

Coupling In Silico and In Vitro Analysis of Peptide-MHC Binding: A Bioinformatic Approach Enabling Prediction of Superbinding Peptides and Anchorless Epitopes

Irini A. Doytchinova,¹ Valerie A. Walshe,¹ Nicola A. Jones,² Simone E. Gloster,³ Persephone Borrow, and Darren R. Flower⁴

The ability to define and manipulate the interaction of peptides with MHC molecules has immense immunological utility, with applications in epitope identification, vaccine design, and immunomodulation. However, the methods currently available for prediction of peptide-MHC binding are far from ideal. We recently described the application of a bioinformatic prediction method based on quantitative structure-affinity relationship methods to peptide-MHC binding. In this study we demonstrate the predictivity and utility of this approach. We determined the binding affinities of a set of 90 nonamer peptides for the MHC class I allele HLA-A*0201 using an in-house, FACS-based, MHC stabilization assay, and from these data we derived an additive quantitative structure-affinity relationship model for peptide interaction with the HLA-A*0201 molecule. Using this model we then designed a series of high affinity HLA-A2-binding peptides. Experimental analysis revealed that all these peptides showed high binding affinities to the HLA-A*0201 molecule, significantly higher than the highest previously recorded. In addition, by the use of systematic substitution at principal anchor positions 2 and 9, we showed that high binding peptides are tolerant to a wide range of nonpreferred amino acids. Our results support a model in which the affinity of peptide binding to MHC is determined by the interactions of amino acids at multiple positions with the MHC molecule and may be enhanced by enthalpic cooperativity between these component interactions. *The Journal of Immunology*, 2004, 172: 7495–7502.

The activation and effector functions of CD4⁺ and CD8⁺ T cells are controlled by molecular recognition events, premier among which is TCR recognition of peptide-bound MHC molecules. Interaction of T cell receptors with peptide-MHC complexes triggers intracellular signals that are required for induction of initial T cell expansion and differentiation and triggering of T cell effector functions or, conversely, may stimulate partial T cell activation or lead to T cell anergy or death (1). The array of peptide-MHC complexes presented by professional APCs thus shapes the specificity of the T cell response, and the peptides displayed by target cell MHC dictate the recognition of these cells by effector T cells. The ability to define and manipulate the recognition of peptide-MHC complexes is thus a principal goal of modern immunology.

Definition of the peptides that are recognized by T cells responding to infections and tumors is of utility for evaluation of immunity during natural responses and also facilitates analysis of T cell responses after vaccination. Identification and manipulation of the peptide epitopes recognized during the natural response to an Ag can also enable the design of peptides that not only mimic, but actually improve upon, the natural immunogen (2). Such het-

eroclitic peptides are now finding application in vaccine design (3). Conversely, understanding of the peptide-MHC interactions involved in immunopathological T cell responses (e.g., in autoimmunity, allergy, or transplant rejection) can enable the design of altered peptide ligands that either antagonize or block undesirable responses. For example, an analog peptide was shown to inhibit the autoimmune demyelinating disease experimental allergic encephalomyelitis by antagonizing the CD4⁺ T cell response to a highly encephalitogenic peptide (4, 5), and a blocking peptide was demonstrated to prevent autoimmune insulin-dependent diabetes mellitus by inhibiting the expansion of autoreactive CTL (6), both in murine models.

Although the utility of manipulating MHC-peptide-TCR interactions has been recognized for some time, many studies continue to identify epitopes or design competitor peptides in a random way, by screening peptide libraries, rather than enhancing affinity in a rational way. By using in silico techniques to understand the basis of activity, affinity can, through a cyclic process, be improved by making incremental changes to the peptide structure. In this paper we exemplify an in silico method for the analysis and prediction of peptide-MHC binding affinity that is able to direct affinity enhancement in a rational or guided manner.

A great variety of methods exist for peptide-MHC binding prediction (for a recent review, see Ref. 7). Some involve the identification of so-called binding motifs (8), which characterize the peptide specificity of MHC alleles in terms of dominant anchor positions with strong preferences for a highly restricted amino acid set. It is well known, for example, that the best-understood human class I allele, HLA-A*0201, has anchor residues at peptide positions P2 (accepting leucine and methionine) and P9 (accepting valine and leucine). Motifs are widely exploited, being simple to use and to understand. There are fundamental problems with motif-based epitope prediction methods, however, as they produce

Edward Jenner Institute for Vaccine Research Compton, Berkshire, United Kingdom
Received for publication December 22, 2003. Accepted for publication March 12, 2004.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¹ I.A.D. and V.W. are joint first authors.

² Current address: Oxon Pharmaccines, Oxford, U.K.

³ Current address: Austin Research Institute, A&RMC, Victoria, Australia.

⁴ Address correspondence and reprint requests to Dr. Darren R. Flower, Edward Jenner Institute for Vaccine Research Compton, High Street, Berkshire, Compton, U.K. RG20 7NN.

significant numbers of both false positives and false negatives and are overly reliant on the choice of anchors. Subsequently, much more sophisticated methods have arisen (7). These have included empirical methods, such as de Groot's EpiVax methodology (9), Artificial Neural Networks (10), Hidden Markov Models (11), Support Vector Machines (12), and Profiles (13).

Recently, we have applied, in a systematic fashion, a novel bioinformatics approach to the problem of affinity prediction based on quantitative structure-activity relationship (QSAR)⁵ methods (14). QSAR analysis is a successful and widely used strategy for designing compounds with desired biological properties and is based on the assumption that the biological activity of a chemical entity, such as a peptide, depends on its structure. Properly used, this strategy can save large amounts of laboratory-based experimental work.

We have developed an additive QSAR method for peptide-MHC binding (15), based on the concept, defined by Free and Wilson (16), that each substituent makes an additive and constant contribution to the biological activity regardless of variation in the rest of the molecule. Parker's hypothesis (17) for the independent binding of side chains (IBS hypothesis) is also based on this concept. According to the additive method, the binding affinity of a peptide can be represented as the sum of amino acid contributions at each position. We extended the classical Free-Wilson model with terms accounting for interactions between amino acid side chains. This method was applied to peptides binding to several class I (18, 19) and class II (20) alleles using data from the literature (www.jenner.ac.uk/JenPep). The derived models are available free on the Internet (www.jenner.ac.uk/MHCPred) and can be used for binding affinity predictions (21).

Working solely with data from the literature has a number of disadvantages. Peptides are highly biased in terms of their position-dependent amino acid composition, often favoring hydrophobic sequences. This arises, in part, from preselection processes that result in self-reinforcement. Binding motifs are often used to reduce the experimental cost of epitope identification. Very sparse sequence patterns are matched, and the corresponding subset of peptides tested, with an enormous resulting reduction in sequence diversity. This bias is more prominent at the anchor positions, which usually have extremely restricted sets of amino acid types. In addition, when working solely with literature data it is not possible to test the predicted binding affinities of newly designed peptides.

In this study we determined the binding affinities of a set of 90 nonamer peptides to the MHC class I allele HLA-A*0201 using an in-house, FACS-based, MHC stabilization assay (22). From these data we then derived an additive QSAR model for peptide interaction with HLA-A*0201. HLA-A*0201 is one of the most frequent class I alleles in many different populations (23). Peptides that bind to this allele are 8–11 aa in length and, as noted above, have two main anchor residues at positions 2 and at the C-terminal end (24). Generally speaking, the presence of anchors is deemed to be necessary, but not sufficient, for high affinity binding. Prominent roles for several other positions (1, 3, and 7), so-called secondary anchor residues, are also well known (25). We used our QSAR model to reassess the preferred amino acids at each position and to design new A2 binding peptides. The top 10 high binders, as predicted, were tested. All showed extremely high binding affinity, 2 orders of magnitude greater than the highest value from the initial training set. Furthermore, the importance of the primary

anchors was tested by systematically evaluating the A2 binding affinities of monosubstituted variants of the best newly designed binder.

Materials and Methods

Peptide selection

The initial training set included 90 peptides (Table I). Eighty-eight of them had known binding affinities (IC_{50}) in the range from 10^{-5} to 10^{-9} M and were originally assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules (25, 26) and presented as $-\log IC_{50}$ (pIC_{50}). Known peptides were selected from JenPep (www.jenner.ac.uk/JenPep) (27, 28). Fifteen of them were low binders ($pIC_{50} < 6.301$; $IC_{50} > 500$ nM; Table I; peptides 1–15), 33 intermediate binders ($6.301 < pIC_{50} < 7.301$; 50 nM $< IC_{50} < 500$ nM; peptides 16–48), and 40 high binders ($pIC_{50} > 7.301$; $IC_{50} < 50$ nM; peptides 48–88). Two variants of the best binder 88 were included as well (peptides 89 and 90). All peptides used in the present study were ordered from Mimotopes (Pensby, U.K.). The test set was designed to include the top 10 predicted by the QSAR model to be high binders. Thirty-eight monosubstituted variants of the best newly designed binder were then tested to assess the importance of anchor positions 2 and 9.

Peptide binding assay

Peptide binding to HLA-A2 was assessed using a FACS-based MHC stabilization assay (29) with modifications as described previously (22). Briefly, T2 cells were incubated in 96-well, flat-bottom plates at 2×10^5 cells/well in a 200- μ l volume of AIM V medium (Life Technologies, Paisley, U.K.) with human β_2 -microglobulin at a final concentration of 100 nM (Scipac, Sittingbourne, U.K.) with and without peptides at concentrations between 200 and 0.04 μ M for 16 h at 37°C. Cells were then washed, and surface levels of HLA-A2 were assessed by staining with FITC-conjugated, A2.1-specific mAb BB7.2 (BD Biosciences, Oxford, U.K.) or an FITC-conjugated isotype control Ab (BD Biosciences). Cells were fixed at 4°C in 4% paraformaldehyde and analyzed on a FACSCalibur (BD Biosciences) using CellQuest software. Results are expressed as fluorescence index (FI) values. These were calculated as the test mean fluorescence intensity (MFI) minus the no peptide isotype control MFI divided by the no peptide HLA-A2-stained control MFI minus the no peptide isotype control MFI. The half-maximal binding level (BL_{50}), which is the peptide concentration yielding the half-maximal FI of the reference peptide in each assay, was calculated and presented as pBL_{50} ($-\log BL_{50}$). The HLA-A2 high binder FLPSDFFPSV ($IC_{50} = 2.6$ nM) (30) was used as a reference peptide.

Additive method

The additive method for binding affinity prediction was described in detail previously (15). Briefly, the binding affinity of a nonamer is represented by equation 1:

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

where the const accounts for the peptide backbone contribution,

$$\sum_{i=1}^9 P_i$$

is the sum of amino acid contributions at each position,

$$\sum_{i=1}^8 P_i P_{i+1}$$

is the sum of adjacent peptide side-chain interactions, and

$$\sum_{i=1}^7 P_i P_{i+2}$$

is the sum of every second side-chain interactions. Our previous studies (18–20) indicated that the interaction terms are only important for sets

⁵ Abbreviations used in this paper: QSAR, quantitative structure-affinity relationship; BL_{50} , half-maximal binding level; FI, fluorescence index; MFI, mean fluorescence intensity; PLS, partial least squares.

Table I. Training set

	Peptides	pIC ₅₀	pBL ₅₀
1	VCMTVDSL	5.146	4.202
2	HLESFLTAV	5.301	3.792
3	TTAEEAAGI	5.380	3.385
4	LLSCLGCKI	5.447	Nonbinder
5	LQTTIHDI	5.501	3.897
6	LTVILGVLL	5.580	Nonbinder
7	AMFQDPQER	5.740	Nonbinder
8	HLLVGSSGL	5.792	3.905
9	SLHVGTCQA	5.842	3.789
10	ALPYWNFAT	5.869	4.659
11	SLNFMGYVI	5.881	4.000
12	NLQSLTNLL	6.000	3.955
13	FVTWHRYHL	5.869	4.211
14	DPKVKQWPL	6.176	Nonbinder
15	ITSQVPFSV	6.196	4.055
16	GLGQVPLIV	6.301	4.762
17	MLDLQPETT	6.335	4.355
18	VLHSFTDAI	6.170	4.541
19	AAAKAAA	6.398	Nonbinder
20	ILTVILGVL	6.419	Nonbinder
21	KLPQLCTEL	6.484	4.504
22	WILRGTSFV	6.556	4.060
23	IISCTCPTV	6.580	5.165
24	ALIHHTHL	6.623	4.302
25	NLSWLSLDV	7.114	4.745
26	YMIMVKCWM	6.663	Nonbinder
27	LLWFHISCL	6.682	4.129
28	GTLGIVCPI	6.714	5.234
29	TLHEYMLDL	6.726	4.937
30	VTWHRYHLL	6.793	4.379
31	PLLPFFCL	7.114	5.320
32	TLGIVCPIC	6.815	4.676
33	CLTSTVQLV	6.832	4.933
34	FLCKQYLN	7.538	5.211
35	QLFHLCLII	6.886	Nonbinder
36	ITDQVPFSV	6.764	4.483
37	LMAVVLASL	6.954	3.994
38	YVITQHWL	6.793	4.393
39	ALCRWGLLL	7.000	4.908
40	ITAQVPFSV	7.020	4.434
41	YLEPGPVT	7.058	5.405
42	YTDQVPFSV	7.066	4.800
43	NLGNLNVSI	7.119	Nonbinder
44	HLYSHPII	7.131	5.407
45	SIISAVVGI	7.159	4.466
46	ITFQVPFSV	7.179	4.418
47	FTDQVPFSV	7.212	4.756
48	GLSRYVARL	7.102	4.780
49	LLAQFTSAI	7.301	4.508
50	YMLDLQPET	7.447	5.281
51	VLLDYQGML	6.945	4.518
52	RLMKQDFSV	7.342	4.966
53	KLHLYSHPI	7.140	4.768
54	YLSPPGPTA	7.383	5.436
55	YMGTMVSQV	7.398	4.673
56	SVYDFFVWL	7.447	5.120
57	ITWQVPFSV	7.457	5.012
58	KIFGLAFL	7.478	4.398
59	VMGTLVALV	7.538	5.029
60	SLDDYNHLV	7.585	5.271
61	VLIQRNPQL	7.644	5.062
62	SLYADSPSV	7.854	5.242
63	RLLQETELV	7.682	4.829
64	GLYSSTVPV	7.699	5.146
65	IMDQVPFSV	7.719	5.712
66	YLYPGPVT	7.772	5.768
67	YAILDPVSV	7.807	5.631
68	YLAPGPVT	7.818	6.002
69	MLGTHTMEV	7.845	5.367
70	LLFGYPVYV	7.886	5.453
71	ILKEPVHGV	7.921	5.589
72	WLDQVPFSV	7.939	5.225
73	KTWQYVWQV	7.959	4.429

	Peptides	pIC ₅₀	pBL ₅₀
74	ALMPYACI	8.000	5.082
75	YLAPGPVTA	8.032	5.740
76	FLLSLGIHL	8.000	5.170
77	LLMGTLGIV	6.681	4.210
78	YLWPGPVT	8.125	5.698
79	FLLTRILTI	8.009	4.952
80	GLLGWSPQA	7.886	5.132
81	ILYQVPFSV	8.310	5.061
82	GILTVILGV	8.337	4.573
83	YLMPGPVTA	8.367	5.272
84	NMVPFFPPV	8.409	5.597
85	ILDQVPFSV	8.481	6.092
86	YLFPGPVT	8.495	6.305
87	FLDQVPFSV	8.658	5.976
88	ILWQVPFSV	8.770	5.913
89	YLWQYIPSV		5.172
90	YLWQYIFSV		4.939

consisting of >200 peptides. For smaller sets, as in the present study, these two terms could be neglected, yielding the final equation:

$$pBL_{50} = \text{const} + \sum_{i=1}^9 P_i$$

The peptide sequence was represented as a set of 180 positions (20 aa × 9 positions), which could take the value 1 (present) or 0 (absent) depending on whether a certain amino acid exists at a certain position, thus generating a matrix with 90 rows by 180 columns. Columns containing only zeros were omitted. The pBL₅₀ values (y or dependent variable) were included as the first column. Partial least squares (PLS) was used to solve this matrix. PLS handles matrices with more variables than observations. It forms new x variables, called principal components, as linear combinations of the old x variables and uses them in correlation with the binding affinities. In the derived regression equation, which we called the additive model, each amino acid at each position has a regression coefficient accounting for its contribution to the affinity. Thus, amino acids with positive coefficients have positive contributions (they increase the binding affinity), and those with negative coefficients decrease the affinity because of their negative contributions. Using the regression coefficients, the binding affinity of a nonapeptide could be calculated easily as a sum of the contributions of the amino acid at each position in the peptide and the const term.

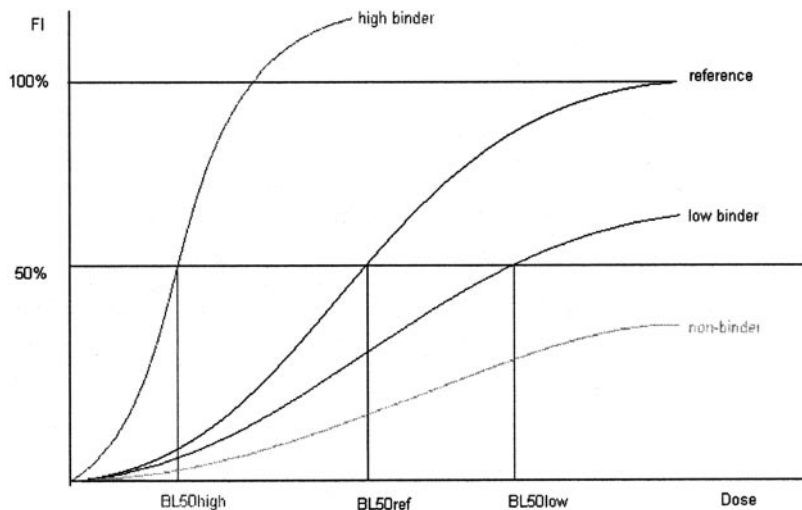
We used PLS as implemented in the QSAR module of SYBYL6.9 (Tripos, St. Louis, MO). The scaling method was set at none. The column filtering was switched off. The optimal number of components was found by leave-one-out cross-validation. The predictive power of the model was assessed by the cross-validated coefficient q². Outliers with residuals above 1 log unit were excluded, and the model was rederived. The non-cross-validated model was assessed by the explained variance r² and was used to predict the binding affinity of newly designed peptides.

Results

IC₅₀/BL₅₀ correlation

Using data solely derived from the literature for development of peptide-MHC binding models has significant limitations. One can explore only the role at each particular position of those amino acids already present in the data. Moreover, the models can only be validated using cross-validation, or arbitrarily selected artificial test sets, rather than by evaluating truly independent, or blind, test sets, and the affinity of newly designed high binding peptides cannot be determined. To overcome these limitations, we established a FACS-based assay for measuring peptide binding affinities experimentally. Initially we used this assay to measure the HLA-A*0201 binding affinities of a set of peptides whose binding to HLA-A*0201, as assessed using radiolabeled competition assays, has been reported in the literature. The peptides were chosen to cover a full range of measurable affinities. Dose/FI curves were created for each peptide and presented as semilogarithmic plots (Fig. 1). The high binders have low BL₅₀ values (high pBL₅₀,

FIGURE 1. Semilogarithmic dose/FI curves. The HLA-A2 high binder FLPSDFFPSV was used as a reference peptide. The high binders have low BL_{50} values (high pBL_{50} , $pBL_{50} = -\log BL_{50}$), and the low binders have high BL_{50} values (low pBL_{50}). Peptides that did not reach 50% of the binding level of the reference peptide were considered nonbinders.



$pBL_{50} = -\log BL_{50}$), and the low binders have high BL_{50} values (low pBL_{50}). Peptides that did not reach 50% of the binding level of the reference peptide were considered nonbinders. There were nine nonbinders in the training set.

A good correlation was found between the literature radiolabeled competition assay IC_{50} values and the BL_{50} values from the present study (Fig. 2). The correlation coefficient is 0.796, indicating an excellent degree of congruity. The competitive binding assay is more sensitive than the fluorescence assay, with IC_{50} values in the nanomolar range, compared with the BL_{50} values in the micromolar range. The corresponding pBL_{50} range for low binding is <4 ($BL_{50}, >10^{-4}$), that for intermediate is between 4 and 5 ($10^{-4} > BL_{50} > 10^{-5}$), and that for high is >5 ($BL_{50}, <10^{-5}$). The highest pBL_{50} value in the training set is 6.305, and it belongs to peptide YLFPGPVTA ($pIC_{50}, 8.495$).

Additive QSAR model

The measured BL_{50} values were used to build an additive QSAR model for peptide binding affinity to the HLA-A*0201 molecule. There were 41 absent amino acids for all nine positions. Most of the missing amino acids were for positions 2 and 9. The initial matrix consisted of 140 columns ($1y + 139x$ variables) and 81 rows (peptides). Peptides 17, 28, and 44 gave residuals between experimental and predicted, by leave-one-out cross-validation, pBL_{50} values of >1 log unit. They were excluded from the training set as outliers. The final model had $q^2 = 0.602$, $r^2 = 0.954$, and number of components = 6. This model was used to analyze the amino acid preferences at each position in the peptide sequences and to design a set of high binders.

The contributions of the amino acids at each of the nine positions according to the additive model are given in Fig. 3. The preferred amino acids for position 1 are Ile and Phe. His is deleterious at this position. Leu and Met are the preferred amino acids for position 2, whereas Thr is deleterious. Asp, Trp, and Phe are favored at position 3, whereas Glu, Ser, Gln, and Met are deleterious in this study. Pro and Asp are well accepted at position 4, whereas Val, Phe, and Ser are not preferred. Phe is the best contributing amino acid at position 5, and Pro, Ile, Leu, Asp, and Arg give small positive contributions to the affinity. At position 6, Pro, Val, and Tyr are preferred, whereas Ile, Ala, Gln, and Leu are deleterious. At position 7, Val and Pro are favored, whereas Thr is disfavored. Glu, Thr, and Asp are well accepted at position 8, whereas Ile, Ala, Val, and Met are not. The only acceptable residue at position 9 is Val.

Design and testing of new peptides

The derived QSAR model was then used to design peptides with very high HLA-A2 binding affinities. For this purpose we combined the preferred amino acids at each position. For certain positions (1, 2, 3, 4, 5, and 9) there were clear leaders, but for other positions (6, 7, and 8) a wider range of amino acids was acceptable. It is well known that peptide positions 2 and 9 (C terminal) are primary anchors for binding to the HLA-A*0201 molecule (24). The side chains of these residues occupy pockets B and F, respectively, in the MHC binding groove (31). Positions 1, 3, 6, and 7 are considered secondary anchors (25) that bind to pockets A, D, C, and E, respectively (31). Positions 4 and 8 are named flag positions because of their solvent-exposed orientation and possible interactions with the TCR (31).

We selected Leu for position 2 and Val for position 9 as anchors. For position 1, Ile and Phe were selected; for position 3, Phe, Asp, and Trp were selected; for position 4, Pro and Asp were selected; for position 5, Phe, Leu, and Ile were selected; for position 6, Pro, Val, and Phe were selected; for position 7, Pro, Val, and Ile were selected; and for position 8, Pro, Glu, Thr, Asp, and Ser were selected. The combination of all preferred amino acids generated 1620 peptides. Their affinities were predicted by the additive model, and the affinities of the top 10 high binders were tested

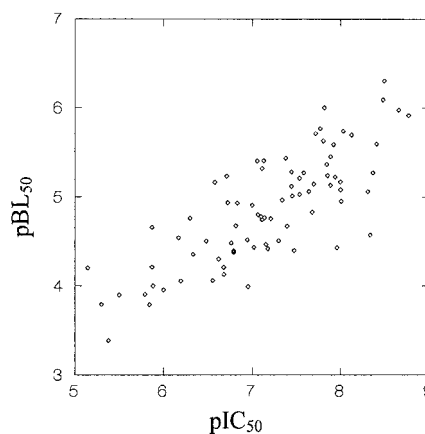


FIGURE 2. Correlation between the literature radiolabeled competition assay IC_{50} values (pIC_{50}) and the BL_{50} values for the training set (pBL_{50}). The correlation coefficient is 0.796.

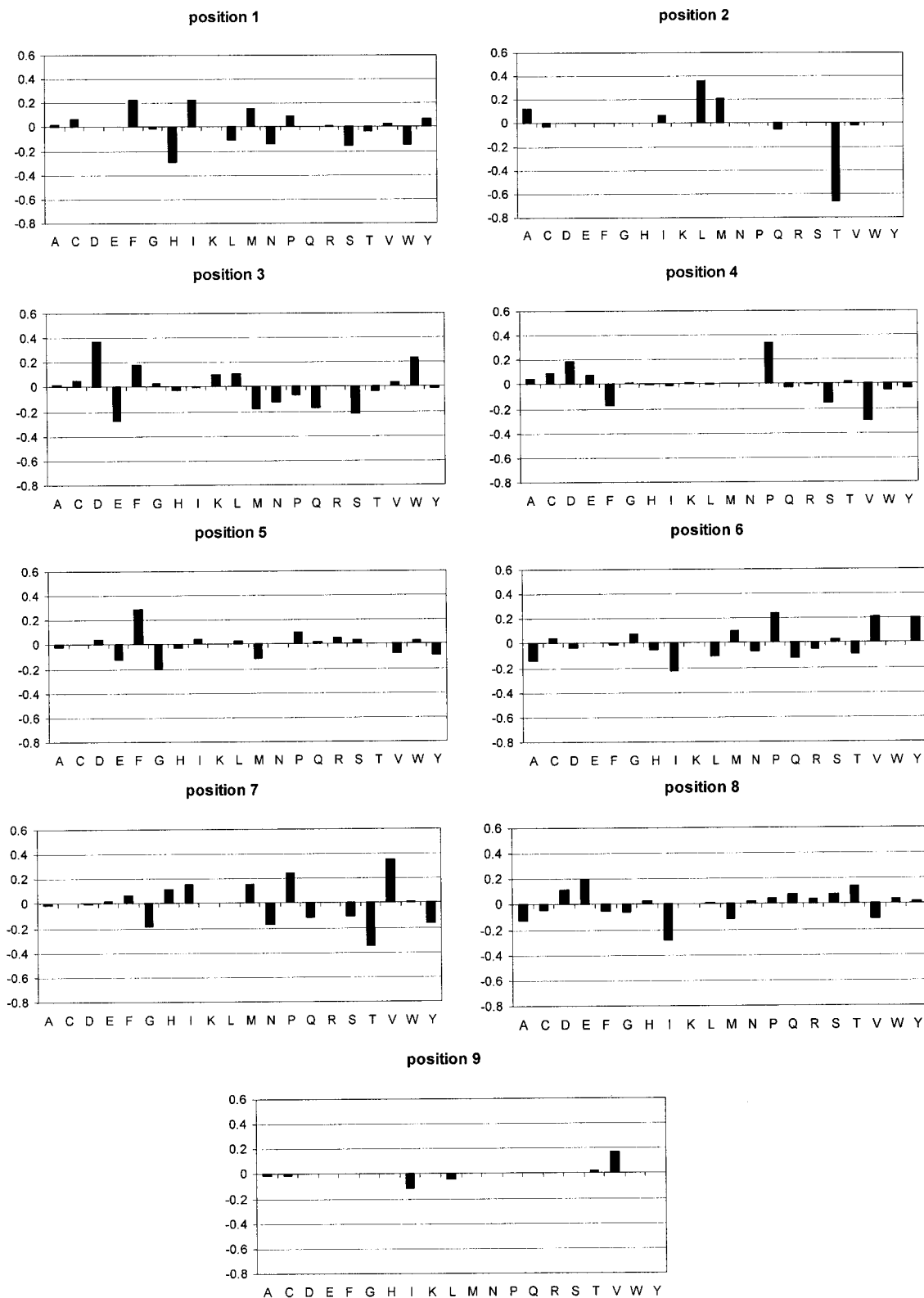


FIGURE 3. Contributions of different amino acids at each of the nine positions to the interaction of a nonameric peptide with the HLA_A*0201 molecule, according to the additive model. The bar heights indicate the strength of the positive or negative contribution made by each amino acid residue at the given position in a peptide.

experimentally. The test peptides and their predicted and experimental affinities are given in Table II (peptides 91–100). A good correlation between both affinities was found, $r_{\text{pred}} = 0.683$ (Fig. 4). Notably, these 10 peptides all had BL_{50} values

higher than those of the best peptides in the training set, with the pre-eminent test peptide 93 having a measured binding affinity >2 orders of magnitude greater than that of the best binder from the training set.

Table II. *Test set*

Peptides	pBL ₅₀		Residuals (experimental – predicted)
	Predicted	Experimental	
91 ILDPFPPTV	6.786	8.170	1.384
92 ILDPIPPTV	6.534	7.296	0.762
93 ILDPFPVTV	6.755	8.654	1.899
94 ILDDFPPTV	6.631	7.083	0.452
95 ILDDLPTV	6.367	7.144	0.777
96 ILDDFPVTV	6.600	7.155	0.555
97 ILDPFPPEV	6.836	7.682	0.846
98 ILDPFPPV	6.685	7.442	0.757
99 ILDPFPITV	6.699	8.139	1.440
100 ILDPLPPTV	6.522	7.145	0.623

Design and testing of monosubstituted variants of the best new binder

We reasoned that an optimized high binder might bear non-preferred amino acids at the anchor positions and retain adequate measurable affinity. To test this hypothesis, a set of variants of the best binding peptide (ILD₅₀FPVTV), monosubstituted at positions 2 and 9, was designed and tested. The experimental pBL₅₀ values are shown in Table III (peptides 101–138). Peptides with pBL₅₀ >5.000 were considered high binders, these with pBL₅₀ between 4.000 and 5.000 as intermediate binders, and those with pBL₅₀ <4.000 as low binders. Peptides with pBL₅₀ <3.000 are nonbinders. Among the 19 variants for position 2 there were 11 high, four intermediate, one low, and three nonbinders. The variants at position 9 gave 11 high, three intermediate, three low, and two nonbinders. This analysis showed that high affinity MHC binding can be achieved in the absence of the amino acids normally preferred at anchor positions.

Discussion

A bioinformatic prediction method, developed previously in our laboratory (15), has been applied to HLA-A*0201 binding data for a set of 90 peptides, and a model for peptide interaction with the HLA-A*0201 molecule has been developed. Using this model, a series of high affinity HLA-A2-binding peptides were designed. Empirical analysis revealed that all the peptides showed high binding affinities to the HLA-A*0201 molecule, significantly higher than the highest previously recorded. By the use of systematic substitution at the principal anchor positions 2 and 9, we have also shown that high binding peptides are tolerant to a wide range of nonpreferred amino acids.

The approach we have developed, which we have called the additive method, is an example of a QSAR technique. QSAR procedures are a powerful, if underused, bioinformatic tool for *in silico* prediction. QSAR has found much application, however, in computational drug design, where it can function as an engine of either interpolation or extrapolation. In interpolation, it can describe the properties of novel or extant molecules, peptides in our case, within a window of measured properties, but it can also be used to explore beyond those boundaries, most often being used to enhance binding affinity. The property of extrapolation into novel property space is the one we have exploited in this study.

In previous papers we have shown that the interpolative powers of our approach work effectively, capturing the essence of predictivity (15, 18–20). We demonstrate in this study that similar techniques can be used to effectively increase binding affinity in a rational and directed manner, allowing us to design a series of so-called superbinders and, in turn, to use these to explore the

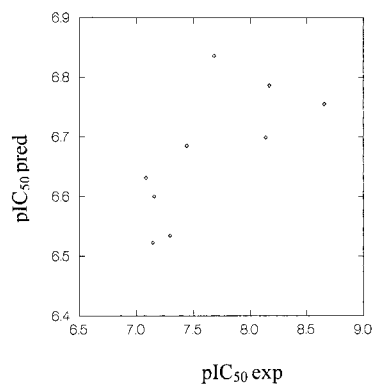


FIGURE 4. Correlation between the predicted and experimental BL₅₀ values for the test set. The correlation coefficient is 0.683.

effect of systematic substitution of dominant anchor positions. The list of tolerated anchors can be extended to a much larger set than has been commonly envisaged. This has implications for both our understanding of the role anchor residues play in peptide binding to MHCs and the relative effectiveness of binding motif and more sophisticated models of binding, such as *in silico* prediction devices.

Although our QSAR method has a tendency to underestimate the predicted BL₅₀ values (i.e., the predicted BL₅₀ values are lower than the experimental), it is able to distinguish accurately the proper amino acid preferences at each position in the peptide. The test peptides have strong amino acid preferences at position 1 (Ile), position 2 (Leu), position 3 (Asp), position 6 (Pro), and position 9 (Val). These positions are well known primary and secondary anchors. They fit into pockets A, B, D, C, and F, respectively, on the MHC molecule. Variations at positions 4, 5, 7, and 8 are allowed. The amino acids at these positions either do not fit any pocket (positions 4, 5, and 8) or bind a shallow pocket (position 7 corresponds to pocket E). The top three high binders (pBL₅₀ >8.000) exhibit variation only at position 7. All the newly designed peptides are high binders with BL₅₀ values higher than the highest BL₅₀ value in the training set. The best binder from the test set is peptide ILDPFPVTV. Its pBL₅₀ value is 8.654, which is >2 orders of magnitude higher than the pBL₅₀ value of the best binder from the training set (peptide YLFPGPVTA with pBL₅₀ of 6.305) and may be the highest binding affinity ever cited in the literature.

The rational design of high or superbinding peptides is a technique with wide application in a variety of immunological settings. Our current results are a further vindication of the utility of our approach in the prediction of peptide-MHC binding affinity, the principal prerequisite for proteinacious epitopes. Peptides presented by HLA-A2, in particular, would be useful from a vaccination standpoint as they would give responses in a high proportion of the HLA-diverse population. Perhaps more important, however, is the demonstrated ability of this approach to engineer epitopes with special properties dependent on enhanced affinity. These might include augmenting the immunogenicity of potential cancer vaccines derived from cancer Ag epitopes or designing high affinity epitopes, responses to which are reported to be less dependent on CD4 help (5). Alternatively, one could design effective and efficacious competitor peptides able to block detrimental responses, as has been done in a murine diabetes model (6).

The experimental pBL₅₀ values of the monosubstituted variants showed that an extremely good binder can tolerate a wide range of amino acids at the anchor positions. Only Glu, Lys, and Arg at position 2 and Asp and Arg at position 9 lead to a total loss of

Table III. Monosubstituted variants of the best binder ILDFPFVTV^a

	Peptides	pBL ₅₀ Experimental
101	IADFPFVTV	5.759
102	ICDFPFVTV	5.447
103	IDDFFPVTV	4.355
104	IEDFPFVTV	Nonbinder
105	IFDFPFVTV	4.888
106	IGDFPFVTV	3.916
107	IHDFPFVTV	4.955
108	IIDFPFVTV	6.308
109	IKDFPFVTV	Nonbinder
110	IMDFPFVTV	7.209
111	INDFPFVTV	4.778
112	IPDFPFVTV	5.103
113	IQDFPFVTV	6.048
114	IRDFFPVTV	Nonbinder
115	ISDFPFVTV	5.496
116	ITDFPFVTV	6.077
117	IVDFPFVTV	6.214
118	IWDFPFVTV	5.126
119	IYDFPFVTV	5.405
120	ILDFFPVTA	6.317
121	ILDFFPVTC	5.649
122	ILDFFPVTD	Nonbinder
123	ILDFFPVTE	3.126
124	ILDFFPVTF	5.674
125	ILDFFPVTG	6.662
126	ILDFFPVTH	3.604
127	ILDFFPVTI	6.688
128	ILDFFPVTK	4.590
129	ILDFFPVTL	7.033
130	ILDFFPVTM	6.126
131	ILDFFPVTN	5.286
132	ILDFFPVTP	5.824
133	ILDFFPVTV	5.277
134	ILDFFPVTR	Nonbinder
135	ILDFFPVTS	4.779
136	ILDFFPVTT	5.543
137	ILDFFPVTV	4.705
138	ILDFFPVTV	3.187

^a Substituted amino acids are shown in bold.

affinity, whereas Gly at position 2, and Glu, His, and Tyr at position 9 give low binding. The most preferred amino acids at positions 2 and 9 are Leu and Val, respectively, followed by Met, Ile, and Val at position 2, and Leu, Ile, Gly, and Ala at position 9. However, many other amino acids, previously thought nonoptimal, are also tolerated. This extensive tolerance at the anchor positions further strengthens our view that peptide-MHC molecule affinity is consequence of the whole peptide structure, not simply of anchor residues. This is manifest as a complicated ensemble of multiple amino acid interactions with the MHC molecule, arising from all parts of the peptide, not just primary and secondary anchors.

Moreover, a possible synergistic effect may also operate between amino acids at different positions on the peptide, this synergism perhaps explaining the underestimation of the predicted affinities of these high binders. Indeed, these superbinders have much higher affinities than a simple sum of amino acid contributions from different positions might suggest. This phenomenon is an example of enthalpic cooperativity or so-called enthalpy-entropy compensation (32). Generally, where multiple weak noncovalent interactions hold a molecular complex together, the enthalpy of all the individual intermolecular bonding interactions is weakened by extensive intermolecular motion. The noncovalent complex between a peptide and a protein is an excellent example of such a system. As additional interaction sites generate a more strongly bound complex, intermolecular motion is dampened, with all individual interactions becoming more favorable. Experimen-

tally, at least for other systems, the trade-off between intermolecular motion and enthalpic interactions accounts has been shown to account for the way in which entropy and enthalpy compensate for each other.

We have demonstrated that systematic monosubstitution of high binding peptides produces peptides that lack traditional anchors, yet retain high affinity. The relative importance of the anchor residues should thus be rethought. One does not require traditional anchors if the rest of the peptide is sufficiently optimized, either artificially, as in this case, or by chance in naturally occurring epitopes (33–35). Instead, one should seek more sophisticated and comprehensive models of binding better able to account for all possibilities. This helps to explain why many epitopes are missed when using only anchor motif-based epitope prediction programs. Flexibility as to which amino acids can be tolerated at the anchor positions increases the effective number of peptides that can be presented by a given HLA allele. This augments the chance that a T cell response can be mounted by every individual to each Ag or pathogen. It also has other implications, e.g., if multiple amino acids in an epitope can influence the peptide-HLA interaction, this may increase opportunities for pathogen escape from CD8 responses via alteration of peptide binding to MHC (36).

However, we must strike a minor note of caution in this study. Although we can now undertake the rational manipulation of peptide MHC affinity, such binding events are, in themselves, only part of the overall process of Ag presentation, albeit ones of paramount importance. Instead, a complex pathway is involved (37). Put at its simplest, proteins are synthesized, cleaved by proteolysis within the proteasome, and transported via TAP to the endoplasmic reticulum, before being exported to the cell surface bound to MHCs. However, the process is complicated by the involvement of other proteases, such as tripeptidyl peptidase II (38) in the cytoplasm and ERAAP in the endoplasmic reticulum (39). To properly predict T cell epitopes we will require not only an understanding of binding, but also a complete dynamic model of the cell biology underlying the Ag presentation pathway.

In conclusion, we have shown that our additive method is of utility for peptide-MHC binding affinity prediction and can be used successfully for the design of novel high binding peptides. Indeed, QSAR is a technique able to optimize molecular structure to deliver enhanced, reduced, or otherwise modulated biological properties of any measurable kind. We could, for example, use it to optimize the MHC binding affinity of weak affinity peptides, such as putative cancer vaccines. Further, it is equally appropriate for the analysis and manipulation of peptide-MHC complex interaction with T cell receptors as for determining peptide affinity for MHC. It is thus a tool of general utility to the immunologist, whether they are looking to design or enhance epitopes, nonimmunogenic competitor peptides, or T cell antagonists. These are themes we will explore in later work.

References

- Lanzavecchia, A., G. Iezzi, and A. Viola. 1999. From TCR engagement to T cell activation: a kinetic view of T cell behavior. *Cell* 96:1.
- Sette, A., M. Newman, B. Livingston, D. McKinney, J. Sidney, G. Ishioka, S. Tangri, J. Alexander, J. Fikes, and R. Chesnut. 2002. Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. *Tissue Antigens* 59:443.
- Tangri, S., G. Y. Ishioka, X. Q. Huang, J. Sidney, S. Southwood, S. Fikes, and A. Sette. 2001. Structural features of peptide analogs of human histocompatibility leukocyte antigen class I epitopes that are more potent and immunogenic than wild-type peptide. *J. Exp. Med.* 194:833.
- Kuchroo, V. K., J. M. Greer, D. Kaul, G. Ishioka, A. Franco, A. Sette, R. A. Sobel, and M. B. Lees. 1994. A single TCR antagonist peptide inhibits experimental allergic encephalomyelitis mediated by a diverse T-cell repertoire. *J. Immunol.* 153:3326.

5. Franco, A., S. Southwood, T. Arrhenius, V. K. Kuchroo, H. M. Grey, A. Sette, and G. Y. Ishioka. 1994. T-cell receptor antagonist peptides are highly effective inhibitors of experimental allergic encephalomyelitis. *Eur. J. Immunol.* 24:940.
6. von Herrath, M. G., B. Coon, H. Lewicki, H. Mazarguil, J. E. Gairin, and M. B. Oldstone. 1998. In vivo treatment with a MHC class I-restricted blocking peptide can prevent virus-induced autoimmune diabetes. *J. Immunol.* 161:5087.
7. Flower, D. R., I. A. Doytchinova, K. Paine, P. Taylor, M. J. Blythe, D. Lamponi, C. Zygouri, P. Guan, H. McSparron, and H. Kirkbride. 2002. Computational vaccine design. In *Drug Design: Cutting Edge Approaches*. D. R. Flower, ed. RSC, Cambridge, p. 136.
8. Sette, A., S. Buus, E. Appella, J. A. Smith, R. Chesnut, C. Miles, S. M. Colon, and H. M. Grey. 1989. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci. USA* 86:329.
9. De Groot, A. S., H. Sbai, C. S. Aubin, J. McMurry, and W. Martin. 2002. Immunoinformatics: mining genomes for vaccine components. *Immunol. Cell Biol.* 80:255.
10. Brusic, V., G. Rudy, and L. C. Harrison. 1994. Prediction of MHC binding peptides using artificial neural networks. In *Complex Systems: Mechanism of Adaptation*. R. J. Stonier and X. S. Yu, eds. IOS Press, Amsterdam; OHMSHA Tokyo, p. 253.
11. Udaka, K., H. Mamitsuka, Y. Nakaseko, and N. Abe. 2002. Prediction of MHC class I binding peptides by a query learning algorithm based on hidden Markov models. *J. Biol. Phys.* 28:183.
12. Donnes, P., and A. Elofsson. 2002. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 3:25.
13. Reche, P. A., J. P. Glutting, and E. L. Reinherz. 2002. Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* 63:701.
14. Doytchinova, I. A., and D. R. Flower. 2002. Quantitative approaches to computational vaccinology. *Immunol. Cell Biol.* 80:270.
15. Doytchinova I. A., M. J. Blythe, and D. R. Flower. 2002. Additive method for the prediction of protein-peptide binding affinity: application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.* 1:263.
16. Free, S. M., Jr., and J. W. Wilson. 1964. A mathematical contribution to structure-activity studies. *J. Med. Chem.* 7:395.
17. Parker, K. C., M. A. Bednarek, and J. E. Coligan. 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chain. *J. Immunol.* 152:163.
18. Guan, P., I. A. Doytchinova, and D. R. Flower. 2003. HLA-A3-supermotif defined by quantitative structure-activity relationship analysis. *Protein Eng.* 16:11.
19. Doytchinova, I. A., and D. R. Flower. 2003. The HLA-A2-supermotif: a QSAR definition. *Org. Biomol. Chem.* 1:2648.
20. Doytchinova, I. A., and D. R. Flower. 2003. Towards the *in silico* identification of class II restricted T cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics* 19:2263.
21. Guan, P., I. A. Doytchinova, C. Zygouri, and D. R. Flower. 2003. MHCpred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res.* 31:3621.
22. Lopes, A. R., A. Jaye, L. Dorrell, S. Sabally, A. Alabi, N. A. Jones, D. R. Flower, A. De Groot, P. Newton, R. M. Lascar, et al. 2003. Greater CD8⁺ TCR heterogeneity and functional flexibility in HIV-2 compared to HIV-1 infection. *J. Immunol.* 171:307.
23. Bodmer, J. 1996. World distribution of HLA alleles and implications for disease. *Ciba Found. Symp.* 197:233.
24. Falk, K., O. Rötzschke, S. Stefanovic, G. Jung, and H.-G. Rammensee. 1991. Allele specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351:290.
25. Ruppert, J., J. Sidney, E. Celis, R. T. Kubo, H. M. Grey, and A. Sette. 1993. Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. *Cell* 74:929.
26. Sette, A., J. Sidney, M.-F. del Guercio, S. Southwood, J. Ruppert, C. Dalberg, H. M. Grey, and R. T. Kubo. 1994. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.* 31:813.
27. McSparron, H., M. J. Blythe, C. Zygouri, I. A. Doytchinova, and D. R. Flower. 2003. JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.* 43:1276.
28. Blythe, M. J., Doytchinova, I. A., Flower, D. R. 2002. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 2002. 18:434.
29. Stuber, G., S. Modrow, P. Hoglund, L. Franksson, J. Elvin, H. Wolf, K. Karre, and G. Klein. 1992. Assessment of major histocompatibility complex class I interaction with Epstein-Barr virus and human immunodeficiency virus peptides by elevation of membrane H-2 and HLA in peptide loading-deficient cells. *Eur. J. Immunol.* 22:2697.
30. Chen, Y., J. Sidney, S. Southwood, A. L. Cox, K. Sakaguchi, R. A. Henderson, E. Appella, D. F. Hunt, A. Sette, and V. H. Engelhard. 1994. 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.* 152:2874.
31. Saper, M. A., P. J. Bjorkman, and D. C. Wiley. 1991. Refined structure of the human class I histocompatibility antigen HLA-A2 at 2.6 Å. *J. Mol. Biol.* 219:277.
32. Calderone, C. T., and D. H. Williams. 2001. An enthalpic component in cooperativity: the relationship between enthalpy, entropy, and noncovalent structure in weak associations. *J Am Chem Soc.* 123:6262.
33. Kurokohchi, K., T. Akatsuka, C. D. Pendleton, A. Takamizawa, M. Nishioka, M. Battegay, S. M. Feinstone, and J. A. Berzofsky. 1996. Use of recombinant protein to identify a motif-negative human cytotoxic T-cell epitope presented by HLA-A2 in the hepatitis C virus NS3 region. *J. Virol.* 70:232.
34. Alexander-Miller, M. A., K. C. Parker, T. Tsukui, C. D. Pendleton, J. E. Coligan, and J. A. Berzofsky. 1996. Molecular analysis of presentation by HLA-A2.1 of a promiscuously binding V3 loop peptide from the HIV-envelope protein to human cytotoxic T lymphocytes. *Int. Immunol.* 8:641.
35. Rongcun, Y., F. Salazar-Onfray, J. Charo, K. J. Malmberg, K. Evrin, H. Maes, K. Kono, C. Hising, M. Petersson, and O. Larsson, et al. 1999. Identification of new HER2/neu-derived peptide epitopes that can elicit specific CTL against autologous and allogeneic carcinomas and melanomas. *J. Immunol.* 163:1037.
36. Borrow, P., and G. M. Shaw. 1998. Cytotoxic T-lymphocyte escape viral variants: how important are they in viral evasion of immune clearance in vivo? *Immunol. Rev.* 164:37.
37. Flower, D. R. 2003. Towards in silico prediction of immunogenic epitopes. *Trends Immunol.* 24:667.
38. Seifert, U., C. Maranon, A. Shmueli, J. F. Desoutter, L. Wesoloski, K. Janek, P. Henklein, S. Diescher, M. Andrieu, H. de la Salle, et al. 2003. An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope. *Nat. Immunol.* 4:375.
39. Serwold, T., F. Gonzalez, J. Kim, R. Jacob, and N. Shastri. 2002. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature* 419:480.