Toward Prediction of Class II Mouse Major Histocompatibility Complex Peptide Binding Affinity: in Silico Bioinformatic Evaluation Using Partial Least Squares, a Robust Multivariate Statistical Technique

Channa K. Hattotuwagama,* Christopher P. Toseland, Pingping Guan, Debra J. Taylor, Shelley L. Hemsley, Irini A. Doytchinova, and Darren R. Flower

The Jenner Institute, University of Oxford, Compton, Berkshire, U. K. RG20 7NN

The accurate identification of T-cell epitopes remains a principal goal of bioinformatics within immunology. As the immunogenicity of peptide epitopes is dependent on their binding to major histocompatibility complex (MHC) molecules, the prediction of binding affinity is a prerequisite to the reliable prediction of epitopes. The iterative self-consistent (ISC) partial-least-squares (PLS)-based additive method is a recently developed bioinformatic approach for predicting class II peptide−MHC binding affinity. The ISC−PLS method overcomes many of the conceptual difficulties inherent in the prediction of class II peptide−MHC affinity, such as the binding of a mixed population of peptide lengths due to the open-ended class II binding site. The method has applications in both the accurate prediction of class II epitopes and the manipulation of affinity for heteroclitic and competitor peptides. The method is applied here to six class II mouse alleles (I-A^b, I-A^d, I-A^k, I-A^s, I-E^d, and I-E^k) and included peptides up to 25 amino acids in length. A series of regression equations highlighting the quantitative contributions of individual amino acids at each peptide position was established. The initial model for each allele exhibited only moderate predictivity. Once the set of selected peptide subsequences had converged, the final models exhibited a satisfactory predictive power. Convergence was reached between the 4th and 17th iterations, and the leave-one-out cross-validation statistical terms—q^2, SEP, and NC—ranged between 0.732 and 0.925, 0.418 and 0.816, and 1 and 6, respectively. The non-cross-validated statistical terms r^2 and SEE ranged between 0.98 and 0.995 and 0.089 and 0.180, respectively. The peptides used in this study are available from the AntiJen database (http://www.jenner.ac.uk/AntiJen). The PLS method is available commercially in the SYBYL molecular modeling software package. The resulting models, which can be used for accurate T-cell epitope prediction, will be made freely available online (http://www.jenner.ac.uk/MHCPred).

INTRODUCTION

The products of the major histocompatibility complex (MHC) play a fundamental role in regulating immune responses. T cells recognize antigens as peptide fragments complexed with MHC molecules, a process requiring antigen degradation through complex proteolytic digestion prior to complexation. The biological role of MHC proteins is, thus, to bind peptides and "present" these at the cell surface for inspection by T-cell antigen receptors (TCRs, or TRs using IMGT nomenclature). MHC genes are grouped into two classes on the basis of their related biological properties and similar secondary and tertiary structures, yet they exhibit important functional differences. Class I molecules are composed of a heavy chain complexed to β2-microglobulin, while class II molecules consist of two chains (α and β) of similar size. Both classes of MHC molecule have similar 3-D structures composed of two domains. The MHC peptide-binding site consists of a β sheet, forming the base, flanked by two α helices, which together form a narrow cleft or groove accommodating bound peptides. The principal difference between the two classes are the dimensions of

the peptide-binding groove: class I is closed at either end and is constrained to bind short peptides (typically 8−11 amino acids in length), while class II is open at both ends, allowing much larger peptides of varying length to be bound.

As previously mentioned, class II MHC molecules are noncovalent heterodimers and are called HLA-DP, HLA-DQ, and HLA-DR in humans and I-A and I-E in mice. Peptides binding to class II MHC molecules are usually 10−25 residues long, with peptide lengths of 13−16 amino acids being the most frequently observed.[1−4] From X-ray crystallographic data of MHC class II and TCR−peptide−MHC class II complexes,[5,6] it is clear that nine amino acids are bound in an extended conformation within the class II binding site. In contrast to class I peptides, they are not anchored at either amino or carboxyl termini but stretch along the binding groove, with residues accommodated by binding pockets along the cleft. Previous interpretations, extant within the literature, suggest that class II peptides have a small number of anchor residues upon which binding depends. These anchors are residues of an appropriate type, which must sit at particular spacings along the peptide in order for MHC allele-restricted binding to occur; residues at other peptide positions are, in terms of peptide specificity, less constrained.

* To whom correspondence should be addressed. E-mail: channa.hattotuwagama@jenner.ac.uk.

The side chain at peptide position P1 binds into a deep pocket while four shallow pockets bind side chains at peptide positions P4, P6, P7, and P9. The side chains at positions P2, P3, P5, and P8 point toward the T-cell receptor.

We have recently developed an iterative self-consistent (ISC) partial-least-squares (PLS)-based extension,[7] to the additive method,[8,9] for prediction of class II peptide-binding affinity and applied it to three human class II alleles. We now address binding to six class II mouse alleles (I-A^b, I-A^d, I-A^k, I-A^s, I-E^d, and I-E^k), for peptides of up to 25 amino acids in length. The ISC additive method assumes that the binding affinity of a large peptide is principally derived from the interaction, with an MHC molecule, of a continuous subsequence of amino acids within it. The ISC is able to factor out the contribution of individual amino acids within the subsequence, which is initially identified in an iterative manner.

SYSTEMS AND METHODS

Peptide Database and Binding Affinities. The information and data based on the peptide sequences and their binding affinities were obtained from the AntiJen database, a development of JenPep[10,11] (URL: http://www.jenner.ac.uk/AntiJen). For each set of class II alleles, peptide lengths of 10−25 were obtained from the database. A total of 44 I-A^b, 145 I-A^d, 55 I-A^k, 81 I-A^s, 69 I-E^d, and 52 I-E^k peptide sequences were found with known binding affinities (IC$_{50}$). Extracted IC$_{50}$ values were first converted to log(1/IC$_{50}$) values [or −log(IC$_{50}$) or pIC$_{50}$] and used as the dependent variables in the QSAR model. The pIC$_{50}$ values were predicted from a combination of the contributions of individual amino acids at each position of the peptide. The binding affinities were originally assessed using a competition assay based on the inhibition of binding of a radio-labeled standard peptide to a detergent-solubilized MHC molecule.[12−13]

Additive Method. Initially, the sequence of each peptide was segmented into a set of nine amino acid subsequences, that is, nonamer sequences at positions 1−9, 2−10, 3−11, 4−12, and so forth. Each nonameric subsequence is then converted into a binary bit string of 180 bins (9 positions × 20 amino acids) to create a "matrix". Within the matrix, a term is set to 1 when an amino acid is present at a particular position and 0 when it is absent. 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

$$pIC_{50} = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2} +$$
$$\sum_{i=1}^{6} P_i P_{i+3} + \sum_{i=1}^{5} P_i P_{i+4} + \sum_{i=1}^{3} P_i P_{i+6} + \sum_{i=1}^{2} P_i P_{i+7} + P_i P_{i+8}$$

$$(1)$$

where the const accounts, at least nominally, 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, $\sum_{i=1}^{7} P_i P_{i+2}$ is the

sum of every second side-chain interaction, $\sum_{i=1}^{6} P_i P_{i+3}$ is the sum of every third side-chain interaction, and so on. As these two models were roughly equivalent in terms of statistical quality, to simplify the matrix, we applied the principle of Occam's razor and sought the simplest explanation: only the amino acid contributions to be considered were chosen, and interactions between side chains at relative positions 1−2 and 1−3 were neglected. The matrix was analyzed using PLS,[14] an extension of multiple linear regression (MLR). Leave-one-out cross-validation was used to indicate the predictive ability of the model.

Cross-Validation Using the "Leave-One-Out" (LOO−CV) Method. The predictive statistical power of the models from the additive method for each allele was carried out using PLS[14] as implemented within SYBYL 6.9.[15] The method works by producing an equation or QSAR (quantitative structure−activity relationship), 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. In our case, CV works by dividing the data set into a set of peptides, developing several parallel models from the reduced data with one or more of the peptides being 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: cross-validated coefficient (q^2) and the standard error of prediction (SEP), which are defined in eqs 2 and 3.

$$q^2 = 1.0 - \frac{\sum_{i=1}[pIC_{50(exp)} - pIC_{50(pred)}]^2}{\sum_{i=1}[pIC_{50(exp)} - pIC_{50(mean)}]^2} \text{ or simplified to } q^2 =$$

$$1.0 - \frac{PRESS}{SSQ} \quad (2)$$

where, in eq 2, pIC$_{50(pred)}$ is a predicted IC$_{50}$ value and

$$SEP = \sqrt{\frac{PRESS}{p - 1}} \quad (3)$$

pIC$_{50(exp)}$ is an experimental IC$_{50}$ 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. In eq 3, p is the number of peptides omitted from the data set. 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 MLR validation terms: explained variance (r^2) and standard error of estimate (SEE), which are defined in eqs 4 and 5

$$r^2 = \frac{\sum_{i=1}^{n}(Y_{obs} - Y_{pred})^2}{\sum_{i=1}^{n}(Y_{obs} - \bar{Y})^2} \qquad (4)$$

where, in eq 4, Y_{pred} is the predicted, Y_{obs} the observed, and

$$SEE = \sqrt{\frac{PRESS}{n - c - 1}} \qquad (5)$$

$\bar{Y}$ the average dependent variable, in this case, IC_{50} values. In eq 5, n is the number of peptides and c is the number of 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, up to an optimal value, improves the fit between target and explanatory properties; the statistic which is optimized in the LOO−CV when extracting the optimal number of components is the best q^2 value (>0.4). 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.

Iterative Self-Consistent (ISC) Algorithm. Each iteration of our ISC PLS-based algorithm generates a set of nonameric subsequences extracted from the parent peptide.[7] The subsequences were extracted by taking segments of nine amino acids starting at positions 1−9, followed by positions 2−10, 3−11, 4−12, and so on, depending on the length of the original parent peptide. Values for pIC_{50} corresponding to this set of peptides were predicted using PLS and compared to the experimental pIC_{50} value for each parent peptide. LOO−CV was then employed to extract the optimal number of components, which was then used to generate the non-cross-validated model. The previous model is used to predict the pIC_{50} values and a new set is extracted. The best predicted nonamer was selected for each peptide; those with the lowest residual between the experimental and predicted pIC_{50} were

chosen for each peptide. Each new model is built from the set of optimally scored nonamers. The method works by comparing the new set of peptide sequences with the old set, and if the new set is different, an extra iteration is begun. The process is repeated, and the new set is compared with the previous one, and if they are the same, the final model is obtained; that is, the model has reached convergence. The resulting coefficients of the final non-cross-validated model describe the quantitative contributions of each amino acid at each of the nine positions. An example coefficient matrix for the I-A^b allele is shown in Table 1.

Additive Model Evaluation. A key step in bioinformatics is validation. We used an independent test data set to evaluate the predictivity of our mouse class II models. Sixty class II epitopes from 21 protein sequences were used in the test set, see Table 2.[16−54] The full protein sequences were retrieved from SWISS−PROT.[55] Complete protein sequences were used as input in this test, and the ability of the algorithm to identify epitopes correctly was assessed. A threshold, based on percentages of returned potential epitopes, was set in the test. In real-life situations, immunologists wish to test a small number of peptides with high binding affinity, as these are more likely to be actual epitopes. The number of binding peptides, if not immunodominant epitopes, is proportional to the sequence length. We assessed the top 5%, 10%, and 15% of predicted peptides, counting the number of epitopes identified. We compared results from our method with those from RANKPEP,[56,57] the only server to offer a significant number of mouse models.

RESULTS

The additive method was applied to peptide binding data for six mouse class II alleles: I-A^b,[58−64] I-A^d, I-A^k, I-A^s, I-E^d, and I-E^k, with the number of peptides for each allele ranging from 44 to 145. The resulting statistical parameters were based exclusively on the amino acid contributions (amino acids only models). The validation results for the final models are shown in Table 3. The initial model for each allele exhibited only moderate predictivity. Once the set of selected

Table 1. Additive Model for the Binding Affinity Prediction to the I-A^b Allele[a]

	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	−0.016	−0.008	0.265	−0.115	0.066	−0.442	0.050	0.447	−0.034
C	0.000	0.083	0.037	−0.051	0.090	0.050	0.216	0.079	−0.139
D	−0.065	0.000	−0.067	0.000	0.000	0.107	−0.077	−0.041	−0.203
E	−0.028	−0.129	0.000	0.000	0.000	0.000	0.000	−0.048	0.000
F	0.000	0.000	0.000	−0.283	0.000	0.000	0.000	0.000	0.000
G	−0.286	−0.039	0.050	−0.011	0.000	0.000	−0.003	0.000	−0.067
H	−0.003	−0.013	0.000	0.000	0.000	0.000	0.000	0.213	0.000
I	−0.043	0.090	−0.364	−0.090	0.000	−0.244	−0.351	0.000	−0.069
K	0.094	0.000	0.000	0.000	−0.069	0.000	0.000	0.000	0.000
L	0.000	−0.215	−0.110	0.094	−0.162	0.000	−0.003	−0.242	0.066
M	0.008	−0.067	0.000	0.258	0.223	0.154	0.017	−0.027	0.082
N	0.000	0.298	0.000	0.042	−0.003	−0.069	0.064	−0.097	−0.455
P	0.100	0.000	0.032	0.090	0.030	0.201	0.080	0.000	0.280
Q	−0.013	0.000	−0.235	0.000	0.000	0.000	0.000	−0.067	−0.051
R	0.164	−0.286	0.066	0.122	−0.233	0.120	0.213	−0.229	0.216
S	−0.051	0.090	0.161	0.036	−0.078	0.041	−0.125	0.000	0.213
T	0.054	0.151	0.079	−0.060	0.233	0.000	−0.079	0.012	0.161
V	−0.069	−0.048	0.000	0.064	0.000	0.000	0.000	0.000	0.000
W	0.000	0.000	−0.029	0.000	−0.097	−0.003	0.000	0.000	0.000
Y	0.155	0.092	0.116	−0.097	0.000	0.085	0.000	0.000	0.000

[a] Constant = 6.044 (the constant accounts, at least nominally, for the peptide backbone contribution). 0.000 represents position where amino acids are absent within the matrix.

Table 2. Sequences of the Epitopes and Their Source Proteins

overlapping epitopes	location	allele	protein	source	reference
KYLEFISDAIIHVLHS	102−117	I-As	myoglobin	equine	16
FISDAIIHVLHSK	106−118	I-Ak	myoglobin	equine	16
FISDAIIHVLHSK	106−118	I-Ad	myoglobin	equine	16
ELFRKDIAAKY	136−146	I-Ed	myoglobin	sperm whale	17
SALAMVYLGAKDSTR	36−50	I-Ad	ova	chicken	18
PKYVKQNTLKLATG	306−319	I-Ed	hemagglutinin H3	influenza virus	19
LKLATGMRNVPEKQT	314−328	I-Ed	hemagglutinin H3	influenza virus	19
ISQAVHAAHAEINEAGR	323−339	I-Ad	ova	chicken	20
ISQAVHAAHAEINEAGR	323−339	I-Ab	ova	chicken	20
ISQAVHAAHAEINEAGR	323−339	I-As	ova	chicken	20
AIWQVEQKASIAGTDSGW	55−72	I-As	PorB	*Neisseria meningitidis*	21
NYKNGGFFVQYGGAYKRH	163−180	I-Ad	PorB	*Neisseria meningitidis*	21
TPRVSYAHGFKGLVDDAD	244−261	I-As	PorB	*Neisseria meningitidis*	21
TLAYRFGNVTPRVSYAHG	235−252	I-Ad	PorB	*Neisseria meningitidis*	21
DKYRSITVRV	127−136	I-Ab	enterotoxin b	*Streptococcus* sp.	22
DENPVVHFF	79−87	I-Ak	myelin basic protein	mouse	23
DENPVVHFF	79−87	I-Ak	myelin basic protein	guinea pig	23
YALKRQGRTLYG	88−99	I-Ad	histone H4	calf thymus	24
GVADPVKVTRSALQN	489−503	I-Ad	HSP65	*Mycobacterium avium*	25
GVADPVKVTRSALQN	489−503	I-Ad	HSP65	*Mycobacterium tuberculosis*	25
GVADPVKVTRSALQN	489−503	I-Ad	HSP65	*Mycobacterium bovis*	25
GVADPVKVTRSALQN	489−503	I-Ad	HSP65	*Mycobacterium leprae*	25
FEVGATYYF	295−303	I-Ed	OMPF	*Escherichia coli*	26
AAKFESNFNTQATNRNT	31−47	I-Ak	HEL	chicken	27
DGSTDYGILQINSRWW	48−63	I-Ak	HEL	chicken	27
ATSLSPFYLRPPSFLR	41−56	I-As	α-B-crystallin	bovine	28
GSTDYGILQINSR	49−61	I-Ed	HEL	chicken	29
TGKICNNPHRILDGIDCTLID	48−68	I-Ak	hemagglutinin HA1	influenza virus	30
LEFITEGFTWTEVTQNGGSNA	118−138	I-Ak	hemagglutinin HA1	influenza virus	30

nonoverlapping epitopes	location	allele	protein	source	reference
AWVAWRNRCK	107−116	I-Ed	HEL	chicken	31
AAKFESNFNTQATNRNT	31−47	I-Ak	HEL	chicken	27
AAKFESNFNTQATNRN	31−46	I-Ak	HEL	chicken	27
DGSTDYGILQINSR	48−61	I-Ak	HEL	chicken	32
DYGILQINSR	52−61	I-Ak	HEL	chicken	33
TEWTSSNVMEERKIKV	265−280	I-Ab	OVA	chicken	34
PKSDNQIKAVPAS	234−246	I-Ak	SM-P40	*Schistosoma mansoni*	35
SKYPNCAYKTTQANKH	90−105	I-Ek	ribonuclease	buffalo	36
LGIWTYDGTKVSISPES	146−162	I-Ab	acetylcholine receptor	*Torpedo californica*	37
RNDGSTDYGILQINSR	46−61	I-Ak	HEL	chicken	38
YIYADGKMVN	173−182	I-Ek	straptococcal nuclease	*Straptococcal aureus*	39
GKKVITAFNEGLK	64−76	I-Ek	hemoglobin	mouse	40
DGSTDYGILQINSRW	48−62	I-Ak	HEL	chicken	41
GRGLAYIYADGKMVN	168−182	I-Ek	straptococcal nuclease	*Straptococcal aureus*	42
SVSSFERFEIFPK	107−119	I-Ed	hemaglutinin	influenza virus	43
CPKYVRSAKLRM	302−313	I-Ed	hemaglutinin	influenza virus	43
NLCNIPCSALLSSDI	74−88	I-Ab	HEL	chicken	44
NLCNIPCSALLSSDI	74−88	I-Ak	HEL	chicken	44
DYGILQINS	52−60	I-Ak	HEL	chicken	45
SSDITASVNCAK	85−96	I-Ek	HEL	chicken	46
WRRQARFK	71−78	I-Ed	nucleocapsid protein	infectious bronchitis virus	47
DGSTDYGILQINSRW	48−62	I-Ak	HEL	chicken	48
IIANDQGNRTTPSY	28−41	I-Ak	HSP70	drosphila simulans	48
WVAWRNRCK	107−116	I-Ek	HEL	chicken	49
WVAWRNRCK	107−116	I-Ed	HEL	chicken	49
TYTEHAKRKTVTAMDVVYALKRQG	71−94	I-Ab	histone H4	mouse	50
TYTEHAKRKTVTAMDVVYALKRQG	71−94	I-Ad	histone H4	mouse	50
TYTEHAKRKTVTAMDVVYALKRQG	71−94	I-As	histone H4	mouse	50
LRDNIQGITKPAIRR	22−36	I-Ab	histone H4	mouse	50
LRDNIQGITKPAIRR	22−36	I-As	histone H4	mouse	50
NIQGITKPAIRRLAR	25−39	I-Ab	histone H4	mouse	50
NIQGITKPAIRRLAR	25−39	I-Ad	histone H4	mouse	50
NIQGITKPAIRRLAR	25−39	I-As	histone H4	mouse	50
LIAYLKQATAK	94−104	I-Ek	cytochrome c	pigeon	51
ISQAVHAAHAEINEAGR	323−339	I-Ad	ova	chicken	52
TQFHPPHIEIQML	48−60	I-Ad	microglobulin	mouse	53
ISQAVHAAHAEINEAGR	323−339	I-Ad	ova	chicken	54

peptide subsequences had converged, the predictive power of the final models was satisfactory. Convergence was reached between the 4th and 17th iteration, and the LOO−

CV terms—q^2, SEP, and NC—ranged between 0.732 and 0.925, 0.418 and 0.816, and 1 and 6, respectively. The non-cross-validated terms r^2 and SEE ranged between 0.98 and

Table 3. Summary of Results for the Final Converged Alleles Generated after the Additive Method

model	I-A^b	I-A^d	I-A^k	I-A^s	I-E^d	I-E^k
number of peptides	44	145	55	81	69	52
number of iterations	7	14	4	17	8	8
Leave-One-Out Cross-Validation						
NC^a	6	6	6	6	6	6
q^{2b}	0.850	0.898	0.790	0.783	0.732	0.925
SEP^c	0.459	0.534	0.816	0.588	0.557	0.418
Non-Cross-Validation						
r^2	0.994	0.993	0.990	0.980	0.992	0.995
SEE^d	0.089	0.136	0.180	0.177	0.096	0.106

^{a} Optimal number of components. ^{b} q^2 obtained after LOO–CV. ^{c} Standard error of prediction. ^{d} Standard estimate of error.

0.995 and 0.089 and 0.180, respectively. The statistical results for each of the six models are summarized in Table 3. Graphical representations of the respective amino acid contributions at each binding position are given in Figure 1. The most favored and most disfavored amino acids at each binding position for each allele are summarized in Table 4.

While I-A^d gives marginally the better statistics, we chose to report results from I-A^b, which converged in fewer iterations, indicating a more stable and self-consistent result. The best PLS model had excellent predictivity, with LOO–CV parameters $q^2 = 0.850$, SEP = 0.459, NC = 6 and the non-cross-validated results being $r^2 = 0.994$ and SEE = 0.089 (Table 3). The final "matrix" of amino acid contributions at each position is shown in Table 1. Convergence was reached at the seventh iteration. The best predicted nonamer found within each of the full sequences is shown in Table 5. The matrix shown in Table 1 is the final converged model used to predict the pIC$_{50}$ values reported in Table 5.

In cross-validation and prediction, optimally scored nonamers can be determined using the ISC algorithm. Leave-one-out cross-validation is used to extract the optimum number of components subsequently used to generate the non-cross-validated model, while the previous model is used to predict pIC$_{50}$ values as a new set of subsequences are extracted. The best predicted nonamers were selected for each peptide (lowest residual difference between the experimental and predicted IC$_{50}$), as shown in Table 5. The validation statistics reported in Table 3 are encouraging with respect to our methods. Though the r^2 values are high, which some might interpret as overfitting, the high values of q^2—the equivalent of r^2 under cross-validation—and its similarity with r^2 is a testament to the lack of significant overfitting.

The present study defines models for peptides binding to a series of mouse class II molecules: I-A^b, I-A^d, I-A^k, I-A^s, I-E^d, and I-E^k molecules. Class II epitopes are believed to be nine or more amino acids in length and to possess primary anchors at positions P1 and P4 and secondary anchors at positions P6 and P9.[65–67] In our models, specificity for amino acid classes in I-A and I-E models are identical at some—but different at other—anchor positions, illustrating the different requirements for peptides presented by different alleles. The anchors with the most restricted specificity are P1 and P9. The results do not show a complete consistency in amino acid residue found within the different models,

although a significant number of residues fall into similar classifications (e.g., hydrophobic, aliphatic, and polar). Table 4 shows that P1 is predominantly aliphatic and hydrophobic for all I-A and I-E molecules, with the exception of I-A^d and I-A^s, where it is aromatic. P9 is similar to those amino acids at P1, in that it is predominantly hydrophobic and aliphatic for I-A molecules and somewhat hydrophobic and aliphatic for the I-E molecules. P4 and P6 (Table 4) show a totally different amino acid preference between the models. Our results for I-E^d and I-E^k binding peptides agree well with X-ray crystallographic data for HLA-DR$_1$ binding pocket peptide specificity,[68] thus confirming the predicted similarity, in terms of function as well as structure, of mouse class II I-E molecules with human DR molecules. The complementarity of MHC pockets and peptide side chains, in respect to the size and shape of the pocket versus those of the side chain, largely determines peptide selectivity. For example, Pro (I-A^b), Phe (I-A^s), or Ile (I-E^k) are found in the large, hydrophobic pocket at P1; Leu (I-A^d), Ala (I-E^d), or Ser (I-A^s) are found in the pocket at P6; Cys (I-A^k) is found in the pocket at P9.

To further benchmark our method, the favored binding anchor positions identified by the ISC additive method were compared to existing literature definitions of anchor motifs, as collated in SYFPEITHI.[69] Table 6 shows favored amino acid residues identified by the ISC–PLS method compared to well-tolerated anchor residues from SYFPEITHI at positions P1, P4, P6, and P9. I-A^b and I-A^s alleles are not represented in the table as no anchor motif is present in SYFPEITHI. Although the correspondences between the ISC method and SYFPEITHI are not exact, they do show similarity between amino acid properties. Peptides binding to certain I-A alleles, such as I-A^d, do not have as clearly identifiable binding motifs,[70] suggesting that their amino acid specificity may not be as stringent as that for other class II molecules, for example, the I-A^d allele, where the underlying amino acid property is polar at position P1 and hydrophobic at positions P4 and P6 in the SYFPEITHI database. The I-A^k allele[71] prefers peptides containing negatively charged, polar amino acids at position P1 (Asp and Asn from SYFPEITHI; Thr from ISC method). The P1 binding pockets of most HLA-DR and I-E molecules are reported to be large and hydrophobic, and these class II molecules prefer to bind peptides with larger aliphatic or aromatic side chains at P1. Consistent with empirical results apparent in SYFPEITHI and elsewhere, our results are suggestive that the I-A^k P1 binding pocket may have an atypical specificity, when compared to the P1 binding pockets of I-E molecules. Results for the ISC method and SYFPEITHI for the I-E molecules reveal differences in specificity for I-E^{d72} and I-E$^{k.49,72–74}$ Table 6 shows that, at P1, the amino acid residues (Met and Ile) for the two I-E alleles are both hydrophobic. In SYFPEIHTI, these residues are also listed as broadly hydrophobic: the I-E^d allele consists mainly of aromatic residues, and the I-E^k allele consists mainly of aliphatic residues. The types of amino acid residues listed in SYFPEIHTI for position P9 are predominantly basic, hydrophilic groups which are different from the small nonpolar residues (alanine and cysteine, respectively) identified by the ISC additive method for the I-E^d and I-E^k alleles.

Among online epitope prediction algorithms, only SYFPEITHI[69] and RANKPEP[56] include mouse class II
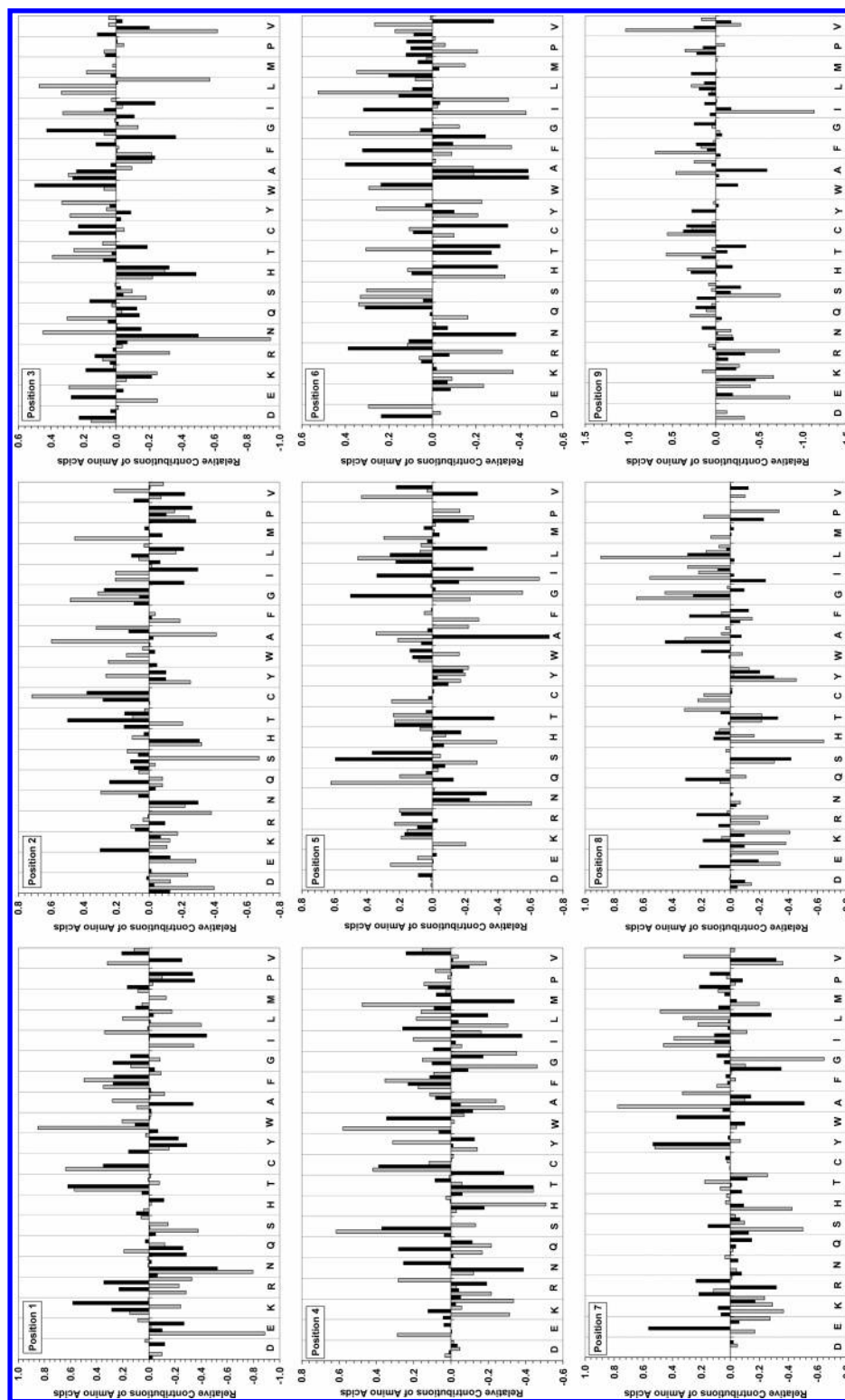
Figure 1. Relative contributions of position-wise amino acids at each binding position 1−9 for the I-A[b], I-A[d], I-A[k], I-A[s], I-E[d], and I-E[k] alleles. The contribution made by different individual amino acids at each position of the 9-mer I-A[b], I-A[d], I-A[k], I-A[s], I-E[d], and I-E[k] binding peptide. The contribution is equivalent to a position-wise amino acid regression coefficient obtained by PLS regression (as described in the text).

models. However, SYFPEITHI has only three class II models, I-A[k], I-E[k], and I-L[d], only two of which are common to the models described here. RANKPEP,[56,57] which is a reliable, robust, if underused, method, does contain models for seven different mouse class II alleles. It is the only server to offer multiple mouse models. We used an independent test set comprising 60 epitopes to assess the performance of

our algorithm and to compare its performance to that of RANKPEP. Complete sequences were input to both methods; this reduces bias, as it mimics the use to which immunologists will put the methods: identifying high binders capable of displaying immunogenicity. The performances are summarized in Table 7. Viewed in isolation, our approach performs well, returning the majority of known epitopes with

Table 4. Summary of the Favored and Disfavored Amino Acid Residues at Each Binding Position for the Six MHC Class II Alleles[a]

	P1	P2	P3	P4	P5	P6	P7	P8	P9
				Favored Binding					
I-A[b][l]	*P*	*K*	*A*	*L*	*L, T*	*M*	*R*	*A*	*M*
I-A[d][l]	C, T, W	A, G, M	*T*	C, M, S, W	Q, L, V	L	A, I, Y	G, I, L	A, C, F, T, V
I-A[k][l]	T	*T*	G	C	G, S	F	E, Y	Q	C
I-A[s][l]	F	C	N, L	F	A	I	I	G	H
I-E[d][l]	M	*Q*	W	W	S	A	W	R	C
I-E[k][l]	*I*	*A*	*Y*	*R*	*R*	*Q*	*L*	*T*	*A*
				Disfavored Binding					
I-A[b][l]	*Q*	*P*	*G*	*C*	*P*	A	*G*	*I*	K
I-A[d][l]	N, E, L	N	N, V	G, T	N, I	I	H, S	H, Y	E, I, K, S
I-A[k][l]	N	*H*	N, H	T	A	A	A	S	A
I-A[s][l]	*K*	A, S	*R*	H	G	*K*	G	*P*	R
I-E[d][l]	I	*I*	*H*	I	*L*	C	*L*	*Y*	*I*
I-E[k][l]	*R*	*R*	*L*	*G*	A, Y	*I*	*E*	K	*E*

[a] The best positively favored and worst negatively disfavored amino acids chosen at each binding position at their various cut-off values as represented in Figures 1−6. A cut-off value of > ±0.4 is applied to favored and disfavored binding amino acid residues. Where no amino acid residue exceeds this threshold (±0.4), the next best residue is chosen (shown in italics).

the top 15%, a conveniently small set of peptides to make. Comparatively, neither method wins hands down. At the 15% level, RANKPEP does not significantly outperform the additive method, except for I-E[k], where our results are poor. However, RANKPEP is based, in part, on the use of epitope data, as opposed to binding data, thus introducing bias into any validation exercise, as there will be an overlap between test and training sets. In light of this, the performance of the ISC method is commendable. Algorithms underlying these servers vary, and often, direct comparison between such servers is problematic.[75] Moreover, relatively little is known about the molecular mechanism that is relevant to epitope selection by CD4+ T cells. Our method appears to predict IC$_{50}$ values well, but there may be other issues in helper T-cell epitope selection which we do not properly account for, for which an epitope-based predictor, such as RANK-PEP, may do implicitly. RANKPEP is, essentially, based on "memorizing" existing epitopes, including ones used to test it. This is unavoidable since the data set, from which the most up-to-date version of RANKPEP is built, is not available. Because it uses epitope data, RANKPEP may, artificially, enhance its prediction on data known to it but be poorer on unseen data. This contention remains to be tested. Moreover, because of the limited number of test sequences available, any comparison is likely to lack statistical rigor. As more epitope data become available, the quality of all models will, doubtless, improve.

DISCUSSION

As is well-known, the prediction of MHC class II epitopes is very much more difficult than that of class I. In the present study, we have examined a recently developed bioinformatics method: the ISC PLS-based additive method, which was applied to the prediction of class II MHC−peptide binding affinity. We have shown previously that ISC is a reliable, quantitative method for binding affinity prediction[7] and have applied it here to peptides binding to six mouse MHC class II molecules (I-A[b], I-A[d], I-A[k], I-A[s], I-E[d], and I-E[k]), developing a series of quantitative, systematic models, based on literature IC$_{50}$ values.

It is seems clear from experimental studies of T-cell epitope analogue binding and data from X-ray crystallography that peptides bind to MHC molecules through the interaction of side chains of certain peptide residues with pockets situated in the MHC class II peptide-binding site: these side chains extend into discreet pockets within the binding groove.[6,76−77] Peptide side chains form favorable (polar, hydrophobic, or steric) interactions with MHC side chains within these pockets;[77] the most critical determinant of binding, other than the presence of appropriate types of side chains, is their relative spacing. Indeed, it has been suggested that different MHC class II molecules may bind the same peptide in multiple binding registers, whereby the peptide is displaced longitudinally within the binding groove with side chains being bound by different pockets.[78−80] Reviewing this concept, two main alternative scenarios are identified: binding of the same peptide in different registers by the same or different alleles.[81] The more common second alternative is well-demonstrated and results from minor polymorphic differences in the amino acid residue composition of the binding groove.[78,80] In the DRB$_5$ complex, the large P1 pocket accommodates Phe from the peptide and Ile occupies the shallow pocket at P4. However, in the DRB$_1$ allele, the small pocket at P1 is occupied by Val, shifting the peptide to the right, while Phe occupies a deeper pocket at P4. This also causes certain peptide side chains, which are orientated toward the TCR, to change.[78] Unequivocal evidence supporting the former alternative is somewhat scarce: there are few, if any, proper examples of exactly the *same* peptide binding in different registers to exactly the *same* MHC molecule. For example, the recent work of Xia et al.[82] shows no evidence of hierarchical binding to HLA-DQ2 even for the long, highly repetitive epitope LQLQPFPQPELPYPQPELPYPQPELPYPQPQPF.

The above observations are easily rationalized by recourse to simple statistical mechanics.[83] The proportions (*P*) of two binding modes (state 1 and state 2) are given by the ratio of their Boltzmann factors:

$$\frac{P(\text{state 1})}{P(\text{state 2})} = \frac{\exp(-E_1/kT)}{\exp(-E_2/kT)}$$

where E_1 and E_2 are the relative energies of the two states. Unless these energies are, by chance, close, the proportions of the lower energy state will prevail by many orders of magnitude, as is seen experimentally.[84] Repetitive sequences

Table 5. List of Peptides Used in This Study of the I-A^b Mouse Allele, Highlighting the Best Predicted 9-mer Sequences[a]

number	epitope	peptide length	exptl IC$_{50}$	highest predicted IC$_{50}$	predicted IC$_{50}$ closest to exptl IC$_{50}$	ref.
1	A**QGALANIA**IE	11	6.959	7.054	7.054	11
2	YD**AQGALANI**A	11	6.260	6.263	6.263	11
3	KPVS**QMRMATP**L	12	4.982	7.218	5.297	10
4	EAI**QPGCIGGP**K	12	3.211	5.913	3.384	13
5	**MRMATPLLM**RPM	12	6.682	7.194	6.767	10
6	EAI**QPGCIAGP**K	12	3.346	6.360	3.186	13
7	AKF**VAAWTLKAA**A	13	7.377	7.384	7.384	17
8	KPVS**QMRMATP**LL	13	5.584	7.218	5.297	10
9	YD**AQGALANI**AIE	13	7.155	7.054	7.054	11
10	Q**MRMATPLLM**RPM	13	6.818	7.194	6.767	10
11	TP**PAYRPPNAP**IL	13	6.593	6.593	6.593	12
12	SQMR**MATPLLM**RPM	14	7.237	7.194	7.194	10
13	KPVS**QMRMATP**LLM	14	7.009	7.218	6.998	10
14	ISQAV**HAAHAEI**NE	14	6.426	6.566	6.417	18
15	QGQMV**HQAISPRTL**N	15	6.815	6.810	6.810	17
16	**KPVSQMRMATP**LLMR	15	7.215	7.218	7.218	10
17	VSQMR**MATPLLM**RPM	15	7.086	7.194	7.089	10
18	KIL**EPFRKYTAF**TIP	15	5.832	6.374	5.780	17
19	**AKRKTVTAM**DVVYAL	15	7.367	7.371	7.371	15
20	KTVTA**MDVVYALKR**Q	15	5.277	6.756	5.276	15
21	LRDNIQ**GITKPAIRR**	15	5.125	6.625	5.169	15
22	NIQ**GITKPAIRR**LAR	15	5.187	6.760	5.169	15
23	PRTLN**GPGPGSPA**IF	15	6.504	6.813	6.240	17
24	EKVYLA**WVPAHKGIG**	15	6.992	6.829	6.829	17
25	**HQAISPRTLN**SPAIF	15	7.854	7.395	7.395	17
26	**SPAIFQSSMTK**ILEP	15	6.231	6.498	6.233	17
27	FRK**YTAFTIPS**INNE	15	6.994	6.972	6.972	17
28	PRTLN**SPAIFQ**SSMT	15	7.125	7.395	7.395	17
29	**HQAISPRTLN**SPAIF	15	7.854	7.395	7.395	17
30	HSNW**RAMASDF**NLPP	15	5.379	6.613	5.217	17
31	PVS**QMRMATP**LLMRPM	16	6.979	7.218	6.998	10
32	NTDGS**TDYGILQIN**SR	16	5.426	6.227	5.394	12
33	**KPVSQMRMAT**PLLMRP	16	7.276	7.218	7.218	10
34	AS**FEQGALANI**AVDKA	16	6.260	6.988	6.273	11
35	KPVSQMR**MATPLLM**RPL	17	7.097	7.218	7.089	10
36	KPVSKMR**MATPLLM**QAM	17	7.585	7.651	7.651	10
37	KPVSKMR**MATPLLM**QAL	17	7.509	7.453	7.453	10
38	KPVSQLR**MATPLLM**RPM	17	7.009	7.411	7.089	10
39	KPVSQMR**MATPLLL**RPM	17	7.114	7.218	7.147	10
40	KPVSQMR**MATPLLM**RPM	17	7.131	7.218	7.089	10
41	KPVSQMR**LATPLLM**RPM	17	6.975	7.194	6.997	10
42	KPVSQMR**MATPLLM**RPM	17	7.131	7.218	7.089	14
43	LGIETYDG**TKVSISPES**	17	7.076	7.066	7.066	16
44	YK**PVSQLRLAT**PLLLRPL	18	7.310	7.348	7.348	10

[a] For the I-A^b allele, this table shows the observed pIC$_{50}$ values for each of the best predicted nonameric sequences originating from the parent peptide. The best 9-mer sequences are highlighted in bold.

Table 6. Comparison of Favored Binding Positions between the ISC Additive Method and the SYFPEITHI Database[a]

	P1		P4		P6		P9	
	additive method	SYFPEITHI	additive method	SYFPEITHI	additive method	SYFPEITHI	additive method	SYFPEITHI
I-A^d	C, T, W	S, T, Y, E	C, S, M, W	V, L, I, A	L	A, V		
I-A^k	T	D, N	C	I, V, L, N	F	E, Q		
I-E^d	M	W, Y, F	W	**K, R**, I	A	**I, L, V**, G	C	K, R
I-E^k	I	I, L, V	R	I, L, V, F, S	C	**Q, N**, A, G	A	K, R

[a] Amino acid residues shown in bold represent well-tolerated anchors, and those residues that are not shown in bold represent less-tolerated anchors.

are an exception as they are much more likely to produce closely spaced energy differences between peptides binding in different registers. However, recognition and binding are not the same. The affinities exhibited by a TCR for a low-affinity, low-abundance register-shifted peptide, and, indeed, the repertoire frequency of T cells expressing such a receptor, may be totally different, dwarfing that of TCRs which bind to the high-affinity register.

Not withstanding the arguments outlined above, the task common to all class II prediction methods is the identification of the binding subsequence: the region of the peptide that actually interacts strongly with the MHC. As discussed above, this search is complicated, conceptually at least, by the ability of MHCs to bind in a degenerative manner. Long peptides, in particular, might exhibit a hierarchy of multiple binding modes. However, as we have said, relatively little

Table 7. Comparison of the Performance of the ISC Additive Method with RANKPEP for the Six Alleles Described in the Paper[a]

			method					
	number of	number of	MHCPred			RANKPEP		
alleles	proteins	epitopes	top 5%	top 10%	top 15%	top 5%	top 10%	top 15%
I-A[b]	5	7	0.57	0.85	0.85	0.57	0.71	0.71
I-A[d]	10	13	0.23	0.46	0.85	0.46	0.54	0.54
I-A[k]	7	16	0.25	0.37	0.56	0.75	0.81	0.88
I-A[s]	5	7	0.28	0.57	0.71	0.71	0.86	0.86
I-E[d]	5	10	0.10	0.20	0.40	0.50	0.60	0.60
I-E[k]	5	7	0.14	0.28	0.28	0.86	0.86	0.86

[a] We have compared the performance, in terms of epitope identification with full sequences, of the additive method with that of RANKPEP, the only sever to offer a similar number of mouse models. Percentages are given of the number of correctly identified epitopes in the top scoring selections of predicted binders. The top 5%, 10%, and 15% are shown.

is known concerning the explicit degeneracy of the binding process. Nonetheless, the fact that the binding groove is open at both ends in class II molecules is consistent with the possibility. Whether this phenomenon actually occurs in reality is, except for repetitive sequences, unlikely on theoretical grounds, as discussed above. Nonetheless, the inability to identify the correct register has confounded attempts to produce accurate models for class II, necessitating our use of iterative techniques. Ideally, we would wish to compare the improvement in predicted binders versus a random selection, that is, a different starting set of nonamers, where the whole proteins had been analyzed for T-cell epitopes using overlapping peptides. However, fully controlled experiments such as this are costly and rarely preformed. So, we concentrated on the prediction of known class II restricted epitopes.

It may be assumed, from a thermodynamic standpoint, that the best binding nonameric subsequence, within the parent peptide, will have the highest pIC_{50} (i.e., the lowest binding energy) among all possible nonamers originating from the long parent peptide. The problem, from the standpoint of prediction, is that the predicted value closest to the experimental pIC_{50} is not always the highest predicted value, as is shown in Table 5. As the predicted pIC_{50} is not always the nearest to the experimental or observed pIC_{50}, it is difficult to differentiate errors arising from multiple binding modes, and so forth, from an incorrectly defined binding sequence, particularly at the early stages of the iterative cycle. Thus, our attempts to account for a possible multiplicity of binding modes, that is, two or more 9mer subsequences, have not yet yielded a stable solution or workable algorithm (data not shown).

Our results are consistent with an emerging view of MHC binding: motifs are an inadequate representation of the underlying processes of binding. As we have clearly demonstrated elsewhere,[85,86] the whole of a peptide contributes to binding, albeit weighted differently at different positions. At least for class I, it is even possible to generate high-affinity peptides using noncanonical anchors, with extra affinity arising from other interactions made by the rest of the peptides. This is also likely to be a feature of class II binding. For example, Liu et al.,[87] showed that, for I-A[b], it was possible for a peptide bearing alanines to bind to its four main pockets—which correspond to positions P1, P4, P6, and P9 and which usually bind larger peptide side chains—with compensatory interactions made by residues at other positions in order to maintain overall affinity. Our class II models suggest that the relative contributions, of particular residues,

to binding are spread more evenly through the peptide than is generally supposed, rather than being concentrated solely in so-called anchor positions. Our iterative method is different from the manual identification of anchor-based motifs by visual inspection. Such methods are intrinsically tendentious, arbitrary, subjective, and potentially inaccurate. Our method, which is, however, by no means perfect, is, by contrast, an objective and unsupervised approach. It is dependent, however, on the quantity and degeneracy of the data itself, and also upon its quality.

In terms of the quantity of data, we have not reported models where the number of peptides is less than 40,[7] as the size of data sets we have used is suitable for motif analysis and comparisons can provide significant insights (see Table 6). Our experience with other additive models, for example, HLA-A*0201, where sufficient data is available, suggests that predictivity is maintained.[8] In general, for MHC−peptide binding experiments, the sequences of peptides studied are very biased in terms of 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 burden of epitope discovery. Very sparse sequence patterns are matched and the corresponding subset of peptides tested, with an enormous reduction in sequence diversity. The intrinsic quality of data is a more vexing issue still. Peptides are often physically large molecules with extreme physical properties, being multiply charged, zwitterionic, or exhibiting huge ranges of hydrophobicity. Moreover, there is also little consensus on the best assay systems for measuring MHC−peptide affinity, and that data is of an intrinsically inferior quality: multiple measurements of the same peptide can vary by several orders of magnitude, mixtures of different standard peptides are used in radioligand assays, and experiments are conducted at different temperatures and over different concentration ranges.[86] Another issue of importance, which we do not address overtly, is the influence of "flanking" residues on affinity and recognition: Arnold et al.,[88] identified residues at +2 or −2, relative to the core nonamer, as important for effective recognition by T cells. We have sought to address this by increasing the core peptide region identified in our model by 2 in both directions, but again this did not, within an iterative framework, yield a stable solution, perhaps suggesting that this phenomenon is a subsidiary one, at least statistically.

As is well-known, the prediction of MHC class II epitopes is considerably more demanding, and generally less successful, than that of class I. Ideally, we would like to address

the validation of the models' performance with a large unbiased blind test set, but unfortunately, only relatively small, on an allele-by-allele basis, test sets are currently available, see Table 2. Moreover, we have already ruminated upon the variety of practical and conceptual difficulties as yet unaddressed for class II epitopes. All this is reflected in the performance of our models. In the present evaluation test, the predictivity of the additive method models varied between models. The relative success of the models is related, in part, to the number of peptides used to train them. The predictivity of the I-A^d model was very high (training set: 145 peptides), while that of I-E^k (training set: 52 peptides) was very poor. As more binding data for mouse class II becomes available, the quality of our models will improve.

The ISC algorithm described above combines an iterative approach to selecting the best predicted binders with PLS, a robust multivariate statistical tool for model generation. The ISC method is universal in that it can be used for any peptide–protein binding interaction where the peptide length is unrestricted but the binding is limited to a fixed, if unknown, part of the peptide. Implementation of the method is straightforward; it is fast to use, and its interpretation is straightforward. The final models derived from these calculations will be included in an updated version of MHCPred.[7,9,89–91]

REFERENCES AND NOTES

(1) Rudensky, A. Y.; Preston-Hurlburt, P.; Hong, S.-C.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional: application to peptide bound to class I MHC class II molecules. *Nature* **1991**, *353*, 622–627.

(2) Hunt, D. F.; Michel, H.; Dickinson, T. A.; Shabanowitz, J.; Cox, A. L.; Sakaguchi, K.; Appella, E. Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science.* **1992**, *256*, 1817–1820.

(3) Chicz, R. M.; Urban, R. G.; Lane, W. S.; Gorga, J. C.; Stern, L. J.; Vignali, D. A. A.; Strominger, J.L. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* **1992**, *358*, 764–768.

(4) Chicz, R. M.; Urban, R. G.; Gorga, J. C.; Vignali, D. A. A.; Lane, W. S.; Strominger, J. L. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J. Exp. Med.* **1993**, *178*, 27–47.

(5) Dessen, A.; Lawrence, C. M.; Cupo, S.; Zaller, D. M.; Wiley, D. C. X-ray crystal structure of HLA-DR4 (DRA*0101, DRB*0401) complexed with a peptide from human collagen II. *Immunity* **1997**, *7*, 473–481.

(6) Hennecke, J.; Wiley, D. C. Structure of a complex of the human α/β T cell receptor (TCR) HA1.7, influenza hemaglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J. Exp. Med.* **2002**, *195*, 571–581.

(7) Doytchinova, I. A.; Flower, D. R. Towards the in silico identification of class II restricted T cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics* **2003**, *19*, 1–8.

(8) Doytchinova, I. A.; Blythe, M. J.; Flower, D. R. Additive method for the prediction of protein–peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.* **2002**, *1*, 263–272.

(9) Guan, P.; Doytchinova, I. A.; Flower, D. R. HLA-A3 supermotif defined by quantitative structure–activity relationship analysis. *Protein Eng.* **2003**, *16*, 11–18.

(10) Blythe, M.; Doytchiniva, I. A.; Flower, D. R. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* **2002**, *18*, 434–439.

(11) 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.

(12) 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.

(13) 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.

(14) Wold, S. PLS for multivariate linear modelling. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195–218.

(15) *SYBYL*, version 6.9; Tripos Inc.: St. Louis, MO.

(16) Brett, S. J.; Cease, K. B.; Ouyang, C. S.; Berzofsky, J. A. Fine specificity of T cell recognition of the same peptide in association with different I-A molecules. *J. Immunol.* **1989***, 143*, 771–779.

(17) Berkower, I.; Buckenmeyer, G. K.; Berzofsky, J. A. Molecular mapping of a histocompatibility-restricted immunodominant T cell epitope with synthetic and natural peptides: implications for T cell antigenic structure. *J. Immunol.* **1986**, *136*, 2498–503.

(18) Nalefski, E. A.; Rao, A. Nature of the ligand recognized by a hapten- and carrier-specific, MHC-restricted T cell receptor. *J. Immunol.* **1993**, *150*, 3806–3816.

(19) Ffrench, R. A.; Tang, X. L.; Anders, E. M.; Jackson, D. C.; White, D. O.; Drummer, H.; Wade, J. D.; Tregear, G. W.; Brown, L. E. Class II-restricted T-cell clones to a synthetic peptide of influenza virus hemagglutinin differ in their fine specificities and in the ability to respond to virus. *J. Virol.* **1989**, *63*, 3087–3094.

(20) Robertson, J. M.; Jensen, P. E.; Evavold, B. D. DO11.10 and OT-II T cells recognize a C-terminal ovalbumin 323–339 epitope. *J. Immunol.* **2000**, *164*, 4706–4712.

(21) Delvig, A. A.; Rosenqvist, E.; Oftung, F.; Robinson, J. H. T-Cell epitope mapping the PorB protein of serogroup B Neisseria meningitidis in B10 congenic strains of mice. *Clin. Immunol. Immunopathol.* **1997**, *85*, 134–142.

(22) Wen, R.; Surman, S.; Blackman, M. A.; Woodland, D. L. The conventional CD4+ T cell response to staphylococcal enterotoxin B is modified by its superantigenic activity. *Cell Immunol.* **1997**, *176*, 166–172.

(23) Targoni, O. S.; Lehmann, P. V. Endogenous myelin basic protein inactivates the high avidity T cell repertoire. *J. Exp. Med.* **1998**, *187*, 2055–2063.

(24) Decker, P.; Le Moal, A.; Briand, J. P.; Muller, S. Identification of a minimal T cell epitope recognized by antinucleosome Th cells in the C-terminal region of histone H4. *J. Immunol.* **2000**, *165*, 654–662.

(25) Nagabhushanam, V.; Purcell, A. W.; Mannering, S.; Germano, S.; Praszkier, J.; Cheers, C. Identification of an I-Ad restricted peptide on the 65-kilodalton heat shock protein of Mycobacterium avium. *Immunol. Cell Biol.* **2002**, *80*, 574–583.

(26) Williams, K. M.; Bigley, E. C, III. Identification of an I-Ed-restricted T-cell epitope of *Escherichia coli* outer membrane protein. *Infect. Immun.* **2004**, *72*, 3907–3913.

(27) Gugasyan, R.; Velazquez, C.; Vidavsky, I.; Deck, B. M.; van der Drift, K.; Gross, M. L.; Unanue, E. R. Independent selection by I-Ak molecules of two epitopes found in tandem in an extended polypeptide antigen. *J. Immunol.* **2000**, *165*, 3206–3213.

(28) van Stipdonk, M. J.; Willems, A. A.; Amor, S.; Persoon-Deen, C.; Travers, P. J.; Boog, C. J.; van Noort, J. M. T cells discriminate between differentially phosphorylated forms of alphaB-crystallin, a major central nervous system myelin antigen. *Int. Immunol.* **1998**, *10*, 943–950.

(29) L: Luescher, I. F.; Allen, P. M.; Unanue, E. R. Binding of photoreactive lysozyme peptides to murine histocompatibility class II molecules. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 871–874.

(30) Mills, K. H.; Burt, D. S.; Skehel, J. J.; Thomas, D. B. Fine specificity of murine class II-restricted T cell clones for synthetic peptides of influenza virus hemagglutinin. Heterogeneity of antigen interaction with the T cell and the Ia molecule. *J. Immunol.* **1988**, *140*, 4083–4090.

(31) Adorini, L.; Sette, A.; Buus, S.; Grey, H. M.; Darsley, M.; Lehmann, P. V.; Doria, G.; Nagy, Z. A.; Appella, E. Interaction of an immunodominant epitope with Ia molecules in T-cell activation. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5181–5185.

PREDICTING MOUSE CLASS II MHC BINDING AFFINITY

J. Chem. Inf. Model., Vol. 46, No. 3, 2006 **1501**

(32) Latek, R. R.; Petzold, S. J.; Unanue, E. R. Hindering auxiliary anchors are potent modulators of peptide binding and selection by I-Ak class II molecules. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11460−11465.

(33) Nelson, C. A.; Roof, R. W.; McCourt, D. W.; Unanue, E. R. Identification of the naturally processed form of hen egg white lysozyme bound to the murine major histocompatibility complex class II molecule I-Ak. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7380−7383.

(34) Maecker, H. T.; Umetsu, D. T.; DeKruyff, R. H.; Levy, S. Cytotoxic T cell responses to DNA vaccination: dependence on antigen presentation via class II MHC. *J. Immunol.* **1998**, *161*, 6532−6536.

(35) Hernandez, H. J.; Stadecker, M. J. Elucidation and role of critical residues of immunodominant peptide associated with T cell-mediated parasitic disease. *J. Immunol.* **1999**, *163*, 3877−3882.

(36) Chen, J. S.; Lorenz, R. G.; Goldberg, J.; Allen, P. M. Identification and characterization of a T cell-inducing epitope of bovine ribonuclease that can be restricted by multiple class II molecules. *J. Immunol.* **1991**, *147*, 3672−3678.

(37) Wall, K. A.; Hu, J.-Y.; Currier, P.; Southwood, S.; Sette, A.; Infante, A. J. A disease-related epitope of Torpedo acetylcholine receptor. Residues involved in I-Ab binding, self−nonself discrimination, and TCR antagonism. *J. Immunol.* **1994**, *152*, 4526−4536.

(38) Svensson, M.; Stockinger, B.; Wick, M. J. Bone marrow-derived dendritic cells can process bacteria for MHC-I and MHC-II presentation to T cells. *J. Immunol.* **1997**, *158*, 4229−4236.

(39) Vacchio, M. S.; Berzofsky, J. A.; Krzych, U.; Smith, J. A.; Hodes, R. J.; Finnegan, A. Sequences outside a minimal immunodominant site exert negative effects on recognition by staphylococcal nuclease-specific T cell clones. *J. Immunol.* **1989**, *143*, 2814−2819.

(40) Hsu, B. L.; Donermeyer, D. L.; Allen, P. M. TCR recognition of the Hb(64−76)/I-Ek determinant: single conservative amino acid changes in the complementarity-determining region 3 dramatically alter antigen fine specificity. *J. Immunol.* **1996**, *157*, 2291−2298.

(41) Parra-Lopez, C. A.; Lindner, R.; Vidavsky, I.; Gross, M.; Unanue, E. R. Presentation on class II MHC molecules of endogenous lysozyme targeted to the endocytic pathway. *J. Immunol.* **1997**, *158*, 2670−2679.

(42) Nikcevich, K. M.; Kopielski, D.; Finnegan, A. Interference with the binding of a naturally processed peptide to class II alters the immunodominance of T cell epitopes in vivo. *J. Immunol.* **1994**, *153*, 1015−1026.

(43) Chianese-Bullock, K. A.; Russell, H. I.; Moller, C.; Gerhard, W.; Monaco, J. J.; Eisenlohr, L. C. Antigen processing of two H2-IEd-restricted epitopes is differentially influenced by the structural changes in a viral glycoprotein. *J. Immunol.* **1998**, *161*, 1599−1607.

(44) Kang, H. K.; Mikszta, J. A.; Deng, H.; Sercarz, E. E.; Jensen, P. E.; Kim, B. S. Processing and reactivity of T cell epitopes containing two cysteine residues from hen egg-white lysozyme (HEL74-90). *J. Immunol.* **2000**, *164*, 1775−1782.

(45) Lovitch, S. B.; Petzold, S. J.; Unanue, E. R. Cutting edge: H-2DM is responsible for the large differences in presentation among peptides selected by I-Ak during antigen processing. *J. Immunol.* **2003**, *171*, 2183-2186.

(46) Viner, N. J.; Nelson, C. A.; Unanue, E. R. Identification of a major I-Ek-restricted determinant of hen egg lysozyme: limitations of lymph node proliferation studies in defining immunodominance and crypticity. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2214−2218.

(47) Boots, A. M.; Kusters, J. G.; van Noort, J. M.; Zwaagstra, K. A.; Rijke, E.; van der Zeijst, B. A.; Hensen, E. J. Localization of a T-cell epitope within the nucleocapsid protein of avian coronavirus. *Immunology* **1991**, *74*, 8−13.

(48) Nelson, C. A.; Viner, N.; Young, S.; Petzold, S.; Benoist, C.; Mathis, D.; Unanue, E. R. Amino acid residues on the I-Ak alpha-chain required for the binding and stability of two antigenic peptides. *J. Immunol.* **1996**, *156*, 176−182.

(49) Leighton, J.; Sette, A.; Sidney, J.; Appella, E.; Ehrhardt, C.; Fuchs, S.; Adorini, L. Comparison of structural requirements for interaction of the same peptide with I-Ek and I-Ed molecules in the activation of MHC class II-restricted T cells. *J. Immunol.* **1991**, *147*, 198−204.

(50) Shi, Y.; Kaliyaperumal, A.; Lu, L.; Southwood, S.; Sette, A.; Michaels, M. A.; Datta, S. K. Promiscuous presentation and recognition of nucleosomal autoepitopes in lupus: role of autoimmune T cell receptor α chain. *J. Exp. Med.* **1998**, *187*, 367−378.

(51) Bhayani, H.; Carbone, F. R.; Paterson, Y. The activation of pigeon cytochrome c-specific T cell hybridomas by antigenic peptides is influenced by non-native sequences at the amino terminus of the determinant. *J. Immunol.* **1988**, *141*, 377−382.

(52) Diment, S. Different roles for thiol and aspartyl proteases in antigen presentation of ovalbumin. *J. Immunol.* **1990**, *145*, 417−422.

(53) Guery, J. C.; Adorini, L. Dendritic cells are the most efficient in presenting endogenous naturally processed self-epitopes to class II-restricted T cells. *J. Immunol.* **1995**, *154*, 536−544.

(54) London, C. A.; Perez, V. L.; Abbas, A. K. Functional characteristics and survival requirements of memory CD4+ T lymphocytes in vivo. *J. Immunol.* **1999**, *162*, 766−773.

(55) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS−PROT protein knowledge-base and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365−370.

(56) Reche, P. A.; Glutting, J. P.; Reinherz, E. L. Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* **2002**, *63*, 701−709.

(57) Reche, P. A.; Glutting, J. P.; Zhang, H.; Reinherz, E. L. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics* **2004**, *56*, 405−419.

(58) Naujokas, M. F.; Southwood, S.; Mathies, S. J.; Appella, E.; Sette, A.; Miller, J. T cell recognition of flanking residues of murine invariant chain-derived CLIP peptide bound to MHC class II. *Cell. Immunol.* **1998**, *188*, 49−54.

(59) Sant'Angelo, D. B.; Robinson, E.; Janeway, C. A., Jr.; Denzin, L. K. Recognition of core and flanking amino acids of MHC class II bound peptides by the T cell receptor. *Eur. J. Immunol.* **2002**, *32*, 2510−2521.

(60) Alexander, J.; Sidney, J.; Southwood, S.; Ruppert, J.; Oseroff, C.; Maewal, A.; Snoke, K.; Serra, H. M.; Kubo, R. T.; Sette, A.; Grey, H. M. Development of high potency universal DR-restricted helper epitopes by modification of high affinity DR-blocking peptides. *Immunity* **1994**, *1*, 751−761.

(61) Pfeiffer, C.; Stein, J.; Southwood, S.; Ketelaar, H.; Sette, A.; Bottomly, K. Altered peptide ligands can control CD4 T lymphocyte differentiation in vivo. *J. Exp. Med.* **1995**, *181*, 1569−1574.

(62) Sette, A.; Southwood, S.; Miller, J.; Appella, E. Binding of major histocompatibility complex class II to the invariant chain-derived peptide, CLIP, is regulated by allelic polymorphism in class II. *J. Exp. Med.* **1995**, *181*, 677−683.

(63) Livingston, B.; Crimi, C.; Newman, M.; Higashimoto, Y.; Appella, E.; Sidney, J.; Sette, A. A rational strategy to design multiepitope immunogens based on multiple Th lymphocyte epitopes. *J. Immunol.* **2002**, *168*, 5499−5506.

(64) Tsunoda, I.; Sette, A.; Fujinami, R. S.; Oseroff, C.; Ruppert, J.; Dahlberg, C.; Southwood, S.; Arrhenius, T.; Kuang, L.-Q.; Kubo, R. T.; Chesnut, R. W.; Ishioka, G. Y. Lipopeptide particles as the immunologically active component of CTL inducing vaccines. *Vaccine* **1999**, *17*, 675−685.

(65) Gregori, S.; Trembleau, S.; Penna, G.; Gallazzi, F.; Hammer, J.; Papadopoulos, G. K.; Adorini, L. A peptide binding motif for I-E^{g7}, the MHC class II molecule that protects Eα-transgenic nonobese diabetic mice from autoimmune diabetes. *J. Immunol.* **1999**, *162*, 6630−6640.

(66) Hill, J. A.; Wang, D.; Jevnikar, A. M.; Cairns, E.; Bell, D. A. The relationship between predicted peptide-MHC class II affinity and T cell activation in a HLA-DRβ1*0401 transgenic mouse model. *Arthritis Res. Ther.* **2002**, *5*, R40−R48.

(67) Cucca, F.; Lampis, R.; Congia, M.; Angius, E.; Nutland, S.; Bain, S. C.; Barnett, A. H.; Todd, J. A. A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.* **2001**, *10*, 2025−2037.

(68) Brown, J. H.; Jardetzky, T. S.; Stern, L. J.; Gorga, J. C.; Strominger, J. L.; Wiley, D. C. Human class II MHC molecule HLA-DR1: X-ray structure determined from three crystal forms. *Acta Crystallogr., Sect D* **1995**, *D51*, 946−961.

(69) Rammesee, H.-G.; Bachmann, J.; Emmerich, N. P. N.; Bachor, O. A.; Stevanovic, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **1999**, *50*, 213−219.

(70) Rammesee, H.-G.; Friede, T.; Stevanovic, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **1995**, *41*, 178−228.

(71) Nelson, C. A.; Viner, N. J.; Young, S. P.; Petzold, S. J.; Unanue, E. R. A negatively charged anchor residue promotes high affinity binding to the MHC class II molecule I-A^k. *J. Immunol.* **1996**, *157*, 755−762.

(72) Schild, H.; Gruneberg, U.; Pougialis, G.; Wallny, H. J.; Keilholz, W.; Stevanovic, S.; Rammensee, H. G. Natural ligand motifs of H-2E molecules are allele specific and illustrates homology to HLA-DR molecules. *Int. Immunol.* **1995**, *7*, 1957−1965.

(73) Reay, P. A.; Kantor, R. M.; Davis, M. M. Use of global amino acid replacement to define the requirements for MHC binding and T cell recognition of moth cytochrome c (93−103). *J. Immunol.* **1994**, *152*, 3946−3957.

(74) Marrack, P.; Ignatowicz, L.; Kappler, J. W.; Boymel, J.; Freed, J. H. Comparison of peptides bound to spleen and thymus class II. *J. Exp. Med.* **1993**, *178*, 2173−2183.

(75) Flower, D. R.; Doytchinova, I. A.; Paine, K.; Taylor, P.; Lamponi, D.; Zygouri, C.; Guan, P.; McSparron, H.; Kirkbride, H. *Computational vaccine design. Drug Design: Cutting Edge Approaches*; Royal Society of Chemistry: Cambridge, U. K., 2002; p 136.

(76) Fremont, D. H.; Monnaie, D.; Nelson, C. A.; Hendrickson, W. A.; Unanue, E. R. Crystal structure of I-Ak in complex with a dominant epitope of lysozyme. *Immunity* **1998**, *8*, 305−317.

(77) Corper, A. L.; Stratmann, T.; Apostolopoulos, V.; Scott, C. A.; Garcia, K. C.; Kang, A. S.; Wilson, I. A.; Teyton, L. A structural framework for deciphering the link between I-A^{g7} and autoimmune diabetes. *Science* **2000**, *288*, 505−511.

(78) Li, Y.; Li, H.; Martin, R.; Mariuzza, R. A. Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR$_2$ proteins. *J. Mol. Biol.* **2000**, *304*, 177−188.

(79) McFarland, B. J.; Sant, A. J.; Lybrand, T. P.; Beeson, C. Ovalbumin (323−339) peptide bind to the major histocompatibility complex class II, I-A(d) protein using two functionally distinct registers. *Biochemistry* **1999**, *38*, 16663−16670.

(80) Vidal, K.; Daniel, C.; Vidavsky, I.; Nelson, C. A.; Allen, P. M. Hb (64−76) epitope binds in different registers and lengths to I-Ek and I-Ak. *Mol. Immunol.* **2000**, *37*, 203−212.

(81) Bankovich, A. J.; Girvin, A. T.; Moesta, A. K.; Garcia, K. C. Peptide register shifting within the MHC groove: theory becomes reality. *Mol. Immunol.* **2004**, *40*, 1033−1039.

(82) Xia, J.; Sollid, L. M.; Khosla, C. Equilibrium and Kinetic Analysis of the Unusual Binding Behavior of a Highly Immunogenic Gluten Peptide to HLA-DQ2. *Biochemistry* **2005**, *44*, 4442−4449.

(83) Hecht, C. E. *Statistical Thermodynamics and Kinetic Theory*; WH Freeman: New York, 1990.

(84) He, X. L.; Radu, C.; Sidney, J.; Sette, A.; Ward, E. S.; Garcia, K. C. Structural snapshot of aberrant antigen presentation linked to autoim-munity: the immunodominant epitope of MBP complexed with I-Au. *Immunity* **2002**, *17*, 83−94.

(85) Doytchinova, I. A.; Walshe, V.; Jones, N.; Gloster, S.; Borrow, P.; Flower, D. R. Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunol.* **2004**, *172*, 7495−7502.

(86) Flower, D. R. Towards in silico prediction of immunogenic epitopes. *Trends Immunol.* **2003**, *24*, 667−674.

(87) Liu, X.; Dai, S.; Crawford, F.; Fruge, R.; Marrack, P.; Kappler, J. Alternate interactions define the binding of peptides to the MHC molecule IA^b. *PNAS* **2002**, *99*, 8820−8825.

(88) Arnold, P. Y.; La Gruta, N. L.; Miller, T.; Vignali, K. M.; Adams, P. S.; Woodland, D. L.; Vignali, D. A. The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. *J. Immunol.* **2002**, *169*, 739−749.

(89) Guan, P.; Doytchinova, I. A.; Zygouri, C.; Flower, D. R. MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl. Bioinformatics* **2003**, *2*, 63−66.

(90) Guan, P.; Doytchinova, I. A.; Zygouri, C.; Flower, D. R. MHCPred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res.* **2003**, *31*, 3621−3624.

(91) Hattotuwagama, C. K.; Guan, P.; Doytchinova, I. A.; Zygouri, C.; Flower, D. R. Quantitative online prediction of peptide binding to the major histocompatibility complex. *J. Mol. Grapics Modell.* **2004**, *22*, 195−207.