PROTEINS: Structure, Function, and Genetics 48:505–518 (2002)

Physicochemical Explanation of Peptide Binding to HLA-A*0201 Major Histocompatibility Complex: A Three-Dimensional Quantitative Structure-Activity Relationship Study

Irini A. Doytchinova* **and Darren R. Flower**
Edward Jenner Institute for Vaccine Research, Compton, Berkshire, United Kingdom

ABSTRACT **A three-dimensional quantitative structure-activity relationship method for the prediction of peptide binding affinities to the MHC class I molecule HLA-A*0201 was developed by applying the CoMSIA technique on a set of 266 peptides. To increase the self consistency of the initial CoMSIA model, the poorly predicted peptides were excluded from the training set in a stepwise manner and then included in the study as a test set. The final model, based on 236 peptides and considering the steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields, had q^2 = 0.683 and r^2 = 0.891. The stability of this model was proven by cross-validations in two and five groups and by a bootstrap analysis of the non-cross-validated model. The residuals between the experimental pIC_{50} ($-logIC_{50}$) values and those calculated by "leave-one-out" cross-validation were analyzed. According to the best model, 63.2% of the peptides were predicted with $|residuals| \leq 0.5$ log unit; 29.3% with $1.0 \leq |residuals| < 0.5$; and 7.5% with $|residuals| > 1.0$ log unit. The mean |residual| value was 0.489. The coefficient contour maps identify the physicochemical property requirements at each position in the peptide molecule and suggest amino acid sequences for high-affinity binding to the HLA-A*0201 molecule. Proteins 2002;48:505–518.** © 2002 Wiley-Liss, Inc.

Key words: 3D-QSAR; CoMSIA; binding affinity prediction

INTRODUCTION

The protective function of T cells depends on their ability to recognize host cells that are harboring pathogens or that have internalized pathogens or their products. T cells do this by recognizing peptide fragments of pathogen-derived proteins in the form of complexes of peptide and MHC molecules expressed on the surface of these cells. Thus, one necessary condition for a peptide to be a T-cell epitope is that it can bind MHC molecules with high affinity. X-ray crystal structures of MHC molecules have revealed the structure of the peptide-protein complex.[1–3] Peptides that bind to the MHC class I molecule are usually 8–11 amino acids long. The binding cleft of this molecule contains six pockets that have been labeled A–F.[3] Some of the pockets are nonpolar and are expected to make hydrophobic interactions, whereas others contain polar atoms. Based on a large number of tested peptides, peptide selectivities of different HLA alleles have been defined[4–7]; these are often expressed in terms of primary and secondary peptide "anchor" positions. A cluster of tyrosine residues common to all MHC class I molecules forms hydrogen bonds to the amino terminus of the bound peptide (pocket A). A second cluster of residues forms hydrogen bonds and makes ionic interactions with the peptide backbone at the carboxy terminus and with the peptide C-terminus itself (pocket F).[8,9] For the HLA-A*0201 molecule, which is the best studied of all class I MHC molecules, Tyr is a preferred amino acid at P1, Leu at P2, and Val at P9.[7–9] In general, the nature of forces involved in the peptide-protein interaction is not sufficiently well understood to allow the delineation of amino acid sequences that are optimal for affinity at a given HLA molecule.

Many methods for the prediction of MHC-peptide binding affinities have been developed. One of the first was developed by Parker et al.[10,11] Based on the hypothesis that each amino acid side-chain binds independently of the rest of the peptide (the so-called IBS hypothesis), the contribution of each amino acid, at each position, to the stability, or half-life, of the HLA-A2 complex was assessed. Only the most important 82 variables were used from the set of 180 possible variables (20 amino acids at 9 positions). Thus, important amino acids at certain positions were neglected, for example, Phe and Tyr at position 1 and Thr and Pro at position 4. The IBS hypothesis also forms the basis of the polynomial method.[12] The contribution of each amino acid at each position to the overall affinity was assessed by the average negative $\log_{10}$ of IC_{50} of all the peptides carrying certain amino acids at certain positions. Neural networks were also used to predict binding affinities to HLA-A2.1 molecule[12–14] and to score them as binders or nonbinders. Altuvia et al.[15] developed an algorithm to evaluate the interactions of peptide amino acids with MHC contact residues, and the resulting energies were used to score the peptide-binding affinities. Recently,

*Correspondence to: Dr. Irini Doytchinova, Bioinformatics Group, Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK. E-mail: irini.doytchinova@jenner.ac.uk

Received 27 July 2001; Accepted 7 March 2002

a free energy scoring function (Fresno) based on ligand docking was developed to predict the binding free energy of peptides to class I MHC proteins.[16,17] Molecular dynamics (MD) simulations have also been used in the examination of peptide/protein interactions. For example, Mata et al.[18] probed the stability of peptide binding to murine MHC class 1 molecule H-2 K^d and identified a binding motif that does not contain an appropriate anchor residue. Meng et al.[19] used MD simulations to identify the water molecule binding sites at the MHC–peptide interface.

The binding affinity of a ligand to a macromolecular receptor is related to the Gibbs free energy of binding ΔG, which is itself composed of an enthalpic and entropic contribution. The steric and electrostatic complementarity described by different potential energy functions cover the enthalpic contributions. The hydrophobic interactions, which compose major contributions to the entropic part of the free energy, are connected with changes in water structure around the ligand and within the active site.[20–22] The hydrogen bond interactions contribute to the entropic part of the free energy as well as the enthalpic part. There are many two-dimensional (2D) and three-dimensional (3D) molecular descriptors that can account for changes in the binding free energy. 3D *Q*uantitative *S*tructure-*A*ctivity *R*elationship (3D-QSAR) analysis uses sophisticated, robust multivariate statistics to correlate molecular descriptors generated in the space around the ligands with their binding affinities. The widely used *Co*mparative *M*olecular *F*ield *A*nalysis (CoMFA) is a 3D-QSAR method that calculates steric and electrostatic properties according to Lenard-Jones and Coulomb potentials.[23] The more recently reported *Co*mparative *M*olecular *S*imilarity *I*ndices *A*nalysis (CoMSIA) method uses fields based on similarity indices that describe similarities and differences between ligands and correlate them with changes in the binding affinity.[20–22] Similarity indices fields describe steric, electrostatic, hydrophobic, and hydrogen bond donor and acceptor properties. The most important contributions responsible for binding affinity are covered by these properties. Each of them can be visualized in a 3D map denoting the areas, within the binding site, that are "favored" or "disfavored" by the presence of a group with a particular physicochemical property.

In our previous work, we applied two 3D QSAR methods (CoMFA and CoMSIA) to a training set of 102 peptides that bound to HLA-A*0201.[24] The predictive power of both methods was assessed by using a test set of 50 peptides. We found that CoMSIA gave much better predictive pIC_{50} values for binding to the HLA-A*0201 molecule than CoMFA and indicated a dominant role for hydrophobic interactions in peptide binding to the MHC molecule. In the present study, we greatly extended our 3D-QSAR analysis by applying the CoMSIA technique on a set of 266 peptides to assess the contributions of the other physicochemical properties besides the hydrophobic field. The best model was used to evaluate the physicochemical requirements at each position in the peptide structure and to define the preferred amino acid sequence for high-affinity binding to HLA-A*0201 molecule. The explanatory power of 3D-QSAR methods is considerable, not only in their ability to accurately predict binding affinities but also in their capacity to display advantageous and disadvantageous interaction potential, in three dimensions, mapped onto the 3D structures of molecules being studied (in this case peptides). We exploited this feature of the methodology to gain a correlation between the physicochemical similarity indices and the affinities of peptides to MHC molecule HLA-A*0201. The data are highly complementary to the sort of very detailed (but peptide specific) information obtained from crystal structures of individual peptide-MHC complexes.

MATERIALS AND METHODS

Molecular Modeling

All molecular modeling and QSAR calculations were performed on a Silicon Graphics octane workstation by using the SYBYL 6.6 molecular modeling software.[25] The X-ray structure of the nonameric viral peptide TLTSCNTSV,[5] was used as a starting conformation. The structures of the remaining peptides were built by using the BIOPOLYMER option in SYBYL and were subjected to fully geometry optimization by using the standard Tripos molecular mechanics force field (Powell method[26] no electrostatics and 0.05 kcal/mol*Å energy gradient convergence criterion). The peptide backbone was kept as a rigid body or aggregate by using the option "Aggregates" in the Minimize Energy menu. The peptide backbone was fixed in the X-ray conformation. The aggregate consists of the α-carbon atoms, the carbonyl carbon and oxygen atoms, and the amide nitrogen and hydrogen atoms. The partial atomic charges used in CoMSIA were computed by using the AM1 semiempirical method[27] available in the MOPAC program. MOPAC V6 was used as implemented in SYBYL. Single-point calculations were performed. A program for automatic building, optimization, and AM1 calculation of the peptides was created and implemented in SYBYL. The program uses a text file containing the peptide sequences and a pdb file of the peptide used as the starting conformation.

Two types of alignment were tested: atom based and field fit. For the first type of alignment, the backbone atoms, defined as an aggregate in the optimization process, were used.

CoMSIA Method

CoMSIA was performed by using the QSAR option of SYBYL. Five physicochemical properties (steric, electrostatic, hydrophobic, and hydrogen bond donor and acceptor) were evaluated with use of a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, hydrogen bond donor and acceptor properties +1. Similarity indices were calculated by using Gaussian-type distance dependence between the probe and the atoms of the molecules of the data set, according to the equation:

$$A_k^q(j) = \sum w_{probe,k} w_{ik} e^{-\alpha_{iq}^2}$$

where A is the similarity index at grid point q, summed over all atoms i of the molecule j, $w_{probe,k}$, is the probe

atom, w_{ik} is the actual value of the physicochemical property k of atom i, r_{iq} is the mutual distance between the probe at grid point q and atom i of the test molecule, α is an attenuation factor. Different values of the attenuation factor α were tested. These values ranged from 0.1 to 0.5 in incremental steps of 0.1.

PLS methodology was used for the 3D QSAR analysis. The grid extended beyond the molecular dimensions by 2.0 Å in all directions. Different resolution steps were tested: 1.0, 1.5, 2.0, 2.5, and 3.0 Å. Different column filterings σ_{min} (2.0, 1.0, and 0.5) were analyzed as well. σ_{min} indicates which grid points to include in the analysis, for example, $\sigma_{min} = 2.0$ indicates that columns (similarity indices) whose variance is <2.0 are omitted.

The predictive power of the models was assessed by the cross-validated coefficient q^2, the standard error of prediction (SEP), and the *residuals* between the experimental and predicted binding affinity expressed in p-units ($-\log IC_{50}$):

$$q^2 = 1 - \frac{PRESS}{SSO}$$

$$SEP = \sqrt{\frac{PRESS}{p-1}}$$

$$residual = pIC_{50}exp - pIC_{50}pred$$

where *PRESS* is the predictive sum of squares ($\Sigma_{i=1}^{n}$ (pIC50 exp − pIC50pred)2), *SSQ* is the sum of squares of pIC_{50}exp corrected for the mean ($\Sigma_{i=1}^{n}$ (pIC50exp − pIC50mean)2), p is the number of the peptides omitted, and pIC_{50}pred is the value predicted by "leave-one-out" cross-validation. Cross-validation (CV) is performed by dividing the data into a number of groups, developing a number of parallel models from the reduced data with one of the groups omitted, and then predicting the biological activities of the excluded compounds. When the number of the groups omitted is equal to the number of the compounds in the set, the procedure is called "leave-one-out" cross-validation (LOO-CV). More robust CV tests, using five and two groups, were performed to estimate the extent of chance correlation in the model. The means of q^2 values are given as q^2_{CV5} (20 runs) and q^2_{CV2} (50 runs).

According to the *residuals* between the experimental and predicted by LOO-CV pIC_{50} values, the peptides were classified into three categories: very well predicted peptides with $|residuals| \leq 0.5$, well-predicted peptides with $|residuals|$ between 0.5 and 1.0, and poorly predicted peptides with $|residuals| > 1.0$. A mean $|residual|$ value ($\Sigma_{i=1}^{n}$ |residual|/n) for the set also was calculated.

The optimum number of components (NC) used do derive the nonvalidated model was defined as the number of components leading to the highest q^2 and the lowest SEP. The non-cross-validated models were assessed by the explained variance r^2, standard error of estimate (SEE), and F ratio. A bootstrap analysis[28] was performed in 20 runs for the best model and the mean r^2 is given as $r^2_{bootstrap}$. This model was used to display the coefficient contour maps.

Because only the combination of all fields provides full insight into the spatial features of the contribution of different fields, only the all-fields models were considered for further analyses.

CoMSIA Maps

The visualization of the results of the CoMSIA analyses was performed by using the "StDev*Coeff" mapping option contoured by actual values. Favored and disfavored levels fixed at +0.01 and −0.01, respectively, were used for the steric and hydrophobic fields; +0.03 and −0.03, respectively, for the electrostatic field. The hydrogen bond donor and acceptor fields are presented in a map with favored levels fixed at +0.01. The contours of the CoMSIA steric map are shown in green (more bulk is favored) and yellow (less bulk is favored). The electrostatic map is in red (negative potential is favored) and blue (positive potential is favored) contours. CoMSIA hydrophobic fields are colored in yellow (hydrophobic amino acids enhance affinity) and white (hydrophilic groups enhance affinity). The hydrogen bond field contours show regions where hydrogen bond acceptors (cyan) and hydrogen bond donors (magenta) on the receptor enhance the binding.

RESULTS

Peptide Database

The peptide sequences and their pIC_{50} ($-\log IC_{50}$) values used in the study are presented in Table I. The binding affinities (IC_{50}) we used were originally assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules.[6,51] More than one IC_{50} value was found for some of the peptides. A mean value of the multiple IC_{50} values was extracted for a peptide with differences in its pIC_{50} values < 0.5 log unit. Otherwise, the peptide was excluded from the set. The final set consisted of 266 peptides. One hundred ten were high-affinity peptides ($IC_{50} \leq 50$ nM, $pIC_{50} \geq 7.301$), 118 peptides had intermediate affinity (50 nM $< IC_{50} \leq 500$ nM, $7.301 > pIC_{50} \geq 6.301$), and 38 were low-affinity peptides ($IC_{50} > 500$ nM, $pIC_{50} < 6.301$).

Two types of alignment were tested: atom based and field fit. The alignment based on the backbone atoms gives better results than the field fit (data not shown). So, the CoMSIA models in the present study were developed by using the alignment based on the corresponding backbone atoms.

CoMSIA Models

Because our intention was to partition the various properties into spatial locations where they play a decisive role in determining the binding affinity, only the all-fields model was used. Three criteria were tested for model calibration: grid spacing, attenuation factor α, and column filtering. The results from the PLS calculations with varying column filtering value ($\sigma_{min} = 2.0$, 1.0, or 0.5) show that the highest q^2 values were obtained with σ_{min} of 2.0 (data not shown). Thus, $\sigma_{min} = 2.0$ was set in all further calculations as a threshold column-filtering value.

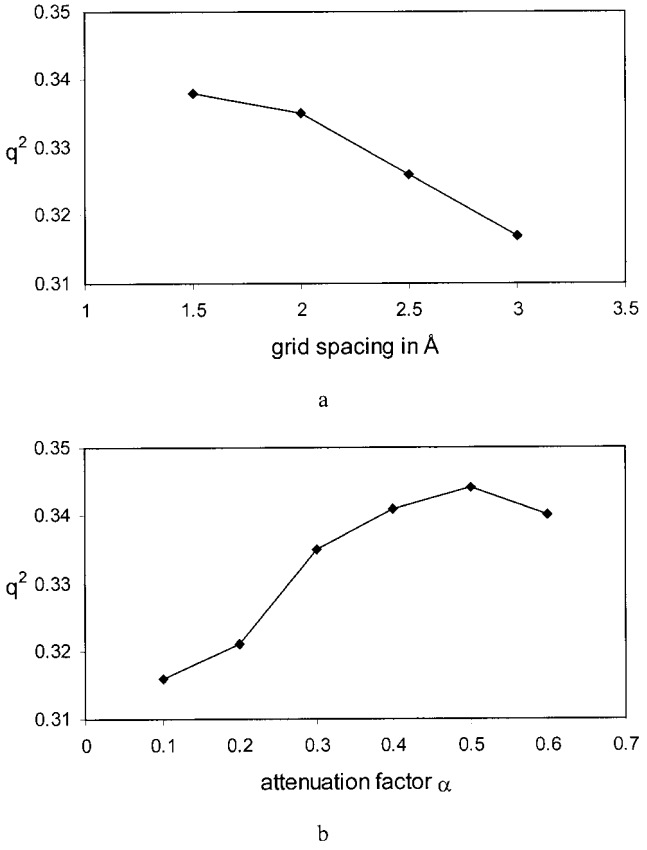

Fig. 1. The q^2 value in function of (**a**) grid spacing and (**b**) attenuation factor α for the CoMSIA(1) model.

Five grid spacings (1.0, 1.5, 2.0, 2.5, and 3.0 Å) were tested [Fig. 1(a)]. The PLS calculation failed for the step of 1.0 Å because of the huge number of points in the grid box. The q^2 values at 1.5 and 2.0 Å were very close ($q^2 = 0.338$ at 1.5 Å and $q^2 = 0.335$ at 2.0 Å). Because the calculation was faster at 2.0 Å, we chose this grid spacing for further analysis. Five attenuating factors α (0.1, 0.2, 0.3, 0.4, and 0.5) were investigated [Fig. 1(b)]. Larger values of α implied a more localized evaluation of similarity. The highest value for q^2 was monitored at $\alpha = 0.5$. Thus, the final calibrated model had the following settings: column filtering 2.0, grid spacing 2.0 Å, and $\alpha = 0.5$.

The initial all-fields model had low q^2 and r^2 values [CoMSIA(1) model, Table II]. This result was not surprising because of the great variety of peptides collecting from many sources. The analysis of residuals between the experimental and predicted pIC_{50} values revealed three groups of predictions: very well predicted peptides with $|residuals| \leq 0.5$ (151 peptides), well-predicted peptides with $|residuals|$ between 0.5 and 1.0 (83 peptides), and poorly predicted peptides with $|residuals| > 1.0$ (32 peptides), and mean $|residual| = 0.553$. To achieve a more self-consistent model, the poorly predicted peptides were excluded in a stepwise manner, beginning with the peptide with the highest residual. At each step, the predictivity of the model increased and the highest residual decreased. This procedure was repeated until the highest residual fell below 1 log unit. The final model [CoMSIA(2), Table II] was based on 236 peptides and had significantly higher characteristics: $q^2 = 0.683$ at seven components and $r^2 = 0.891$. This model was used to predict the binding affinities of the excluded peptides (i.e., we used the excluded peptides as a test set). Twenty of the 30 peptides excluded were poorly predicted, 8 were well predicted, and 2 were very well predicted. So, the number of poorly predicted peptides decreased significantly from 32 to 20, the number of very well predicted increased significantly as well from 151 to 168, and the number of well predicted slightly decreased from 83 to 78. The mean $|residual|$ value for this model was 0.489. Only 10 poorly predicted peptides were common for the two CoMSIA models.

TABLE II. All-Fields CoMSIA Models

Parameters	CoMSIA(1) model	CoMSIA(2) model
n	266	236
q^2LOO	0.344	0.683
q^2CV5	—	0.656
q^2CV2	—	0.558
SEP	0.655	0.443
NC	3	7
r^2	0.585	0.891
r^2 bootstrap	—	0.924
SEE	0.521	0.260
Fractions		
Steric	0.164	0.145
Electrostatic	0.291	0.320
Hydrophobic	0.225	0.210
Hydrogen bond donor	0.182	0.161
Hydrogen bond acceptor	0.138	0.164
Predictions		
Very well predicted	151	168
$\|residuals\| \leq 0.5$ (%)	56.8	63.2
Well predicted	83	78
$1.0 \leq \|residuals\| < 0.5$ (%)	31.2	29.3
Poorly predicted	32	20
$\|residuals\| > 1.0$ (%)	12.0	7.5
Mean absolute residual	0.553	0.489

The stability of the CoMSIA(2) model was tested by cross-validation with two and five groups. The mean q^2 value of 20 runs for a cross-validation in five groups was 0.656, very close to the "leave-one-out" 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 field and the steric field.

CoMSIA Coefficient Contour Maps

The five physicochemical properties involved in the interaction between the peptide and the MHC molecule are presented in coefficient contour maps, generated by the CoMSIA(2) model (Fig. 2). Peptide FLYGALALA ($pIC_{50} = 8.620$, one of the very well predicted high binders) is shown inside the different fields. The peptide is posi-

TABLE I. Peptides Used in the Study

No.	Peptide sequence	Reference	pIC_{50} ($-\log IC_{50}$)	Residuals according to CoMSIA(1)	Residuals according to CoMSIA(2)
1.	AAAKAAAAV	29	6.398	−0.027	0.207
2.	AIAKAAAAV	29	6.176	−0.800	−0.315
3.	AIIDPLIYA	30	6.623	−0.537	−0.618
4.	AIYHPQQFV	31	6.504	0.003	0.146
5.	ALAKAAAAA	29	6.947	0.265	0.470
6.	ALAKAAAAI	29	6.211	−0.555	−0.106
7.	ALAKAAAAL	29	6.511	−0.190	0.061
8.	ALAKAAAAM	29	7.398	0.970	1.915
9.	ALAKAAAAV	29, 32	6.597[a]	−0.620	−0.214
10.	ALCRWGLLL	33	7.000	0.585	0.657
11.	ALIHHNTHL	33	6.623	0.208	0.029
12.	ALLAGLVSL	30	7.117	−0.228	−0.286
13.	ALLSDWLPA	31	7.025	0.083	−0.262
14.	ALMDKSLHV	30, 34	7.767[a]	0.508	0.411
15.	ALMPLYACI	35	8.000	0.933	0.302
16.	ALPYWNFAT	36	5.820	−1.615	−0.756
17.	ALSTGLIHL	37	6.505	−0.082	−0.350
18.	ALTVVWLLV	38	6.893	0.088	−0.272
19.	ALVGLFVLL	36	7.553	0.511	0.228
20.	ALVLLMLPV	31	7.506	−0.003	0.024
21.	ALYGALLLA	38	8.143	0.374	0.225
22.	AMFQDPQER	39	5.740	−1.168	−0.524
23.	AMKADIQHV	31	6.777	0.270	0.471
24.	AMLQDMAIL	31	7.009	0.665	0.055
25.	AMVGAVLTA	30	7.122	0.218	−0.142
26.	AVAKAAAAV	29	6.495	−0.295	0.267
27.	AVIGALLAV	30	7.747	0.599	0.688
28.	CLALSDLLV	30	6.447	−0.263	−0.076
29.	CLTSTVQLV	33	6.832	0.407	0.292
30.	DLMGYIPLV	37	7.097	−1.563	−1.278
31.	DMWEHAFYL	31	6.879	−0.780	−0.314
32.	DPKVKQWPL	29	6.176	0.280	−0.617
33.	FAFRDLCIV	35, 39	6.963[a]	0.432	−0.142
34.	FLAGALLLA	38	6.223	−1.563	−1.069
35.	FLCWGPFFL	30	7.415	0.634	0.644
36.	FLDQVPFSV	40	8.658	0.463	0.262
37.	FLEPGPVTA	40	6.898	−0.131	−0.200
38.	FLGGTPVCL	41	6.623	−0.471	−0.561
39.	FLLLADARV	37	7.747	0.502	−0.099
40.	FLLPDAQSI	31	6.415	−0.989	−0.606
41.	FLLRWEQEI	30	7.592	0.680	0.097
42.	FLLSLGIHL	35, 41	8.053[a]	0.715	0.646
43.	FLLTRILTI	41, 42	8.073[a]	0.610	0.250
44.	FLPWHRLFL	30	6.950	0.304	−0.185
45.	FLWGPRALV	30	7.215	−1.105	−0.929
46.	FLYGAALLA	38	8.469	0.525	0.561
47.	FLYGALALA	38	8.620	0.397	0.070
48.	FLYGALLAA	38	8.201	0.006	−0.403
49.	FLYGALLLA	38	8.585	0.298	0.071
50.	FLYGALRLA	38	8.149	0.343	−0.309
51.	FLYGALVLA	38	7.409	−1.038	−1.002
52.	FLYGGLLLA	38	8.959	0.894	0.817
53.	FLYNRPLSV	43	7.212	−1.294	−0.870
54.	FMGAGSKAV	43	6.200	−0.936	−0.817
55.	FTDQVPFSV	40	7.212	0.048	−0.158
56.	FVDYNFTIV	43	6.620	0.948	0.523
57.	FVNHDFTVV	43	6.523	−0.014	−0.527
58.	FVNHRFTVV	43	6.523	0.472	0.106
59.	FVTWHRYHL	36	5.869	−0.044	−0.172
60.	FVVALIPLV	31	8.119	−0.002	−0.035

TABLE I. (Continued)

No.	Peptide sequence	Reference	pIC_{50} ($-\log IC_{50}$)	Residuals according to CoMSIA(1)	Residuals according to CoMSIA(2)
61.	FVWLHYYSV	30, 36	7.821[a]	−0.457	−0.568
62.	GIGILTVIL	34	6.000	−0.461	−0.686
63.	GILTVILGV	30, 34	8.342[a]	0.888	0.306
64.	GIRPYEILA	43	7.481	0.352	0.238
65.	GLACHQLCA	33	6.380	0.262	0.578
66.	GLCFFGVAL	38	5.380	−1.309	−0.895
67.	GLFLTTEAV	31	7.509	0.490	0.371
68.	GLGQVPLIV	44	6.301	−0.908	−0.572
69.	GLIMVLSFL	45	7.658	0.827	0.281
70.	GLLGNVSTV	45	7.620	0.283	−0.168
71.	GLLGWSPQA	41, 42	8.027[a]	0.078	−0.494
72.	GLMTAVYLV	31	8.051	0.501	−0.045
73.	GLQDCTMLV	37	7.638	1.004	0.707
74.	GLSRYVARL	41, 42, 46, 47	7.174[a]	0.664	0.624
75.	GLVDFVKHI	31	6.663	−0.693	0.041
76.	GLYGAQYDV	31	6.602	−0.771	−0.331
77.	GLYLSQIAV	31	7.017	−0.178	0.441
78.	GLYRQWALA	31	6.733	−0.764	−0.432
79.	GLYSSTVPV	41, 46	7.577[a]	−0.804	−0.267
80.	GLYYLTTEV	31	7.682	0.401	0.604
81.	GTLGIVCPI	35, 39	6.666[a]	0.490	0.134
82.	HLAVIGALL	30, 44	6.986[a]	1.097	0.535
83.	HLESLFTAV	46	5.301	−2.464	−2.415
84.	HLLVGSSGL	46	5.792	−0.573	−0.066
85.	HLYQGCQVV	33	6.832	−0.291	−0.879
86.	HLYSHPIIL	46	7.131	0.751	0.080
87.	HMWNFISGI	48	7.818	1.413	0.447
88.	IAATYNFAV	38	7.032	−0.005	−0.553
89.	IAGGVMAVV	43	6.708	0.305	−0.267
90.	IIDQVPFSV	40	7.398	−0.188	−0.002
91.	IISCTCPTV	41	6.580	−0.213	0.049
92.	ILAGYGAGV	35	6.937	−0.740	−0.335
93.	ILAQVPFSV	40	7.939	0.088	0.121
94.	ILDEAYVMA	33	6.623	−0.670	−0.248
95.	ILDQVPFSV	40	7.284	−0.674	−0.506
96.	ILFTFLHLA	31	8.268	0.320	−0.086
97.	ILHNGAYSL	33	7.127	1.286	1.967
98.	ILLLCLIFL	41	6.845	−0.551	−0.351
99.	ILLSIARVV	31	6.342	−1.443	−0.552
100.	ILMQVPFSV	40	8.125	0.007	−0.225
101.	ILSPFMPLL	41	7.347	−0.081	−0.082
102.	ILSQVPFSV	40	7.699	0.099	0.103
103.	ILSSLGLPV	31	7.301	0.079	0.523
104.	ILTVILGVL	34	6.419	−0.175	−0.607
105.	ILWQVPFSV	40	8.770	0.381	0.244
106.	ILYQVPFSV	40	8.310	−0.319	−0.159
107.	IMDQVPFSV	40	7.719	0.025	0.232
108.	IMPGQEAGL	30	7.188	0.663	0.750
109.	ITAQVPFSV	40	7.020	0.300	0.336
110.	ITDQVPFSV	40, 44, 49	6.947[a]	0.209	0.400
111.	ITFQVPFSV	40	7.179	−0.152	−0.225
112.	ITMQVPFSV	40	7.398	0.447	0.244
113.	ITSQVPFSV	40	6.196	−0.375	−0.374
114.	ITWQVPFSV	40	7.457	0.111	−0.038
115.	ITYQVPFSV	40	7.480	−0.011	0.166
116.	IVGAETFYV	29	8.456	1.869	2.829
117.	IVMGNGTLV	43	6.001	−1.114	−0.465
118.	KIFGSLAFL	33	7.478	0.054	0.531
119.	KILSVFFLA	45	8.301	1.534	0.786

TABLE I. (Continued)

No.	Peptide sequence	Reference	pIC_{50} ($-\log IC_{50}$)	Residuals according to CoMSIA(1)	Residuals according to CoMSIA(2)
120.	KLAGGVAVI	31	6.447	−0.575	−0.319
121.	KLFPEVIDL	43	6.693	−1.321	−0.524
122.	KLTPLCVTL	47	6.991	−0.381	−0.337
123.	KTWGQYWQV	44, 49	7.957[a]	1.149	1.657
124.	LIGNESFAL	36	6.380	0.034	0.257
125.	LLACAVIHA	31	6.602	−0.166	−0.063
126.	LLAGLVSLL	30	7.021	0.085	−0.443
127.	LLAQFTSAI	35	7.301	0.462	0.499
128.	LLAVGATKV	44	6.477	0.184	0.520
129.	LLAVLYCLL	30	7.478	1.227	0.867
130.	LLCLIFLLV	41	6.996	0.016	−0.355
131.	LLDVPTAAV	32	7.770	1.306	0.804
132.	LLFGYPVYV	29	7.886	−0.055	0.365
133.	LLFLGVVFL	31	7.301	−0.003	0.125
134.	LLFLLLADA	48	6.663	−0.591	−0.880
135.	LLFRFMRPL	31	7.447	0.609	0.354
136.	LLGCAANWI	46	5.301	−0.810	−0.384
137.	LLGRNSFEV	6	6.447	−0.514	0.409
138.	LLLCLIFLL	41	7.585	0.149	−0.159
139.	LLLEAGALV	30	8.174	1.106	1.801
140.	LLLLGLWGL	43	7.658	0.626	0.548
141.	LLPLGYPFV	31	6.477	−1.498	−0.418
142.	LLPSLFLLL	43	6.903	−0.187	−1.036
143.	LLSCLGCKI	36	5.342	−1.131	−0.674
144.	LLSSNLSWL	46	6.342	−0.257	−0.171
145.	LLVFACSAV	38	6.342	−0.117	−0.131
146.	LLVVMGTLV	36	5.869	−1.113	−0.803
147.	LLWFHISCL	41	6.682	−0.286	−0.597
148.	LLWQDPVPA	31	7.343	0.032	0.093
149.	LLWSFQTSA	30	7.818	1.012	0.694
150.	LMAVVLASL	44	6.954	0.371	−0.038
151.	LMIGTAAAV	31	7.102	0.158	0.584
152.	LMLPGMNGI	31	6.623	0.835	0.346
153.	LQTTIHDII	39	5.501	1.184	0.448
154.	LTVILGVLL	34	5.580	−0.164	0.410
155.	LVSLLTFMI	38	5.716	−0.897	−0.550
156.	MALLRLPLV	31	7.279	0.032	−0.207
157.	MLASTLTDA	31	6.602	0.174	−0.144
158.	MLGNAPSVV	31	6.644	−0.086	−0.378
159.	MLGTHTMEV	44	7.845	0.298	0.768
160.	MLLAVLYCL	49	6.478	−1.430	−1.105
161.	MLQDMAILT	31	6.777	0.332	−0.223
162.	MMWYWGPSL	41	7.921	0.581	0.167
163.	MTYAAPLFV	31	7.860	1.066	1.819
164.	NLGNLNVSI	46	7.119	0.431	0.783
165.	NLQSLTNLL	46	6.000	−0.497	−0.554
166.	NLYVSLLLL	46	7.114	−0.598	−0.360
167.	NMVPFFPPV	30, 36	8.403[a]	0.792	0.442
168.	PLLPIFFCL	35, 41	6.926[a]	−0.760	−1.245
169.	QLFEDNYAL	33	7.764	1.093	1.437
170.	QLFHLCLII	41	6.886	−0.286	0.225
171.	QMTFHLFIA	38	5.778	−0.518	−0.034
172.	QVMSLHNLV	36	6.025	−0.309	−0.777
173.	RIWSWLLGA	30	7.000	−0.993	1.513
174.	RLLDDTPEV	31	7.017	−1.006	−0.433
175.	RLLGSLNST	30	6.778	0.881	0.490
176.	RLLQETELV	33	7.682	−0.275	0.180
177.	RLMIGTAAA	31	6.644	−0.248	−0.263
178.	RLMKQDFSV	30, 44	7.338[a]	−0.311	0.161

TABLE I. (Continued)

No.	Peptide sequence	Reference	pIC_{50} ($-\log IC_{50}$)	Residuals according to CoMSIA(1)	Residuals according to CoMSIA(2)
179.	RLPLVLPAV	31	8.292	1.323	0.730
180.	RLTEELNTI	43	6.060	0.204	0.344
181.	RLVSGLVGA	31	6.818	−0.237	−0.465
182.	RMFAANLGV	31	7.447	0.030	0.012
183.	RMPAVTDLV	31	6.903	0.265	0.351
184.	RMYGVLPWI	38	7.538	−0.662	−0.328
185.	SAANDPIFV	36	5.342	−0.915	−0.418
186.	SIIDPLIYA	30	6.342	−0.574	−0.624
187.	SIISAVVGI	33	7.159	0.435	0.299
188.	SLADTNSLA	30	6.342	0.198	−0.484
189.	SLAGFVRML	31	6.954	0.401	0.212
190.	SLDDYNHLV	30, 36	7.583[a]	0.544	0.518
191.	SLHVGTQCA	34	5.842	0.300	0.355
192.	SLLEIGEGV	31	7.009	−0.279	−0.385
193.	SLLPAIVEL	32	7.620	0.255	0.164
194.	SLLTFMIAA	38	8.027	0.935	0.274
195.	SLNFMGYVI	46	5.881	−0.161	0.547
196.	SLSRFSWGA	38	6.041	−1.034	−0.972
197.	SLYADSPSV	41, 46	7.658[a]	−0.166	0.498
198.	SLYFGGICV	31	7.975	1.613	2.613
199.	SVMDPLIYA	30	7.079	0.690	0.393
200.	SVYDFFVWL	30, 36	7.289[a]	0.119	0.654
201.	SVYVDAKLV	43	6.991	0.705	2.182
202.	TLDSQVMSL	36	6.580	−0.739	−0.850
203.	TLGIVCPIC	35, 39	6.964[a]	0.976	0.089
204.	TLLVVMGTL	36	5.580	−1.667	−1.246
205.	TTAEEAAGI	34	5.380	−0.391	0.299
206.	TVILGVLLL	34	6.072	−0.496	−0.326
207.	TVLRFVPPL	31	7.114	−0.172	−0.353
208.	VALVGLFVL	36	5.079	−1.411	−0.745
209.	VCMTVDSLV	36	5.146	−1.658	−1.085
210.	VIHAFQYVI	38	5.914	−0.069	0.197
211.	VILGVLLLI	34	6.785	−0.296	−0.762
212.	VLAGLLGNV	45	7.721	0.858	0.533
213.	VLAKDGTEV	43	7.174	0.954	2.750
214.	VLHSFTDAI	36	6.170	−0.648	−0.798
215.	VLIQRNPQL	33	7.644	0.225	0.508
216.	VLLDYQGML	35, 41	7.095[a]	0.455	0.107
217.	VLLLDVTPL	31	7.301	0.962	0.731
218.	VLLPSLFLL	43	7.444	0.252	−0.066
219.	VLTALLAGL	30	7.086	0.313	−0.215
220.	VLVGGVLAA	48	6.732	−0.160	−0.408
221.	VMGTLVALV	30, 36	7.547[a]	0.317	−0.202
222.	VVHFFKNIV	38	4.301	−1.497	−0.142
223.	VVLGVVFGI	33	7.845	0.915	0.419
224.	VVMGTLVAL	30, 36	7.069[a]	0.653	0.307
225.	WILRGTSFV	41	6.556	−0.295	−0.042
226.	WLDQVPFSV	40	7.939	0.500	0.258
227.	WLEPGPVTA	40	6.082	−0.328	−0.426
228.	WLLIDTSNA	31	6.447	0.459	0.421
229.	WLSLLVPFV	29, 35, 41, 47	8.164[a]	1.131	0.683
230.	WMNRLIAFA	48	6.914	0.085	0.042
231.	WTDQVPFSV	40	6.145	−0.454	−0.699
232.	YAIDLPVSV	30, 36	7.801[a]	0.067	0.287
233.	YALTVVWLL	38	6.924	−1.160	−1.007
234.	YLAPGPVTA	40	8.032	0.637	0.618
235.	YLAPGPVTV	40	7.818	−0.055	0.123
236.	YLDLALMSV	31	8.260	0.338	0.584
237.	YLDQVPFSV	40	8.638	0.148	0.153

TABLE I. (Continued)

No.	Peptide sequence	Reference	pIC_{50} ($-\log IC_{50}$)	Residuals according to CoMSIA(1)	Residuals according to CoMSIA(2)
238.	YLEPGPVTI	40	7.187	−0.148	0.265
239.	YLEPGPVTL	40	7.058	0.206	0.422
240.	YLEPGPVTV	40	7.342	1.018	1.837
241.	YLFPGPVTA	40	8.495	−0.050	0.096
242.	YLFPGPVTV	40	8.237	−0.222	−0.159
243.	YLLALRYLA	31	8.000	0.655	0.018
244.	YLLPAIVHI	32	7.745	−0.734	−0.620
245.	YLMPGPVTA	40	8.367	0.761	0.547
246.	YLMPGPVTV	40	7.932	−0.217	−0.228
247.	YLSEGDMAA	31	6.532	−0.220	−0.452
248.	YLSPGPVTA	40	7.383	0.203	0.129
249.	YLSPGPVTV	40	7.642	0.016	0.174
250.	YLSQIAVLL	31	7.917	0.636	0.545
251.	YLVAYQATV	35, 37, 50	7.304[a]	−0.712	−0.737
252.	YLVSFGVWI	41	8.721	1.211	1.894
253.	YLVTRHADV	48	6.342	−1.018	−0.177
254.	YLWPGPVTA	40	8.495	0.149	−0.129
255.	YLWPGPVTV	40	8.125	−0.334	−0.419
256.	YLYPGPVTA	40	7.772	−0.480	−0.410
257.	YLYPGPVTV	40	8.051	−0.637	−0.338
258.	YLYVHSPAL	31	8.268	0.307	0.288
259.	YMDDVVLGA	35	6.699	−1.095	−0.644
260.	YMDDVVLGV	47	8.301	0.547	0.366
261.	YMIMVKCWM	33	6.663	−0.788	−0.595
262.	YMLDLQPET	35, 39	7.373[a]	−0.107	−0.779
263.	YMNGTMSQV	49	7.398	−0.494	−0.252
264.	YTDQVPFSV	40	7.066	−0.423	−0.426
265.	YTYKWETFL	31	7.538	0.400	0.041
266.	YVITTQHWL	30, 36	6.877[a]	−0.613	−0.037
Mean residual				−0.003	0.036
Standard deviation				0.703	0.672
Minimum				−2.464	−2.415
Maximum				1.869	2.829
Absolute mean residual				0.553	0.489

[a]The IC_{50} value is an average value of all the cited IC_{50} values.

tioned with the N-terminus to the left and the C-terminus to the right as it is oriented in the binding cleft on the HLA-A*0201 molecule.[52]

Steric bulk

The steric bulk is well tolerated at position 1 (P1) on the plane of the aromatic ring of Phe and Tyr (Fig. 2, upper left). There are also other significant areas, colored green, at position 2 (P2), position 3 (P3), position 5 (P5), and at position 6 (P6). Yellow disfavored areas exist at position 4 (P4), position 7 (P7), and position 8 (P8).

Electrostatic potential

In the electrostatic map, the areas of favored negative potential are colored red, and areas of favored positive potential are colored blue (Fig. 2, upper right). Negative potentials are favored at almost every position. Positive potentials are required between P3 and P5, and at P8.

Local hydrophobicity

Areas of favorable hydrophobicity are shown as yellow polyhedra (Fig. 2, lower left). Areas of hydrophobicity exist at P1, P3, P5, P6, P8, and P9. The favored hydrophilic groups are shown as white polyhedra. They can be found at P4 and P9.

Hydrogen bond donor and acceptor abilities

In CoMSIA, the hydrogen bond donor field describes areas where hydrogen bond acceptor groups should be located on the receptor[53] (Fig. 2, lower right). These areas are colored cyan. Such groups are favored around P3 and at P4. The hydrogen bond acceptor field shows areas where hydrogen bond donor groups should be located on the receptor[53] and these areas are colored magenta. These groups should be at P4, P6, and P8.

DISCUSSION

Predictability of the CoMSIA Models

Unlike artificial intelligence methods such as neural networks and hidden markov models, robust multivariate statistics, principally Partial Least Squares, has received little attention in the bioinformatics literature. Apart from its application to the prediction of subcellular location,[54,55]

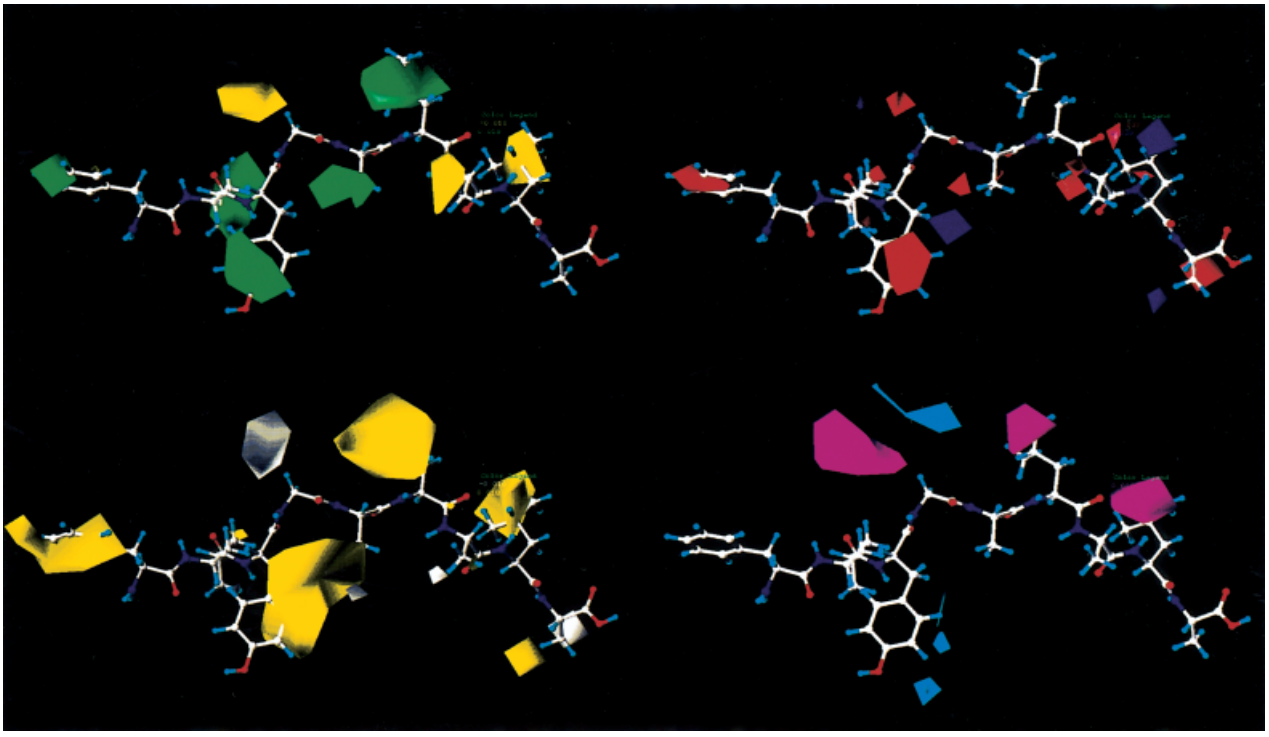

Fig. 2. CoMFA stdev*coeff contour plots. Peptide FLYGALALA is shown inside the field. **Upper left:** Steric map. Green and yellow polyhedra indicate regions where more steric bulk or less steric bulk, respectively, will enhance the affinity. **Upper right:** Electrostatic map. Red and blue polyhedra indicate regions where negative potential or positive potential, respectively, will enhance the affinity. **Lower left:** Hydrophobic map. Yellow and white polyhedra indicate regions where hydrophobic or hydrophilic groups, respectively, will enhance the affinity. **Lower right:** H-bond abilities. Cyan and magenta polyhedra indicate regions where hydrogen bond acceptor or donor groups, respectively, on the receptor will enhance the binding.

and our application of PLS to the prediction of secondary structure content on the basis of amino acid composition,[56] most work in this area has focused on the development of peptide QSAR for the analysis of small peptides.[57,58] However, we can find only one article that applies these methods in the area of MHC binding.[59] Artificial intelligence methods, such as neural networks, have several disadvantages that are addressed by PLS methods. The problem with neural networks, as with all such nonlinear methods, including nonlinear PLS, is that they perform much better in fitting existing data (the phenomenon of memorization or overfitting), than in predicting new data[56]; their ability to interpolate is greater than their capacity to accurately extrapolate. Moreover, neural networks also suffer from problems of overtraining, the influence of chance effects, and significant problems of interpretation.[60,61] They may, on occasion, perform usefully as a prediction engine but offer very limited explanatory powers compared with 3D QSAR methods, such as CoMFA or CoMSIA.

As a modeling process, 3D-QSAR is based on a series of assumptions. The binding of a ligand L to a macromolecule R can be expressed as:

$$\Delta G_{bind}^{L-R} = \Delta G_{sol} + \Delta G_{vac}$$

where $\Delta G_{bind}^{L\text{-}R}$, the free energy of binding between a ligand and its receptor, can be separated into solvation effects (ΔG_{sol}) and effects in a vacuum (ΔG_{vac}). The process in a vacuum can be described by two components: enthalpic (ΔH_{vac}) and entropic ($T\Delta S_{vac}$). 3D-QSAR methods make the assumption that ΔG_{sol} and $T\Delta S_{vac}$ have similar values for congeneric series. In CoMSIA, part of the entropic component also is taken into account. The enthalpy of the process in vacuum ΔH_{vac} can be expressed as:

$$\Delta H_{vac} = \Delta H^{L-R} + \Delta H_{distort}^{L} + \Delta H_{distort}^{R}$$

where $\Delta H^{L\text{-}R}$ is the enthalpy of interaction, $\Delta H_{distort}^{L}$ and $\Delta H_{distort}^{R}$ are the distortions induced by this interaction in ligand and receptor, respectively. In all grid based methods, $\Delta H_{distort}^{R}$ is not considered because the receptor is approximated by a rigid grid, whereas $\Delta H_{distort}^{L}$ is indirectly included by the choice of a conformer.

Another assumption is that the modeled conformation is the bioactive one. For very flexible ligands, this assumption is seldom correct, even if the binding conformations of one or five congeners in the training set are known. We may be looking at two or more series with fundamentally different conformations, which we are forcing to act as a single series of molecules for the purposes of our analysis. This may be manifest as the division of the data set in a small number of well-populated groups or as a set of conformational outliers, each significantly different from the core group. Although the predictivity of our model would not necessarily support this view, only crystallo-

graphic data are likely to categorically resolve this issue. The practical outcome of this assumption is the reduced predictability and explanatory power of the models.

A number of factors differentiate the current analysis from the study of organic small molecules within a pharmaceutical context, the more usual domain of QSAR analyses. These include the size of the peptide molecules being studied; the shear number of molecules being investigated, perhaps 10-fold greater than a typical small molecule study; and the great diversity of physicochemical properties associated with the amino acids being examined at each position of the peptide. We have also avoided the issue of molecular alignment by assuming a constant backbone structure for all peptides. We know from X-ray analyses,[1–3] that there are some small differences in backbone conformation for nonamer peptides, but these are much larger for decamers. Given the number of peptides under study, allowing for conformational flexibility in the backbone is not an obviously tractable problem. Because 88.0% [CoMSIA(1) model] or 92.5% [CoMSIA(2) model] of the peptides are well and very well predicted, the variations in the binding conformation do not seem to be significant for most of the peptides.

Another possibility for the bad predictions could be the great variety of sources (about 20) used in the present study. But no relationship between the poorly predicted peptides and sources they were collected from was apparent. Finally, the poorly predicted peptides could exhibit properties significantly different from those in our set. This is true only for peptides PLLPIFFCL and VCMTVDSLV. The former is the only one in the set carrying Pro at position 1, and the latter is the only one with Cys at position 2.

The CoMSIA(2) model increases significantly the number of very well predicted peptides (from 151 to 168), decreases considerably the number of poor predictions (from 32 to 20), and weakly affects the well-predicted peptides (from 83 to 78). According to the current QSAR practice, predictions within 1.0 log unit are considered good.[20,62–64] This would result in mean residuals of around 0.5 log unit. The mean absolute value for the residuals according to CoMSIA(1) model is 0.553; according to CoMSIA(2), it is even better: 0.489. The experimental, or biological, error in the measurements is, in terms of logs, probably much great than 0.3. In this context, average predictions more accurate than these values cannot be realized.

3D-QSAR Analysis of the Peptide Structure

It was found that all nine side-chains of the bound peptides contact the HLA-A*0201 molecule.[5] The antigen-binding groove has a 30 Å long surface accessible to a solvent probe. There are six pockets in the surface denoted A–F.[7] Some of them are nonpolar and can form hydrophobic contacts, but others contain polar atoms and can make hydrogen bonds with the side-chains.

As a statistical approach, CoMSIA seeks to correlate relative differences of discriminating molecular descriptors with a dependent property (e.g., the binding affinity). In that respect, CoMSIA is a method able to map similarities or dissimilarities between molecules. The five physicochemical properties (steric bulk, electrostatic potential, local hydrophobicity, and hydrogen bond donor and acceptor properties) assessed by CoMSIA were considered below for each position in the peptide structure. Some of these properties were analyzed in our previous article based on a lower number of peptides.[24] In the present analysis, we highlight the correlations between all the five descriptors and the binding affinities of twice the number of peptides. Where experimental evidences are available to support our results, we have provided them.

Hydrophobic steric bulk with negative potential is well tolerated at P1 (Fig. 2). In our first investigation,[24] many areas of hydrogen bond donor groups were found near the N-terminus. These areas are absent in the present map (Fig. 2, lower right) because no changes in the hydrogen atoms positions at N-terminal exist because of the automatic binding of peptides. Topologically, P1 corresponds to pocket A.[3] The surface of this pocket is predominantly polar: five Tyr hydroxyl groups (Tyr7, Tyr59, Tyr99, Tyr159, and Tyr171), a carboxyl group (Glu63), and 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).[2] The most suitable amino acids for this position seem to be Phe and Tyr. Independently of our studies, it was recently reported that the substitution of Ile at P1 with Phe or Tyr in the HIV reverse transcriptase (RT) peptide (309–317) (ILKEPVHGV) increased by threefold the cell surface half-life of complexes.[8,9] A π-π stacking interactions between Trp167 and the aromatic P1 residues was proposed to account for this change.[8] Moreover, Tourdot et al.[9] report that the P1Y substitution in 10 nonimmunogenic 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.

The steric map at P2 indicates that long side-chains like Leu, Ile, and Met are well tolerated here (Fig. 2, upper left). This is in good agreement with many experimental data.[6,10,65,66] 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.[3]

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,[7] and it belongs to the so-called secondary binding pockets. It is a hydrophobic cavity located between the aromatic rings of Tyr99 and Tyr159, including also residues 155, 156, and 160.[1] This pocket prefers large hydrophobic residues like Phe and Trp.[67] The hydrogen-bonding ability map indicates that amino acids able to form hydrogen bonds also will be well accepted here.

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.[8] suggest hydrogen bond formation between Tyr at P1 and Glu at P4 bridged by a

water molecule, which should make the bound peptide more rigid and easily recognized by T-cell receptors. The side-chain at P4 is called the "flag" residue[7] because it is solvent exposed in the complex with MHC molecule and, therefore, it can contact the TCR.

The maps indicate that amino acids with hydrophobic branched or aromatic side-chains are well tolerated at P5. It seems that Phe and Trp are suitable for this position. The small hydrophilic area near the end of the side-chain indicates that Tyr or His also are favored here.

Amino acids with long hydrophobic side-chains are preferred at P6. This side-chain falls into pocket C.[3] This pocket is predominantly polar, made up of Thr73, His70, His74, and Arg97. Hydrogen bond ability is an additional priority. Leu, Ile, Thr, and Tyr are well accommodated here.

Short side-chains are favored sterically at P7. 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.[3] 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 the "part-time" pocket,[7] and it belongs to the class of secondary binding pockets. Pro, His, Ala, and Val will be well accepted at this position.

The side-chain at P8 should be short, with a hydrophobic core and an end capable of forming hydrogen bonds. Ser or Thr seem to be preferred here. Trp147 hydrogen bonds to the P8 carbonyl oxygen.[7] P8 is a "flag" position like P4.

Amino acids with hydrophobic short side-chains like Ala and Val are required for P9. The side-chain of Tyr116 occupies the end of the pocket F and is uncharged, so that the binding site is complementary to small hydrophobic side-chains.[3,5] It is of interest that a small hydrophilic area carrying negative potential appears near P9. It is due obviously to Thr introduced here by the intermediate binder MLQDMAILT and the high binder YMLDLQPET (Table I). The Tyr116 side-chain hydroxyl group forms a hydrogen bond to Asp77 on the α_1 helix, stabilizing it in this orientation.[7] Tyr84, Thr143, and positively charged Lys146 bind to the carboxyl group of the C-terminal.[7]

CONCLUSIONS

The coefficient contour maps obtained by CoMSIA show how substitutions of particular functional groups or residues, which exhibit favorable or unfavorable physicochemical properties, will affect affinity. The CoMSIA method is also predictive, allowing the binding affinity of candidate epitopes to be estimated accurately. CoMSIA maps can also be used in the design of new T-cell receptor agonist or antagonist peptidomimetic compounds.[68,69] Our results are highly complementary to the analysis of X-ray crystal structures but contains data on hundreds of binding peptides. The maps summarize these data, allowing it to be visualized and interpreted easily. The present study shows the utility of the CoMSIA method as an aid in the analysis of structure-activity relationships of both peptide-protein complexes in general and MHC-peptide interactions in particular.

REFERENCES

1. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 1987;329:506–512.
2. 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;329:512–518.
3. 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.
4. Falk K, Rötzschke O, Stefanovic S, Jung G, Rammensee H-G. Allele specific motifs revealed by sequencing of self peptides eluted from MHC molecules. Nature 1991;351:290–296.
5. 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.
6. 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;74:929–937.
7. Madden DR. The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 1995;13:587–622.
8. 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;274:37259–37264.
9. Tourdot S, Scardino A, Saloustrou E, Gross DA, Pascolo S, Cordopatis P, Lemonnier FA, Kosmatopoulos KA. 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;30:3411–3421.
10. Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chain. J Immunol 1994;152:163–175.
11. 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.
12. Gulukota K, Sidney J, Sette A. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 1997;267:1258–1267.
13. Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. J Immunol Methods 1995;185:181–190.
14. Brusic V, Rudy G, Honeyman M, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using evolutionary algorithm and artificial neural network. Bioinformatics 1998;14:121–130.
15. Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. Hum Immunol 1997;58:1–11.
16. Rognan D, Lauemøller SL, Holm A, Buus S, Tschinke, 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;42:4650–4658.
17. Logean A, Sette A, Rognan D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. Bioorg Med Chem Lett 2001;11:675–679.
18. Mata M, Travers PJ, Liu Q, Frankel FR, Paterson Y. The MHC class I-restricted immune response to HIV-gag in BALB/c mice selects a single epitope that does not have a predictable MHC-binding motif and binds to K^d through interactions between a glutamine at P3 and pocket D. J Immunol 1998;161:2985–2993.
19. Meng WS, von Grafenstein H, Haworth IS. Water dynamics at the binding interface of four different HLA-A2-peptide complexes. Int Immunol 2000;12:949–957.
20. 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;37:4130–4146.
21. Klebe G, Abraham U. Comparative Molecular Similarity Index Analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput Aid Mol Des 1999;13:1–10.

22. 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: 458–477.
23. Crammer RD III, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110:5959–5967.
24. Doytchinova I, Flower DR. Towards the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity to class I MHC molecule HLA-A*0201. J Med Chem 2001;44:3572–3581.
25. SYBYL 6.6. Tripos Inc, 1699 Hanley Road, St Louis, MO 63144.
26. Powell MJD. Restart procedures for the conjugate gradient method. Math Progr 1977;12:241–254.
27. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 1985;107:3902–3909.
28. Cramer RD III, 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.
29. Del Guercio M-F, Sidney J, Hermanson G, Perez C, Grey HM, Kubo RT, Sette A. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J Immunol 1995;154:685–693.
30. Reynolds SR, Celis E, Sette A, Oratz R, Shapiro RL, Johnston D, Fotino M, Bystryn JC. HLA-independent heterogeneity of CD8+ T cell responses to MAGE-3, Melan-A/MART-1, gp100, tyrosinase, MC1R, and TRP-2 in vaccine-treated melanoma patients. J Immunol 1998;161:6970–6976.
31. Cho S, Mehra V, Thoma-Uszynski S, Stenger S, Serbina N, Mazzaccaro RJ, Flynn JAL, Barnes PF, Southwood S, Celis E, Bloom BR, Modlin RL. Antimicrobial activity of MHC class I-restricted CD8+ T cells in human tuberculosis. Proc Natl Acad Sci USA 2000;97:12210–12215.
32. Chen Y, Sidney J, Southwood S, Cox AL, Sakaguchi K, Henderson RA, Appella E, Hunt DF, Sette A, Engelhard VH. Naturally processed peptides longer than nine amino acid residues bind to the class I MHC molecule HLA-A2.1 with high affinity and in different conformations. J Immunol 1994;152:2874–2881.
33. 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, Kiessling R. Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas. J Immunol 1999;163:1037–1044.
34. Rivoltini L, Kawakami Y, Sakaguchi K, Southwood S, Sette A, Robbins PF, Marincola FM, Salgaller ML, Yannelli JR, Appella E, Rosenberg SA. Induction of tumor-reactive CTL from peripheral blood and tumor-infiltrating lymphocytes of melanoma patients by in vitro stimulation with an immunodominant peptide of the human melanoma antigen MART-1. J Immunol 1995;154:2257–2265.
35. Wentworth PA, Vitiello A, Sidney J, Keogh E, Chesnut RW, Grey H, Sette A. Differences and similarities in the A2.1-restricted cytotoxic T cell repertoire in humans and human leukocyte antigen-transgenic mice. Eur J Immunol 1996;26:97–101.
36. Parkhurst MR, Fitzgerald EB, Southwood S, Sette A, Rosenberg SA, Kawakami Y. Identification of a shared HLA-A*0201-restricted T-cell epitope from the melanoma antigen tyrosinase-related protein 2 (TRP2). Cancer Res 1998;58:4895–4901.
37. Battegay M, Fikes J, Di Bisceglie AM, Wentworth PA, Sette A, Celis E, Ching WM, Grakoui A, Rice CM, Kurokohchi K. Patients with chronic hepatitis C have circulating cytotoxic T cells which recognize hepatitis C virus-encoded peptides binding to HLA-A2.1 molecules. J Virol 1995;69:2462–2470.
38. Dressel A, Chin JL, Sette A, Gausling R, Höllsberg P, Hafler DA. Autoantigen recognition by human CD8 T cell clones: enhanced agonist response induced by altered peptide ligands. J Immunol 1997;159:4943–4951.
39. Kast WM, Brandt RMP, Sidney J, Drijfhout J-W, Kubo RT, Grey HM, Melief CJM, Sette A. Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. J Immunol 1994;152:3904–3911.
40. Parkhurst MR, Salgaller ML, Southwood S, Robbins PF, Sette A, Rosenberg SA, Kawakami Y. Improved induction of melanoma-reactive CTL with peptides from the melanoma antigen gp100 modified at HLA-A*0201-binding residues. J Immunol 1996;157: 2539–2548.
41. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief CJM, Oseroff C, Yuan L, Ruppert J, Sidney J, del Guercio M-F, Southwood S, Kubo RT, Chesnut RW, Grey HM, Chisari FV. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J Immunol 1994;153:5586–5592.
42. Bertoni R, Sidney J, Fowler P, Chesnut RW, Chisari FV, Sette A. Human histocompatibility leukocyte antigen-binding supermotifs predict broadly cross-reactive cytotoxic T lymphocyte responses in patients with acute hepatitis. J Clin Invest 1997;100:503–513.
43. Wizel B, Palmieri M, Mendoza C, Arana B, Sidney J, Sette A, Tarleton R. Human infection with *Trypanosoma cruzi* induces parasite antigen-specific cytotoxic T lymphocyte responses. J Clin Invest 1998;102:1062–1071.
44. Tsai V, Southwood S, Sidney J, Sakaguchi K, Kawakami Y, Appella E, Sette A, Celis E. Identification of subdominant CTL epitopes of the gp100 melanoma-associated tumor antigen by primary in vitro immunization with peptide-pulsed dendritic cells. J Immunol 1997;158:1796–1802.
45. Doolan DL, Hoffman SL, Southwood S, Wentworth PA, Sidney J, Chesnut RW, Keogh E, Appella E, Nutman TB, Lal AA, Gordon DM, Oloo A, Sette A. Degenerate cytotoxic T cell epitopes from P. falciparum restricted by multiple HLA-A and HLA-B supertype alleles. Immunity 1997;7:97–112.
46. Vitiello A, Sette A, Yuan L, Farness P, Southwood S, Sidney J, Chesnut RW, Grey HM, Livingston B. Comparison of cytotoxic T lymphocyte responses induced by peptide or DNA immunization: implications on immunogenicity and immunodominance. Eur J Immunol 1997;27:671–678.
47. Ishioka GY, Fikes J, Hermanson G, Livingston B, Crimi C, Qin M, del Guercio M F, Oseroff C, Dahlberg C, Alexander J, Chesnut RW, Sette A. Utilization of MHC class I transgenic mice for development of minigene DNA vaccines encoding multiple HLA-restricted CTL epitopes. J Immunol 1999;162:3915–3925.
48. Scognamiglio P, Accapezzato D, Casciaro MA, Cacciani A, Artini M, Bruno G, Chircu ML, Sidney J, Southwood S, Abrignani S, Sette A, Barbara V. Presence of effector CD8+ T cells in hepatitis C virus-exposed healthy seronegative donors. J Immunol 1999;162: 6681–6689.
49. Kawakami Y, Eliyahu S, Jennings C, Sakaguchi K, Kang X, Southwood S, Robbins PF, Sette A, Appella E, Rosenberg SA. Recognition of multiple epitopes in the human melanoma antigen gp100 by tumor-infiltrating T lymphocytes associated with in vivo tumor regression. J Immunol 1995;154:3961–3968.
50. Bertoni R, Sette A, Sidney J, Guidotti LG, Shapiro M, Purcell R, Chisari FV. Human class I supertypes and CTL repertoires extend to chimpanzees. J Immunol 1998;161:4447–4455.
51. Sette A, Sidney J, del Guercio M-F, Southwood S, Ruppert J, Dalberg C, Grey HM, Kubo RT. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. Mol Immunol 1994;31:813–822.
52. 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;88:11325–11329.
53. The Tripos implementation of CoMSIA uses a nomenclature opposite to that used in Ref. 21, in accordance with modifications made by the original authors.
54. Sjöström M, Wold S, Wieslander A, Rilfors L. Signal peptide amino acid sequences in Escherichia coli contain information related to final protein localization: a multivariate data analysis. EMBO J 1987;6:823–831.
55. Edman M, Jarhede T, Sjüstrüm M, Wieslander A. Different sequence patterns in signal peptides from Mycoplasmas, other Gram-positive bacteria, and Escherichia coli: multivariate data analysis. Proteins 1999;35:195–205.
56. Clementi M, Clementi S, Cruciani G, Pastor M, Davis A M, Flower DR. Robust multivariate statistics and the prediction of protein secondary structure content. Protein Eng 1997;10:747–749.
57. Hellberg S, Sjöström M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships: a multivariate approach. J Med Chem 1987;30:1126–1135.
58. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New

chemical descriptors relevant for the design of biologically active peptides: a multivariate characterization of 87 amino acids. J Med Chem 1998;41:2481–2491.
59. Rovero P, Riganelli D, Fruci D, Vigano S, Pegoraro S, Revoltella R, Greco G, Butler R, Clementi S, Tanigaki N. The importance of secondary anchor residue motifs of HLA class I proteins: a chemometric approach. Mol Immunol 1994;31:549–554.
60. Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks: advantages and limitations. J Comput Aid Mol Des 1997;11:135–142.
61. Manallack DT, Livingstone DJ. Neural networks in drug discovery: have they lived up to their promise. Eur J Med Chem 1999;34:195–208.
62. Sicsic S, Serraz I, Andrieux J, Bremont B, Mathé-Allainmat M, Poncet A, Shen S, Langlois M. Three-dimensional quantitative structure-activity relationship of melatonin receptor ligands: a Comparative Molecular Field Analysis Study. J Med Chem 1997; 40:739–748.
63. 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;41:1815–1826.
64. Ducrot P, Legraverend M, Grierson DS. 3D-QSAR CoMFA on cyclin-dependent kinase inhibitors. J Med Chem 2000;43:4098–4108.
65. Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, Zhu N-Z, Arnott D, Sherman N, Shabanowitz J, Michel H, Bodnar WM, Davis TA, Hunt DF. Definition of specific peptide motifs for four major HLA-A alleles. J Immunol 1994;152:3913–3924.
66. Parker KC, Bednarek MA, Hull LK, Utz U, Cunningham B, Zweerink HJ, Biddison WE, Coligan JE: sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. J Immunol 1992;149:3580–3587.
67. 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;102:1239–1248.
68. Hin S, Zabel C, Bianco A, Jung G, Walden P. Cutting edge: N-hydroxy peptides: a new class of TCR antagonists. J Immunol 1999;163:2363–2367.
69. Poenaru S, Lamas JR, Folkers G, Lopez de Castro JA, Seebach D, Rognan D. Nonapeptide analogues containing (R)-3-hydroxybutanoate and beta-homoalanine oligomers: synthesis and binding affinity to a class I major histocompatibility complex protein. J Med Chem 1999;42:2318–2331.