technical notes Journal of proteome research

Additive Method for the Prediction of Protein−Peptide Binding Affinity. Application to the MHC Class I Molecule HLA-A*0201

Irini A. Doytchinova,* Martin J. Blythe, and Darren R. Flower

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

Received December 4, 2001

Abstract: A method has been developed for prediction of binding affinities between proteins and peptides. We exemplify the method through its application to binding predictions of peptides with affinity to major histocompatibility complex class I molecule HLA-A*0201. The method is named "additive" because it is based on the assumption that the binding affinity of a peptide could be presented as a sum of the contributions of the amino acids at each position and the interactions between them. The amino acid contributions and the contributions of the interactions between adjacent side chains and every second side chain were derived using a partial least squares (PLS) statistical methodology using a training set of 420 experimental IC_{50} values. The predictive power of the method was assessed using rigorous cross-validation and using an independent test set of 89 peptides. The mean value of the residuals between the experimental and predicted pIC_{50} values was 0.508 for this test set. The additive method was implemented in a program for rapid T-cell epitope search. It is universal and can be applied to any peptide−protein interaction where binding data is known.

Keywords: QSAR • PLS • Free−Wilson • MHC • HLA-A*0201

Introduction

Proteomics is concerned with the diverse functions and interactions of the protein products arising from a genome, both with each other and with other macromolecular and micromolecular components of the cell. One of its most important goals is to gain a proper understanding of protein ligand affinities. We have developed a method that accurately predicts protein−peptide interactions. We exemplify it here through application of the method to the well-studied immunological problem of peptide−MHC binding. As this may not be familiar to all readers, we put this application into context with a brief description of antigen presentation and recognition within the immune system.

The T cell is a specialized type of immune cell that mediates cellular immunity.[1] T cells constantly patrol the body searching for foreign proteins originating from pathogenic organisms (parasites, bacteria, fungi, or viruses). These cells express a special form of receptor: the T-cell receptor or TCR, which exhibits a broad range of selectivities and affinities. TCRs bind major histocompatibility complex proteins (MHCs), which are presented on other cell surfaces. MHCs bind small peptide fragments, or epitopes, derived from both pathogen and host protein. Recognition of such peptide−MHC complexes lies at the heart of the cellular immune response. The process that leads to the presentation of epitopes at the cell surface is both complex and poorly understood. There are two main presentation pathways, which are referred to as class I and class II. Most nucleated cells express class I MHCs, which are recognized by T cells with surfaces highly express CD8 co-receptors. Class II MHCs are only expressed on so-called "professional antigen presenting cells" and are recognized by T cells whose surfaces highly express CD4 co-receptors.

Class I peptides are typically, but not exclusively, derived from intracellular proteins, which are targeted to the proteasome. There they are cleaved into short peptides of 8−11 amino acids. These peptides are then bound by the transmembrane peptide transporter TAP, which translocates them into the endoplasmic reticulum (ER). Here they are bound by MHC protein. For class II, cells, through a process of receptor-mediated endocytosis, take up pathogen-derived extracellular proteins before these are targeted to endosomal compartments. Here these proteins are cleaved by cathepsins to produce peptides of 15−20 amino acids, which are bound by class II MHCs. Peptide−MHC complexes are presented on the cell surface where they are recognized, as T-cell epitopes, by TCRs. MHC proteins are polymorphic, each exhibiting different peptide selectivities. The combination of MHC and TCR selectivities determines the power and scope of peptide recognition in the immune system and thus the recognition of foreign and self-antigenic peptides.

Determination of peptides eliciting a T-cell response in vivo is important for identifying autoimmune and CTL epitopes[2,3] and for peptide vaccine design.[4,5] With respect to MHC molecule binding, it was found that an affinity threshold of approximately 500 nM (preferably 50 nM or less) determines the capacity of a peptide epitope to elicit a CTL response.[6] During the past decade, many methods for binding affinity prediction have been developed. Parker's method[7,8] is based on the hypothesis that each amino acid side chain binds independently of the rest of the peptide (the IBS hypothesis). The IBS hypothesis is also implemented in the polynomial method.[9] Neural networks were also used to predict binding affinities to HLA-A2.1 molecules[9−11] and to score them as binders or nonbinders. Altuvia and colleagues[12] developed an

* To whom correspondence should be addressed. Tel: +44 1635 577968. Fax: +44 1635 577908. E-mail: irini.doytchinova@jenner.ac.uk.

10.1021/pr015513z CCC: $22.00

Published on Web 04/12/2002

algorithm to evaluate the interactions of peptide amino acids with MHC contact residues, and the resulting energies were used to score the peptide binding affinities. Molecular dynamics simulations also take place in the examination of peptide−protein interactions. Using this approach, Mata and colleagues[13] probed the stability of peptide binding to murine MHC class I molecule H-2 K^d and identified a binding motif that does not contain an appropriate anchor residue.

Recently, we proposed a quantitative method for binding affinities prediction based on the 3D-QSAR method—comparative molecular similarity indices analysis (CoMSIA).[14,15] During the process of peptide modeling, it was clear that the conformation of a certain amino acid side chain at a certain position strongly depends on the neighboring amino acids. This means that the IBS hypothesis is not sufficient to explain the binding abilities of the peptides.

In the present study, we propose a method overcoming this insufficiency through the inclusion of terms accounting for neighboring amino acids. We name this method the "additive" method because it is based on the additivity concept, developed by Free and Wilson,[16] whereby each substituent makes an additive and constant contribution to the biological activity regardless of substituent variation in the rest of the molecule

$$\text{biological activity} = \sum_{ij} G_{ij}X_{ij} + \mu$$

in which μ is the overall average of biological activity values and G_{ij} is the activity contributions of the substituent X_i in position j ($X_{ij} = 1$ if the substituent X_i is in position j; otherwise, $X_{ij} = 0$). The values of the individual group contributions are calculated by multiple linear regression (MLR) analysis. Other models based on the additivity concept are different modifications of the Free−Wilson model. The Fujita−Ban modification[17] is a simple linear transformation of the Free−Wilson model, where μ is the activity of the unsubstituted compound predicted by the least-squares method. In Cammarata's model[18] (Cammarata and Yau, 1970), μ is the experimental activity of the unsubstituted compound (all X_{ij} = H). The models based on the additivity concept are simple to perform and easy to interpret. Because of that they have found a wide application in molecular design over the years.[19−25]

We extended the classical Free−Wilson model with terms accounting for the possible interactions between the amino acids side chains. Thus, the binding affinity of a nonamer expressed in p-units (negative decimal logarithm of IC_{50} values) could be presented by eq 1

$$pIC_{50} = \text{const} + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_iP_{i+1} + \sum_{i=1}^{7} P_iP_{i+2} + \sum_{i=1}^{6} P_iP_{i+3} + \sum_{i=1}^{5} P_iP_{i+4} + \sum_{i=1}^{4} P_iP_{i+5} + \sum_{i=1}^{3} P_iP_{i+6} + \sum_{i=1}^{2} P_iP_{i+7} + P_iP_{i+8} \quad (1)$$

where the const accounts, at least nominally, for the peptide backbone contribution, $\Sigma_{i=1}^{9}P_i$ is the sum of amino acids contributions at each position, $\Sigma_{i=1}^{8}P_iP_{i+1}$ is the sum of adjacent peptide side-chain interactions, $\Sigma_{i=1}^{7}P_iP_{i+2}$ is the sum of every second side-chain interaction, $\Sigma_{i=1}^{6}P_iP_{i+3}$ is the sum of every third side-chain interaction, and so on. The contributions of the last six terms are negligibly small, although a hydrogen bond formation between Tyr at P1 and Glu at P4 bridged by a water molecule was suggested by Kirksey and colleagues,[26] making the bound peptide more rigid and easily recognized by T-cells. The binding affinity will depend significantly on the contributions of the amino acid side chains at each position and the interactions between the adjacent and every second side-chain:

$$pIC_{50} = \text{const} + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_iP_{i+1} + \sum_{i=1}^{7} P_iP_{i+2} \quad (2)$$

Initially, the predictive power of the final equation was assessed by a "leave-one-out" cross-validation[27] on the training set. In the meantime, new entries were compiled from our database and we used them as a test set to further assess the predictability of the additive method.

Table 1. Frequency of the Amino Acid Residues at Different Positions in the Peptides in the Data Set

aa	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	39	14	41	24	44	26	64	38	69
R	17	0	3	18	12	6	6	8	1
N	10	0	5	21	8	14	10	12	0
D	3	0	27	31	18	8	3	11	0
C	3	1	11	6	3	11	12	9	3
Q	4	1	6	37	14	20	11	9	0
E	0	0	12	16	7	5	6	12	0
G	33	1	20	55	60	24	12	26	0
H	12	0	12	4	13	4	11	15	0
I	40	29	20	13	19	29	29	21	38
L	41	291	67	46	51	59	53	65	124
K	13	0	5	18	6	2	3	2	0
M	11	31	27	4	3	15	5	7	2
F	48	0	23	8	21	18	54	22	1
P	3	1	10	36	16	64	29	19	0
S	43	0	30	32	11	22	22	48	2
T	15	22	10	21	32	21	19	49	8
W	11	0	18	7	8	3	7	10	0
Y	51	0	49	4	24	12	13	13	0
V	23	29	24	19	50	57	51	24	172

Methods

Peptide Database and Binding Affinities. The peptide sequences and their binding affinities were extracted from our recently developed JenPep database.[28] The database is freely accessible (http://www.jenner.ac.uk/Jenpep). The training set consists of 420 IC_{50} values for 340 nonamer peptides. Eighty IC_{50} values are higher than 500 nM (low binders), 182 values are between 50 and 500 nM (intermediate binders), and 158 values are less than 50 nM (high binders). The frequency of the amino acid residues at different positions in the peptides in the data set is given in Table 1. More than one IC_{50} value was found for some of the peptides. The binding affinities (IC_{50}) we used were originally assessed by a quantitative assay based on the inhibition of binding of a radiolabeled standard peptide to detergent-solubilized MHC molecules.[29,30] In the present study, and as is common practice among QSAR practitioners, the IC_{50} values were converted to p-units. The magnitude of measured binding affinity ranges over almost 5 orders: from 4.301 to 8.959 in log units. Many amino acids are presented only once at a certain position. This presumes their contributions and the contributions of the interactions are spurious, achieving significance by chance. However, by disregarding these single amino acids one runs the risk of eliminating legitimate predictors. This problem will be minimized as the database grows. The test set consisted of 89 peptides: 19 low binders, 50 intermediate binders, and 20 high binders.

Matrix Construction. A program was developed to transform the nine amino acid peptide sequence into a row of the table

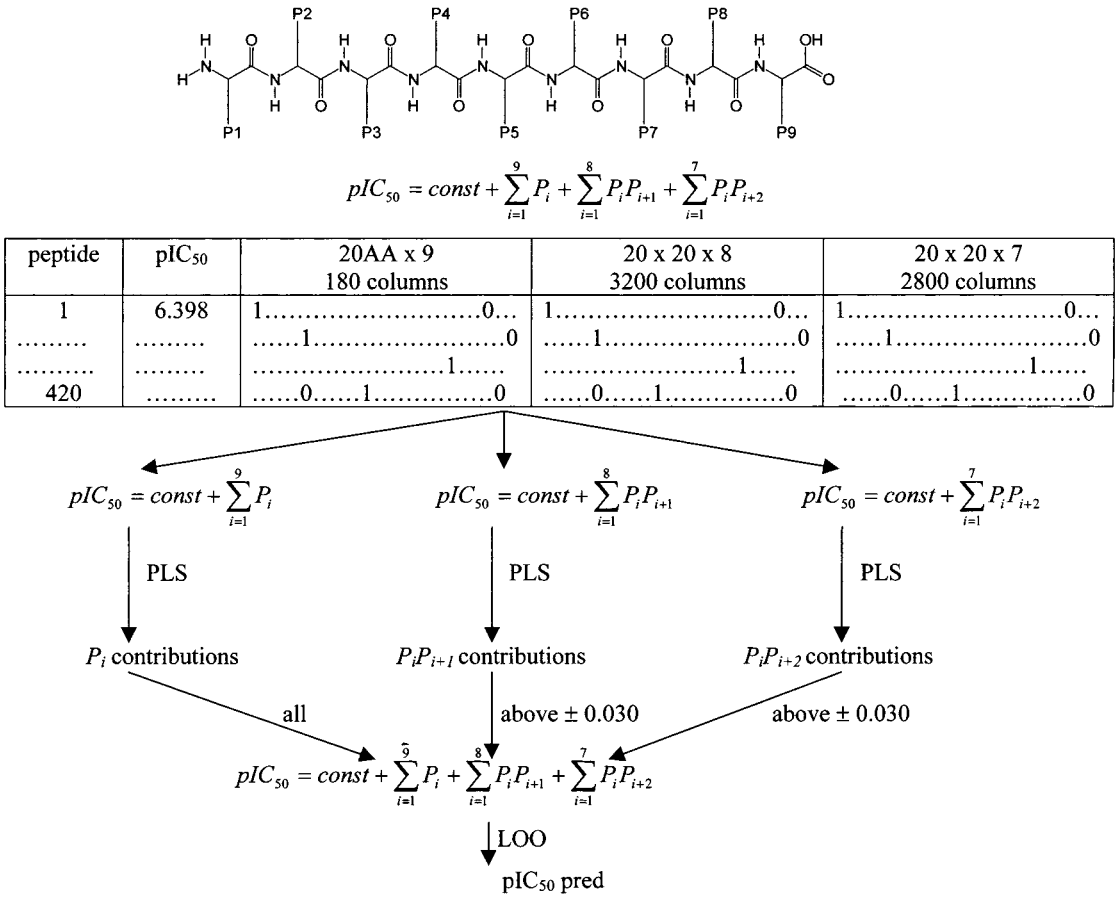


Figure 1. Protocol of the additive method.

presented in Figure 1. A term is equal to 1 when a certain amino acid at a certain position or a certain interaction between two side chains exists and 0 when they are absent. Thus, a matrix of 420 rows and 6180 columns was generated. One hundred and eighty columns account for the amino acids contributions (20 aa × 9 positions); 3200 columns account for the adjacent side chains, or 1−2 interactions (20 × 20 × 8); and 2800 columns account for every second side chain, or 1−3 interactions (20 × 20 × 7). To reduce the column number, the program omits columns that contain only 0s. The final matrix consisted of 420 rows and 2158 columns.

Multiple Linear Regression by the Partial Least Squares Method. The partial least squares (PLS) method belongs to so-called projection methods. These methods handle data matrixes with more variables than observations very well, and the data can be both noisy and highly collinear. In this situation, conventional statistical methods such as multiple regression produce a formula that fits the training data but is unreliable for prediction. PLS forms new *x* variables, named *principal components*, as linear combinations of the old ones and then uses them as predictors of the biological activity.[27]

We used the PLS method as implemented in the QSAR module of SYBYL6.7.[31] pIC_{50} was set as a dependent variable. The scaling method was set to "none". The column filtering was switched off. The optimal number of components (NC) was found by cross-validation using SAMPLS.[32] The non-cross-validated models were assessed by the MLR parameters as explained by variance r^2, standard error of estimate (S), and F ratio. A cross-validation using the "leave-one-out" procedure assessed the predictive power of the models.

Cross-Validation Using the "Leave-One-Out" Procedure. Cross-validation (CV) is a practical and reliable method for testing the predictive power of the models. It has become a standard in PLS analysis and is incorporated in all available PLS software.[27] In principle, CV is performed by dividing the data into a number of groups, developing a number of parallel models from the reduced data with one of the groups omitted, and then predicting the biological activities of the excluded compounds. When the number of the groups omitted is equal to the number of the compounds in the set, the procedure is named "leave-one-out" (LOO). The predictive power of the models was assessed by the cross-validated coefficient q^2, the standard error of prediction (SEP), and the *residuals* between the experimental and predicted binding affinity:

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

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

$$\text{residual} = \text{pIC}_{50}(\text{exp}) - \text{pIC}_{50}(\text{pred})$$

where PRESS is the predictive sum of squares ($\Sigma_{i=1}^{n}(\text{pIC}_{50}(\text{exp}) - \text{pIC}_{50}(\text{pred}))^2$), SSQ is the sum of squares of pIC_{50}(exp) corrected for the mean ($\Sigma_{i=1}^{n}(\text{pIC}_{50}(\text{exp}) - \text{pIC}_{50}(\text{mean}))^2$), p is the number of the peptides omitted, and pIC_{50}(pred) is that predicted by the CV−LOO value. The CV−LOO procedure does not give reliable results for the peptides in the set expressed by more than one pIC_{50} value because "omit one value" is not equivalent to "omit one peptide". For these peptides, the pIC_{50}-

(pred) were calculated by omitting all the available pIC_{50} values. The *residuals* between the experimental and predicted pIC_{50} values were classified into three categories: below |0.5|, from |0.5| to |1.0|, and above |1.0|. A mean |residual| was extracted as well. The same categorization was applied for the test set.

Results and Discussion

The resolution of the matrix consisting of 420 rows and 2158 columns requires a long computing time and a high computer specification. To make this analysis tractable, we divided eq 2 into three equations:

$$pIC_{50} = \text{const} + \sum_{i=1}^{9} P_i \quad (3)$$

$$pIC_{50} = \text{const} + \sum_{i=1}^{8} P_i P_{i+1} \quad (4)$$

$$pIC_{50} = \text{const} + \sum_{i=1}^{7} P_i P_{i+2} \quad (5)$$

Equation 3 generates the amino acid contributions on the basis of the IBS hypothesis. Equations 4 and 5 give the contributions of the side-chain interactions. The contributions of the interactions below ±0.030 were omitted, and the three equations were combined into one again. As the columns are much more than the rows in all equations they were solved using the partial least squares method.

The final equation consisted of 1815 terms including the constant. It is available as Supporting Information. The regression equation contains the amino acid contributions and the contributions of the significant side-chain interactions. Its MLR parameters are $r^2 = 0.898$, $S = 0.285$, $F = 588.883$, number of components (NC) = 5, $n = 340$. The "leave-one-out" cross-validation (CV−LOO) gives $q^2 = 0.337$, SEP = 0.726, NC = 5, $n = 340$. The low q^2 value was surprising at this high value of explained variance (near 90%). In the cases of multiple pIC_{50} values for one peptide, the pIC_{50}(pred) were recalculated omitting all available pIC_{50}(exp) values. For each peptide, the experimental pIC_{50} value that was closest to the predicted one was considered. For 172 peptides (50.5%), the residuals were below |0.5|, for 128 peptides (37.5%) the residuals were between |0.5| and |1.0|, and for 40 (12.0%) peptides the residuals were above |1.0|. The mean absolute value was 0.573 with standard deviation 0.442. The analysis of the peptides with residuals above |1.0| log unit revealed that as more terms are absent in eq 2 the predicted value is poorer. According to the current QSAR practice, predictions within 1.0 log unit are considered good.[33−35] This would result in mean residuals of around 0.5 log unit. In ideal cases, QSAR methods allowing extrapolation in their predictions (e.g., 3D-QSAR) give an acceptable extrapolation up to 0.3 log units.[36] However, in our work, which is far from an ideal case, the experimental measurements we are trying to predict are much less accurate than those obtained for the much smaller datasets typical in pharmaceutical applications. The experimental, or biological, error in these measurements is, in terms of logs, probably much great than 0.3. In this context, in the additive method, as a method not using any extrapolation, average predictions less than 0.5 log unit cannot be realized.

The regression equation was used to predict the binding affinities of an independent test set of 89 peptides. Fifty-five peptides (61.8%) had residuals below |0.5|, 26 (29.2%) had residuals between |0.5| and |1.0|,and only eight (9.0%) had residuals above |1.0|. The mean |residual| was 0.508 with standard deviation 0.430. The plot of predicted vs experimental pIC_{50}'s using the test set is shown in Figure 2. There is one striking outlier among the poorly predicted peptides. Peptide TLQDIVLHL has unexplainable high pIC_{50}(exp) = 9.000[37] being out of the binding affinity range of the training set. Its predictive pIC_{50}(pred) is 6.313. The poorly predicted peptides have many absent values (from 20 to 38%) with a detrimental concomitant effect on predictive power. Obviously, this is a weakness of the investigated set but not of the method. The growth of the database will decrease the number of missing amino acids at particular positions and likewise missing interactions they are involved in, and thus, the number of absent values in eq 2 will decrease. The "ideal" dataset should comprise 6000 ($20 \times 20 \times 8 + 20 \times 20 \times 7$) nonamer peptides, designed to cover all the possible 1−2 and 1−3 interactions. As this is an unachievable goal even for combinatorial chemistry, one is obliged to deal with the available "real" databases.

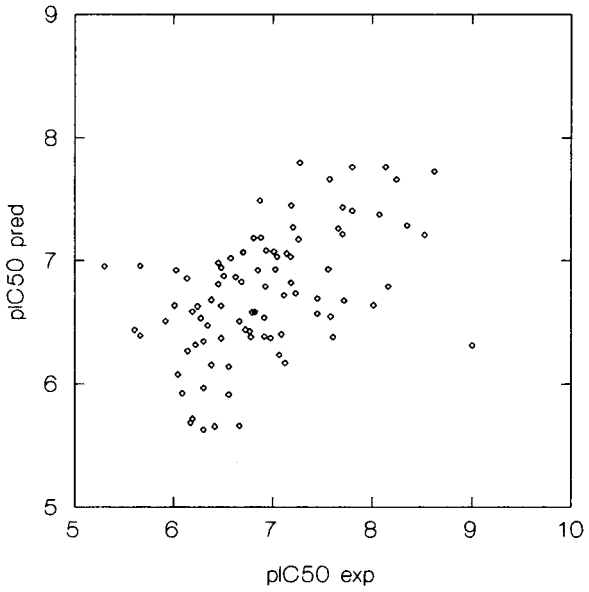


Figure 2. Predicted vs experimental pIC_{50} values of the test set ($n = 89$).

The contributions of the amino acids at different positions are presented in Figure 3. The contributions of the adjacent side chains interactions are plotted in Figure 4, and those of the every second side chain are shown in Figure 5. Because of limited space, only contributions above ±0.050, ±0.070, or ±0.075 are presented in the last two figures.

Amino Acid Contributions. Tyr is the favorite amino acid for P1 (Figure 3). This is in a good agreement with many experimental data.[26,38] Topologically, this position corresponds to pocket A of the cleft of the peptide-binding site on HLA-A*0201, as described by Sapper and colleagues.[39] The pocket surface is predominantly polar, and there is a network of hydrogen bonds that directly involve two hydrogen bonds to the peptide N-terminus. Phe and Lys are also well tolerated at P1. Arg, His, and Thr are not preferred at P1. The remaining amino acids make negligibly small contributions.

Met and Leu are the favorite amino acids at P2. It is well-known that Leu falls into pocket B in the HLA-A*0201 binding cleft.[7,29,40−42] This pocket has a polar rim and hydrophobic inner

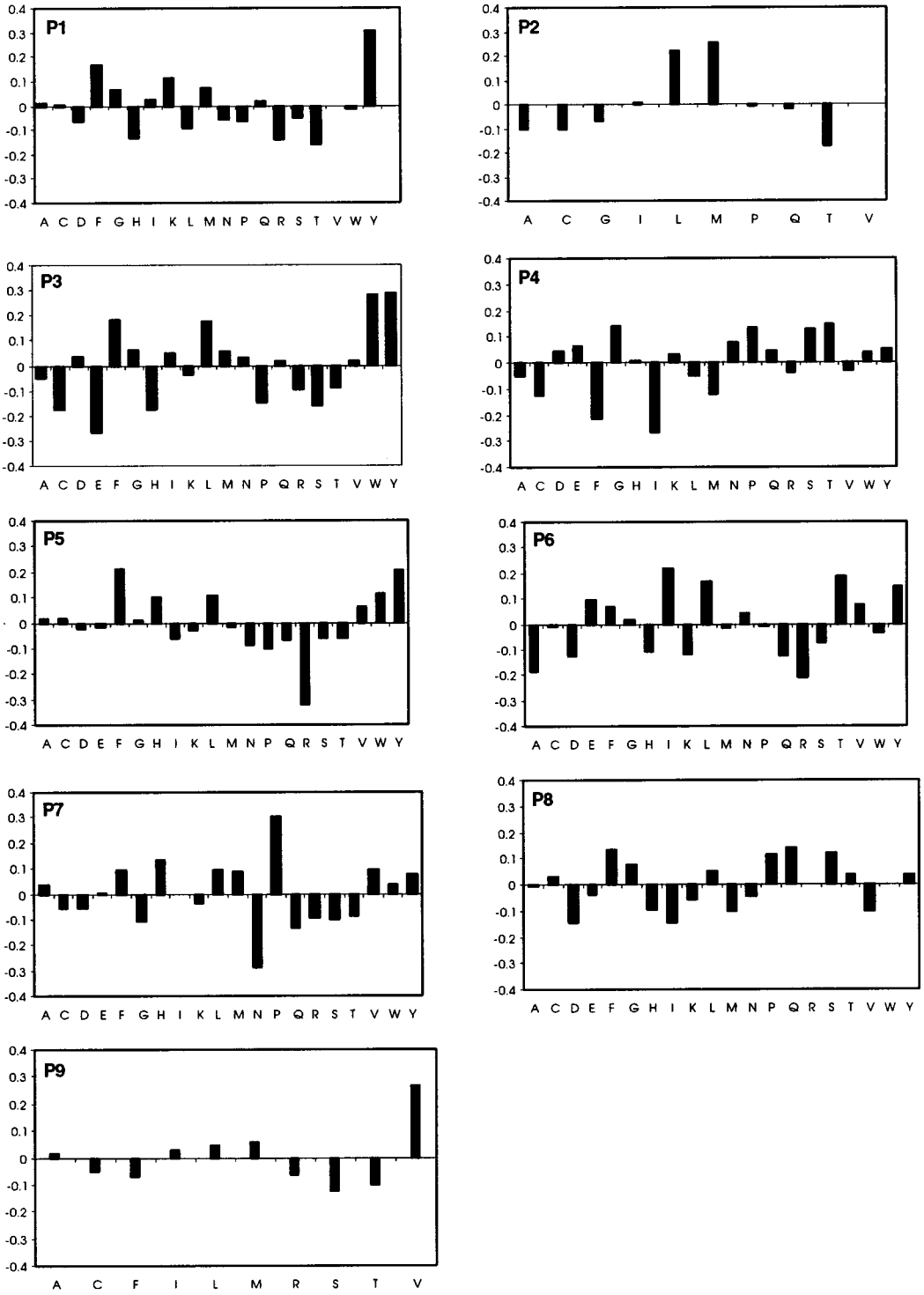
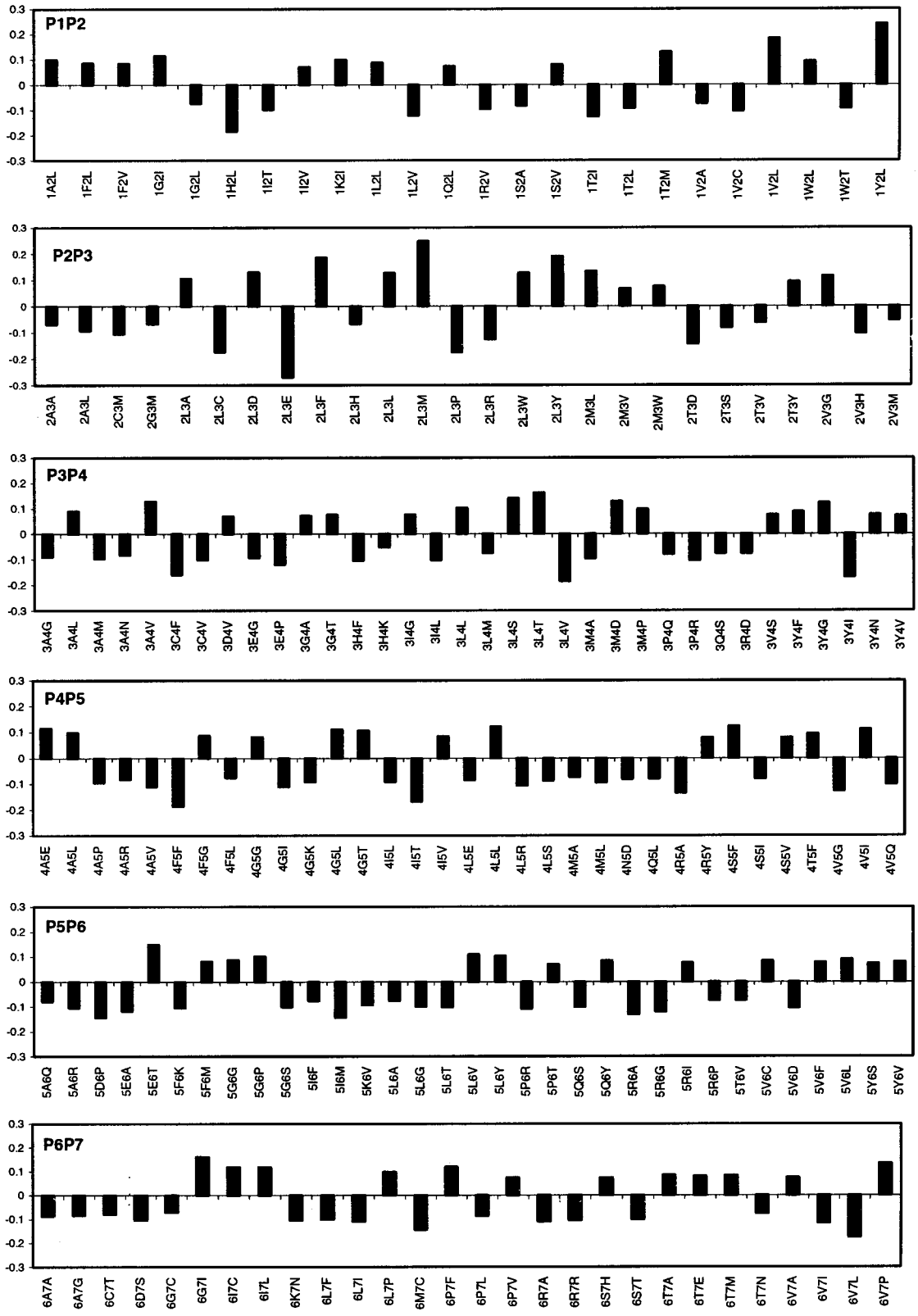


Figure 3. Amino acid contributions.

walls. The contribution of Met is surprisingly higher than that of Leu. Ala, Cys, Gly, and Thr make negative contributions. The contributions of the other amino acids are, again, negligibly small.

Trp and Tyr are the favorite amino acids for P3, but Leu and Phe are also well tolerated. Glu is deleterious for affinity. Cys, His, Pro, and Ser contribute negatively. The side chains of the amino acids at this position fall into pocket D. Pocket D has been called the "loose" pocket[43] and is a secondary binding pocket. It is a hydrophobic cavity located between the aromatic rings of Tyr99 and Tyr159, including also residues 155, 156, and 160.[44] This pocket prefers large hydrophobic residues, like Phe and Trp,[45] in good agreement with their positive contributions.

There is not a favorite amino acid at P4. Ile and Phe are deleterious, Cys and Met have significant negative contributions. Gly, Pro, Ser, and Thr are well accepted here. The side chain at P4 has been called the "flag" side chain[43] because it is solvent-exposed in the complex with MHC molecule and can


P1P2
1A2L 1F2L 1F2V 1G2I 1G2L 1H2L 1I2T 1I2V 1K2I 1L2L 1L2V 1Q2L 1R2V 1S2A 1S2V 1T2I 1T2L 1T2M 1V2A 1V2C 1V2L 1W2L 1W2T 1Y2L
P2P3
2A3A 2A3L 2C3M 2G3M 2L3A 2L3C 2L3D 2L3E 2L3F 2L3H 2L3L 2L3M 2L3P 2L3R 2L3W 2L3Y 2M3L 2M3V 2M3W 2T3D 2T3S 2T3V 2T3Y 2V3G 2V3H 2V3M
P3P4
3A4G 3A4L 3A4M 3A4N 3A4V 3C4F 3C4V 3D4V 3E4G 3E4P 3G4A 3G4T 3H4F 3H4K 3I4G 3I4L 3L4L 3L4M 3L4S 3L4T 3L4V 3M4A 3M4D 3M4P 3P4Q 3P4R 3Q4S 3R4D 3V4S 3Y4F 3Y4G 3Y4I 3Y4N 3Y4V
P4P5
4A5E 4A5L 4A5P 4A5R 4A5V 4F5F 4F5G 4F5L 4G5G 4G5I 4G5K 4G5L 4G5T 4I5L 4I5T 4I5V 4L5E 4L5L 4L5R 4L5S 4M5A 4M5L 4N5D 4Q5L 4R5A 4R5Y 4S5F 4S5I 4S5V 4T5F 4V5G 4V5I 4V5Q
P5P6
5A6Q 5A6R 5D6P 5E6A 5E6T 5F6K 5F6M 5G6G 5G6P 5G6S 5I6F 5I6M 5K6V 5L6A 5L6G 5L6T 5L6V 5L6Y 5P6R 5P6T 5Q6S 5Q6Y 5R6A 5R6G 5R6I 5R6P 5T6V 5V6C 5V6D 5V6F 5V6L 5Y6S 5Y6V
P6P7
6A7A 6A7G 6C7T 6D7S 6G7C 6G7I 6I7C 6I7L 6K7N 6L7F 6L7I 6L7P 6M7C 6P7F 6P7L 6P7V 6R7A 6R7R 6S7H 6S7T 6T7A 6T7E 6T7M 6T7N 6V7A 6V7I 6V7L 6V7P

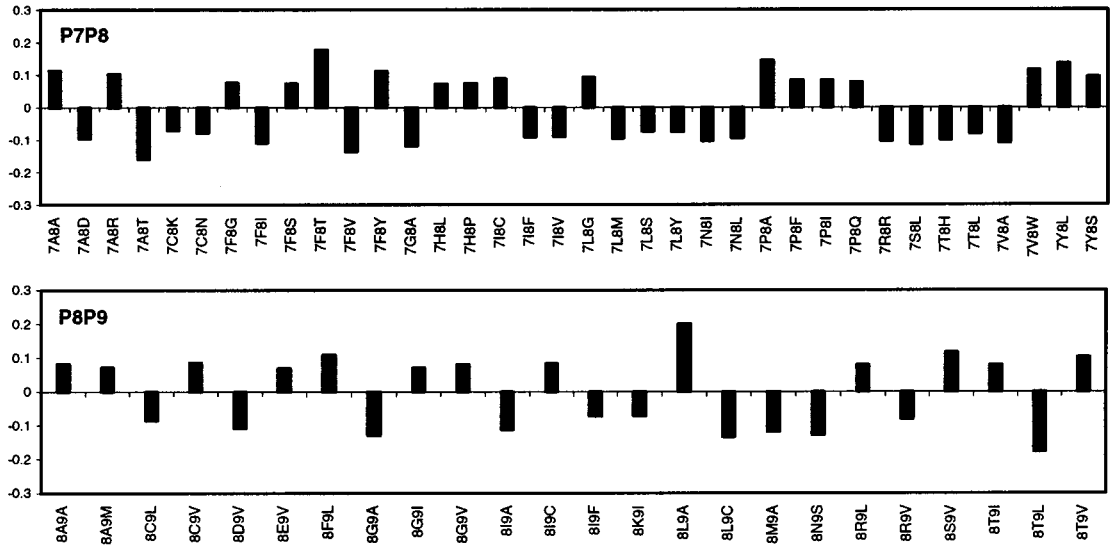


Figure 4. Contributions of the adjacent side-chain interactions. The presented contributions for P2P3 are above ±0.050, for P1P2, P3P4, P5P6, P6P7, P7P8 and P8P9 above ±0.070, and for P4P5 above ±0.075.

contact the TCR. Favorite amino acids for P5 are Phe and Tyr. His, Leu, and Trp also contribute positively. Arg should be avoided at this position.

Ile, Leu, Thr, and Tyr are well accepted at P6. Ala, Arg, Asp, Gln, His, and Lys here contribute negatively to the affinity. This side chain falls into pocket C.[39] This pocket is predominantly polar with a shallow depression.[39] This explains the acceptance of the hydrophilic Thr and Tyr, but it cannot explain the preference of the hydrophobic Ile and Leu.

Pro is the favorite amino acid at P7. His makes a good contribution as well. Asn is deleterious here; Arg, Gln, Gly, Ser, and Thr are not preferred. Pocket E corresponds topologically to P7. Two-thirds of the surface area in the pocket is hydrophobic, one-third is polar.[39] This pocket can accommodate a variety of complementary peptide side chains, but an incompatible side chain need not prevent complex formation. This pocket has been called the "part-time" pocket[43] and belongs to the secondary binding pockets.

There is not a favorite amino acid at P8, though Gln, Phe, Pro, and Ser are well accepted here. The presence of Asp, Ile, His, Met, or Val is not desirable. P8 is a "flag" position like P4.[39] Val is the preferred amino acid at P9. Ser and Thr should be avoided here. The C-terminal of the peptide falls into pocket F of the binding site.[39]

Contributions of the Side-Chain Interactions. Peptides bound to the HLA-A*0201 molecule assume extended but twisted conformations.[41] As a result, the adjacent side chains protrude in largely opposite directions, and in practice, interactions between them are not likely to exist. The interaction between the adjacent side chains may be considered as a change in the backbone conformation caused by a certain amino acid at a certain position that can produce changes in the conformation of the adjacent amino acid side chain. The twisted conformation, however, makes possible the interactions between every second amino acid side chain. These interactions might have steric, electrostatic, hydrophobic, or hydrogen-bond formatting nature. A conformational change also is possible here. The additive method cannot explain the nature of forces involved in such interaction, but it can assess quantitatively its significance for the affinity.

Among the adjacent side-chain interactions, 1Y2L, 2L3M, and 8L9A (Figure 4) are particularly strong. The high contributions of the last two combinations are very unexpected as 3M, 8L, and 9A are not among the favorite or highly positively contributed amino acids. The only reasonable explanation is a conformational change favoring the binding. The only delirious combination is 2L3E. Glu is not a favored amino acid for P3.

The combinations 1G3L, 2L4Q, 4S6G, and 7F9V have the highest contributions in the binding affinity to HLA-A*0201 among the 1−3 side-chain interactions (Figure 5). 2L and 9V are favorite amino acids, 3L and 4S make significant positive contributions, and 1G, 4Q, 6G, and 7F make negligibly small positive contributions. It is possible to suppose conformational changes and steric interaction rather than electrostatic or hydrophobic interactions or intramolecular hydrogen-bond formation. Deleterious combinations are 1I3L, 2T4Q, and 4I6V. The first of them seems counterintuitive because both 1I and 3L make positive contributions to the affinity. Furthermore, 3L, in combination with 1G, makes a highly valued positive contribution. Obviously, the steric bulk of 1I causes an inappropriate change in the conformation of the 3L side chain. The intolerance of 2T4Q and 4I6V is probably due to the high negative contributions of 2T and 4I.

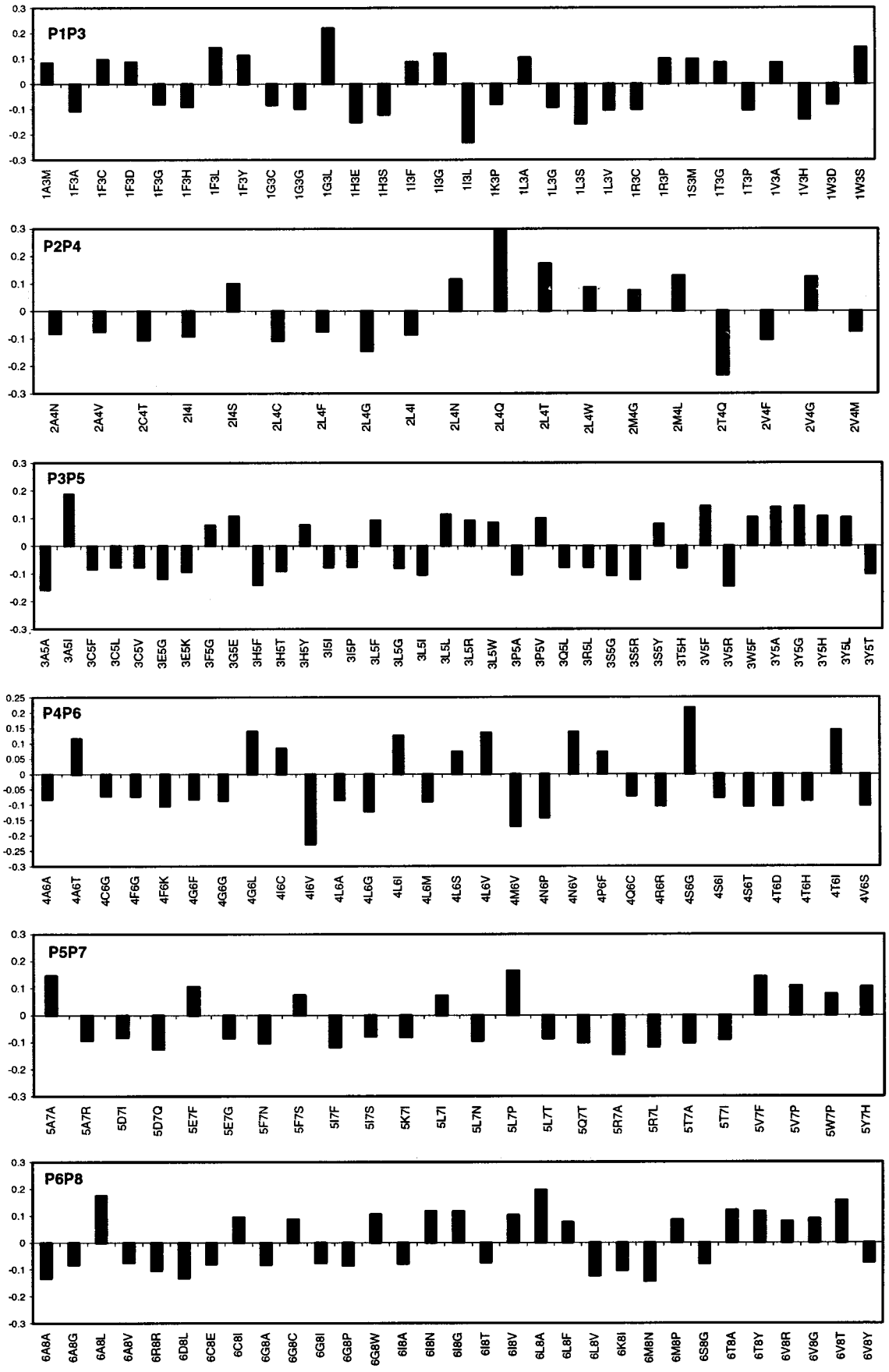
How Does the Additive Method Work? Although the contributions of the amino acids and the interactions between them were implemented into a program for rapid T-cell epitope search, the binding affinity of a peptide to HLA-A*0201 molecule can be calculated, by hand, both quickly and easily. For example, eq 2 for the binding affinity of the nonamer YLSPGPVTV can be presented in the following way:

$$\begin{aligned} pIC_{50} = \text{const} + 1Y + 2L + 3S + 4P + 5G + 6P + 7V + 8T + \\ 9V + 1Y2L + 2L3S + 3S4P + 4P5G + 5G6P + 6P7V + 7V8T + \\ 8T9V + 1Y3S + 2L4P + 3S5G + 4P6P + 5G7V + 6P8T + 7V9V \end{aligned}$$

Substituting each term by its quantitative value, the final calculated pIC_{50} value is 7.700:

$$\begin{aligned} pIC_{50} = 6.213 + 0.304 + 0.219 - 0.164 + 0.135 + 0.013 - 0.008 + \\ 0.096 + 0.035 + 0.263 + 0.240 - 0.015 + 0 + 0 + 0.101 + \\ 0.075 + 0.059 + 0.102 + 0.031 + 0.044 - 0.107 + 0.046 + \\ 0.011 + 0.008 - 0.001 = 7.700 \end{aligned}$$

The pIC_{50} value of the peptide YLSPGPVTV in our database is 7.642. There are two absent values: one for 3S4P and one


P1P3
1A3M 1F3A 1F3C 1F3D 1F3G 1F3H 1F3L 1F3Y 1G3C 1G3G 1G3L 1H3E 1H3S 1I3F 1I3G 1I3L 1K3P 1L3A 1L3G 1L3S 1L3V 1R3C 1R3P 1S3M 1T3G 1T3P 1V3A 1V3H 1W3D 1W3S
P2P4
2A4N 2A4V 2C4T 2I4I 2I4S 2L4C 2L4F 2L4G 2L4I 2L4N 2L4Q 2L4T 2L4W 2M4G 2M4L 2T4Q 2V4F 2V4G 2V4M
P3P5
3A5A 3A5I 3C5F 3C5L 3C5V 3E5G 3E5K 3F5G 3G5E 3H5F 3H5T 3H5Y 3I5I 3I5P 3L5F 3L5G 3L5I 3L5L 3L5R 3L5W 3P5A 3P5V 3Q5L 3R5L 3S5G 3S5R 3S5Y 3T5H 3V5F 3V5R 3W5F 3Y5A 3Y5G 3Y5H 3Y5L 3Y5T
P4P6
4A6A 4A6T 4C6G 4F6G 4F6K 4G6F 4G6G 4G6L 4I6C 4I6V 4L6A 4L6G 4L6I 4L6M 4L6S 4L6V 4M6V 4N6P 4N6V 4P6F 4Q6C 4R6R 4S6G 4S6I 4S6T 4T6D 4T6H 4T6I 4V6S
P5P7
5A7A 5A7R 5D7I 5D7Q 5E7F 5E7G 5F7N 5F7S 5I7F 5I7S 5K7I 5L7I 5L7N 5L7P 5L7T 5Q7T 5R7A 5R7L 5T7A 5T7I 5V7F 5V7P 5W7P 5Y7H
P6P8
6A8A 6A8G 6A8L 6A8V 6R8R 6D8L 6C8E 6C8I 6G8A 6G8C 6G8I 6G8P 6G8W 6I8A 6I8N 6I8G 6I8T 6I8V 6L8A 6L8F 6L8V 6K8I 6M8N 6M8P 6S8G 6T8A 6T8Y 6V8R 6V8G 6V8T 6V8Y

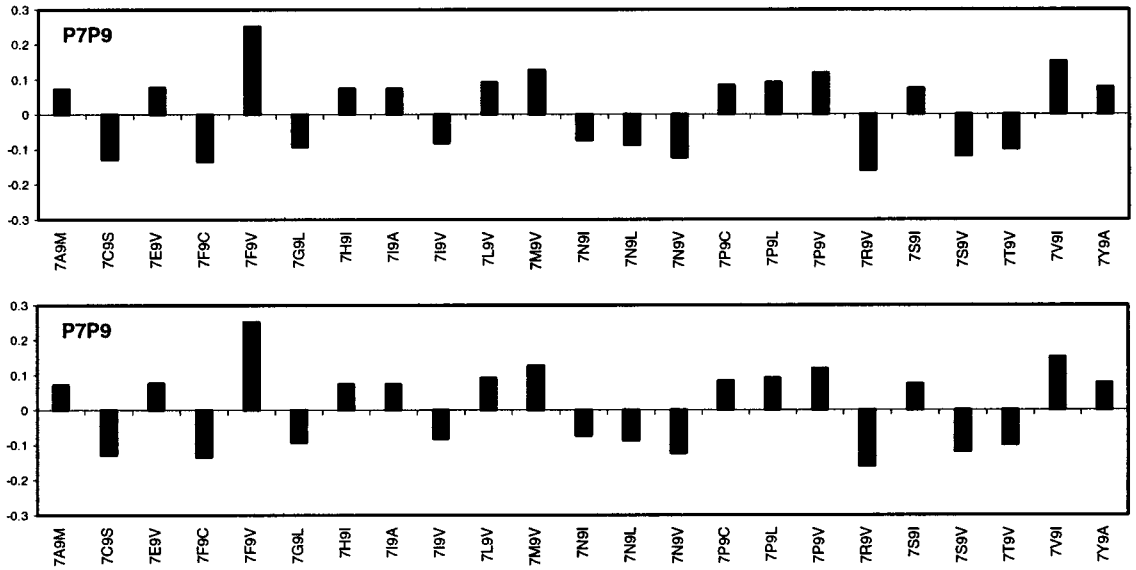


Figure 5. Contributions of every second side-chain interaction. The presented contributions for P1P3, P2P4, P4P6, P5P7, and P7P9 are above ±0.070 and for P3P5 and P6P8 above ±0.075.

for 4P5G. It is important to note that as the number of absent values in the equation increases there is greater possibility for incorrect prediction.

In developing these methods, we have encountered problems that are only rarely associated with QSAR analyses of small molecules. These include the size of the peptide molecules being studied; the number of molecules being investigated, perhaps 10 times greater than a small molecule study; and the great diversity of physicochemical properties associated with each position being examined. As indicated, some of our results may contain minor statistical anomalies. These should decrease as the number and diversity of peptides we study increases. However, substituting residues with high contributions at each position allows us to design sets of very high affinity peptides, each one capable of acting as a T-cell epitope. Our preliminary experimental work in this regard gives promising results. Internet access to the additive method will be forthcoming.[46] Expansion of the method to other alleles is in progress.

In conclusion, the additive method for quantitative binding affinity prediction proposed here has many advantages in comparison with other methods. First, it is easy and fast to use. Second, it gives a quantitative value for the binding affinity with very good predictive power—the mean |residual| value is about 0.5. Finally, although in this paper we have concentrated on the immunological problem of MHC peptide binding, the method we have described is universal. As is well known, many proteins bind peptide ligands including, for example, SH2 domains, which bind phosphotyrosine containing peptides; SH3 domains, which recognize polyproline peptides; EH domains, which recognize proteins containing Asn-Pro-Phe (NPF) sequences; PDZ domains, which prefer peptides with a free carboxy termini; and WW domains, among others. Thus, the additive method can address affinity prediction in any type of peptide−protein interaction system where sets of known binders, of a known length, are available.

Supporting Information Available: The final regression equation containing 1815 terms including the constant: 159 terms account for the amino acid contributions, 879 terms account for the contributions of the adjacent side-chain interactions, and 776 terms account for the contributions of every second side-chain interaction. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Janeway, C. A., Jr.; Travers, P.; Walport, M.; Capra, J. D. *Immunobiology*; Elsevier Science Ltd.: New York, 1999; pp 115−162.
(2) Nijman, H. W.; Houbiers, J. G.; Vierboom, M. P.; van der Burg, S. H.; Drijfhout, J. W.; D'Amaro, J.; Kenemans, P.; Melief, C. J.; Kast, W. M. Identification of Peptide Sequences That Potentially Trigger HLA-A2.1 Restricted Cytoxic T Lymphocytes. *Eur. J. Immunol.* **1993**, *23*, 1215−1219.
(3) Kast, W. M.; Brandt, R. M. P.; Sidney, J.; Drijfhout, J.-W.; Kubo, R. T.; Grey, H. M.; Melief, C. J. M.; Sette, A. Role of HLA-A Motifs in Identification of Potential CTL Epitopes in Human Papiloma Virus Type 16 E6 and E7 Proteins. *J. Immunol.* **1994**, *152*, 3904−3912.
(4) Arnon, R.; Horwitz, R. J. Synthetic Peptides as Vaccines. *Curr. Opin. Immunol.* **1992**, *4*, 449−453.
(5) Naruse, H.; Ogasawara, K.; Kaneda, R. A Potential Peptide Vaccine Against Two Different Strains of Influenza Virus Isolated at Intervals of About 10 Years. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 9588−9592.
(6) Sette, A.; Vitiello, A.; Reherman, B.; Fower, P.; Nayersina, R.; Kast, W. M.; Melief, C. J. M.; Oseroff, C.; Yuan, L.; Ruppert, J.; Sidney, J.; del Guercio, M.-F.; Southwood, S.; Kubo, R. T.; Chesnut, R. W.; Grey, H. M.; Chisati, F. V. The Relationship Between Class I Binding Affinity and Immunogenicity of Potential Cytotoxic T-Cell Epitopes. *J. Immunol.* **1994**, *153*, 5586−5592.
(7) 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 Chain. *J. Immunol.* **1994**, *152*, 163−175.
(8) Parker, K. C.; Shields, M.; DiBrino, M.; Brooks, A.; Coligan, J. E. Peptide Binding to MHC Class I Molecules: Implications for Antigenic Peptide Prediction. *Immunol. Res.* **1995**, *14*, 34−57.
(9) Gulukota, K.; Sidney, J.; Sette, A. Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules. *J. Mol. Biol.* **1997**, *267*, 1258−1267.
(10) Adams, H. P.; Koziol, J. A. Prediction of Binding to MHC Class I Molecules. *J. Immunol. Methods* **1995**, *185*, 181−190.
(11) Brusic, V.; Rudy, G.; Honeyman M.; Hammer, J.; Harrison, L. Prediction of MHC Class II-binding Peptides Using Evolutionary Algorithm and Artificial Neural Network. *Bioinformatics* **1998**, *14*, 121−130.
(12) Altuvia, Y.; Sette, A.; Sidney, J.; Southwood, S.; Margalit, H. A Structure-based Algorithm to Predict Potential Binding Peptides to MHC Molecules with Hydrophobic Binding Pockets. *Hum. Immunol.* **1997**, *58*, 1−11.

(13) Mata, M.; Travers, P. J.; Liu, Q.; Frankel, F. R.; Paterson, Y. The MHC Class I-restricted Immune Response to HIV-gag in BALB/c Mice Selects a Single Epitope That Does Not Have a Predictable MHC-binding Motif and Binds to K^d Through Interactions Between a Glutamine at P3 and Pocket D. *J. Immunol.* **1998**, *161*, 2985−2993.
(14) Doytchinova, I. A.; Flower, D. R. Toward the Quantitative Prediction of T-Cell Epitopes: CoMFA and CoMSIA Studies of Peptides with Affinity for the Class I MHC Molecule HLA-A*0201. *J. Med. Chem.* **2001**, *44*, 3572−3281.
(15) 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**, in press.
(16) Free, S. M., Jr.; Wilson, J. W. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(17) Fujita, T.; Ban, T. Structure−Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters. *J. Med. Chem.* **1971**, *14*, 148−152.
(18) Cammarata, A.; Yau, S. Predictability of Correlations Between *in vitro* Tetracycline Potencies and Substituent Indices. *J. Med. Chem.* **1970**, *13*, 93−97.
(19) Tomic, S.; Nilsson, L.; Wade, R. C. Nuclear Receptor−DNA Binding Specificity: A COMBINE and Free-Wilson QSAR Analysis. *J. Med. Chem.* **2000**, *43*, 1780−1792.
(20) Terada, Y.; Nanya, K. Free-Wilson Analysis of the Antibacterial Activity of Fluoronaphthyridines Against Various Microbes. A New Application of Indicator Variables. *Pharmazie* **2000**, *55*, 133−135.
(21) Tmej, C.; Chiba, P.; Huber, M.; Richter, E.; Hitzler, M.; Schaper, K. J.; Ecker, G.. A Combined Hansch/Free−Wilson Approach as Predictive Tool in QSAR Studies on Propafenone-type Modulators of Multidrug Resistance. *Arch Pharm (Weinheim)* **1998**, *331*, 233−240.
(22) Dalpiaz, A.; Gessi, S.; Varani, K.; Borea, P. A. De Novo Analysis of Receptor Binding Affinity Data of 8-ethenyl-xantine Antagonists to Adenosine A1 and $A2_A$ Receptors. *Arzneimittelforschung* **1997**, *47*, 591−594.
(23) Nisato, D.; Wadnon, J.; Callet, G.; Mettefeu, D.; Assens, J. L.; Plouzane, C.; Tonnerre, B.; Pliska, V.; Fauchère J. L. Renin Inhibitors. Free-Wilson and Correlation Analysis of the Inhibitory Potency of a Series of Pepstatin Analogues on Plasma Renin. *J. Med. Chem.* **1987**, *30*, 2287−2291.
(24) Gombar, V. Quantitative Structure−Activity Relationships. Fujita-Ban Analysis of Beta-adrenergic Blocking Activity of 1-phenoxy-3-([(substituted amido)alkyl] amino)-2-propanols. *Arzneimittelforschung* **1986**, *36*, 1014−1018.
(25) Bindal, M. C.; Singh, P.; Gupta, S. P. Structure−Activity Studies on Hallucinogenic Phenylalkylamines Using Fujita−Ban Approach. *Arzneimittelforschung* **1982**, *32*, 719−721.
(26) Kirksey, T. J.; Pogue-Caley, R. R.; Frelinger, J. A.; Collins, E. J. The Structural Basis for the Increased Immunogenicity of Two HIV− Reverse Transcriptase Peptide Variant/Class I Major Histocompatibility Complexes. *J. Biol. Chem.* **1999**, *274*, 37259−37264.
(27) Wold, S. PLS for Multivariate Linear Modeling. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 195−218.
(28) Blythe, M. J.; Doytchinova, I. A.; Flower, D. R. JenPep: A Database of Quantitative Functional Peptide Data for Immunology. *Bioinformatics* **2002**, in press.
(29) 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-A*0201 Molecules. *Cell* **1993**, *74*, 929−937.
(30) Sette, A.; Sidney, J.; del Guercio, M.-F.; Southwood, S.; Ruppert, J.; Dalberg, C.; Grey, H. M.; Kubo, R. T. Peptide Binding to the Most Frequent HLA-A Class I Alleles Measured by Quantitative Molecular Binding Assays. *Mol. Immunol.* **1994**, *31*, 813−822.
(31) SYBYL 6.7. Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
(32) Bush, B. L.; Nachbar, R. B., Jr. Sample−Distance Partial Least Squares: PLS Optimized for Many Variables, with Application to CoMFA. *J. Comput.-Aid. Mol. Des.* **1993**, *7*, 587−619.
(33) 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.
(34) Sicsic, S.; Serraz, I.; Andrieux, J.; Bremont, B.; Mathé-Allainmat, M.; Poncet, A.; Shen, S.; Langlois. M. Three-Dimensional Quantitative Structure−Activity Relationship of Melatonin Receptor Ligands: A Comparative Molecular Field Analysis Study. *J. Med. Chem.* **1997**, *40*, 739−748.
(35) Ducrot, P.; Legraverend, M.; Grierson, D. S. 3D-QSAR CoMFA on Cyclin-dependent Kinase Inhibitors. *J. Med. Chem.* **2000**, *43*, 4098−4108.
(36) Ligand-Based Design Manual, SYBYL 6.6. Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144.
(37) Rudolf, M. P.; Man, S.; Melief, C. J. M.; Sette, A.; Kast, M. Human T-Cell Responses to HLA-A-restricted High Binding Affinity Peptides of Human Papilomavirus Type 18 Proteins E6 and E7. *Clin. Cancer Res.* **2001**, *7*, 788s−795s.
(38) Tourdot, S.; Scardino, A.; Saloustrou, E.; Gross, D. A.; Pascolo, S.; Cordopatis, P.; Lemonnier, F. A.; Kosmatopoulos, K. A. General Strategy to Enhance Immunogenicity of Low-affinity HLA-A2.1-associated Peptides: Implication in the Identification of Criptic Tumor Epitopes. *Eur. J. Immunol.* **2000**, *30*, 3411−3421.
(39) Saper, M. A.; Bjorkman, P. J.; Wiley: D. C. Refined Structure of the Human Class I Histocompatibility Antigen HLA-A2 at 2.6 Å. *J. Mol. Biol.* **1991**, *219*, 277−319.
(40) Falk, K.; Rötzschke, O.; Stefanovic, S.; Jung, G.; Rammensee, H.-G. Allele Specific Motifs Revealed by Sequencing of Self-Peptides Eluted from MHC Molecules. *Nature* **1991**, *351*, 290−296.
(41) Madden, D. R.; Garboczi, D. N.; Wiley: D. C. The Antigenic Identity of Peptide-MHC Complexes: A Comparison of the Conformations of Five Viral Peptides Presented by HLA-A2. *Cell* **1993**, *75*, 693−708.
(42) Madden, D. R. The Three-dimensional Structure of Peptide-MHC Complexes. *Annu. Rev. Immunol.* **1995**, *13*, 587−622.
(43) Bjorkman, P. J.; Saper, M. A.; Samraoui, B.; Bennett, W. S.; Strominger, J. L.; Wiley, D. C. The Foreign Antigen Binding Site and T-Cell Recognition Regions of Class I Histocompatibility Antigens. *Nature* **1987**, *329*, 512−518.
(44) 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.
(45) Sarobe, P.; Pendleton, C. D.; Akatsuka, T., D.; Engelhard, V. H.; Feinstone, S. M.; Berzofsky, J. A. Enhanced *in vitro* Potency and *in vivo* Immunogenicity of a CTL Epitope from Hepatitis C Virus Core Protein Following Amino Acid Replacement at Secondary HLA-A2.1 Binding Positions. *J. Clin. Invest.* **1998**, *102*, 1239−1248.
(46) Guan, P.; Doytchinova, I. A.; Zygouri, C.; Flower, D. R. MHCPred: Bringing a Quantitative Dimension to the Online Prediction of MHC Binding. Manuscript in preparation.

PR015513Z