In Silico Prediction of Peptide Binding Affinity to Class I Mouse Major Histocompatibility Complexes: A Comparative Molecular Similarity Index Analysis (CoMSIA) Study

Channa K. Hattotuwagama,* Irini A. Doytchinova, and Darren R. Flower

Edward Jenner Institute for Vaccine Research, Compton, Berkshire, U.K. RG20 7NN

Received November 4, 2004

Current methods for the in silico identification of T cell epitopes (which form the basis of many vaccines, diagnostics, and reagents) rely on the accurate prediction of peptide−major histocompatibility complex (MHC) affinity. A three-dimensional quantitative structure−activity relationship (3D-QSAR) for the prediction of peptide binding to class I MHC molecules was established using the comparative molecular similarity index analysis (CoMSIA) method. Three MHC alleles were studied: H2-D^b, H2−K^b, and H2−K^k. Models were produced for each allele. Each model consisted of five physicochemical descriptors—steric bulk, electrostatic potentials, hydrophobic interactions, and hydrogen-bond donor and hydrogen-bond acceptor abilities. The models have an acceptable level of predictivity: cross-validation leave-one-out statistical terms q^2 and SEP (standard error of prediction) ranged between 0.490 and 0.679 and between 0.525 and 0.889, respectively. The non-cross-validated statistical terms r^2 and SEE (standard error of estimate) ranged between 0.913 and 0.979 and between 0.167 and 0.248, respectively. The use of coefficient contour maps, which indicate favored and disfavored areas for each position of the MHC-bound peptides, allowed the binding specificity of each allele to be identified, visualized, and understood. The present study demonstrates the effectiveness of CoMSIA as a method for studying peptide−MHC interactions. The peptides used in this study are available on the Internet (http://www.jenner.ac.uk/AntiJen). The partial least-squares method is available commercially in the SYBYL molecular modeling software package.

INTRODUCTION

T cells play a central role in the immune system, searching for the presence of intracellular pathogens, as infected cells exhibit on their surface peptide fragments derived from pathogen proteins.[1] Specialized host cell glycoproteins, known as major histocompatibility complex (MHC) molecules, transport these foreign peptides to the cell surface, where T cells detect peptide−MHC complexes and kill the infected cells. The antigen peptides presented to T cells are known as epitopes. There are two classes of MHC molecules. MHC class I molecules deliver peptides from the cytosol to the cell surface and are recognized by CD8+ T cells, destroying, for example, cancer and virally infected cells. MHC class II molecules deliver peptides from the interstitial space to the cell surface and are recognized by CD4+ T cells, which are "helper" cells that facilitate and coordinate the immune response to infection. MHCs bind both endogenous, also known as self-peptides, and exogenous, or pathogen-derived, peptides, but as a general rule, MHC class I proteins present predominantly endogenous peptides of 8−11 amino acids and, occasionally, longer peptides, whereas MHC class II proteins mainly present exogenous peptides, which have a much wider distribution of lengths.

The mouse, the primary experimental animal in immunology, has not received as much attention as its pre-eminent position as an object of immunological investigation might warrant. The H2 genes are part of the mouse MHC and form a multigene cluster containing three major gene classes: class I located in the H2-L, H2-D, H2-K, Qa, and H2-T18 regions; class II located in the H2-I region; and class III in the H2-S region. MHC class I gene products of the H2-D and H2-K regions are found on most cells, except in very early embryos, and function in cytolytic immune responses. Differences at these loci can induce vigorous graft rejection and strong primary in vitro cytotoxic responses.

Class I molecules usually bind nonapeptides in an extended conformation with a kink near P4. Octapeptides bind with a less acute kink, while decapeptides and longer peptides bind with much more pronounced structural deviations. Some peptide side chains are deeply bound in "pockets" within the peptide-binding groove of MHC molecules and are often called "anchors". Crystallographic studies of the mouse MHC class I molecules reveals that the amino acid and carboxyl termini of high affinity 8−10-mers peptides are bound in the groove by conserved hydrogen-bond networks.[2−4] The side chains of the bound peptides occupy various pockets (A−F) within the binding groove formed between the long α-1 and α-2 helices and the β-sheet platform.[5] A pocket within the binding site exhibits affinity for the corresponding peptide side chain. Pockets differ significantly in their specificity for peptide side chains. Anchors for most mouse class I alleles are found at position P2 (pocket B) and at the C-terminal residue position P9 (pocket F), and for certain alleles, other anchor residues may bind in the center of the groove. Other peptide side chains also contact the binding site, but are not bound within specific pockets. Pockets A, B, C, and F are viewed as deep pockets, while pockets D and E are shallow.

* Author to whom correspondence should be addressed. Tel.: 44-(0)-1635-577954. Fax: 44-(0)-1635-577901. E-mail: channa.hattotuwagama@jenner.ac.uk.

10.1021/ci049667l CCC: $30.25
Published on Web 08/06/2005

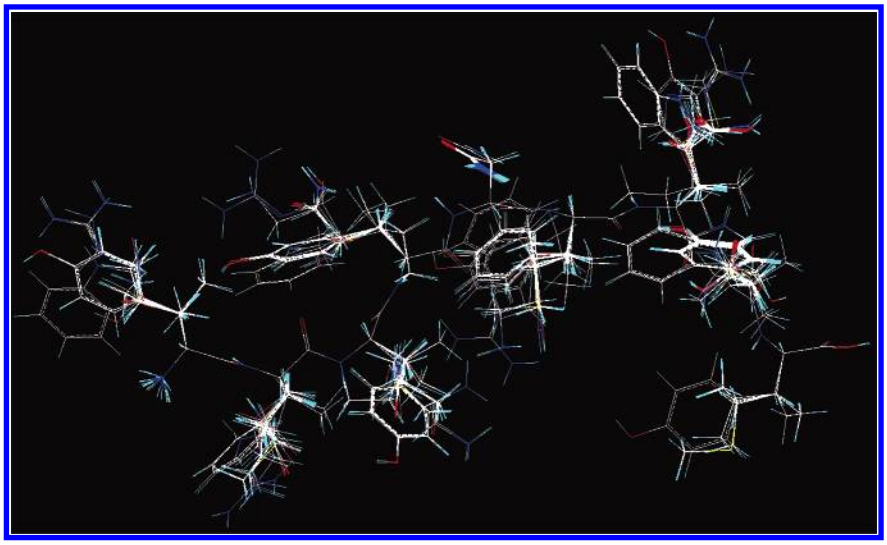

Figure 1. Superimposed alignment of peptide molecules for the H2-D^b allele.

Three-dimensional quantitative structure−activity relationship (3D-QSAR) analysis uses robust multivariate statistics to identify correlations between molecular descriptors generated in the space surrounding series of ligand structures and their binding affinities. The value of 3D-QSAR studies is often greatly enhanced when analyzed in the context of high-resolution ligand−receptor structures. In such cases, enthalpic changes—van der Waals and electrostatic interactions—and entropic changes—conformational and solvent-mediated interactions—in ligand binding can be compared with structural changes in both ligand and macromolecule, providing insight into the binding mechanism.[6,7] 3D-QSAR techniques have become pre-eminent because of their robustness and interpretability.[8] The widely used comparative molecular field analysis (CoMFA) method calculates steric and electrostatic properties according to Lennard-Jones and Coulomb potentials. The more recently reported comparative molecular similarity index analysis (CoMSIA) method uses fields based on the intermolecular interactions (steric, electrostatic, hydrophobic, hydrogen-bond formation) within a molecular binding site. The most important contributions responsible for binding affinity are covered by these properties. A Gaussian-type functional form is used so that no arbitrary threshold is required and interactions can be calculated at all grid points. The obtained relationships are evaluated using partial least-squares (PLS) analysis.[9] CoMSIA allows each physicochemical descriptor to be visualized in 3D using a map, which denotes the areas within the binding site that are either "favored" or "disfavored" by the presence of a group with a particular physicochemical property.

Recently, CoMSIA has been used to produce predictive models for peptide binding to human MHCs: HLA-A*0201[10] and the HLA-A2 and HLA-A3 supertypes.[11,12] In this study, we have applied CoMSIA to three mouse class I MHC alleles: H2-D^b, H2−K^b, and H2−K^k. These models were used to evaluate the physicochemical requirements for binding. The explanatory power of such a 3D-QSAR method is considerable, not only in its direct prediction accuracy but also in its ability to map advantageous and disadvantageous interaction potentials onto the structures of the peptides being studied. The data are highly complementary to the detailed information obtained from crystal structures of individual peptide−MHC complexes.

RESULTS

CoMSIA Models. For each of the H2-D^b, H2-K^b, and H2-K^k alleles, all peptides were built and their geometries optimized and then aligned on the basis of their backbone atoms in three dimensions (Figure 1). The AM1 force field was used within SYBYL 6.9[13] for geometry optimization. The final aligned peptides were placed in three separate 3D lattices (Figure 2). In terms of q^2, the generated models have acceptable predictive ability and high explained variance for all three alleles. It is possible to overemphasize the usefulness of cross-validation and q^2 as measures of performance: high values of q^2_{LOO} are a necessary, but not a sufficient, condition for a model to possess significant predictive power. The cross-validation leave-two-out (q_{CV2}^2) and cross-validation leave-five-out (q_{CV5}^2) values were, thus, also generated for each allele. These values are relatively close to the corresponding cross-validation leave-one-out (q_{LOO}^2) values, indicating a good stability of the model, with the exception of the H2-K^b allele. All statistical results are summarized in Table 1.

Progressive scrambling of the CoMSIA models was also carried out to analyze the stability of the models. Each allele was subjected to a progressive scrambling by 10× (the default value) and 30× in order to test the strength of the models compared to the cross-validation leave-one-out (CV-LOO) results. The statistical results of the scrambling can be seen in Table 2. When comparing the results from Table 2 to those in Table 1 (CV-LOO), we can see some degree of stability within the models.

External test sets and randomization of the training data are also important criteria for assessing model quality. The regression equations for each allele were used to predict the binding affinities of an external, independent test set of peptides by leaving out a random group of compounds that are not used in the cross-validation process (i.e., in the model building) and making predictions for this external set. For the H2-D^b and H2-K^b alleles, too few peptides were available to allow the construction of valid training and test sets. However, with the H2-K^k allele, we were able to select

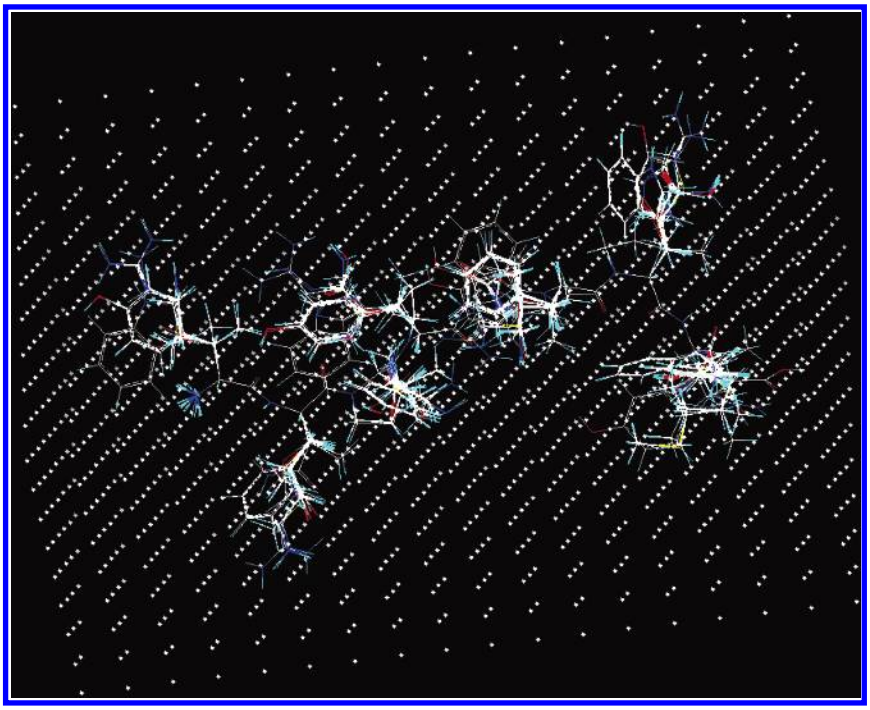

Figure 2. Superimposed H2-D^b peptide molecules placed within a 3D grid lattice.

Table 1. Summary of CoMSIA Models

	H2-D^b	H2-K^b	H2-K^k
N^a	64	62	154
NC^b	6	6	6
q^2_{LOO} [c]	0.679	0.490	0.611
q^2_{CV2} [d]	0.466	0.223	0.521
q^2_{CV5} [e]	0.617	0.385	0.581
SEP[f]	0.651	0.889	0.525
r^2	0.979	0.962	0.913
SEE[g]	0.167	0.244	0.248
Training Set			
N			112
NC			5
q^2_{LOO}			0.405
SEP			0.635
SEE			0.314
r^2			0.855
Test Set			
N			42
r			0.882
r^2			0.778
grid size (Å)	18 × 14 × 11	19 × 13 × 11	18 × 13 × 12
grid spacing (Å)	2	2	2
grid steps (Å)	0.5	0.5	0.5

[a] Number of peptides. [b] Number of components. [c] q^2 obtained after cross validation-leave-one-out. [d] q^2 obtained after cross validation-leave-two-out. [e] q^2 obtained after cross validation-leave-five-out. [f] Standard error of prediction. [g] Standard error of estimate.

random training and test sets consisting of 112 and 42 peptides, respectively, which approximates a 70:30 ratio. The validation of the external test set ($r^2 = 0.778$) showed significant predictive ability, see Table 1. Ultimately, however, we are limited by the data itself. In light of this, we have attempted and succeeded in producing useful, if imperfect, models with clear utilitarian value.

CoMSIA Contour Maps. To generate CoMSIA coefficient contour maps for each allele, which describe the relationship between binding affinity and each physicochemical descriptor, three non-cross-validated "all fields" models

Table 2. Progressive Scrambling: Pertinent Statistics as a Function of the Number of Components for the CoMSIA Models

	scrambling	10×		30×	
	NC[a]	q^2	SEP	q^2	SEP
H2-D^b	2	0.356	0.886	0.333	0.902
	3	0.478	0.806	0.447	0.829
	4	0.531	0.772	0.499	0.797
	5	0.516	0.792	0.483	0.819
	6	0.469	0.840	0.440	0.863
H2-K^b	2	0.212	1.058	0.216	1.056
	3	0.201	1.075	0.201	1.076
	4	0.232	1.064	0.227	1.068
	5	0.249	1.062	0.242	1.068
	6	0.254	1.069	0.244	1.076
H2-K^k	2	0.275	0.706	0.288	0.700
	3	0.383	0.653	0.391	0.649
	4	0.387	0.652	0.398	0.646
	5	0.408	0.643	0.417	0.638
	6	0.397	0.651	0.402	0.648

[a] Number of components.

were created on the basis of the five physicochemical descriptors. Coefficient contour maps are given in Figures 3A−C, 4A−C, 5A−C, 6A−C, and 7A−C for the H2-D^b, H2-K^b, and H2-K^k alleles. For simplicity, the interaction between only one peptide and its respective contour map is shown. All three alleles show the peptide positioned with the N terminus to the left and the C terminus to the right. Table 3 summarizes the preferences for different physicochemical fields at each peptide position. Where no entries are shown, there is no significant interaction between the peptide and MHC for that physicochemical descriptor.

DISCUSSION

Our study has identified favored and disfavored regions that are consistent with both the properties of peptide positions and those of pockets (designated by A−F) within the MHC binding groove. It is well-known that each class I

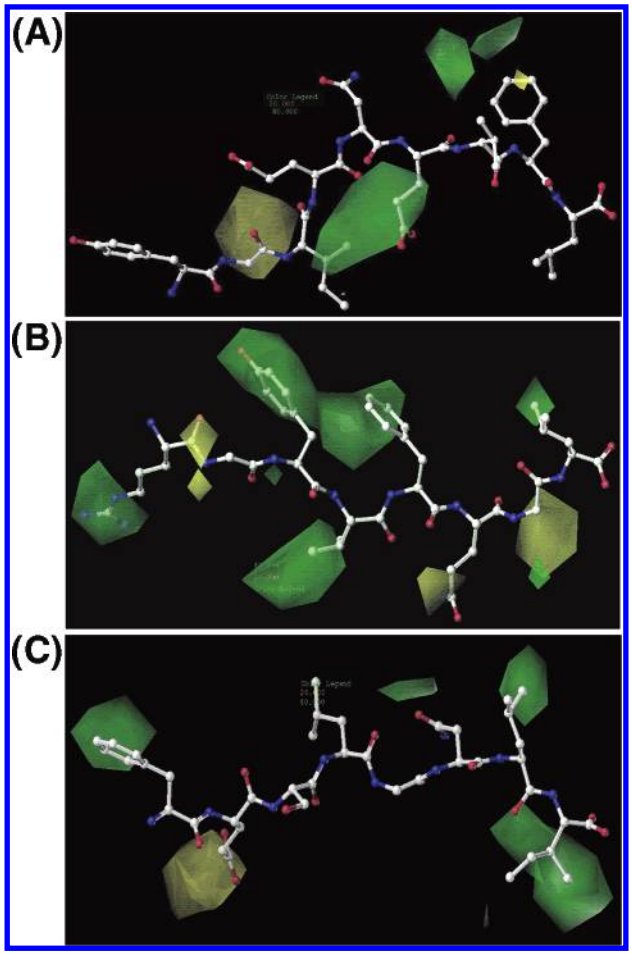

Figure 3. (A) H2-D^b, (B) H2-K^b, and (C) H2-K^k Steric Bulk Maps

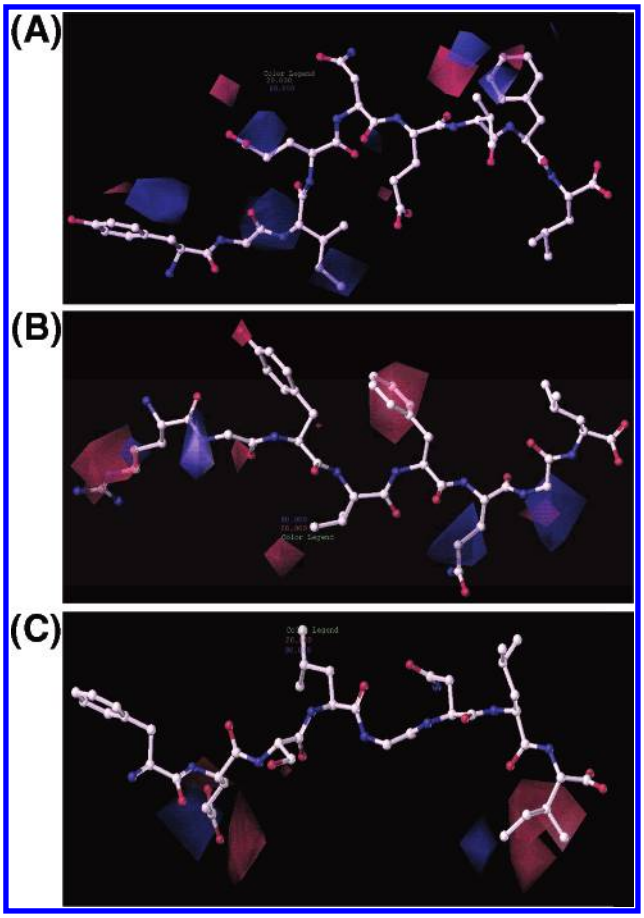

Figure 4. (A) H2-D^b, (B) H2-K^b, and (C) H2-K^k Electrostatic Potential Maps.

mouse MHC allele binds a mixture of structurally diverse peptides, typically 8−10 amino acids in length, and that each allele possesses a defined peptide specificity. The crystal structures of several mouse class I molecules[3,14−18] have helped to rationalize observed peptide binding. Such results show that the peptide binding cleft is closed at both ends, that the cleft has the same length in all class I molecules, that the carboxyl-terminal peptide position is deeply buried in the F pocket, and that there is little restriction on amino acids bound by pocket A.[14,16−19] The crystal structure of the antigenic peptide SIINFEKL, in complex with the mouse MHC class I H2-K^b molecule, shows that bound peptides have a strong preference for octamers with Tyr or Phe at position 5, Leu or Met at position 8, and to a lesser extent Tyr at position 3.[20] Zhang et al.[14] show that residues at positions 5 and 8 of VSV-8 (RGYVYQGL) and positions 6 and 9 of SEV-9 (FAPGNYPAL) are deeply buried in the central and C-terminal pockets C and F of H2-K^b. In VSV-8, Tyr at position 3 acts as a secondary anchor and is located in pocket D of H2-K^b, which is significantly altered in structure in the SEV-9 complex (with Pro at position 3). Although most peptides bound by the class I MHC molecules are eight or nine amino acids long, longer peptides will bulge from the middle of the groove.[21] Other studies on human class I MHC molecules show that amino acids at position 3 fall into pocket D,[5] which has been called a "loose" pocket,[22] while positions 4 and 8 are known as "flag" residues because they are solvent-exposed and contact the T-cell receptor.

The steric map for the H2-D^b allele (Figure 3A) indicates volume is favored at positions 3, 6, and 8 and disfavored at positions 2 and 8 (although this interaction is weak). The side chains of favored positions 3 and 6 fall into pockets D and C, respectively. When Figure 4A is inspected, the disfavored electrostatic potential (blue) is found at positions 1, 2, 3, 4, 7, and 8. Minor affinity-enhancing electrostatic potential (red) is found near positions 4, 6, 7, and 8. For the electrostatic potential field, the alkyl side chain of position 1 falls into pocket A, which consists of valine and serine residues.[5] At position 2, where the side chain falls into pocket B, electrostatic interaction is disfavored.[5] In the remaining positions, there are no favorable electrostatic potential interactions. Figure 5A shows the favored hydrophobic interaction positions to be 1, 2, 3, 6, and 8, with a very strong interaction being found at position 8, while the disfavored positions are 4 and 7. There is a strongly favored hydrophobic interaction at position 8 where the side chain is solvent-exposed and contacts the T cell. Areas of favored hydrogen-bond donor fields (cyan) and hydrogen-bond acceptor fields (magenta) for H2-D^b (Figures 6A and 7A) are at positions 1, 3, 4, and 8 and positions 2, 5, and 8, respectively. Specifically, the major favored interactions of the hydrogen-bond donor fields are found at position 1 and lie across the peptide backbone between positions 3 and 4. The hydrogen-bond acceptor map shows position 2 to be favored and, to a lesser extent, positions 5 and 7 as well.

For the H2-K^b allele, the steric map (Figure 3B) shows volume is favored at positions 1, 3, 4, and 5 and disfavored at positions 2, 6, and 7. Figure 4B shows that the electrostatic

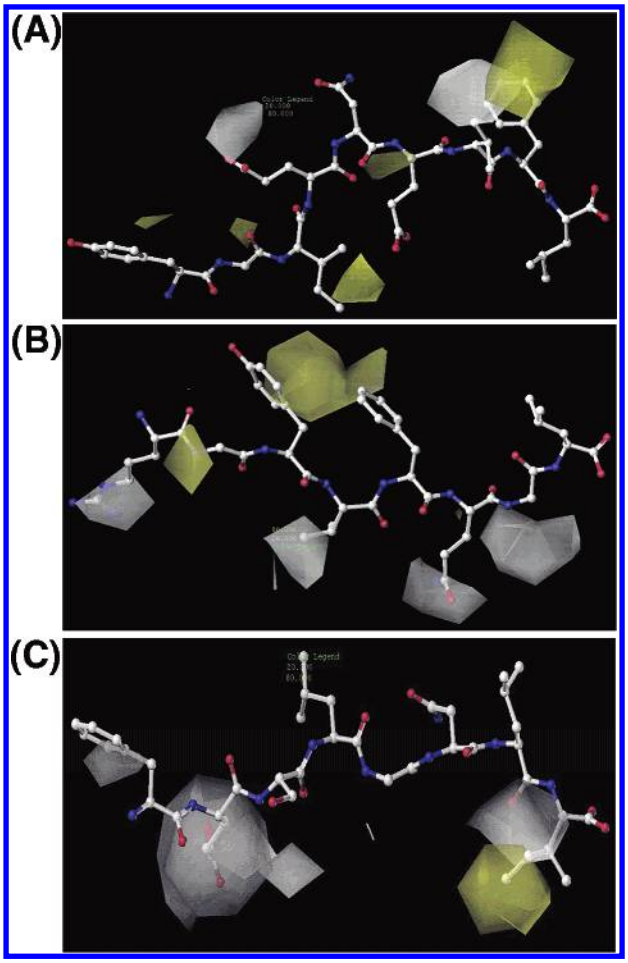

Figure 5. (A) H2-D^b, (B) H2-K^b, and (C) H2-K^k Hydrophobic Interaction Maps.

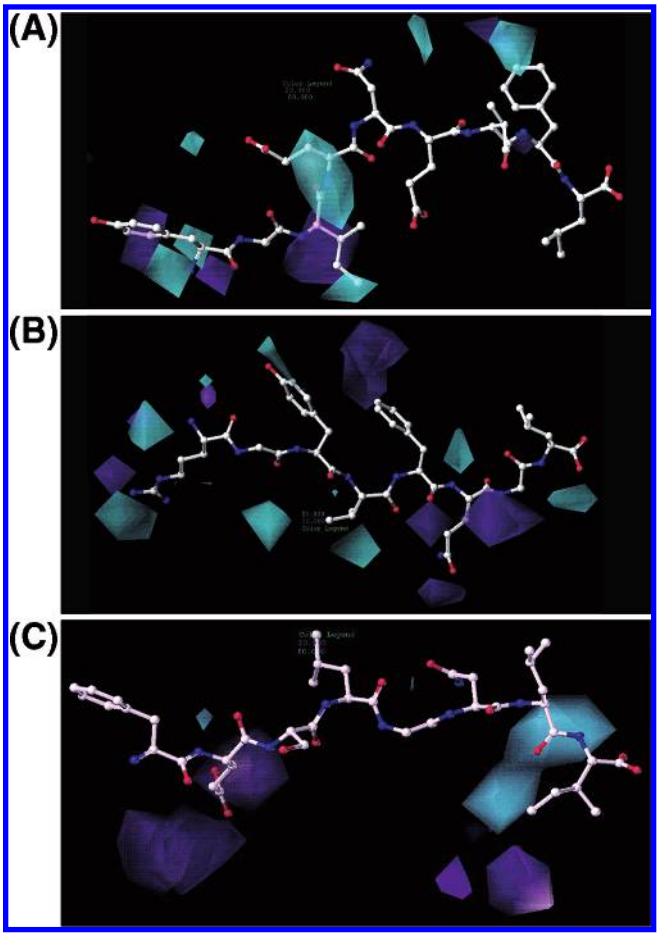

Figure 6. (A) H2-D^b, (B) H2-K^b, and (C) H2-K^k H-Bond Donor Maps.

interaction is favored at positions 1 and 5 and disfavored at positions 1, 2, 6, and 7, although these negative interactions are weak. The side chain at position 2 of the electrostatic potential map indicates that aromatic-type residues, such as Tyr and Phe, are well-tolerated. This is in good agreement with experimental data.[23,24] There is no major electrostatic interaction between side chains at position 3 (falling into pocket D) indicated by our model, and in the remaining positions, there are no clear favorable electrostatic interactions. Figure 5B shows a favored hydrophobic interaction at positions 2, 3, and 5, with the field spreading between positions 3 and 5, indicating some interaction with both side chains. The disfavored positions are at 1, 4, 6, and 7. Pocket D is a hydrophobic cavity, and amino acids such as Tyr and Ile are well-tolerated here, which would significantly deepen the depth and volume of pocket D.[19] When Figures 6B and 7B are looked at, the main favored positions for hydrogen-bond donor interaction are seen to be 1, 3, and 4 (pockets A, D, and the "flag" pocket, respectively)[5,] although some interactions are favored between positions 2 and 4, 5 and 6, and 7 and the C terminus. The favored hydrogen-bond acceptor positions are found at position 4 and between binding positions 1 and 2, with a major disfavored interaction found at position 6 (pocket C) and between the side chain positions 3 and 5.

The steric bulk map for H2-K^k (Figure 3C) shows volume is favored at positions 1, 6, 7, and 8 and disfavored at position 2. Both the disfavored and favored negative electrostatic potential field interactions for the H2-K^k allele (Figure 4C) are found near positions 2 and 8. In the remaining positions, there seems to be no discernibly favored or disfavored interactions. The hydrophobic map (Figure 5C) shows the only major interaction to be at position 8, with positions 1, 2, 5, and 8 being disfavored. Finally, when Figures 6C and 7C are inspected, the main favored hydrogen-bond donor fields are seen at positions 7 and 8, with hydrogen-bond acceptor fields being favored at position 8. The only favored interaction in the hydrogen-bond donor map in the H2-K^k allele lies between positions 7 and 8. Within the hydrogen-bond acceptor map, there is a strong disfavored interaction between the side chains at positions 2 and 3.

By comparing the CoMSIA maps for the three alleles, we can gain some insight into the character of anchor residues versus that of nonanchor residues and also into the nature of the data itself. For both the H2-D^b and H2-K^b alleles, we see less interaction at the "anchor" positions relative to the "nonanchor" positions, although the difference between the two types of positions is by no means significant. The H2-K^k allele differs from the other alleles as all affinity-affecting variation is centered on anchor positions (positions 2 and 8 and, to a lesser extent, position 3). As changes to the binding affinity at anchor positions 2 and 8 are associated with substitution by noncanonical residues, this shows how monosubstituents in a single sequence can hide variation at other positions. The H2-K^k data set is atypical in being a systematic substitution of a single sequence. Sets for H2-D^b and H2-K^b are much more usual. The relative nature of the

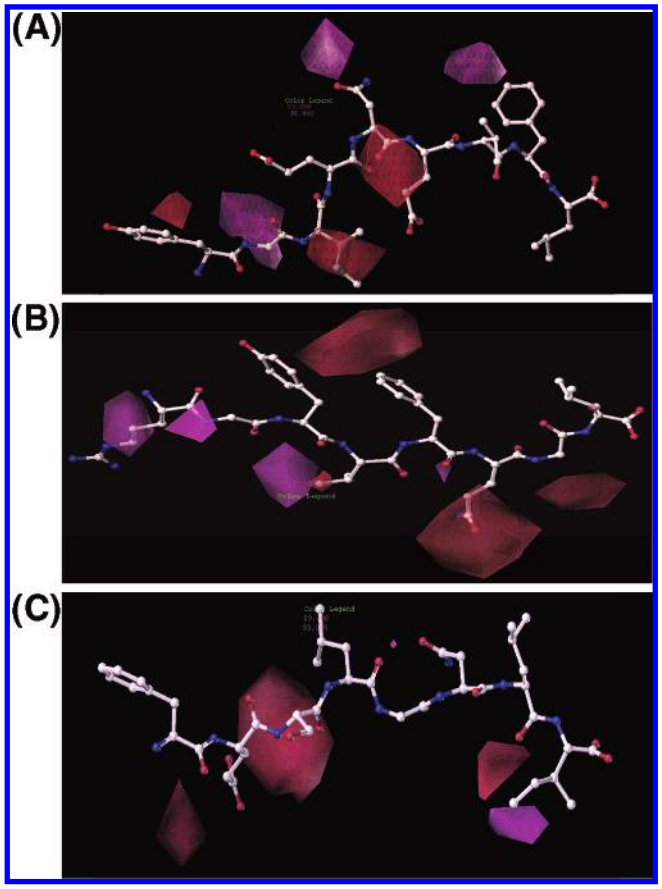

Figure 7. (A) H2-D^b, (B) H2-K^b, and (C) H2-K^k H-Bond Acceptor Maps.

resulting CoMSIA maps is, thus, much as expected. Compared to more familiar CoMSIA studies, the peptide sets we use are larger than those typically found in the pharmaceutical literature and the peptides are physically large and their physical properties are extreme; some may be zwitterionic or multiply charged, and some may exhibit large ranges of hydrophobicity. Moreover, the sequences and properties of the peptides are very biased in our data sets. This occurs through a process of preselection, which can result in significant self-reinforcement. Very simple "motifs" based on a few anchor positions are frequently used to decrease the experimental burden of identifying epitopes. Sparse sequence patterns are matched and the corresponding peptides tested, with an enormous diminution of peptide variety. The failure of our methods, and those of others, has as much to do with problems relating to the underlying data as it has to do with minor methodological flaws. Currently, most data sets show several key characteristics, as is well-exemplified by the current sets: they either concentrate too much on anchor positions (H2-K^k) or concentrate too little (H2-D^b and H2-K^b), or they are either overly (H2-K^k) or insufficiently (H2-D^b and H2-K^b) systematic and, thus, do not balance the need for variation to be spread evenly through each peptide without over-representing particular positions. With a properly designed training set, containing diverse sequences yet with systematic exploration of all amino acids at all positions, most of these issues would be resolved.

Unlike other QSAR methods, the interpretation of CoMSIA results is straightforward and uncomplicated. Through the display of property maps, CoMSIA is well-suited to molecular design.[25,26] Contouring the regions in space that are important with respect to a particular property suggests how to modify the known structures in order to improve binding affinity. When the present study is looked at, our coefficient contour maps are useful in visualizing how substitutions of particular functional groups or residues are favored or disfavored with respect to individual physicochemical properties and how this affects binding affinity. This method may also prove important in the design of new T-cell receptor agonist or antagonist peptidomimetic compounds and enhanced binders.[27,28]

As we have seen here, the CoMSIA study has proved successful in the analysis of the structure−activity relationships of peptide−protein complexes and, in particular, MHC−peptide interactions. Post-translational modified peptides, including glycosylated and phosphorylated peptides, can form pMHCs and, thus, can be recognized by T-cell receptors. Chemically modified peptides and peptidomimetics also bind MHCs.[29] MHCs are not limited to peptides as ligands; indeed, many drug-like molecules that bind to MHCs exhibit pathological effects through this mechanism.[30] Moreover, small molecule−MHC interactions are important in behavior-altering pheromone reception in mice.[31] Peptides are not currently favored as leads or drugs, and it has been suggested that the formation of a MHC−drug−T-cell-receptor complex may prove a useful alternative approach to the immunotherapeutic inhibitor design for T-cell-mediated processes.[30] The resulting complex would resemble the kind of complex formed by FK506 and the FK binding protein, which together form an inhibitor of calcinerin.[32] The visual interpretation of the COMSIA maps generated here, and previously,[10−12] will allow us to directly address the design of small molecule ligands of MHC molecules.

Here, we have shown how CoMSIA has complemented the crystallographic study of mouse MHC peptide complexes. X-ray crystallography, like all techniques, has its own limitations: for example, crystals represent a single conformation averaged in time and space. Methods, such as CoMSIA, can compensate for certain of these limitations by exploring dynamic and thermodynamic properties inaccessible to crystallography. CoMSIA can characterize the contribution to peptide binding to class I mouse alleles in terms of sequence-dependent side-chain binding at each peptide position, thus providing a reliable method for epitope prediction in the ongoing search for new epitope-based vaccines.

EXPERIMENTAL SECTION

Peptide Database. Peptide sequences and their binding affinities (IC_{50}) were extracted from the AntiJen database, formerly JenPep.[33,34] The database is freely available at the URL http://www.jenner.ac.uk/AntiJen. Both nonamer and octamer peptide sequences were studied. Binders to the H2-D^b allele consisted of 64 nonameric peptides,[35−42] the H2-K^b allele peptide set consisted of 62 octameric peptides,[37,39,42−45] and the H2-K^k allele set consisted of 154 octameric peptides[46−48] (Tables 4A−C, respectively, in the Supporting Information). IC_{50} values were used to quantify peptide−MHC interactions, which were originally calculated by a quantitative assay based on the inhibition of the binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules.[49] The IC_{50} values were converted to log-

Table 3. Summary of CoMSIA Position Specificities for Class I Mouse H2-K^b, H2-K^k, and H2-D^b

	class I mouse (H2-D^b, H2-K^b, and H2-K^k)				
position	steric bulk	electron density	hydrophobicity	H-bond donor	H-bond acceptor
P1 side chain falls into pocket A	favored	disfavored	disfavored	favored	favored
P2 side chain falls into pocket B	disfavored	disfavored	favored		favored
P3 side chain falls into pocket D	favored		favored	favored	
P4 side chain is solvent-exposed and can contact T cell			disfavored	favored	favored
P5			favored		favored
P6 side chain falls into pocket C	favored				
P7 side chain falls into pocket E		disfavored	disfavored		
P8 side chain is solvent-exposed and can contact T cell	favored	favored	favored	favored	favored
P9 side chain falls into pocket F					

($1/IC_{50}$), $-\log(IC_{50})$, or pIC_{50} and used as a dependent variable in the QSAR regression.

Molecular Modeling. All QSAR and molecular modeling calculations were carried out on a Silicon Graphics octane workstation using the SYBYL 6.9 molecular modeling package (Tripos Inc).[13] The X-ray structures of the nonameric peptide FAPGVFPYM[50] bound to the H2-D^b allele and the octameric peptide RGYVYQGL[15] bound to the H2-K^b and H2-K^k alleles were used as starting conformations. When the X-ray peptide was used as a template, all the studied peptides were built and then subjected to an initial geometry optimization, within SYBYL 6.9, using the Tripos molecular force field and charges derived using the MOPAC AM1 Hamiltonian semiempirical method.[51] Molecular alignment was based on the backbone atoms of the peptides, which was defined as an aggregate during optimization.

CoMSIA Method. CoMSIA was performed within SYBYL 6.9. Five physicochemical descriptors (electrostatic, steric, hydrophobic, and hydrogen-bond donor and acceptor) were evaluated using a probe atom placed within a 3D grid. The atom had a radius of 1 Å and charge, hydrophobic interaction, and hydrogen-bond donor and acceptor properties all equal to +1. The grid was extended beyond the molecular dimensions by 4.0 Å in the *x*, *y*, and *z* directions. The spacing between probe points within the grid was set at 2.0 Å and was increased in steps of 0.5 Å.

Cross-Validation Using the "Leave-One-Out" (LOO-CV) Method. The predictive power of the models from the CoMSIA analysis for each allele was carried out using PLS[52] as implemented within SYBYL 6.9. The method works by producing an equation or QSAR, which relates one or more dependent variables to the values of descriptors and uses them as predictors of the dependent variables (or biological activity). The IC_{50} values (the dependent variable *y*) were represented as negative logarithms (pIC_{50}). The predictive ability of the model was validated using cross-validation (CV), which is a reliable technique for testing the predictivity of models. With QSAR analysis in general and PLS methods in particular, CV is a standard approach to validation. CV works by dividing the data set into a set of groups, developing several parallel models from the reduced data with one or more of the groups excluded, and then predicting the activities of the excluded peptides. When the number of each excluded peptide is the same as the number in the set, the technique is called leave-one-out cross-validation (LOO-CV). The predictive power of the model is assessed using the following parameters: the cross-validated coefficient (q^2) and the standard error of prediction (SEP), which is defined in eqs 1 and 2:

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

Where $pIC_{50(pred)}$ is a predicted value and $pIC_{50(exp)}$ is an actual or experimental value. The summations are over the same set of pIC_{50} values. PRESS is the predictive error sum of squares, and SSQ is the sum of squares of $pIC_{50(exp)}$ corrected

for the mean. Where p is the number of the peptides omitted from the data set,

$$\mathrm{SEP} = \sqrt{\frac{\mathrm{PRESS}}{p-1}} \tag{2}$$

A more robust cross-validation test was also performed, dividing the sets into two and five groups, developing a number of parallel models from the reduced data with one of the groups of two and five randomly omitted, and then predicting the affinities of the excluded peptides. The means of the q^2 values from 20 runs are given as $q_{\mathrm{CV2}}{}^2$ and $q_{\mathrm{CV5}}{}^2$, respectively.

The optimal number of components (NC) resulting from the LOO−CV is then used in the non-cross-validated model, which was assessed using standard multiple linear regression validation terms, explained by variance r^2 and the standard error of estimate (SEE), which are defined in eqs 3 and 4, respectively.

$$r^2 = \frac{\mathrm{PRESS}}{\mathrm{SSQ}} \tag{3}$$

Where n is the number of peptides and c is the number of

$$\mathrm{SEE} = \sqrt{\frac{\mathrm{PRESS}}{n-c-1}} \tag{4}$$

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

The predictive power of the models from the CoMSIA analysis for each allele was carried out using PLS[52] as implemented within SYBYL 6.9. These models were then used to display the coefficient contour maps for each allele with respect to the five physicochemical descriptors.

Progressive Scrambling. Progressive scrambling is a technique developed by Clark et al.[53] to assess how sensitive a QSAR model is to random correlations and is particularly useful for large data sets containing redundant information. Progressive scrambling applies small perturbations to the data. Reductions in predictivity for unstable models are greater than those for robust models. Standard cross-validation, that is, LOO or leave-several-out, can be insensitive to redundancy when used for large data sets. Most compounds, peptides in our case, in a large set will have one or more close "twins", a molecule with similar descriptor values. In such cases, cross-validation will often achieve satisfactory predictions, as a "twin" of the omitted left-out molecule will probably remain in the training data. For redundant datasets, therefore, $q_{\mathrm{LOO}}{}^2$ may be unrealistic.

Progressive scrambling works by sorting rows (in our case, peptides) with respect to the dependent variable. Rows are then partitioned into bins. Within each bin, the dependent variables are shuffled several times. For each such shuffle, the correlation of the scrambled responses is assessed relative to the unperturbed data. PLS is applied to the perturbed data to obtain SEP and q^2 values. This process is repeated, decreasing the number of bins by one per iteration, until the number of bins reaches two.

Using the data calculated from the five descriptors, the models of three alleles were subjected to both scramblings of 10 and 30 in order to test the stability of the models. The number of components was increased consecutively from two to six, keeping all default values for the other settings. The statistics, q^2 and SEP, are reported once the scrambling is complete. The optimum statistics are seen when q^2 is at a maximum and SEP is at a minimum for the corresponding number of components.

CoMSIA Maps. The results of the non-cross-validated CoMSIA models were displayed as contour maps, with each physicochemical descriptor highlighted in different colors, reflecting favorable or disfavorable changes in the peptide structure and its influence on MHC binding. These maps were created using the standard deviation coefficient option based on actual values. The CoMSIA steric bulk map is shown using green (more bulk is favored) and yellow (less bulk is disfavored) contours. The electrostatic potential map is shown with blue (negative potential is disfavored) and red (negative potential is favored) contours. CoMSIA hydrophobic interaction fields are colored yellow (where hydrophobic interaction enhances affinity) and white (where hydrophilic interactions enhance affinity). The hydrogen-bond donor map is shown using cyan (donors on the ligand are preferred) and purple (donors are disfavored) contours. Finally, in the hydrogen-bond acceptor map, favored areas are shown in magenta and disfavored in yellow.

Supporting Information Available: Tables of peptides used in the study of the H2-D^b, H2-K^b, and H2-K^k mouse alleles. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Janeway, C. A.; Travers, P.; Walport, M.; Capra, J. D. *Immunobiology*; Elsevier Science Ltd.: New York, 2001; p 115.

(2) Madden, D. R.; Gorga, J. C.; Stominger, J. L.; Wiley, D. C. The three-dimensional structure of HLA-B27 at 2.1A resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* **1992**, *70*, 1035−1048.

(3) Fremont, D. H.; Stura, E. A.; Matsumara, M.; Peterson, P. A.; Wilson, I. A. Crystal structure of an H-2Kb-ovalbumin peptide complex reveals the interplay of primary and secondary anchor positions in the major histocompatibility complex binding groove. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2479−2483.

(4) Matsumara, M.; Fremont, D. H.; Peterson, P. A.; Wilson, I. A. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* **1992**, *257*, 927−934.

(5) Saper, M. A.; Bjorkman, P. J.; Wiley, D. C. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. *J. Mol. Biol*. **1991**, *219*, 277−319.

(6) 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.

(7) 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.

(8) Bohm, M.; Sturzebecher, 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.

(9) Stahle, L.; Wold, S. Multivariate data analysis and experimental design in biomedical research. *Prog. Med. Chem.* **1988**, *25*, 291−338.

(10) Doytchinova, I. A.; Flower, D. R. Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex:

a three-dimensional quantitative structure−activity relationship study. *Proteins* **2002**, *48*, 505−518.
(11) Doytchinova, I. A.; Flower, D. R. A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 535−544.
(12) Guan, P.; Doytchinova, I. A.; Flower, D. R. A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. *Bioorg. Med. Chem.* **2003**, *11*, 2307−2311.
(13) *Sybyl 6.9*; Tripos Inc.: St. Louis, MO.
(14) Zhang, W.; Young, A. C.; Imarai, M.; Nathenson, S. G.; Sacchettini, J. L. Crystal structure of the major histocompatibility complex class I H-2Kb molecule containing a single viral peptide: implications for peptide binding and T cell receptor recognition. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8403−8407.
(15) Bjorkman, P. J.; Saper, M. A.; Samraoui, B.; Bennett, W. S.; Strominger, J. L.; Wiley, D. C. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **1987**, *329*, 506−512.
(16) Young, A. C.; Zhang, W.; Sacchettini, J. C.; Nathenson, S. G. The three-dimensional structure of H-2D^b at 2.4A resolution: implications for antigen-determinant selection. *Cell* **1994**, *76*, 39−50.
(17) Smith, K. J.; Reid, S. W.; Stuart, D. I.; McMichael, A. J.; Jones, E. Y.; Bell, J. I. An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. *Immunity* **1996**, *4*, 203−213.
(18) Smith, K. J.; Reid, S. W.; Harlos, A. J.; McMichael, A. J.; Stuart, D. I.; Bell, J. I.; Jones, E. Y. Bound water structure and polymorphic amino acids act together to allow the binding of different peptides to MHC class I HLA-B53. *Immunity* **1996**, *4*, 215−228.
(19) Fremont, D. H.; Matsumura, M.; Stura, E. A.; Peterson, P. A.; Wilson, I. A. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b. *Science* **1992**, *257*, 919−927.
(20) Falk, K.; Rotzschke, O.; Stevanovic, S.; Jung, G.; Rammensee, H. G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **1991**, *351*, 290−296.
(21) Parham, P. Immunology. Deconstructing the MHC. *Nature* **1992**, *360*, 300−301.
(22) Madden, D. R. The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.* **1995**, *13*, 587−622.
(23) Ruppert, J.; Sidney, J.; Celis, E.; Kubo, R. T.; Grey, H. M.; Sette, A. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **1993**, *74*, 929−937.
(24) Parker, K. C.; Bednarek, M. A.; Coligan, J. E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* **1994**, *152*, 163−175.
(25) Oprea, T. I.; Waller, C. L. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Volume 11, pp 127.
(26) Greco, G.; Novellino, E.; Martin, Y. C. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Volume 11, pp 183.
(27) 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.
(28) Poenaru, S.; Lamas, J. R.; Folkers, G.; Lopez de Castro, J. A.; Seebach, D.; Rognan, N. 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.
(29) Krebs, S.; Rognan, D. From peptides to peptidomimetics: design of nonpeptide ligands for major histocompatibility proteins. *Pharm. Acta Helv.* **1998**, *73*, 173−181.
(30) Pichler, W. J. Modes of presentation of chemical neoantigens to the immune system. *Toxicology* **2002**, *181*, 49−54.
(31) Brennan, P. A. The nose knows who's who: chemosensory individuality and mate recognition in mice. *Horm. Behav.* **2004**, *46*, 231−240.
(32) Griffith, J. P.; Kim, J. L.; Kim, E. E.; Sintchak, M. D.; Thomson, J. A.; Fitzgibbon, M. J.; Fleming, M. A.; Caron, P. R.; Hsiao, K.; Navia, M. A. X-ray structure of calcineurin inhibited by the immunophilin-immunosuppressant FKBP12−FK506 complex. *Cell* **1995**, *82*, 507−522.
(33) Blythe, M.; Doytchiniva, I. A.; Flower, D. R. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* **2002**, *18*, 434−439.
(34) McSparron, H.; Blythe, M. J.; Zygouri, C.; Doytchinova, I. A.; Flower, D. R. JenPep: A novel computational information resource for immunology and vaccinology. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1276−1287.
(35) Hudrisier, D.; Mazarguil, H.; Laval, F.; Oldstone, M. B. A.; Gairin, J. E. Binding of viral antigens to major histocompatibility complex class I H-2Db molecules is controlled by dominant negative elements at peptide nonanchor residues. Implications for peptide selection and presentation. *J. Biol. Chem.* **1996**, *271*, 17829−17836.
(36) Price, G. E.; Ou, R.; Jiang, H.; Huang, L.; Moskophidis, D. Viral escape by selection of cytotoxic T cell-resistant variants in influenza A virus pneumonia. *J. Exp. Med.* **2000**, *191*, 1853−1867.
(37) Vitiello, A.; Yuan, L.; Chesnut, R. W.; Sidney, J.; Southwood, S.; Farness, P.; Jackson, M. R.; Peterson, P. A.; Sette, A. Immunodominance analysis of CTL responses to influenza PR8 virus reveals two new dominant and subdominant Kb-restricted epitopes. *J. Immunol.* **1996**, *157*, 5555−5562.
(38) Ostrov, D. A.; Roden, M. M.; Shi, W.; Palmieri, E.; Christianson, G. J.; Mendoza, L.; Villaflor, G.; Tilley, D.; Shastri, N.; Grey, H.; Almo, S. C.; Roopenian, D.; Nathenson, S. G. How H13 histocompatibility peptides differing by a single methyl group and lacking conventional MHC binding anchor motifs determine self-nonself-discrimination. *J. Immunol.* **2002**, *168*, 283−289.
(39) Wizel, B.; Starcher, B. C.; Samten, B.; Chroneos, Z.; Barnes, P. F.; Dzuris, J.; Higashimoto, Y.; Appella, E.; Sette, A. Multiple Chlamydia pneumoniae antigens prime CD8+ Tc1 responses that inhibit intracellular growth of this vacuolar pathogen. *J. Immunol.* **2002**, *169*, 2524−2535.
(40) Gairin, J. E.; Mazarguil, H.; Hudrisier, D.; Oldstone, M. B. A. 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*, 2297−2305.
(41) Hudrisier, D.; Mazarguil, H.; Oldstone, M. B. A.; Gairin, J. E. Relative implication of peptide residues in binding to major histocompatibility complex class I H-2Db: application to the design of high-affinity, allele-specific peptides. *Mol. Immunol.* **1995**, *32*, 895−907.
(42) Van der Most, R. G.; Murali-Krishna, K.; Whitton, J. L.; Oseroff, C.; Alexander, J.; Southwood, S.; Sidney, S.; Chesnut, R. W.; Sette, A.; Ahmed, R. Identification of Db- and Kb-restricted subdominant cytotoxic T cell responses in lymphocytic choriomeningitis virus-infected mice. *Virology* **1998**, *240*, 158−167.
(43) Rudolph, M. G.; Speir, J. A.; Brunmark, A.; Mattsson, N.; Jackson, M. R.; Peterson, P. A.; Teyton, L.; Wilson, I. A. The crystal structures of K(bm1) and K(bm8) reveal that subtle changes in the peptide environment impact thermostability and alloreactivity. *Immunity* **2001**, *14*, 231−242.
(44) Franco, A.; Yokoyama, T.; Huynh, D.; Thomson, C.; Nathenson, S. G.; Grey, H. M. Fine specificity and MHC restriction of trinitrophenyl-specific CTL. *J. Immunol.* **1999**, *162*, 3388−3394.
(45) Sette, A.; Oseroff, C.; Sidney, J.; Alexander, J.; Chesnut, R. W.; Kakimi, K.; Guidotti, L. G.; Chisari, F. W. Overcoming T cell tolerance to the hepatitis B virus surface antigen in hepatitis B virus-transgenic mice. *J. Immunol.* **2001**, *166*, 1389−1397.
(46) Nielsen, H. V.; Lauemoller, S. L.; Christiansen, L.; Buus, S.; Fomsgaard, A.; Petersen, E. Complete protection against lethal Toxoplasma gondii infection in mice immunized with a plasmid encoding the SAG1 gene. *Infect. Immun.* **1999**, *67*, 6358−6363.
(47) Stryhn, A.; Anderson, P. S.; Pederson, L. O.; Svejgaard, A.; Holm, A.; Thorpe, C. J.; Fugger, L.; Buus, S.; Engberg, J. Shared fine specificity between T cell receptors and an antibody recognizing a peptide/major histocompatibility class I complex. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10338−10342.
(48) Lauemoller, S. L.; Holm, A.; Hilden, J.; Brunak, S.; Nissen, M. H.; Stryhn, S.; Pederson, L. O.; Buus, S. Quantitative predictions of peptide binding to MHC class I molecules using specificity matrices and anchor-stratified calibrations. *Tissue Antigens* **2001**, *57*, 405−414.
(49) Sidney, J.; Grey, H. M.; Southwood, S.; Celis, E.; Wentworth, P. A.; del Guercol, M. F.; Kubo, R. T.; Chestnut, R. W.; Sette, A. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Hum. Immunol.* **1996**, *45*, 79−93.
(50) Zhao, R.; Loftus, D. J.; Appella, E.; Collins, E. J. Structural evidence of T cell xeno-reactivity in the absence of molecular mimicry. *J. Exp. Med.* **1999**, *189*, 359−370.
(51) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(52) Young, D. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley Inter-Science: New York, 2001; pp 243.
(53) Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating models based in large data sets. In *Rational Approaches to Drug Design*; Prous Science SA: Barcelona, Spain, 2001; pp 475−485.

CI049667L