations in the binding conformation do not seem significant. Poorly predicted peptides might exhibit properties significantly different from those in our training set. This is certainly true for peptides PLLPIFFCL and VCMTVDSLV. However, it should be noted that according to QSAR convention, predictions within 1.0 log unit are considered acceptable.[36,53–55] This would result in mean residuals of approximately 0.5 log units. The mean absolute values for the residuals for the additive method and CoMSIA are 0.573 and 0.489, respectively.

In conclusion, the proposed methods for binding affinity prediction (additive and CoMSIA) have many advantages in comparison with other methods. The combination of the two methods leads to very reliable results. They are complementary because they are based on totally different approaches, yet give similar results. Finally, a set of high-affinity peptides can be designed or optimized using these methods. Our initial experimental work gives promising results in this regard. Internet access to these methods is forthcoming.

As part of its ambitious programme, the Edward Jenner Institute for Vaccine Research is committed to the rapid development of computational vaccinology as a vital component in the fight against global disease. We would expect computational vaccinology to have a similar effect on the search for new vaccines as molecular modelling and other informatics strategies have had on the discovery of novel drugs.

Acknowledgements

We should like to thank Martin Blythe, Christianna Zygouri, PingPing Guan and Paul Taylor for their contributions. We should also like to thank Dr Vladimir Brusic and Professor Peter Beverley for encouraging discussions.

References

1 Janeway CA Jr, Travers P, Walport M, Capra JD. *Immunobiology*. London: Elsevier Science, 1999; 115–62.
2 Sette A, Vitiello A, Reherman B *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* 1994; **153**: 5586–92.
3 Flower DR, Doytchinova IA, Paine K *et al.* Computational vaccine design. In: Flower DR (ed.) *Drug Design: Cutting Edge Approaches*. Cambridge: RSC Publications, 2002.
4 D'Amaro J, Houbiers JGA, Drijfhout JW *et al.* A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum. Immunol.* 1995; **43**: 13–18.
5 Meister GE, Roberts CGP, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-

binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. *Vaccine* 1995; **13**: 581–91.

6 Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 1994; **152**: 163–75.

7 Gulukota K, Sidney J, Sette A, DeLisi C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* 1997; **267**: 1258–67.

8 Honeyman MC, Brusic V, Stone NL, Harrison LC. Neutral network-based prediction of candidate T-cell epitopes. *Nature Biotechnol.* 1998; **16**: 966–9.

9 Rognan D, Lauemoller SL, Holm A, Buus S, Schinke V. Predicting binding affinities of protein ligands from three-dimensional models: Application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem*. 1999; **42**: 4650–8.

10 Walden P. T-cell epitope determination. *Curr. Opin. Immunol.* 1996; **8**: 68–74.

11 Udaka K, Wiesmüller K-H, Kienle S *et al.* An automated prediction of MHC class I – binding peptides based on positional scanning with peptide libraries. *Immunogenetics* 2000; **51**: 816–28.

12 Doytchinova I, Blythe M, Flower DR. An additive method for the prediction of binding affinity. Application MHC Class I Molecule HLA-A*0201. *J. Proteome. Res.* 2002 (in press).

13 Doytchinova IA, Flower DR. Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J. Med. Chem*. 2001; **44**: 3572–81.

14 Doytchinova IA, Flower DR. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex. A Three-Dimensional Quantitative Structure – Activity Relationship Study. *Proteins* 2002, (in press).

15 Blythe MJ, Doytchinova IA, Flower DR. JenPep: A database of quantitative functional peptide data for immunology. *Bioinformatics* 2002; **48**: 434–9.

16 Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A. Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell* 1993; **74**: 929–37.

17 Sette A, Sidney J, del Guercio MF *et al.* Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.* 1994; **31**: 813–22.

18 Sidney J, Oseroff C, del Guercio MF *et al.* Definition of a DQ3.1-specific binding motif. *J. Immunol.* 1994; **152**: 4516–23.

19 Marshall KW, Liu AF, Canales J *et al.* Role of the polymorphic residues in HLA-DR molecules in allele-specific binding of peptide ligands. *J. Immunol.* 1994; **152**: 4946–57.

20 van der Burg SH, Visseren MJ, Brandt RM, Kast WM, Melief CJ. Immunogenicity of peptides bound to MHC class I molecules depends on the MHC-peptide complex stability. *J. Immunol.* 1996; **156**: 3308–15.

21 ten Bosch GJ, Kessler JH, Joosten AM *et al.* A BCR-ABL oncoprotein p210b2a2 fusion region sequence is recognized by HLA-DR2a restricted cytotoxic T lymphocytes and presented by HLA-DR matched cells transfected with an Ii(b2a2) construct. *Blood* 1999; **94**: 10381–7.

22 Shiga H, Shioda T, Tomiyama H *et al.* Identification of multiple HIV-1 cytotoxic T cell epitopes presented by human leukocyte antigen B35 molecules. *AIDS* 1996; **10**: 1075–85.

23 Sakaguchi T, Ibe M, Miwa K *et al*. Predominant role of N-terminal residue of nonamer peptides in their binding to HLA-B* 5101 molecules. *Immunogenetics* 1997; **46**: 245–55.

24 Sobao Y, Tsuchiya N, Takiguchi M, Tokunaga K. Overlapping peptide-binding specificities of HLA-B27 and B39: evidence for a role of peptide supermotif in the pathogenesis of spondylarthropathies. *Arthritis Rheum.* 1999; **42**: 175–81.

25 Van den Eynde B, Mazarguil H, Lethe B, Laval F, Gairin JE. Localization of two cytotoxic T lymphocyte epitopes and three anchoring residues on a single nonameric peptide that binds to H-2Ld and is recognized by cytotoxic T lymphocytes against mouse tumor P815. *Eur. J. Immunol.* 1994; **24**: 2740–5.

26 Gairin JE, Mazarguil H, Hudrisier D, Oldstone MB. Optimal lymphocytic choriomeningitis virus sequences restricted by H-2Db major histocompatibility complex class I molecules and presented to cytotoxic T lymphocytes. *J. Virol.* 1995; **69**: 2297–305.

27 Ayyoub M, Mazarguil H, Monsarrat B, Van den Eynde B, Gairin JE. A structure-based approach to designing non-natural peptides that can activate anti-melanoma cytotoxic T cells. *J. Biol. Chem.* 1999; **274**: 10227–34.

28 Parker KC, DiBrino M, Hull L, Coligan JE. The beta 2-microglobulin dissociation rate is an accurate measure of the stability of MHC class I heterotrimers and depends on which peptide is bound. *J. Immunol.* 1992; **149**: 1896–903.

29 DiBrino M, Parker KC, Shiloach J *et al.* Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8. *J. Immunol.* 1994; **152**: 620–9.

30 Free SM Jr, Wilson JW. A mathematical contribution to structure–activity studies. *J. Med. Chem.* 1964; **7**: 395–9.

31 Parker KC, Shields M, DiBrino M, Brooks A, Coligan JE. Peptide binding to MHC class I molecules: Implications for antigenic peptide prediction. *Immunol. Res.* 1995; **14**: 34–57.

32 *SYBYL 6.7* (computer program) St Louis, MO: Tripos, 2001.

33 Oprea TI, Waller CL. Theoretical and practical aspects of three-dimensional Quantitative Structure – Activity Relationships. In: Lipkowitz KB, Boyd DB (eds). *Reviews in Computational Chemistry*, Vol. 11. New York: John Wiley and Sons, 1997; 127–82.

34 Greco G, Novellino E, Martin YC. Approaches to three-dimensional Quantitative Structure – Activity Relationships. In: Lipkowitz KB, Boyd DB (eds). *Reviews in Computational Chemistry*, Vol. 11. New York: John Wiley and Sons, 1997; 183–240.

35 Klebe G. Comparative molecular similarity indices analysis: CoMSIA. *Perspectives Drug Discovery Design* 1998; **12/13/14**: 87–104.

36 Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 1994; **37**: 4130–46.

37 Klebe G, Abraham U. Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aided Mol. Des.* 1999; **13**: 1–10.

38 Böhm M, Stürzebecher J, Klebe G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* 1999; **42**: 458–77.

39 Madden DR, Garboczi DN, Wiley DC. The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* 1993; **75**: 693–708.

40 Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* 1985; **107**: 3902–9.

41 Cramer RD, Bunce JD, Patterson DE. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct. Act. Relat.* 1988; **7**: 18–25.

42 Latron F, Moots R, Rothbard JB, Garrett TPJ, Strominger JL, McMichael A. Positioning of a peptide in the cleft of HLA-A2 by complementing amino acid changes. *Proc. Natl. Acad. Sci. USA* 1991; **88**: 11325–9.

43 Madden DR. The three-dimensional structure of peptide–MHC complexes. *Annu. Rev. Immunol.* 1995; **13**: 587–622.

44 Saper MA, Bjorkman PJ, Wiley DC. Refined structure of the human class histocompatibility antigen HLA-A2 at 2.6 Å. *J. Mol. Biol.* 1991; **219**: 277–319.

45 Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 1987; **329**: 512–18.

46 Kirksey TJ, Pogue-Caley RR, Frelinger JA, Collins EJ. The structural basis for the increased immunogenicity of two HIV-reverse transcriptase peptide variant/class I major histocompatibility complexes. *J. Biol. Chem*. 1999; **274**: 37259–64.

47 Tourdot S, Scardino A, Saloustrou E *et al.* General strategy to enhance immunogenicity of low-affinity HLA-A2.1-associated peptides: Implication in the identification of criptic tumor epitopes. *Eur. J. Immunol.* 2000; **30**: 3411–21.

48 Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* 1994; **74**: 929–34.

49 Kubo RT, Sette A, Grey HM *et al.* Definition of specific peptide motifs for four major HLA-A alleles. *J. Immunol.* 1994; **152**: 3913–24.

50 Parker KC, Bednarek MA, Hull LK *et al.* Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. *J. Immunol.* 1992; **149**: 3580–7.

51 Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 1987; **329**: 506–12.

52 Sarobe P, Pendleton CD, Akatsuka TD, Engelhard VH, Feinstone SM, Berzofsky JA. Enhanced *in vitro* potency and *in vivo* immunogenicity of a CTL epitope from hepatitis C virus core protein following amino acid replacement at secondary HLA-A2.1 binding positions. *J. Clin. Invest.* 1998; **102**: 1239–48.

53 Sicsic S, Serraz I, Andrieux J *et al.* Three-dimensional quantitative structure – activity relationship of melatonin receptor ligands: a Comparative Molecular Field Analysis Study. *J. Med. Chem.* 1997; **40**: 739–48.

54 Pajeva I, Wiese M. Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: a Comparative Molecular Field Analysis study. *J. Med. Chem.* 1998; **41**: 1815–26.

55 Ducrot P, Legraverend M, Grierson DS. 3D-QSAR CoMFA on cyclin-dependent kinase inhibitors. *J. Med. Chem.* 2000; **43**: 4098–108.