



Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules

Björn Peters¹, Weiwei Tong², John Sidney³, Alessandro Sette³ and Zhiping Weng^{2,4,*}

¹Institut für Biochemie, Charite, Humboldt Universität Berlin, Monbijoustr. 2, 10117 Berlin, Germany, ²Department of Biomedical Engineering, 44 Cummington Street, Boston University, Boston, MA 02215, USA, ³La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, La Jolla, CA 92121, USA and ⁴Bioinformatics Program, 48 Cummington Street, Boston University, Boston, MA 02215, USA

Received on October 22, 2002; revised on February 13, 2003; accepted on April 25, 2003

ABSTRACT

Motivation: Various methods have been proposed to predict the binding affinities of peptides to Major Histocompatibility Complex class I (MHC-I) molecules based on experimental binding data. They can be classified into two groups: (1) *AIB methods* that assume independent contributions of all peptide positions to the binding to MHC-I molecule (e.g. scoring matrices) and (2) *general methods* which can take into account interactions between different positions (e.g. artificial neural networks). We aim to compare the prediction accuracies of these methods, and quantify the impact of interactions between peptide positions.

Results: We compared several previously published and widely used methods and discovered that the best AIB methods gave significantly better predictions than three previously published general methods, possibly due to the lack of a sufficient training data for the general methods. The best results, however, were achieved with our newly developed general method, which combined a matrix describing independent binding with pair coefficients describing pair-wise interactions between peptide positions. The pair coefficients consistently but only slightly improved prediction accuracy, and were much smaller than the matrix entries. This explains why neglecting them—as is done in AIB methods—can still lead to good predictions.

Availability: The new prediction model is implemented at <http://zlab.bu.edu/SMM>. The underlying matrix and pair coefficients are also available as supplementary materials.

Contact: zhiping@bu.edu

INTRODUCTION

Major Histocompatibility Complex class I (MHC-I) molecules present endogenous antigenic peptides on a cell surface. This is a prerequisite for stimulating cytotoxic T cell response—a crucial mechanism against viral infections and certain tumors [for a recent review see (Shastri *et al.*, 2002)]. Thus, the ability to identify the peptides that can bind to MHC molecules is of practical immunological importance. Furthermore, studies of MHC–peptide binding can broadly impact our predictive understanding of molecular interaction.

Various prediction methods have been applied to MHC–peptide binding. One class of approaches is based on structural data and binding free energy calculations (Vajda *et al.*, 1994; Zhang *et al.*, 1997; Schueler-Furman *et al.*, 2000). Typically such methods are limited to MHC molecules whose crystal structures have been determined, and are too computationally intensive to be applied to a large number of peptides. In this paper, we limit ourselves to methods based on experimental binding data. These methods try to identify the functional relationship between a peptide sequence and its binding affinity to a particular MHC molecule. It might be optimistic to assume that such a relationship is discernable, since each amino acid encapsulates a multitude of physical properties such as charge, hydrophobicity, size, hydrogen bonding potential etc., which can contribute to binding in a number of different ways, while only the overall affinity is measured experimentally. Studies have shown that simple predictions using scoring matrices yield reasonably good results (Sette *et al.*, 1989; Ruppert *et al.*, 1993; Hammer *et al.*, 1994; Parker *et al.*, 1994; Rammensee, 1995; Schönbach *et al.*, 1995; Brusica *et al.*, 1997; Rammensee *et al.*, 1999; Borrás-Cuesta *et al.*, 2000), even with little experimental data (typically hundreds of peptides) compared to the size of the sequence space (20^9 for 9-mers). This indicates that the relationship between sequence and affinity can be roughly

*To whom correspondence should be addressed.

approximated by the independent binding assumption, i.e. amino acids at different positions of a peptide contribute independently to the overall binding affinity of the peptide. This assumption is further supported by the extended peptide conformation in the MHC groove, the homogeneity of peptide sizes and the good alignment of end positions of all peptides bound to the same MHC.

In this paper, we develop a new matrix-based algorithm called the Stabilized Matrix Method (SMM). It differs from previous approaches by taking into special consideration the errors inevitably contained in experimental data. We applied SMM to a set of 9-mer peptides bound to the HLA-A2 MHC molecule. We compare the performance of SMM with three widely used matrix-based *AIB methods* that assume independent binding: BIMAS (Parker *et al.*, 1994), SYFPEITHI (Rammensee *et al.*, 1999) and the polynomial method (PM) (Gulukota *et al.*, 1997). SMM considerably outperforms the other methods on three independent test sets.

In order to investigate whether including interactions between peptide positions can improve prediction accuracy, we compare SMM to three *general methods* that do not assume independent binding: an artificial neural network (ANN) (Gulukota *et al.*, 1997), a classification tree (CART) (Segal *et al.*, 2001) and the additive method (Doytchinova *et al.*, 2002). Surprisingly, SMM substantially outperforms all of these methods, indicating that there is not enough data to reliably train the general methods.

We then set out to investigate ways of incorporating interactions between pairs of peptide positions into SMM, without compromising the already achieved prediction accuracy. Our strategy was to start with the SMM predictions based on the independent binding assumption and then derive pair interactions from the systematic difference between thus predicted binding affinities and the experimental measurements. We limited ourselves to those pair interactions that we could reliably determine from the training data. The combined SMM + pair coefficient predictions led to consistent improvements over the plain matrix-based SMM.

The pair coefficients quantify the impact of interactions between peptide positions for peptide–MHC binding. We analyze how this impact depends on the distance of the involved positions in a peptide.

ALGORITHM

Stabilized matrix method

A scoring matrix quantifies residue contributions to binding. The matrix element $s_{a,i}$ corresponds to amino acid a at position i of the peptide. The total score S_k for a given peptide k with the amino acids $a_k(i)$ at positions i is then given by the summation:

$$S_k = s_0 + \sum_i s_{a_k(i),i} \quad (1)$$

where s_0 is a constant offset. For the SMM method, the values for $s_{a,i}$ and s_0 are determined by minimizing the distance between predicted scores S_k and the affinities measured for the peptides in the training set (described below):

$$\Phi(\{s_{a,i}\}) = \sum_k \|S_k - \text{measured}_k\| \quad (2)$$

For a peptide with a measurable IC50 value, the norm in Equation (2) has the form:

$$\|S_k - \text{measured}_k\| = (S_k - \ln(\text{IC50}_k))^2 \quad (3)$$

Some peptides in the training set have too low affinities for their IC50 to be measurable. We set their IC50 values equal to or greater than the largest experimentally measurable value of $\ln(50\,000\text{ nM})$ (=bound). Accordingly, for these ‘heavy non-binders’, the norm in Equation (2) is:

$$\|S_k - \text{measured}_k\| = \begin{cases} 0 & \text{if } S_k > \text{bound} \\ (\text{bound} - S_k)^2 & \text{if } S_k < \text{bound} \end{cases} \quad (4)$$

To avoid over-fitting, a second term is added to the minimization function in (2):

$$\Psi(\{s_{a,i}\}, \lambda) = \Phi(\{s_{a,i}\}) + \lambda \sum_{a,i} s_{a,i}^2 \quad (5)$$

By minimizing Equation (5) with a non-zero λ value, a tradeoff is introduced between optimally reproducing the experimental values (including their inevitable experimental error) and minimizing parameters $s_{a,i}$. This forces all $s_{a,i}$ that do not significantly lower the distance Φ towards zero. The optimal value for λ , $\lambda_{\text{opt}} = 1$, was determined by minimizing the distance defined by Equation (2), via a 10-fold cross-validation on the training set.

A similar mathematical concept is used to solve ‘inverse problems’, where λ is called the regularization parameter. A short introduction to inverse problems is given in (Press *et al.*, 1992), chapter 18. To minimize Equations (5), we applied a commercial non-linear optimizer (Frontline Systems, 1999) using a generalized-reduced-gradient method.

Pair coefficients

We introduce pair coefficients $s'_{a,i,a',i'}$ to quantify the impact of interactions between amino acids a at position i and amino acid a' at positions i' on binding. For example, coefficient $s'_{A3,L7}$ (**A***L**) describes the difference in binding between the following two scenarios: (1) an Ala is at the 3rd AND a Leu at the 7th position of the peptide and (2) the sum of the average contributions of having an Ala at the 3rd position, described by matrix value s_{A3} (**A*****), and having a Leu at the 7th position, described by s_{L7} (*****L**). The total

score for a given peptide k is then given by

$$S'_k = S_k + \sum_i \sum_{i'} s'_{a_k(i),i,a_k(i'),i'} \quad (6)$$

with S_k being the matrix score defined in Equation (1), and the sum including all pairs of amino acids found in the peptide. For 9-mer peptides, this would result in $20 * 20 * 36 = 14\,400$ different pair coefficients. Since it is impossible to determine so many coefficients given limited data, we go through a two-stage selection process: in the first step we eliminate all coefficients for which fewer than $N_{\min} = 10$ peptides exist in the training set, which indicates the lack of sufficient experimental information. The optimal values for the remaining 269 pair coefficients can then be calculated by minimizing

$$\Psi'(\{s'_{a,i,a',i'}\}, \lambda') = \Phi'(\{s'_{a,i,a',i'}\}) + \lambda' \sum_{a,i,a',i'} s_{a,i,a',i'}^2 \quad (7)$$

where Φ' is the same as in Equation (2), except that scores S'_k as defined in Equation (6) are used instead of S_k . When optimizing pair coefficients $s_{a,i,a',i'}$ in Equation (7), the matrix coefficients $s_{a,i}$ are frozen at their optimal value determined from Equation (5).

In the second selection step, we randomly split the training set into 10 equal-size non-overlapping subsets and determine 10 different optimal values for each pair coefficient, leaving out one subset at a time. If a pair coefficient contains both positive and negative optimal values, it cannot be estimated reliably using our training set. Therefore, we discard these pair coefficients. Setting all discarded coefficients to zero, we minimize Equation (7) to determine the values of the remaining 124 pair coefficients. The optimal value for λ' is again determined using cross-validation, similar to λ in Equation (5). The cross-validated distance as calculated using Equation (6) for various values of λ' is shown in Figure 1. For very high values of λ' , the pair coefficients are forced to stay around zero, and the prediction accuracy is equal to that of the plain matrix approach. For $\lambda' = 0$, there is no restriction on the value of pair coefficients, leading to over-fitting, and the resulting performance is much worse than the plain matrix. In between, there is a maximum in distance improvement at the optimal value $\lambda'_{\text{opt}} = 5$.

SYSTEM AND METHODS

Experimental binding data

Four non-overlapping sets of 9-mer peptides binding to the HLA-A2 allele are used in this paper. All data sets are separated into binders and non-binders. The cutoff for making this separation is not critical for our measure of prediction accuracy (see below) as long as ‘binders’ show higher affinities than ‘non-binders’. All sets have similar sequence composition: at the anchor positions 2 and 9, amino acids A, I, L, M, T and V are over-represented; at other positions, all 20 amino acids

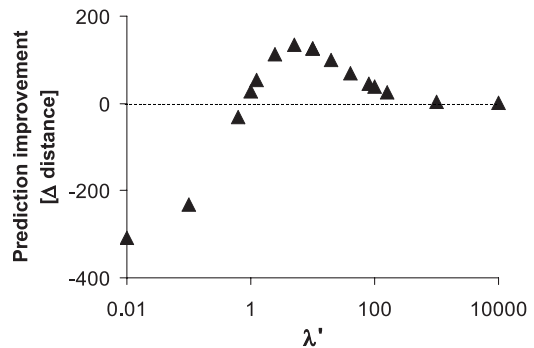


Fig. 1. Using cross-validation to determine optimal λ' . The cross-validated improvement in distance between SMM + pair coefficient predictions and experimental values over that of the plain SMM on the Training-set are plotted against varying values of λ' . The optimal value $\lambda'_{\text{opt}} = 5$ corresponds to the peak in the figure.

are roughly equally represented. For the SYFPEITHI-set, this reflects the distributions of amino acids at peptide positions in naturally occurring epitopes. For the other three *in vitro* data sets, experimentalist chose similarly biased peptides in order to include a large proportion of binders. Although it would be better from a mathematical point of view to train and test on randomly selected peptides, this would require far more measurements as only 1 in about 200 peptides bind (Udaka *et al.*, 2000). Since the described bias is present in all training and test sets, no particular prediction method is favored.

- **Training-set:** Using an IC50 cutoff of 500 nM, we obtain 127 binders and 406 non-binders. This data set has been used in several previous studies (Ruppert *et al.*, 1993; Gulukota *et al.*, 1997; Doytchinova *et al.*, 2002). The 500 nM cutoff lies within the intermediate-affinity range, and it was discovered that majority of known HLA-restricted T-cell epitopes binding with $\text{IC}_{50} < 500$ nM (Ruppert *et al.*, 1993).
- **Blind-set:** Using an IC50 cutoff of 500 nM, we obtain 67 binders and 108 non-binders. This data set has not been published before. It includes 9-mer peptides from viral and human proteins, with IC50 to HLA-A2 measured using the same competition assay as for the peptides in the Training-set.
- **BIMAS-set:** We extracted all peptides in (Parker *et al.*, 1994), for which $\beta 2$ microglobulin dissociation half-lives have been measured. Four of the peptides overlap with the Training-set. Assuming a linear correlation between half-life and IC50 values, a half-life cutoff of 653 min corresponds to the $\text{IC}_{50} = 500$ nM cutoff in the Training-set. Excluding the four overlapping peptides, the BIMAS-set has 25 binders and 105 non-binders.

- *SYFPEITHI-set*: We obtain all 143 known 9-mer epitopes presented by HLA-A2 from the SYFPEITHI database with unambiguous sequences as binders. For non-binders, we include those 59 heavy non-binders from the Blind-set for which IC50 values are too large to be measured.

Matrix-based AIB methods

BIMAS: This is a widely used method based on experimentally measured β 2-microglobulin dissociation half-lives (Parker *et al.*, 1994). Its training data included the BIMAS-set. We obtained half-life predictions using the BIMAS web server (http://bimas.dcrt.nih.gov/molbio/hla_bind/).

SYFPEITHI: Its matrix scores reflect the abundance of amino acids in natural MHC ligands, T-cell epitopes, or binding peptides (Rammensee *et al.*, 1999). Its training data included the SYFPEITHI-set. We obtained SYFPEITHI predictions from the SYFPEITHI web server (<http://www.uni-tuebingen.de/uni/kxi>).

Polynomial method (PM): The contribution of an amino acid type at a particular peptide position is calculated as the mean of the $\ln(\text{IC}_{50})$ values of all peptides in the training set that have the amino acid type at that position (Gulukota *et al.*, 1997). We computed the PM matrix using the Training-set.

General methods

Classification and regression trees (CART): Classification trees were described in detail in a textbook (Breiman *et al.*, 1984), and were first used by Segal *et al.* to predict MHC-peptide binding (Segal *et al.*, 2001). Briefly, a classification tree is built by introducing splits in a set of peptides according to what amino acid is at a certain position of a peptide. The splitting is repeated, leading to a tree shaped classification scheme. Each split is chosen so that it maximizes the homogeneity of the peptides in both daughter nodes. A perfectly homogenous node contains only binders or only non-binders. A very large tree is first built so that all nodes are perfectly homogenous. It is then pruned back to an optimal size determined by cross-validation. Each terminal node in the optimal tree is assigned a binding score, computed as the percentage of non-binders the node contains. We used two commercial software packages (SPSS and CART), leading to identical trees. Switching from the classification trees described in this paper, which use binary experimental binding data, to regression trees, which use quantitative IC50 values, improves the prediction performance only slightly.

ANN: Following the work in (Gulukota *et al.*, 1997), we used a feed-forward neural network with three layers: an input layer with 180 neurons, a hidden layer with 50 neurons and an output layer with one neuron. The Aspirin/MIGRAINES software package from the MITRE

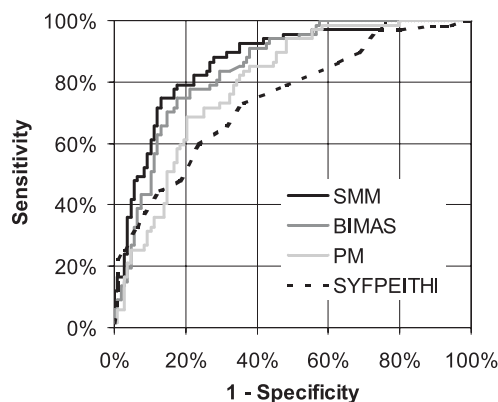


Fig. 2. ROC curves for matrix-based methods on the Blind-set. For each method, the cutoff was varied from the lowest to the highest predicted value of peptides in the Blind-set. For each cutoff value, the sensitivity and specificity were calculated, and plotted in the graph.

Corporation (<http://www.emsl.pnl.gov:2080>) was used to simulate the network.

Additive method: This method consists of a scoring matrix + pair coefficients, similar to our SMM + pair coefficients approach. It has been recently introduced by (Doytchinova *et al.*, 2002). They performed a partial least square fit of a linear equation to a set of 340 measured $\ln(\text{IC}_{50})$ values, a subset of our Training-set. The linear equation contains 1815 coefficients, some of which correspond to the entries of a scoring matrix, while the others are pair coefficients describing interactions between the immediate and next neighbor positions. We used the coefficient values determined by Doytchinova *et al.*

Measuring prediction accuracy

We need to compare the prediction accuracies of diverse methods on equally diverse data sets. To ensure a fair comparison, we use ROC curves (Bradley, 1997) to measure prediction accuracy. For a given cutoff value, which separates peptides by their predicted scores into potential binders and non-binders, two variables are calculated: *sensitivity* (true positives/total positives) and *1-specificity* (false positives/total negatives = false alarm rate). By systematically varying the cutoff from the lowest to the highest predicted score, a ROC curve such as Figure 2 is generated. Prediction accuracy is measured by the Area Under the Curve (AUC), which is 0.5 for random predictions and 1.0 for perfect predictions. The AUC is equivalent to the probability that the score of a randomly chosen binder is higher than that of a randomly chosen non-binder. This measure has the advantage of not relying on a single arbitrarily chosen cutoff for the prediction score, and can be equally applied to diverse data sets and prediction methods.

Table 1. Comparison of various methods on HLA-A2

Prediction method	Independent binding assumption	AUC on test set		
		Blind	SYFPEITHI	BIMAS
SMM, $\lambda = 1$	Yes	0.869	0.848	0.866
SMM, $\lambda = 0$	Yes	0.856	0.846	0.865
BIMAS	Yes	0.846	0.829	(0.875)
PM	Yes	0.795	0.792	0.757
SYFPEITHI	Yes	0.745	(0.865)	0.754
SMM + pair coef.	No	0.873	0.852	0.869
ANN	No	0.796	0.788	0.762
Additive method	No	0.820	0.770	0.830
CART	No	0.708	0.620	0.539

RESULTS

Comparison of AIB methods: SMM, PM, BIMAS and SYFPEITHI

Among the four matrix-based methods, PM and SMM were trained on the Training-set. BIMAS and SYFPEITHI were trained on data sets which included the respectively named test sets described above. Figure 2 depicts ROC curves for all prediction methods on the Blind-set, which is the only test set truly ‘blind’ to all methods. The figure indicates that the performance ranks in the order of SMM>BIMAS>PM over almost the entire range. SYFPEITHI is the worst method for sensitivities above 0.33 and becomes the best at specificity above 0.97. This may be due to the fact that SYFPEITHI predictions reflect other components of the antigen presentation pathway in addition to MHC binding, leading to a decrease in sensitivity.

The area under the ROC curve (AUC) gives a single number describing prediction accuracy of a method. Figure 2 translates to AUC values of 0.869 for SMM, 0.846 for BIMAS, 0.745 for SYFPEITHI, and 0.795 for PM (Table 1). The AUC values for all methods on the BIMAS-set and SYFPEITHI-set are also listed in Table 1. The numbers in parantheses denote the overlap between training and test sets. For both sets, the SMM predictions achieve the best results of all truly blind prediction methods.

Comparison of general methods: CART, ANN, the additive method and SMM + pair coefficients

We built a classification tree for the Training-set (Fig. 3), as described by (Segal *et al.*, 2001). For the ANN, we used the same network as described in (Gulukota *et al.*, 1997), and also trained the network on the Training-set. For the additive method, we used the coefficients determined by (Doytchinova *et al.*, 2002). Figure 4 shows the ROC curves for these general

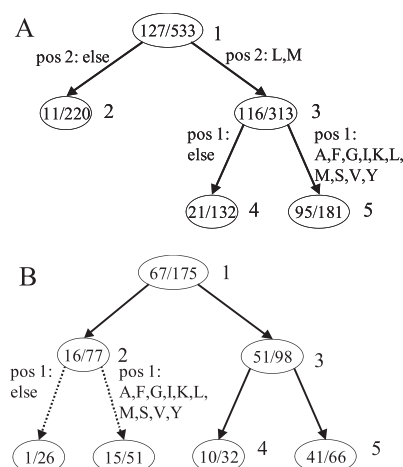


Fig. 3. CART tree for peptides binding to HLA-A2. Each ellipse denotes a node corresponding to a set of peptides, with the first number indicating the total number of binders, and the second number indicating the total number of peptides in the node. The splits in the nodes are symbolized by arrows, which lead to daughter nodes. To the right of each node is a reference number (1–5) used in the text to indicate the node. (A) The optimal tree generated for the Training-set. (B) The tree in (A) is used to classify peptides in the Blind-set. The additional split on the lower left with dotted arrows, which is taken from the split of node 3 in (A), makes mostly correct predictions on non-binders.

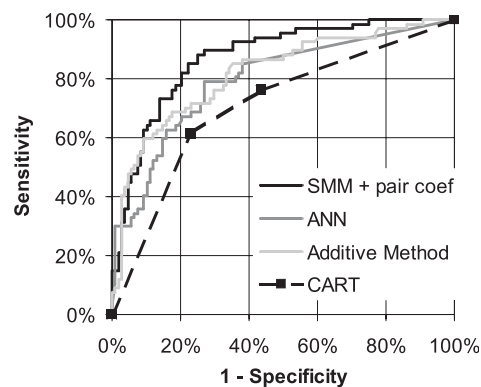


Fig. 4. ROC curves for general methods on the HLA-A2 Blind-set. The figure contains ROC curves described in the legend of Figure 2. Since there are only three terminal nodes in the CART tree, corresponding to three different scores, its ROC curve consists of only two non-trivial points.

methods on the Blind-set. The AUC values for all test sets are listed in Table 1. ANN and the additive method made consistently better predictions than the CART tree. None of the methods reached the prediction accuracy of SMM or BIMAS, which made the independent binding assumption. For all test sets, the ‘SMM matrix + pair coefficients’ approach further improved slightly upon the results achieved by the SMM matrix alone.

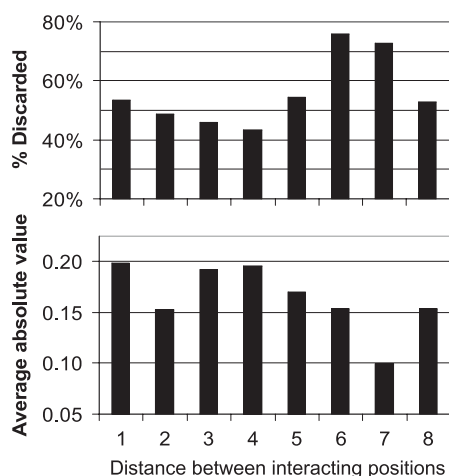


Fig. 5. Distance distribution of pair coefficients. Percentage of discarded pair coefficients (top panel) and the average values of retained pair coefficients (bottom panel) are plotted against the distance between two interacting peptide positions.

Distribution of pair coefficient values

We analyzed if the distance between two peptide positions ($i - i'$) influences the extent they interact with each other, quantified as the absolute values of the corresponding pair coefficients. In Figure 5, we plot two quantities that reflect the influence of position distance: the percentage of pair coefficients discarded due to conflicting information and the average absolute value of retained pair coefficients at the distance. Distances 6 and 7 have the highest percentages of discarded coefficients and distance 7 has the lowest average value of retained coefficients, indicating weak or no interactions between positions at such distances. To a lesser extent, the levels of interaction at distances 2, 5, 6 and 8 are also weaker than those at distances 1, 3 and 4.

DISCUSSION

We have developed a new matrix-based method (SMM) for predicting the binding affinity of a peptide to an MHC molecule. Similar to other matrix-based methods, SMM makes the independent binding assumption, and expresses the binding affinity of a peptide as the sum of individual residue contributions. We derive the matrix entries of SMM by minimizing the distance between predicted scores and measured values for a set of training peptides. Compared to three other well-established matrix-based methods (BIMAS, SYFPEITHI and PM), SMM achieved the highest accuracy on all three test sets.

The SMM approach has two novelties. First, we incorporate the experimental information of heavy non-binders precisely into the distance defined in Equation (2). In contrast, previous approaches either left them out entirely or tried to fit them exactly to the lower bound, since only the lower bound of the

IC50 values for heavy non-binders can be determined. Second, we use the regularization technique. With errors in experimental measurements, there can be multiple sets of matrix coefficients that can reproduce the experimental data within the range of measurement accuracy. Choosing the set of coefficients that gives the minimum distance may mean to overfit the problem. By incorporating a regularization parameters [λ in Equation (5)], we choose a set of coefficients that reproduces the experimental results reasonably while keeping the parameter values small. This effectively prevents overfitting. Table 1 indicates that the AUC values at $\lambda_{\text{opt}} = 1$ are better than those at $\lambda = 0$ for all test sets.

Dropping the independent binding assumption, we supplemented the matrix-based SMM method with pair coefficients describing possible pair-wise interactions between peptide positions. On all test sets, this led to a slight improvement in prediction accuracy compared to SMM alone. One advantage of this approach is that we determine the pair coefficients by systematically investigating differences between matrix predictions and experimental values. Since SMM is highly accurate, it is a better starting point than trying to determine both position contributions and position interactions simultaneously. Another novelty of our approach is that we limit the interactions under investigations to those supported by sufficient amount of consistent training data. An important aspect in our approach is again the use of a regularization parameter [λ' in Equation (7)], which prevents the pair coefficients from overfitting the data. Its importance can be seen in Figure 1: without damping ($\lambda' = 0$) pair coefficients reduce the prediction accuracy below that of the matrix alone.

We have shown that the application of three previously published general methods (ANN, CART and the additive method) to peptide binding to MHC-I led to worse predictions than SMM + pair coefficients or even SMM alone, with the latter assuming that each peptide position contributes independently to the binding. At first glance this may seem surprising, as general methods should be able to describe all binding mechanisms, including the one assuming independent binding. Why does the more restrictive matrix approach perform better? This can best be seen for the CART algorithm. As shown in Figure 3, CART suggests splitting node 3, but not node 2. If this is a truthful depiction of the reality, peptides in node 3 would bind differently than peptides in node 2, signifying an interaction between positions 2 and 1 only for peptides with an L or M at position 2. In contrast, if the independent binding assumption is true and there are no interactions between positions 2 and 1, the split described for node 3 should also be applicable to node 2. Figure 3B shows that performing the same split on node 2 makes mostly correct predictions (25 out of 26) on non-binders in the Blind-set. CART cannot identify this split, because it can only use the peptides in node 2 to decide about splits at node 2, and there are only 11 binders left in that node. This shows that general methods

simply require much more training data than AIB methods, which can lead to inferior performance for the former.

In case of the additive method and the ANN, lack of data has led to overfitting. The additive method has 1850 free parameters, and the ANN architecture we took from (Gulukota *et al.*, 1997) has more than 9000 neurons. So many parameters cannot be determined reliably using our Training-set with 533 experimental data points. For the ANN, choosing a different architecture with fewer free parameters might improve prediction accuracy. For the additive method, the assumption was made that neighboring amino acids should display stronger interactions. While this is a reasonable assumption to limit the number of interactions considered by the method, the total number of parameters is still too large. Interestingly, when we neglected the coefficients describing interactions between neighboring amino acids, and kept only those compatible with the independent binding assumption, the average prediction accuracy actually improved (data not shown).

In a comparison of several methods to predict peptide binding to MHC (Yu *et al.*, 2002), it was reported that the optimal choice of a prediction method depends on the number of peptides available for training: an ANN was outperformed by scoring matrices when the training data consisted of 234 peptides, while the ANN outperformed scoring matrices when trained on over a thousand peptides. Our approach of using a scoring matrix + pair coefficients should maintain good performance over a wide range of training set sizes. If little data is available, few or no pair coefficients will meet the criteria for inclusion, and the method is reduced to the SMM matrix. With more training data available, more pair coefficients are included, thus adjusting the complexity of the method to the available training data.

Another advantage of our pair coefficients over ANN is that the extracted rules for binding are easy to interpret. The values determined for the pair coefficients provide direct information about the MHC-peptide binding mechanism. Since peptides bind in an extended conformation, one would expect the absolute values of the pair coefficients to be lower if their associated amino acids are farther apart. Our results generally support this expected trend (Fig. 5), but to a much lesser extent than expected. Also, our study did not confirm the suspected trend that $i - (i + 2)$ neighbors influence each other more strongly than $i - (i + 1)$ neighbors, with the former facing the same direction thus allowing direct interactions between the side chains.

While we have shown that interactions between binding sites do exist, the values of the pair coefficients describing them are roughly an order of magnitude lower than the entries of the scoring matrix. This could be due to two reasons: (1) the values for these coefficients are indeed small, or (2) the noise level in the training set is high and as a result the coefficients are forced to low values by the regularization parameter λ' . Since the plain SMM performs very well, we believe

that the true values of the pair coefficients are significantly lower than the values of the matrix entries. Otherwise, the general methods such as CART or ANN would have had a considerable advantage over the matrix based approaches.

We have implemented SMM + pair coefficients as a web server at <http://zlab.bu.edu/SMM>. In the future, we plan to incorporate all other MHC-I alleles for which sufficient binding data are available for training the algorithm.

ACKNOWLEDGEMENTS

We thank Dr I. Doytchinova *et al.* for sending us the coefficients for the additive method. This work was funded by Whitaker grant RG-00-0426 and NIAID contract no1-ai-95362.

REFERENCES

- Borras-Cuesta,F., Golvano,J., Garcia-Granero,M., Sarobe,P., Riezu-Boj, J., Huarte,E. and Lasarte,J. (2000) Specific and general HLA-DR binding motifs: comparison of algorithms. *Hum. Immunol.*, **61**, 266–278.
- Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.
- Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. CRC Press, Boca Raton, FL.
- Brusic,V., Schönbach,C., Takiguchi,M., Ciesielski,V. and Harrison,L.C. (1997) Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. *Ismb*, **5**, 75–83.
- Doytchinova,I.A., Blythe,M.J. and Flower,D.R. (2002) Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201. *J. Proteome Res.*, **1**, 263–272.
- Frontline Systems, I. (1999) Solver DLL.
- Gulukota,K., Sidney,J., Sette,A. and DeLisi, C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, **267**, 1258–1267.
- Hammer,J., Bono,E., Gallazzi,F., Belunis,C., Nagy,Z. and Sinigaglia,F. (1994) Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J. Exp. Med.*, **180**, 2353–2358.
- Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.
- Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C*. Cambridge University Press, Cambridge, MA.
- Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
- Rammensee, H.G. (1995) Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.*, **7**, 85–96.

- Ruppert,J., Sidney,J., Celis,E., Kubo,R.T., Grey,H.M. and Sette,A. (1993) Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell*, **74**, 929–937.
- Schönbach,C., Ibe,M., Shiga,H., Takamiya,Y., Miwa,K., Nokihara,K. and Takiguchi,M. (1995) Fine tuning of peptide binding to HLA-B*3501 molecules by nonanchor residues. *J. Immunol.*, **154**, 5951–5958.
- Schueler-Furman,O., Altuvia,Y., Sette,A. and Margalit,H. (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.*, **9**, 1838–1846.
- Segal,M.R., Cummings,M.P. and Hubbard,A.E. (2001) Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics*, **57**, 632–642.
- Sette,A., Buus,S., Appella,E., Smith,J.A., Chesnut,R., Miles,C., Colon,S.M. and Grey,H.M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl Acad. Sci. USA*, **86**, 3296–3300.
- Shastri,N., Schwab,S. and Serwold,T. (2002) Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules. *Ann. Rev. Immunol.*, **20**, 463–493.
- Udaka,K., Wiesmuller,K.H., Kienle,S., Jung,G., Tamamura,H., Yamagishi,H., Okumura,K., Walden,P., Suto,T. and Kawasaki,T. (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*, **51**, 816–828.
- Vajda,S., Weng,Z., Rosenfeld,R. and DeLisi,C. (1994) Effect of conformational flexibility and solvation on receptor–ligand binding free energies. *Biochemistry*, **33**, 13977–13988.
- Yu,K., Petrovsky,N., Schonbach,C., Koh,J.Y. and Brusica,V. (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med.*, **8**, 137–148.
- Zhang,C., Cornette,J.L. and Delisi,C. (1997) Consistency in structural energetics of protein folding and peptide recognition. *Protein Sci.*, **6**, 1057–1064.