

Application of Genetic Search in Derivation of Matrix Models of Peptide Binding to MHC Molecules

Vladimir Brusic*, Christian Schönbach†§, Masafumi Takiguchi†, Vic Ciesielski‡ and Leonard C. Harrison*

* The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.

† Department of Tumor Biology, Institute of Medical Science, University of Tokyo, Tokyo, Japan.

§ Current address: Chugai Research Institute for Molecular Medicine, Inc. Ibaraki, Japan.

‡ Department of Computer Science, RMIT University, Melbourne, Victoria, Australia.

Correspondence to:

Vladimir Brusic, The Walter and Eliza Hall Institute of Medical Research,
P.O. Royal Melbourne Hospital, Victoria 3050, Australia.
Ph. +61 3 9345 2588; Fax +61 3 9347 0852;
Email: vladimir@wehi.edu.au

Keywords: application specific modeling, classification, genetic search, machine learning, major histocompatibility complex, MHC, motif, patterns, peptide binding.

Abstract¹

T cells of the vertebrate immune system recognise peptides bound by major histocompatibility complex (MHC) molecules on the surface of host cells. Peptide binding to MHC molecules is necessary for immune recognition, but only a subset of peptides are capable of binding to a particular MHC molecule. Common amino acid patterns (binding motifs) have been observed in sets of peptides that bind to specific MHC molecules. Recently, matrix models for peptide/MHC interaction have been reported. These encode the rules of peptide/MHC interactions for an individual MHC molecule as a 20×9 matrix where the contribution to binding of each amino acid at each position within a 9-mer peptide is quantified. The artificial intelligence techniques of genetic search and machine learning have proved to be very useful in the area of biological sequence analysis. The availability of peptide/MHC binding data can facilitate derivation of binding matrices using machine learning techniques.

We performed a simulation study to determine the minimum number of peptide samples required to derive matrices, given the pre-defined accuracy of the matrix model. The matrices were derived using a genetic search. In addition, matrices for peptide binding to the human class I MHC molecules, HLA-B35

and -A24, were derived, validated by independent experimental data and compared to previously-reported matrices. The results indicate that at least 150 peptide samples are required to derive matrices of acceptable accuracy. This result is based on a maximum noise content of 5%, the availability of precise affinity measurements and that acceptable accuracy is determined by an area under the Relative Operating Characteristic curve (Aroc) of >0.8. More than 600 peptide samples are required to derive matrices of excellent accuracy (Aroc>0.9). Finally, we derived a human HLA-B27 binding matrix using a genetic search and 404 experimentally-tested peptides, and estimated its accuracy at Aroc>0.88. The results of this study are expected to be of practical interest to immunologists for efficient identification of peptides as candidates for immunotherapy.

Introduction

A major function of the immune system is to discriminate 'self' from 'non-self' and induce or regulate responses to cells expressing 'non-self' molecules. 'Non-self' includes microorganisms, foreign cells and tissues (eg transplants), or altered 'self' (eg tumour cells). In vertebrates, immune recognition of protein antigens is mediated by major histocompatibility complex (MHC) protein molecules. MHC molecules bind peptides derived from intracellular (MHC class I) or extracellular (MHC class II) proteins and display them on the surface of the cell for recognition by T cells,

1. Copyright © 1997, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

reviewed in (Rammensee *et al.*, 1993; Cresswell, 1994). Binding of peptides to MHC molecules is necessary for immune recognition, but only certain peptides are able to bind to particular MHC molecules. Determining which peptides are able to bind to a specific MHC molecule is of major importance for understanding the basis of immunity.

MHC molecules form a peptide-binding groove which in protein structure terms, comprises two parallel alpha helices on top of a beta sheet. Peptides that bind to MHC class I molecules are usually 8 to 11 amino acids long; the majority are 9 amino acids (Rammensee *et al.*, 1995). Class I peptides generally contain two primary 'anchor' amino acids, whose contribution to binding is the most important, and several secondary anchors. Class II peptides that bind MHC class II molecules are 12-25 amino acids long and extend beyond the MHC groove, but contain a minimum 9-mer core (Jardetzky *et al.*, 1996). Class II peptides contain a single primary anchor, which is necessary for binding, and several secondary anchors. The number of reported human MHC molecules is greater than 100 for both class I and class II (Bodmer *et al.*, 1994).

The prediction of MHC binding peptides is a well-defined classification problem. Experimental methods for measuring binding of peptides to MHC molecules have been developed but they require the costly synthesis of multiple peptides and are labour-intensive. High performance prediction methods would therefore facilitate research in this field. Among the particular difficulties that must be addressed are: a) the variable lengths of reported binding peptides, b) under-terminated core regions for individual peptides, c) the multiplicity of amino acids permissible as primary anchors, d) different experimental methods employed for determining binding affinity, e) a lack of general rules for predicting peptide structure-interaction, and f) experimental and reporting errors. Application of artificial intelligence techniques to the MHC binding problem may overcome these difficulties and help define refined rules for MHC/peptide interactions, as well as to define requirements for the prediction of MHC binding peptides. This approach has the potential to be combined with laboratory experimentation in order to achieve significant savings in cost and time.

Several methods have been used to predict MHC binding peptides, including those based on binding motifs, binding matrices and artificial neural networks. Binding motifs specify which residues at given positions within the peptide are necessary or favourable for binding to a specific MHC molecule. Binding motifs for various human and mouse MHC class I and class II molecules have been reported (Rammensee *et al.*, 1995). Binding motifs for most common MHC class I molecules are relatively well-defined. On the other hand, binding motifs for fewer MHC class II molecules have been defined (Hammer *et al.*, 1994; Harrison *et al.*, 1997). There are many examples of peptides conforming to the binding motif that do not bind as predicted; conversely,

peptides shown to bind a given MHC molecule do not always contain a reported binding motif (Brusic *et al.*, 1996).

Binding matrices provide coefficients for each amino acid/position that can be used with appropriate formulae to calculate scores predictive of binding. The assumptions are a) that each position within the peptide contributes independently to binding to the MHC molecule, and b) that a residue when located at a given position in the peptide contributes the same amount to binding, even within different sequences. A number of binding matrices have been defined for MHC class I (Parker *et al.*, 1994; Schönbach *et al.*, 1995,1996; Kondo *et al.*, 1995) and for class II molecules (Rothbard *et al.*, 1994; Hammer *et al.*, 1994; Davenport *et al.*, 1995). The matrix of (Hammer *et al.*, 1996) for the human class II molecule HLA-DR4(B1*0401) was derived from a systematic analysis of binding, after each position in a 9-mer peptide was replaced individually by each of the other 19 amino acids. Other matrices have been derived either theoretically by fitting matrix coefficients to existing binding data or by combining binding experiments with theoretical data fitting. A large collection of peptide binding data is available in the MHCPEP database (Brusic *et al.*, 1996) and therefore the derivation of matrices can be defined as an artificial intelligence problem.

The definition of binding matrix models draws on three areas of artificial intelligence: classifier development, machine learning theory and genetic algorithms. A classifier can be viewed as a function, implemented by a computer program, which maps a vector of data to one of a number of categories, eg the peptide YRATATTWQ to the class BINDS-B27 or DOES-NOT-BIND-B27. A classifier is usually developed by taking a set of data vectors and their classifications and using a machine learning algorithm to induce a classifier. When presented with a data vector whose classification is unknown, say peptide of unknown binding affinity, the classifier will assign a class. In this work a classifier is a binding matrix, together with its associated binding score calculation.

A key question is how accurately the classifier will perform on fresh (previously unseen) cases. An empirical estimate of the error rate is usually obtained by dividing the available data vectors and their known classifications into the training set and the test or validation set. The classifier is developed from the training data only and its performance is evaluated on the test set. The error rate on the test set is then taken to be an estimate of the error rate that will be achieved on fresh cases.

Machine learning theory is concerned with the relationship between the number of examples used in developing a classifier and its error rate. Specifically, two questions are of interest: 'If I have n examples how accurate will the classifier be?' and 'If I want the classifier to have at most a certain error rate, how many examples are needed in the training

set?’ The ‘probably approximately correct’ or PAC approach (Valiant, 1984) seeks to determine the number of examples for which the probability $P(\text{ERC} > \epsilon) < \delta$, where ERC is the error rate of the classifier. Typical values of ϵ are around 0.1 and of δ are 0.05 for biological data. Thus we would like to know the number of examples for which we can be 95% certain that the error rate of the classifier will be less than 10%. Using PAC learning theory, the number of required samples is calculated by formula (Kearns and Vazirani, 1994):

$$m \geq Co \times \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

For $\epsilon=0.1$ (maximum learning error 10%), $\delta=0.05$ (95% confidence), $d=180$ (number of variables in a matrix of size 20×9 representing every possible amino acid at each position in a 9-mer peptide) and constant $Co=1$ (the best case), the number of required samples is $m \geq 1813$ in a noiseless environment. The costs of laboratory experimentation curtails the number of peptide samples that can be tested for binding to particular MHC molecule, usually to several hundred at best. The PAC learning model provides an extremely conservative estimate as it caters for the worst-case prediction risk. Empirical results suggest that the number of training samples required for learning is much smaller (Holden and Niranjana, 1995). The alternative approach is to determine the required number of samples in a simulation study, using a computer model aided by heuristics acquired from the laboratory experiments.

A genetic algorithm is an approach to solving certain kinds of search and optimisation problems using a process inspired by Darwinian evolution. The approach involves maintaining a population of potential solutions and then generating new solutions by the use of genetic operators such as reproduction, crossover and mutation. A ‘fitness function’ is a measure of the quality of the solution. As the genetic algorithm proceeds, individuals of improved fitness appear in the population. We are seeking the best/optimal binding matrix so that each individual in the population represents a different matrix and the fitness function is a measure of how well that matrix classifies the known data. As the genetic algorithm proceeds, the binding matrices improve.

The computational study was performed in a simulated environment which mimicked the real one. In this environment, peptide binding was encoded by a template matrix. A number of training sets of different sizes, and a sufficient number of validation sets, were created and used in simulated learning. The learning was performed utilising a genetic search (Goldberg, 1989). Partial solutions were assessed for their prediction accuracy using validation sets. These results were analysed to suggest required numbers of samples for the derivation of binding matrices of certain accuracy. The effect of noise in the training set was also

studied by corrupting training sets with 5% or 10% of noise and applying simulated learning.

The results of the simulation study were further validated by deriving binding matrices using experimental data and genetic search for two human class I molecules HLA-A24(A*2401/A*2402) and HLA-B35(B*3501). These matrices were validated using the independent sets of experimental data and their performance compared with the reported binding matrices for these molecules. Finally, a matrix for the human class I molecule HLA-B27(B*2705) was derived using the learning process described here and sets of available binding data.

In this study, we combined the results of experimental studies and computer simulations to clarify the accuracy and usefulness of machine learning in the prediction of peptide/MHC binding. These results are expected to be of practical interest to immunologists for efficient identification of peptides as candidates for immunotherapy.

Materials and Methods

Simulated environment

A simulated computational environment was created to determine the relationship between the number of peptide samples and the accuracy of the model that can be derived by machine learning. Training sets containing 150, 300, 600 and 900 samples were created using a program for random peptide generation and a template matrix. Three different sets were created for each sample size. Five validation sets each comprising 1200 samples were also created. Binding affinity of the peptides in data sets was described as high-, moderate-, low- or zero-affinity. The respective proportions of the high-, moderate-, low- and zero-affinity binders in the data sets were 1:1:1:3. Frequencies of amino acids used for random peptide generation reflected codon frequencies in Genbank database (Benson *et al.*, 1996). The template matrix (Table 1) was modified from that of (Hammer *et al.*, 1994) and used to determine peptide binding affinities in the simulated environment.

A genetic search was used to derive binding matrices. Five search runs were performed with each of the data sets and five best matrices found in each run were saved. Independent validations were performed on each saved matrix using five different validation sets. The validation results were used to assess the expected accuracy of matrices derived by the genetic search relative to the number of samples in the training set. The schematic representation of this experiment is shown (Figure 1).

Additional simulated experiments were performed with noisy data sets. The training data sets described above were randomly corrupted with 5% or 10% noise. The accuracy of matrices derived by the genetic search was also assessed, relative to the number of samples in the noisy training sets.

Position	Amino acid																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	-250	-250	-250	-250	0	-250	-250	-10	-250	-10	-10	-250	-250	-250	-250	-250	-250	-10	0	0
P2	0	0	-13	1	8	5	8	11	11	10	11	8	-5	12	22	-3	0	21	-1	9
P3	0	0	-13	-12	8	2	2	15	0	10	14	5	3	0	7	2	0	5	0	8
P4	0	0	17	8	-8	-15	8	8	-22	-6	14	5	-21	11	-15	11	8	5	-12	-10
P5	0	0	-2	-1	3	2	-1	1	3	1	3	2	5	1	0	4	6	4	-1	-2
P6	0	0	0	-12	-13	-11	-16	-2	-23	-13	-13	17	1	-12	-22	17	19	13	-9	-11
P7	0	0	-11	-2	-8	-15	-8	-2	-12	4	7	-1	-3	-5	-12	-4	-2	5	-13	-7
P8	0	0	-11	-2	1	-5	0	-1	9	6	4	7	-2	16	7	6	5	4	6	13
P9	0	0	-25	-18	-8	-2	3	-4	-9	-13	-4	-11	-16	7	-9	12	-3	5	-3	-15

Table 1: The template matrix that encodes peptide binding rules. Each non-anchor amino acid at position P1 was assigned a value of -250. The coefficients at positions P2 to P9 were constrained between -25 and +25. A binding score was calculated by summing coefficients for amino acids that match query peptide. For example, the binding score of the peptide YRAFATTWQ is $0+22+0-8+0+19-2+6+7=44$. Scores of ≥ 40 correspond to high-, 30-39 to moderate- and 20-29 to low-affinity binding, and scores of < 20 to non-binding.

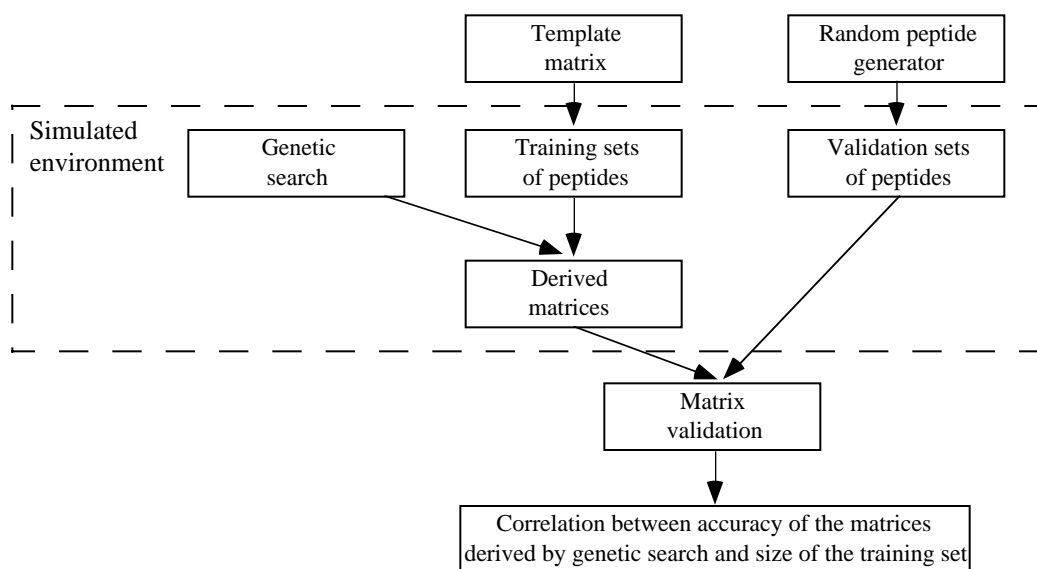


Figure 1: Procedure for determining the relationship between the accuracy of matrices derived by genetic search and the number of training samples.

Genetic search

Genetic search was performed using GAucsd package (Schraudolph and Grefenstette, 1992). A 20×9 matrix comprising coefficients for each of 20 amino acids at each position in the 9-mer peptide was encoded as a 1080-bit long binary string. Individual coefficients were encoded as 6-bit binary strings. A population of individual structures was evolved towards increased capacity to discriminate between binding and non-binding peptides. The following parameters of the genetic algorithm were selected by fast prototyp-

ing: population size of 100 individuals, crossover rate of 5.0 and mutation rate of 0.001. The structures were evolved for 3×10^4 function evaluation cycles. The GAucsd package was used for function minimisation and the fitness function $F=1/(TP+3 \times TN)$ to maximise the number of true negatives (TN) and true positives (TP). The numbers of TP and TN were determined when the individual is used to discriminate between positives (binders) and negatives (non-binders) in the training set. The values for coefficients were constrained between -25 and +25 except for the primary anchors. Matrices in the simulated environment encoded peptide binding

with a single anchor and the coefficients at position P1 were constrained between -250 and +250. Peptides binding human class I molecules HLA-B35, -A24 and -B27 contain two primary anchors at positions P2 and P9 (Rammensee *et al.*, 1995) and the coefficients at those positions were constrained between -125 and +125.

Derivation of HLA-B35, -A24 and -B27 binding matrices

Nine amino acid long peptides reported as binders to HLA-B35, HLA-A24 or HLA-B27 were extracted from the MHCPEP database. A number of non-binding peptides to these molecules was also collected from the literature. Some of the reported non-binder peptides were longer than 9 amino acids and each 9-mer substring from within these peptides was assumed to be a non-binder. The B35 data set comprised 72 binders and 193 non-binders, the A24 set 59 binders and 170 non-binders and the B27 set 221 binders and 183 non-binders. Genetic search was performed as described above and a single binding matrix for each MHC molecule was saved. The accuracy of the B35 and A24 matrices derived by the genetic search were assessed against experimental data.

Binding matrices derived using the genetic search were compared to previously-reported binding matrices. A B35 binding matrix (B35S) was reported by (Schönbach *et al.*, 1995) and an A24 binding matrix (A24K) by (Kondo *et al.*, 1995). The parameters of B35S were calculated by comparing ranks of each residue at one position. The parameters of A24K were calculated as the relative average binding affinity of peptides carrying particular residues. We have assessed the ability of these matrices to predict peptide binding affinity. The value "1" was assigned to anchor residues at position 2 (P2) in the B35S and to anchor residues at the P2 and position 9 (P9) in A24K. Binding scores for peptides were calculated by multiplying the coefficients corresponding to the query peptide.

Of peptides reported as B27 binders, finer specificity was known for only 29; the remaining 192 peptides were reported as binders of unknown binding affinity. The B27 data set was split into a training set (192 binders and 153 non binders) and a test set (5 high-, 17 moderate-, 7 low-, and 30 zero-affinity binders). A single binding matrix derived by the genetic search was validated using the test set. Finally, the B27 binding matrix was derived using all available B27 binding data.

Assessment of the accuracy of binding matrices

The accuracy of binding matrices was determined using Relative Operating Characteristic (ROC) analysis (Swets, 1988). This analysis uses a plot of the true positive proportion vs. the false positive proportion for the various thresholds of the decision criterion. It provides a single measure,

Aroc, which is a proportion of the area under the ROC curve. This measure removes biases due to an overwhelming proportion of negative events (non-binding peptides). By utilising several thresholds for decision criterion it removes biases due to optimistic (over-prediction) or pessimistic (under-prediction) outcomes. The evaluated thresholds were those for high-, moderate- and low- binding. The quality of solution was assessed as good for Aroc>0.8 or excellent for Aroc>0.9, as suggested by (Swets, 1988). Value of Aroc=0.5 indicates that the accuracy of the prediction is equivalent to that of the random guessing. The points on ROC curves were calculated using thresholds for high-, moderate- or low- affinity binding of peptides for both experimental and predicted binding affinities.

Laboratory Experimentation

Thirty-two peptides (3 high-, 2 moderate-, 3 low-, and 24 zero-affinity binders) were tested for binding to HLA-B35 and subsequently used to validate the HLA-B35 binding matrix derived by genetic search. Fifty one peptides (16 high-, 13 moderate-, 9 low-, and 13 zero-affinity binders) were tested for binding to HLA-A24 and used to validate the HLA-A24 binding matrix. The 51 peptides were experimentally tested as binders to HLA-A24(A*2402) which is identical to (A*2401) except for the position 182 (the last residue of α -2 domain) which plays no role in peptide interactions. The peptide binding to these two variants of HLA-A24 should therefore be identical, which has also been suggested by (Chelvanayagam, 1996).

Peptides. Sequences derived from human immunodeficiency virus, Epstein-Barr virus and hepatitis C virus were searched for anchor residues at position 2 and 9 of the B35 (Falk *et al.*, 1993, 1994) and A24 (Maier *et al.*, 1994, Kubo *et al.*, 1994) peptide motif. Peptides were synthesised from matched sequences using F-moc strategy (Nokihara *et al.*, 1992).

Peptide-HLA class I binding assay. The affinity of peptides for B35 and HLA-A24 was determined by measuring peptide-dependent cell surface stabilisation of the HLA molecules on a mouse T-cell lymphoma cell line RMA (Ljunggren *et al.*, 1990) transfected with human β -2 microglobulin (RMA-S) and either B*3501 (RMA-S-B*3501) or A*2402 (RMA-S-A*2402). As RMA-S transfectants lack functional TAP-2 proteins required to transport cytosolic peptides into the endoplasmic reticulum to interact with nascent class I, they express stable, but empty, HLA molecules at 26°C which become unstable when the temperature is increased to 37°C. If exogenous peptides that fit into the peptide binding groove of HLA molecules are added, the latter remain stable on the cell surface. The assay was performed as previously described by (Takamiya *et al.*, 1994). Briefly, RMA-S transfectants cultured at 26°C were loaded with peptides at concentrations of 10⁻⁴ to 10⁻⁸ M and

incubated at 37°C for 3h. The cell surface stabilisation of HLA molecules by bound peptides was measured by binding of fluorescinated HLA class I-specific antibodies, detected as mean linear fluorescence intensity with a FACS-can flow cytometer.

Analysis of peptide binding. Peptides at a high, non-physiological concentration of 10^{-4} M giving more than 25% of the mean fluorescence intensity of RMA-S-B*3501 or RMA-S-A*2402 transfectants not loaded with peptides were evaluated as non-binders. Peptide/B35 binding was quantitatively compared (Schönbach *et al.*, 1995) with the binding of a control and a high affinity self-peptide 37F (LPFDFT-PGY). Binders were classified into three categories according to the ratio of the half-maximal mean fluorescence intensity of 37F (BL5037F) vs. the tested peptide (BL50TEST); Binding index $BI=BL5037F/BL50TEST$. The values of BI that define high, moderate and low binders are $BI \geq 0.1$, $0.1 > BI \geq 0.01$ and $0.01 > BI \geq 0.001$, respectively. Binding to A24 was semi-quantitatively determined (Ibe *et al.*, 1996), because a suitable control self-peptide was not available. Therefore, the classification into high binders ($BL50 \leq 10^{-5}M$), moderate binders ($10^{-5}M < BL50 \leq 10^{-4}M$) and low binders ($BL50 > 10^{-4}M$) relied only on the BL50 value of the tested peptide.

Results

Simulation study

Binding matrices were derived using training sets of various sizes containing noiseless data, 5% noise or 10% noise. The relationship between the quality of solution assessed by calculating Aroc values, and the number of samples in the training set is shown (Figure 2). The results of the simulation study indicate that 600 samples of binding peptides are sufficient as a training set for derivation of excellent solutions. If

the noise content is below 5%, 900 samples are sufficient for derivation of excellent solutions. Furthermore, 150 samples are sufficient for derivation of good solutions using noiseless training sets, whereas 300 samples are required in the presence of noise. Simulations using noiseless data sets were of excellent reproducibility (maximal standard deviation of $SD=0.03$). Reproducibility of searches using 5% noisy data sets were acceptable for training sets of size 300 or more.

The implication for derivation of binding matrices using machine learning and experimental data sets is that assuming a maximum noise content of 5% in experimental data, 300 samples should be sufficient for derivation of good solutions. More than 600 and less than 900 samples are sufficient for derivation of binding matrices of excellent classification ability.

Prediction of peptide binding to HLA-B35 and HLA-A24

Prediction of peptide binding affinities was performed by classifying peptides into four groups: high-, moderate-, low-, and zero-affinity. Binding matrices were derived using genetic search and tested with independent peptide sets whose binding affinities were experimentally measured. The B35 and A24 test sets consisted exclusively of peptides possessing two anchor residues (P or A at P2 and L, I, M or F at P9 for B35; F or Y at P2 and I, L, F or W at P9 for A24). The test results are given in Table 2. Aroc values for B35 and A24 binding matrices were 0.74 and 0.73, respectively. The accuracies of the derived binding matrices cannot be described as acceptable, which is consistent with the results of the simulation study indicating that 300 samples are required for derivation of good solutions ($Aroc > 0.8$). The training set for the B35 binding matrix comprised 265 and for A24 matrix 229 peptides. The number of binders in the

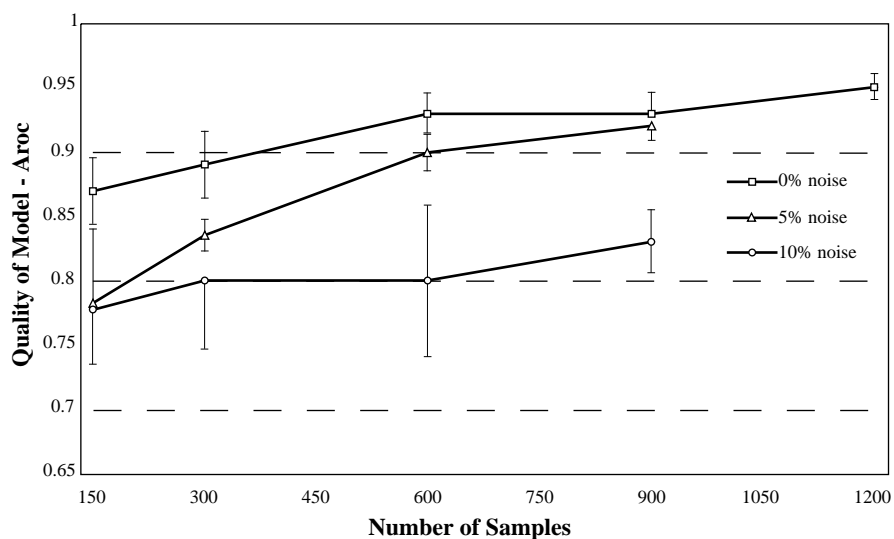


Figure 2: The relationship between the quality of derived matrices (assessed by the value of Aroc) and the number of samples in the training set. Three levels of noise content in the training set were analysed. For clarity of presentation, only one of the upper or lower standard deviation bars are shown at some points.

training sets for both B35 and A24 is relatively low (72 and 59, respectively), suggesting that more binders need to be added to training sets to produce good binding matrices

A) Predicted affinity	Experimental affinity			
	High	Mod.	Low	Zero
High	2	1	1	3
Moderate	0	0	0	1
Low	1	0	0	3
Zero	0	2	0	18

B) Predicted affinity	Experimental affinity			
	High	Mod.	Low	Zero
High	12	7	4	4
Moderate	0	2	1	0
Low	1	1	0	0
Zero	2	4	3	10

Table 2: Prediction performance of binding matrices derived using genetic search: A) B35 binding matrix, and B) A24 binding matrix.

The predictive ability of the previously reported matrices, B35S and A24K, was assessed using the same test sets of peptides as for matrices derived using genetic search (Figure 3). The accuracy of the B35S matrix was assessed as Aroc=0.82. Therefore, it is a good model for predicting peptide binding to HLA-B35. By combining experimental measurements that provide accurate classification of peptides into four binding classes, with statistical rank analysis, it is possible to define good binding matrices from relatively small numbers of peptides (65 binders and 53 non-binders were used in derivation of B35S matrix). The A24K binding matrix did not perform well in predicting binding affinities of test peptides. Possible reasons for the poor performance are a) the limits for values of the coefficients in the matrix were too high, and b) small number of non-binders (15) compared to binders (126) were used for derivation of the matrix.

Both the B35S and A24K matrices were developed to help refine binding motifs and not necessarily to predict peptide binding affinity. The results of validating of these matrices indicate that it is possible to use the combination of precise experimental measurements and statistical rank analysis to derive good binding matrices using smaller number of peptides than required in a genetic search (which uses only binder/non-binder classification of peptide binding affinity). The essential requirements that need to be satisfied are: a) the definition of a reasonable upper limit for the values of coefficients in the binding matrix, and b) the sufficient representation of both binders and non-binder peptides in a training set.

Some of the peptides reported as T-cell epitopes or binding peptides to either B35 or A24 contain amino acids at the

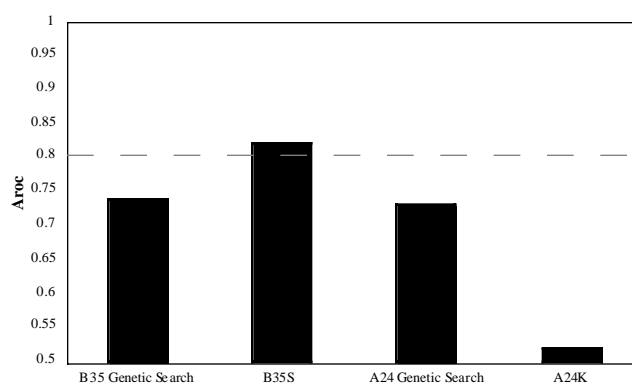


Figure 3: Comparison of the ROC areas of binding matrices derived by genetic search and of previously reported matrices.

anchor positions other than those assigned as primary anchors in the B35S matrix. Querying of MHCPEP database revealed that residues P,V,A or S were observed at the position P2 and F,Y,W,I,L,V,M,K,R,H,D,S or N at the P9 of 9-mer peptides reported as B35 binders. In the B35S matrix, only residue P was allowed as an anchor at P2 and Y,L,I,F or M at P9. Primary anchors as defined by B35S matrix were present in 52, but not in 20, of HLA-B35 binding peptides used in genetic search. The refinement of the B35S matrix will therefore require re-defining of allowed anchor residues and consequently more peptide samples.

HLA-B27 matrix

A test B27 binding matrix was derived using genetic search with a training set of 345 peptides and then validated using a test set of 59 peptides. The ROC analysis of binding affinity predictions of the test peptides gave an Aroc=0.88, demonstrating that this matrix is a good solution. The predictive performance of the test B27 matrix is shown in Table 3. Finally, all 404 peptides were used to derive a B27 binding matrix (Table 4).

Predicted affinity	Experimental affinity			
	High	Mod.	Low	Zero
High	3	11	5	2
Moderate	2	1	0	3
Low	0	2	1	0
Zero	0	3	1	25

Table 3: Prediction performance of the preliminary B27 binding matrix derived using genetic search.

Discussion

Determining which peptides bind a particular MHC molecule critical to understanding the basis of immune responses, and has potential applications in the design of

Position	Amino acid																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
P1	-11	11	-14	-20	23	15	2	19	-7	-16	-9	14	-21	-6	11	6	-2	-24	20	-22
P2	-90	-20	-120	-110	-25	-85	-45	-75	-90	-110	-60	-120	-115	25	115	-115	-90	-110	-80	-60
P3	5	13	-24	12	2	-23	-10	4	-1	-7	-18	16	-16	-3	-23	-19	-14	-10	-20	14
P4	-19	-5	16	10	23	-11	-6	-15	11	15	23	7	-4	18	-15	-22	-3	-15	11	4
P5	9	-11	1	5	-3	-22	22	-2	4	-13	20	3	-21	-4	9	-13	-5	10	15	-20
P6	3	2	-20	-7	-17	-1	-22	-6	-8	10	-24	24	-24	-6	-14	8	-21	-7	-6	17
P7	22	-5	0	17	-7	-21	0	21	2	-17	24	24	-16	2	-20	-14	-3	-16	22	1
P8	-24	-14	-20	-2	-18	-16	0	-11	1	-11	-22	-20	-4	-8	21	-15	-4	-9	12	-22
P9	-20	-75	-40	35	-40	-105	-55	-25	10	-40	5	-80	-70	-35	-25	-110	-60	10	-65	-50

Table 4: The HLA-B27(*2705) binding matrix derived using genetic search. A predictive binding score is calculated by summing coefficients for amino acids that match the query peptide. Predictive scores of ≥ 40 correspond to high-, 30-39 to moderate- and 20-29 to low-affinity binding, and < 20 to non-binding.

peptide vaccines and other therapeutics. Tools to facilitate prediction of peptide/MHC binding have practical utility in minimising the number of direct binding experiments in the laboratory. Binding motifs (Rammensee *et al.*, 1995) can be used to identify potentially immunogenic peptides, but they do not generalise well. Binding matrices represent linear approximations of peptide/MHC interactions and can provide predictions of higher accuracy. Artificial neural networks (ANNs) have been used for prediction of peptide/MHC binding (Brusic *et al.*, 1994; Adams and Koziol, 1995) and demonstrate superior predictive ability. However, ANNs require even larger numbers of training peptides than matrices, which are not available except for a few MHC molecules. There are more than 100 variants of both human class I and class II MHC molecules and it is likely that matrices will remain primary tools for predicting peptide binding affinity to a broad range of MHC molecules.

In this contribution to the field, we developed binding matrices for several MHC molecules using peptide data and applied validation techniques as required for classifier systems. Insufficient numbers of peptides were used to properly validate previously reported matrices. This study should provide a basis for future definition or further refinement of binding matrices. Binding matrices can be derived experimentally, by combining binding experiments and statistical rank analysis to define matrix coefficients. The matrix coefficients should be contained between values of 0 and 2. The number of peptides required to derive good matrices (Aroc >0.8) should be at least 150 with adequate representations of both binding and non-binding peptides. In addition, a sufficient number of peptides should be used for validating the matrix and these peptides should represent a variety of amino acids allowed as primary anchors. Peptides used as a test set should preferably be specifically designed as a representative set.

When peptide binding data are available, but exact peptide binding affinities are lacking, a genetic search can be applied to derive binding matrices. Genetic search requires larger amounts of data; at least 300 peptides are required for the derivation of acceptable binding matrices. Proper validation of such derived binding matrices is also required. Preferably, the data set will contain a sufficient number of the peptides, whose finer binding affinity is known, to define a reasonable test set. The HLA-B27 matrix reported here is an example of such an approach. The available data included 404 peptides, with binders and non-binders almost equally represented. Binding affinity was known for only 29 binders (12.5%), but enough to create a reasonably-sized test set.

A further caveat on the present study is the representativeness of the test sets used. The test peptides for B35 and A24 binding matrices consisted of peptides which contain not all, but the commonly observed, anchor residues. The vast majority of peptides not containing primary anchor residues will not bind a specific MHC molecule and it is likely that the validation results for B35 and A24 are conservative estimates of matrix accuracy.

Comparison of the simulation and experimental results provides an interesting insight into the machine learning aspects of this study. It is well known that PAC estimates of the number of examples needed for learning are at least an order of magnitude higher than empirical estimates. However, in this work these numbers are relatively close, 1,800 and 600, assuming Aroc >0.9 . Further work may reveal why PAC estimates tend to be overly high.

This study illustrates the power of intersecting laboratory experimentation and computer modelling. This enables a cyclical flow of information between laboratory experiments and computer models, refining both in the process.

Acknowledgments

The help of Margaret Thompson and John Wilkins with the manuscript is gratefully acknowledged.

References

- Adams, H.P., and Koziol, J.A. 1995. Prediction of binding to MHC class I molecules. *Journal of Immunological Methods* 185:181–190.
- Benson, D.A., Boguski, M., Lipman, D.J., and Ostell, J. 1996. GenBank. *Nucleic Acids Research* 24:1–5.
- Bodmer, J.G., Marsh, S.G.E., Albert, E.D. *et al.* 1995. Nomenclature for factors of the HLA system, 1994. *Tissue Antigens* 46:1–18.
- Brusic, V., Rudy, G., and Harrison L.C. 1994. Prediction of MHC binding peptides using artificial neural networks. In Stonier, R., and Yu, X.H. eds. *Complex Systems: Mechanism of Adaptation*. pp. 253–260. Amsterdam: IOS Press.
- Brusic, V., Rudy, G., Kyne, A.P., and Harrison L.C. 1996. MHCPEP - a database of MHC-binding peptides: update 1995. *Nucleic Acids Research* 24:242–244.
- Chelvanayagam, G. 1996. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities, *Immunogenetics* 45:15–26.
- Cresswell, P. 1994. Assembly, transport, and function of MHC class II molecules, *Annual Review In Immunology* 12:259–293.
- Davenport, M.P, Shon, I.A.P.H., and Hill, A.V.S. 1995. An empirical method for the prediction of T-cell epitopes. *Immunogenetics* 42: 392–397.
- Falk, K., Rötzschke, O., Grahovac, B., Schendel, D., Stevanovic, S., Jung, G., and Rammensee, H.G. 1993. Peptide motifs of HLA-B35 and -B37 molecules. *Immunogenetics* 38:161–162.
- Falk, K., Rötzschke, O., Grahovac, B., Schendel, D., Stevanovic, S., Jung, G., and Rammensee, H.G. 1994. Peptide motifs of HLA-B35 and -B37 molecules (erratum). *Immunogenetics* 39:379.
- Goldberg, D.E. 1989. *Genetic algorithms in search, optimization and machine learning*. Mass.:Addison-Wesley
- Hammer, J., Bono, E., Gallazzi, F., Belunis, C., Nagy, Z., and Sinigaglia, F. 1994. Precise prediction of MHC class II-peptide interaction based on peptide side chain scanning. *Journal of Experimental Medicine* 180:2353–2358.
- Harrison, L.C., Honeyman, M.C., Tremblau, S., Gregori, S., Gallazzi, F., Augstein, P., Brusic, V., Hammer, J., and Adorini, L.A. 1997. A peptide binding motif for I-A^{g7}, the class II MHC molecule of NOD and Biozzi ABH mice. *Journal of Experimental Medicine* . Forthcoming.
- Holden, S.B., and Niranjan, M. 1995. On the practical applicability of VC dimension bounds. *Neural Computation* 4:1265–1288.
- Ibe, M., Ikeda Moore, Y., Miwa, K., Kaneko, Y., Yokota, S., and Takiguchi, M. 1996. Role of strong anchor residues in the effective binding of 10-mer and 11-mer peptides to HLA-A*2402 molecules. *Immunogenetics* 44:233–241.
- Jardetzky, T.S, Brown, J.H., Gorga, J.C., Stern, L.J., Urban, R.G., Strominger, J.L., and Wiley, D.C. 1996. Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proceedings of the National Academy of Sciences of the USA* 93:734–738.
- Kearns, M.J., and Vazirani, U.V. 1994. *An introduction to computational learning theory*. Mass.:MIT Press.
- Kondo, A., Sidney, J., Southwood, S., *et al.* 1995. Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class molecules. *Journal of Immunology* 155:4307–4312.
- Ljunggren, H.G., Stam, N.J., Ohlen, C. *et al.*, 1990. Empty MHC class I molecules come out in the cold, *Nature* 346:476–480.
- Maier, R., Falk, K., Rötzschke, O., Maier, B., Gnau, V., Stefanovic, S., Jung, G., Rammensee, H.G., and Meyerhans, A. 1994. Peptide motif of HLA-A3, -A24 and -B7 molecules as determined by pool sequencing. *Immunogenetics* 40:306–308.
- Nokihara, K., Yamamoto, R., Hazama, M., Wakizawa, O., and Nakamura, S. 1992. Design and applications of a novel simultaneous multiple solid-phase peptide synthesizer. In Epton, R., ed. *Innovation and Perspectives in Solid-Phase Synthesis*, pp. 445, Intercept:Andover.
- 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. *Journal of Immunology* 152:163–175.
- Rammensee, H.G., Falk, K., and Rötzschke, S. 1993. Peptides naturally presented by MHC class I molecules, *Annual Review In Immunology* 11:213–244.
- Rammensee, H.G., Friede, T., and Stevanovic, S. 1995. MHC ligands and peptide motifs: first listing. *Immunogenetics* 41:178–228.
- Rothbard, J.B., Marshall, K., Wilson, K.J., Fugger, L., and Zaller, D. 1994. Prediction of peptide affinity to HLA DRB1*0401. *International Archives of Allergy and Immunology* 105:1–7.
- Schönbach, C., Miwa, K., Ibe, M., Shiga, H., Nokihara, K., and Takiguchi, M. 1996. Fine tuning of peptide binding to HLA-B*3501 molecules by nonanchor residues. *Journal of Immunology* 154:5951–5958.
- Schönbach, C., Ibe, M., Shiga, H., Takamiya, Y., Miwa, K., Nokihara, K., and Takiguchi, M. 1995. Refined peptide-HLA-B*3501 binding motif reveals differences in 9-mer to 11-mer peptide binding. *Immunogenetics* 45:121–129.
- Schraudolph, N.N. and Grefenstette, J.J. 1992. A User's guide to GAUCSD 1.4. Technical Report, CS92-249, University of California, San Diego.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240:1285–1293.
- Takamiya, Y., Schönbach, C., Nokihara, K., Ferrone, S., Yamaguchi, M., Kyochi, K., Egawa, K., and Takiguchi, M. 1994. Role of anchor residues of peptides in their binding to HLA-B*3501 molecules. *International Immunology* 6:255–261.
- Valiant, L.G. 1984. A theory of the learnable. *Communications of the ACM* 27:1134–1142.