Molecular Medicine 8(3): 137–148, 2002

Methods for Prediction of Peptide Binding to MHC Molecules: A Comparative Study

Kun Yu,[1] Nikolai Petrovsky,[2] Christian Schönbach,[1,3] Judice L.Y. Koh,[1] and Vladimir Brusic[1]

[1]BIC-KRDL, Kent Ridge Digital Labs, Singapore
[2]National Bioinformatics Centre, University of Canberra, The Canberra Hospital, Woden, Australia
[3]Presently at Genomic Sciences Center RIKEN, Tsurumi-ku, Yokohama-shi, Kanagawa-ken, Japan

Accepted March 12, 2002. Contributed by R. Bucala.

Abstract

Background: A variety of methods for prediction of peptide binding to major histocompatibility complex (MHC) have been proposed. These methods are based on binding motifs, binding matrices, hidden Markov models (HMM), or artificial neural networks (ANN). There has been little prior work on the comparative analysis of these methods.
Materials and Methods: We performed a comparison of the performance of six methods applied to the prediction of two human MHC class I molecules, including binding matrices and motifs, ANNs, and HMMs.
Results: The selection of the optimal prediction method depends on the amount of available data (the number of peptides of known binding affinity to the MHC molecule of interest), the biases in the data set and the intended purpose of the prediction (screening of a single protein versus mass screening). When little or no peptide data are available, binding motifs are the most useful alternative to random guessing or use of a complete overlapping set of peptides for selection of candidate binders. As the number of known peptide binders increases, binding matrices and HMM become more useful predictors. ANN and HMM are the predictive methods of choice for MHC alleles with more than 100 known binding peptides.
Conclusion: The ability of bioinformatic methods to reliably predict MHC binding peptides, and thereby potential T-cell epitopes, has major implications for clinical immunology, particularly in the area of vaccine design.

Introduction

Major histocompatibility complex (MHC) molecules play a critical role in immune system function by binding short peptides and presenting them for recognition by T-cell receptors (TCR). Peptides presented by MHC class I molecules mostly originate from the cytoplasm (endogenous proteins or intracellular pathogens) whereas peptides presented by MHC class II molecules are mainly derived from exogenous antigens. Peptide-MHC complexes activate T cells to destroy abnormal (infected or neoplastic) or foreign (transplanted) cells.

The MHC has the highest level of polymorphism among all known functional loci in the human genome. The number of human leukocyte antigen (HLA) alleles characterized at the three classical HLA class I loci (A, B, and C) is shown in Table 1. Polymorphism in HLA molecules is concentrated around nucleotides encoding peptides adjoining the HLA peptide-binding groove. Each distinct HLA allele thereby encodes a slightly different peptide-binding domain. A discrete, albeit large, spectrum of peptides bind each particular HLA molecule and HLA-binding peptides are characterized by specific binding motifs (1). Peptides that bind HLA class I are 7–12 amino acids long (2). Because peptide binding to HLA is a prerequisite for peptide presentation and T-cell recognition, T-cell epitopes comprise a subset of HLA-binding peptides.

HLA class I restricted T-cell epitopes are potential vaccine candidates for use in immunization against cancer (3) or infectious disease (4,5). The ascertainment of which peptides bind to a given HLA molecule is a useful first step in the identification of T-cell epitopes. Although the number of peptides that can bind to a specific MHC molecule is large, it is approximately two orders of magnitude smaller than the number of peptides that can be generated by the degradation of protein antigens. T-cell epitope mapping, including HLA peptide-binding studies, is one of the most intensively researched areas of molecular and cellular immunology. Because of extensive HLA allelic variation, a systematic laboratory approach to T-cell epitope mapping, even of a single protein antigen, requires a large number of experiments. Consequently, prediction of peptide-MHC binding is useful for preselection of potential T-cell epitopes. For maximum benefit, computer models must be treated as experiments analogous to more typical wet-lab experimental procedures. This requires the definition of standards and protocols for application of predictive models (6).

The most important requirements for predictive modeling are accuracy, broad coverage and

Correspondence and reprint requests should be addressed to:
V. Brusic, 21 Heng Mui Keng Terrace, Singapore 119613.
Phone +65 874 7920; fax +65 774 8056.

Table 1. Number of allelic variants of classical HLA Class I Molecules*

Locus	Number of Functional Allelic Variants
HLA-A	200
HLA-B	416
HLA-C	92
Total	708

*As of October, 2001. Available from URL; http://www.anthonylan.com/HIG/index.html.

generalizability. Accuracy requires low levels of false-positive and false-negative predictions. Broad coverage refers to the ability of the model to predict different subsets of binding peptides and not only those that belong to the most numerous groups. For example, 9-mer peptides that are "ideal" binders to HLA-B15 molecules (with primary anchors at positions P2 and P9 in concordance with experimentally determined motifs) have been estimated to represent only 29% of all HLA-B15 binders (7). The majority of peptides for which experimental binding data are available are motif concordant, implicating strong biases in peptide preselection. Generalizability is an important quality of predictive models, which must be able to accurately generate new data rather than simply fit existing data to the model.

Peptide Preselection Using Theoretical Predictions

Several methods have been used for prediction of peptide binding to MHC molecules, including data binding motifs identified by pool sequencing (1), quantitative matrices (8–12), artificial neural networks (ANNs) (13–16), hidden Markov models (HMMs) (17), and molecular modeling (18–20). The majority of the methods predict whether peptides are binders to a given MHC, whereas some of the later developments focus on predicting peptide binding affinity to MHC receptors (19,20). The former are more suitable for a large-scale screening of potential T-cell epitopes, and the latter are computationally intensive and are better suited for detailed analysis of short immunogenic regions of antigens.

Binding motifs indicate the positions and the amino acids of primary anchors, secondary anchors, and other preferred or observed amino acids at specific positions within a peptide. Quantitative matrices provide coefficients for amino acids at each position within the peptide and an appropriate formula to calculate the peptide binding scores. Binding motifs and quantitative matrices assume the independent contribution of individual amino acids to peptide binding. ANN- and HMM-based predictions use more sophisticated computational algorithms that allow capturing of the complex patterns that define peptide binding. Molecular modeling uses detailed knowledge of the crystal structure of MHC molecules and of protein-peptide interactions. Molecular modeling has been useful for visualization and detailed analysis of pocket interactions in clefts of various MHC molecules (21), but this methodology is currently less useful for large-scale screening of potential HLA-binding peptides. Quantitative matrices based on amino acid yields during Edman degradation from multiple independent pool sequences of eluted self-peptides have been reported for several HLA class II molecules (22,23). This method quantifies amino acids adjoining the peptide binding core, which may also be involved in antigen processing.

Predictive models have been successfully used in the discovery of novel T-cell epitopes involved in cancer immunity (24–26), autoimmunity (27), infectious diseases (28,29), and allergies (30). Theoretical T-cell epitope predictions have been shown to minimize the time and cost of epitope mapping. However, before theoretical prediction methods can be used as standard methodology for T-cell epitope discovery and mapping, it is essential to first assess the accuracy, coverage, and potential biases these methods may introduce.

Early T-cell epitope prediction methods based on identifying amphipathic helices or generalized motifs performed poorly (31). Comparative performance of various predictive methods has been reported in several studies and these have shown binding motifs to be amongst the least accurate of all predictive methods. Gulukota et al. (15) compared the performances of quantitative matrices and ANNs for prediction of peptide binding to the class I molecule HLA-A*0201 and reported that quantitative matrices have high specificity and that ANNs have high sensitivity. Brusic et al. (16), compared the performance of an experimentally derived matrix and ANN for prediction of peptides that bind HLA-DRB1*0401, and found the ANN to be more accurate. This finding has been independently confirmed by Borrás-Cuesta et al. (32), who compared the performance (using HLA-DRB1*0401) of four different matrix-based methods and ANN. Mamitsuka (17) reported the superior predictive power of HMM compared to ANN models for prediction of peptide binding to HLA-A*0201. Finally, Andersen et al. (33) compared predictions of two publicly available prediction programs (SYFPEITHI and BIMAS) with the results of experimental peptide binding. They reported poor correspondence between predicted and experimental binding of peptides to two common HLA class I molecules. Although these findings are indicative of the relative strengths of individual methods, a systematic comparison of prediction methods for peptide–MHC binding has not been previously performed.

Here we report the results of a systematic comparison of six methods for prediction of peptide

binding to two HLA class I molecules, HLA-A*0201 and HLA-B*3501. These methods include binding motifs (SYFPEITHI), experimentally derived quantitative matrices (BIMAS), data-derived matrices (YK and YKW), ANNs, and HMMs. This comparison clarifies the strengths and deficiencies of these prediction methods. It also serves as a practical guide for the use of predictive models in identification of potential T-cell epitopes. To compare predictive performance, we considered multiple factors, including multiple measures of the goodness of prediction, influence of the number of peptides available for building the predictive system, effects of biases in the data set, validation method, scaling of prediction values, and inclusion of expert knowledge for the refinement of predictions. The performances of various methods were also tested for prediction of subsets of MHC-binding peptides, namely sets of reported T-cell epitopes, naturally processed peptides, poly-Ala peptides, and other synthetic peptides.

Materials and Methods

Experimental Peptide Binding Data

Binding sequences were extracted from MHCPEP, a database of MHC binding peptides (2). Nonbinding peptides (available from the authors upon request) were extracted from a collection of MHC experimentally determined nonbinding data retrieved from the same literature sources as the MHCPEP entries (V. Brusic, unpublished). Binding data have been derived using a variety of experimental methods and have been assigned descriptive values of high, moderate, low, and nonbinding as defined in the MHCPEP database (2). The initial HLA-A*0201 data set is composed of 1146 peptides and the expanded set includes 1230 peptides. The expanded HLA-A*0201 data set included an additional 60 T-cell epitopes and 24 naturally processed peptides. These peptides were added to the HLA-A*0201 set of the SYFPEITHI database after analysis of the initial set was completed. The HLA-B*3501 data set is composed of 234 peptides. The binders in each data set comprised reported T-cell epitopes, eluted peptides (naturally processed), poly-Alanine variants, and other synthetic peptides (Table 2). Peptide reported only as binders (eluted peptides) were assigned the moderate binding affinity for predictive modeling. Peptides reported as T-cell epitopes, but for which binding affinity was not determined, were also assigned moderate binding affinity. All binders were nine amino acids long; nonbinders also included longer peptides.

Quantitative Matrices and Motifs

SYFPEITHI (www.uni-tuebingen.de/uni/kxi/) is a predictive method based on scoring binding motifs. The SYFPEITHI matrix is based on the selection of positive or negative scores (based on the observed amino acid frequencies and the positions of primary anchors) to each amino acid at each position in the peptide. The scoring reflects the frequency of the respective amino acid in natural MHC ligands, T-cell epitopes, or binding peptides (12).

We have also assessed the predictive performance of several binding matrices. The HLA-A*0201 matrices include BIMAS (8; bimas.dcrt.nih.gov/molbio/hla_bind/), and two new matrices derived in this work, YK0201 and YKW0201. The BIMAS HLA-A*0201 matrix was derived experimentally from the measurements of half-time dissociation rates of peptide-HLA complexes. The YK0201 matrix was generated using logarithmic equations based on the frequency of amino acids at specific positions within the training set of peptides using the following formulae:

$$S_{AA} = \log_2\left(\frac{F_{AAp} - F_{AAn}}{F_{AAp} + F_{AAn}} + 1\right) if\ F_{Aap} > F_{AAn}$$

$$S_{AA} = -\log_2\left(\frac{F_{AAn} - F_{AAp}}{F_{AAp} + F_{AAn}} + 1\right) if\ F_{Aap} < F_{AAn}$$

Table 2. Data sets used for building prediction methods and the partitions used for assessing predictive performances

		Binders				
HLA	Nonbinders	T-cell Epitopes	Naturally Processed Peptides	Poly-Ala	Other Synthetic	Total Binders
A*0201a*	787	123	33	44	159	359
A*0201b†		183	57	44	159	443
B*3501	128	24	—	—	82	106

*Initial data set.
†Expanded data set.

where S_{AA} is the matrix coefficient for a particular amino acid at a particular position in the peptide, F_{AAp} refers to the proportion of a particular amino acid at the observed position within binders, and F_{AAn} refers to the proportion of a particular amino acid at the observed position within nonbinders. We weighted the columns of YK0201 matrices according to the reported binding motifs to derive the YKW0201 matrix. The respective division coefficients for columns 1–9 were 4, 1, 4, 4, 4, 5, 2, 3, and 1. The division coefficients were derived using a search program for the best fitting of binding data.

The B35CS matrix for predicting peptide binding to HLA-B*3501 was constructed from rank scores of amino acid frequencies in HLA-B*3501 binding peptides identified in a single study (10). We modified the coefficients from the original publications: non-anchor amino acids at the primary anchor positions P2 and P9 were assigned 0 values, and the proline at the anchor position P2 was assigned value of 1 (Table 3). The B35CS binding scores were calculated by multiplication of the coefficients for each peptide.

Artificial Neural Networks

ANNs are connectionist models commonly used for classification (34) and pattern recognition (35) tasks. In this study, ANN models were built as previously described (13,27). We trained a fully connected, three-layer, feed forward ANN using PlaNet software (36). The training set consisted of binding and nonbinding 9-mer peptides (Table 2). The ANN architecture is composed of 180 input units (corresponding to the binary representation of 9-mer peptides), two hidden layer units, and a single output unit. The learning algorithm was error back propagation (37). The ANN training was performed for 300 cycles. The values for momentum and learning rate were 0.5 and 0.2, respectively. Each prediction result was calculated as the average of four independent prediction runs. Each amino acid was encoded as a binary string of length 20 with a unique position set to "1" and all other positions set to "0." A 9-mer peptide was represented as a sparse binary string of length 180, sequentially combining representations of each amino acid. The output value was scaled 0–10, representing a range from no affinity to very high binding affinity. Binding scores used for ANN training were 1, 4, 6, and 8 for no-, low-, moderate-, and high-affinity binders, respectively.

Hidden Markov Models

HMM is a statistical model that can learn generalized probabilistic rules from data sets. The most

Table 3. B35CS matrix for prediction of peptide binding to HLA-B*3501

	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	0.27	0	0.88	0.91	1.30	0.86	1.14	1.06	0.00
C	1.29	0	1.31	1.53	1.32	1.23	1.09	1.37	0.00
D	1.54	0	0.91	0.75	1.50	1.25	0.88	1.18	0.00
E	1.54	0	0.91	0.75	1.50	1.25	0.88	1.18	0.00
F	1.31	0	0.25	0.91	0.25	1.00	0.60	0.40	1.30
G	0.27	0	0.88	0.91	1.30	0.86	1.14	1.06	0.00
H	1.31	0	0.25	0.91	0.25	1.00	0.60	0.40	0.00
I	1.42	0	1.65	1.04	1.43	1.06	2.00	1.11	0.64
K	0.83	0	0.54	1.00	0.39	0.67	0.38	0.44	0.00
L	1.42	0	1.65	1.04	1.43	1.06	2.00	1.11	1.04
M	1.42	0	1.65	1.04	1.43	1.06	2.00	1.11	0.50
N	0.56	0	0.86	0.38	1.25	1.36	1.22	1.17	0.00
P	0.00	1	1.00	0.91	0.88	1.25	0.75	1.00	0.00
Q	0.56	0	0.86	0.38	1.25	1.36	1.22	1.17	0.00
R	0.83	0	0.54	1.00	0.39	0.67	0.38	0.44	0.00
S	1.29	0	1.31	1.53	1.32	1.23	1.09	1.37	0.00
T	1.29	0	1.31	1.53	1.32	1.23	1.09	1.37	0.00
V	1.42	0	1.65	1.04	1.43	1.06	2.00	1.11	0.00
W	1.31	0	0.25	0.91	0.25	1.00	0.60	0.40	0.00
Y	1.31	0	0.25	0.91	0.25	1.00	0.60	0.40	1.56

common use of HMM in molecular biology is as a probabilistic profile of a protein family. HMMs are commonly used to model a family of unaligned sequences or a common motif within a set of unaligned sequences (38,39). The aligned peptides reported as binders to HLA-A*0201 or HLA-B*3501 were used to train first-order profile HMM models. HMM models were built using the HMMER package (40). The training set consisted of the alignment of the binding 9-mer peptides (Table 2). We used the program *HMMbuild* to create a profile HMM using the training set, *HMMcalibrate* to refine the model, and the *HMMsearch* to score the test data. *HMMcalibrate* was used with the fixed length of sequences of 9 amino acids; the cutoff expectation value in *HMMsearch* was set to $E = 60$. The observed range of output values was between -0.3 and -23. Higher values corresponded to predicted binders, lower values to predicted nonbinders.

Model Training and Validation

The accuracy of matrices (YK0201 and YKW0201) and HMM-based predictions was assessed by the leave-one-out cross-validation (34). The variation of the method used in this study included several steps: a) a single peptide (test peptide) was removed from the peptide set, b) all peptides that differed by only a single amino acid from the test peptide were removed, c) a predictive model was generated using all the remaining peptides, d) the binding of the test peptide was predicted, e) test peptide was returned into the peptide set, f) steps a–e were repeated until all peptides are exhausted, and g) the accuracy of the model was assessed using prediction values for all test peptides. Because of the long time required for training ANN models, leave-one-out cross-validation could not be used, and the accuracy of the ANN models was assessed by a 10-fold cross-validation (34). The variation of the method applied in this study consisted of the following steps: a) a data set was randomly divided into 10 partitions, each of which had mutually exclusive training and test sets; the 10 test sets were also mutually exclusive, b) in each partition, all peptides in the training set that differed by only a single amino acid from any of the test peptides of the same partition were removed from the corresponding training set, c) a training set was used to build a predictive model, d) binding of the peptides from the test set was predicted with the corresponding model, e) steps a–d were repeated for each partition, and f) the accuracy of the model was assessed using prediction values for each peptide. For SYFPEITHI and BIMAS methods, all peptides from the HLA-A*0201 set were tested and these results were used for assessing the respective predictive performances. All peptides from the HLA-B*3501 set were used to assess the predictive performance of the B35CS matrix. The testing criteria slightly favored SYFPEITHI and BIMAS methods because the testing set included some of the peptides used to derive their predictive models. The testing criteria were neutral for all other methods using leave-one-out method, and slightly unfavorable for ANN predictions (because ANN models were trained on data sets composed of 90% of the total number of peptides). The exclusion from the training test of all peptides highly similar to the test peptide was done to minimize bias and enhance the assessment of generalization ability of the tested models.

Nonbinding peptides were decomposed into overlapping (by one amino acid) 9-mers that were used for model training, as previously reported (41). For testing, each nonbinding peptide was considered as a single data point. If any one of the subsequences within a nonbinding peptide was predicted as a binder, the whole peptide was considered a false positive.

Measures of Predictive Accuracy

The accuracy of predictive models was assessed using common statistical measures of sensitivity (SE) and specificity (SP). SE indicates the quantity of predictions (the proportion of correctly predicted true positives). SP indicates the quality of predictions, namely the proportion of correctly predicted true negative examples. These measures are calculated by the formulae $SE = TP/(TP + FN)$ and $SP = TN/(TN + FP)$, where TP stands for true positives (peptides that are both predicted and experimentally measured as binders), TN for true negatives (both predicted and measured as nonbinders), FP for false positives (predicted as binders but measured as nonbinders), and FN for false negatives (predicted as nonbinders but measured as binders).

To better understand the overall performance of studied predictors, we also used receiver operating characteristic (ROC) curve analysis (42). ROC analysis provides the A_{roc} measure by integration of the function $SE = f(1 - SP)$ for various decision thresholds. The area under the ROC curve (A_{roc}) provides a single measure of the accuracy of predictive models. Values of $A_{roc} = 50\%$ indicate random choice; $A_{roc} > 80\%$, moderate accuracy; and $A_{roc} > 90\%$, high prediction accuracy (42).

Comparison of Predictive Performances

Predictive performance was assessed for all peptides and various subgroups: T-cell epitopes, naturally processed peptides, synthetic binders, or poly-Ala peptides (Table 2). We compared the overall performance for each method by calculating A_{roc} values. High specificity predictions of peptide–MHC binding are needed because high-affinity binding peptides occur at low frequency in naturally processed antigens. A broad estimate (6) is that between 0.1% and 5% of the overlapping peptides derived from a typical protein may bind to a given MHC molecule. In practice, it means that a low threshold for predicted binders results in high sensitivity but at the expense of low specificity due to the overwhelming

number of false positive predictions. In addition to the assessment of the overall accuracy of predictions, we compared the performance of individual predictive methods for values of specificity > 0.8.

Results

*Overall Prediction of Peptide Binding to HLA-A*0201*

The A_{roc} values for the studied methods (Fig. 1) ranged from 0.81–0.87. These values indicate good overall performance for each method. The A_{roc} values of predictions using ANN, HMM, and YKW0201 were slightly higher than those using BIMAS and SYFPEITHI methods. The frequency-based matrix YK0201 produced the lowest A_{roc} value of all studied methods. The 23% increase in the number of peptides available for the HLA-A*0201 model refinement (Table 2) was expected to improve the accuracy of the data-driven prediction methods (YK0201, YKW0201, ANN, and HMM), but not the accuracy of experimentally derived matrix (BIMAS) or motif-based prediction (SYFPEITHI) methods. Unexpectedly, except for the ANN method, predictions performed using the expanded set of HLA-A*0201 binders (443 binding peptides) produced similar or lower A_{roc} values than predictions using the initial data set (359 binding peptides). To clarify the discrepancy between the expected and observed results, we reassessed the predictive performance of all methods using subsets of HLA-A*0201 binding peptides.

*Binding of Peptide Subsets to HLA-A*0201*

The A_{roc} values of predictions by peptide subsets are shown in Figure 2. The predictions of the subset of peptides corresponding to the reported T-cell epitopes (Fig. 2a) were similar to the overall results shown in Figure 1. The A_{roc} values of the T-cell epitope subset ranged from 0.81–0.87. For T-cell epitope predictions, BIMAS, ANN, or HMM predictions had a slightly higher accuracy than other three studied methods.

The A_{roc} values of predictions of the subset of naturally processed peptides (33 in the initial set and 57 in the expanded set) showed a significant difference for predictions using the initial set of peptides (787 nonbinders and 359 binders) and the expanded set (787 nonbinders and 443 binders). The highest accuracy was achieved by the SYFPEITHI method, followed by the ANN- and BIMAS-based predictions (Fig. 2b). The A_{roc} values for predictions of naturally processed peptides using the extended set were significantly lower ($p = 0.0008$, paired t-test) than the predictions using the initial set across all predictive models.

The prediction of artificial poly-Ala sequences proved to be of high accuracy ($A_{roc} > 0.9$) for all methods except in BIMAS. These sequences have little relevance for the prediction of biologically relevant subsets of T-cell epitopes or naturally processed peptides. Indeed, the removal of the poly-Ala peptides from the training sets did not affect the prediction of biologically relevant peptide sets (data not shown).

The A_{roc} values for predictions of the subset of other synthetic peptides were generally lower than for other subsets, with the ANN, HMM, and YKW models yielding the highest accuracy. There was no difference of A_{roc} values between predictions using the initial data set and those using the expanded data set, except for the ANN-based predictions. The ANN predictions using the expanded data set improved the accuracy of predictions of synthetic peptides consistent with ANN-based methods being

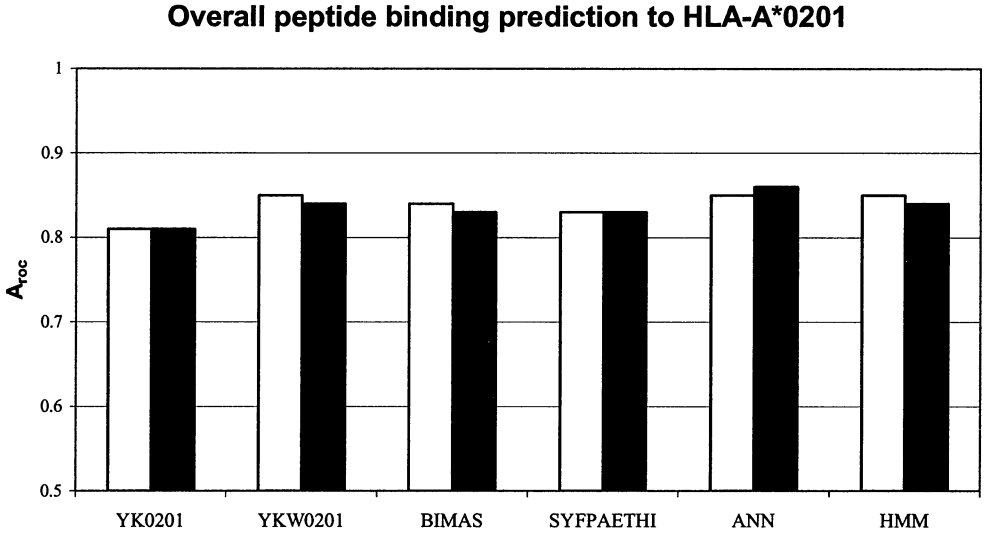

Fig. 1. A_{roc} values of six models cross-validated for prediction of peptide binding to HLA-A*0201. White bars represent predictions with the initial HLA-A*0201 data set, black bars represent predictions with the expanded data set, as described in Materials and Methods.

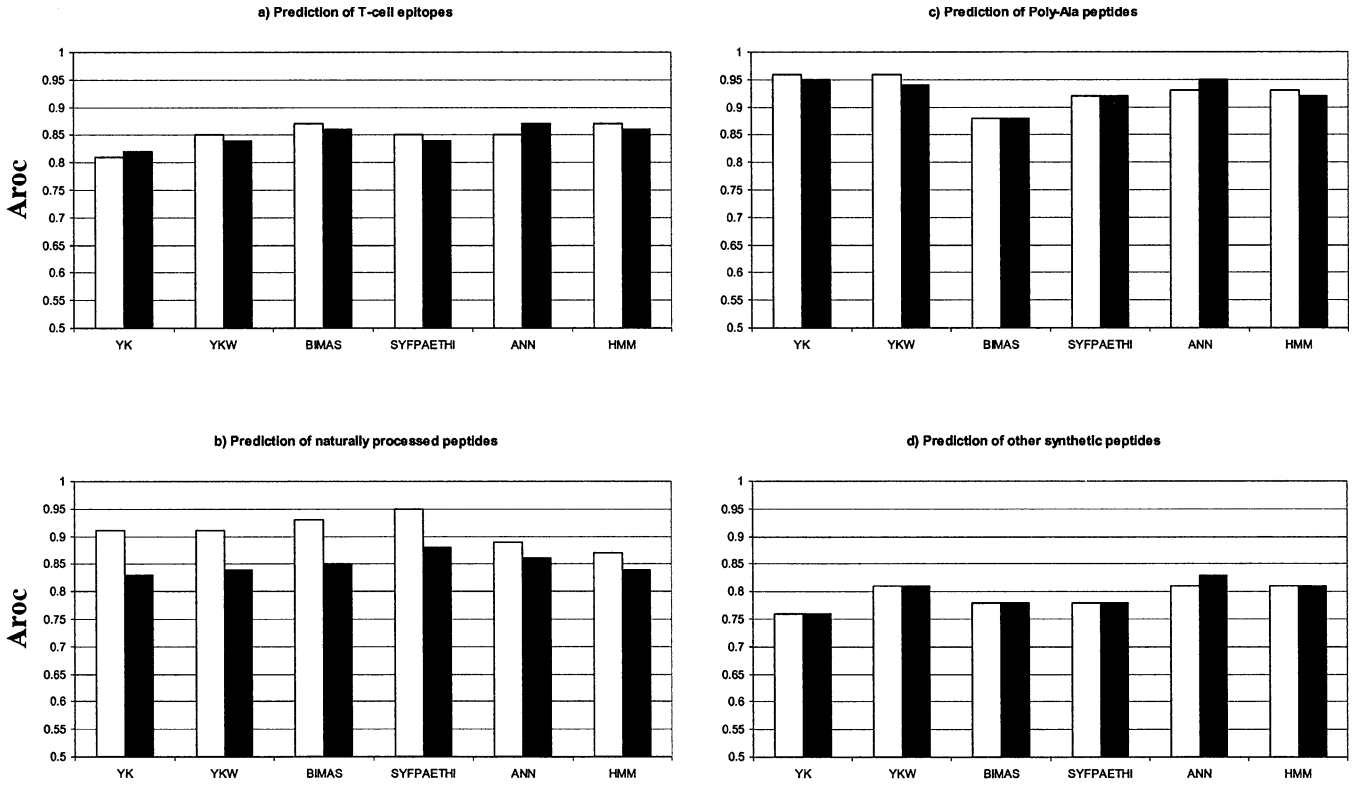

Fig. 2. A_{roc} **values of six models cross-validated for prediction of binding of peptide subsets to HLA-A*0201.** Peptide represented in subsets are exclusively (as reported) T-cell epitopes **(a),** naturally processed peptides peptide **(b),** poly-A **(c),** and synthetic peptides **(d).** White bars represent predictions with the initial HLA-A*0201 data set, and dark bars represent predictions with the expanded data set, as described in Materials and Methods.

robust and capable of being refined by addition of further peptide binding data. Heuristic rules describing relative importance of specific positions within peptides are also useful. These rules, extracted from binding motifs, can be used for modification of frequency-based quantitative matrices for higher accuracy.

*Analysis of HLA-A*0201 Binding Motifs*

Historically, the identification of MHC-binding peptides has often involved the preselection of peptides based on the presence of binding motifs. This has resulted in an overrepresentation of motif-containing peptides in the available data sets. We have therefore analyzed the subsets of peptide binding data for the presence of binding motifs (Table 4). The proportion of non-motif containing peptides in the initial data set (T-cell epitopes and naturally processed peptides) is lower than in the expanded data set.

The analysis of the peptide sets for presence of basic and extended motifs established that there is a positive correlation between the overall accuracy of the predictive models and the number of peptides containing the extended motif x(T,A,V,I,L,M)xxxxxx (T,A,M,I,V,L) (41). Representative correlation plots are shown in Figure 3. The correlation coefficients in ANN- and HMM-based prediction were lower than in other methods (Table 5). The r^2 coefficient indicates that for HMM or ANN methods approximately 70% of predictions can be explained by the presence of the extended motif. For remaining methods, the presence of the extended motif can explain approximately 80% of predictions. Of the six methods studied, the accuracy of ANN predictions was the highest and that of HMM equal second (Fig. 1). Together with the correlation analysis, this indicates that ANN and HMM models tend to capture binding patterns of higher complexity than other methods studied here. Higher complexity patterns and stronger generalization properties have been emphasized in ANN models, which showed a general increase in the accuracy of predictions with the increase of the size of the training set of peptides.

*High Specificity Predictions of Peptide Binding to HLA-A*0201*

The comparison of performance of the studied methods for values of SP $>$ 0.8 is shown in Figure 4. The best performance using the initial data set (Fig. 4A and 4B) was achieved by the ANN and BIMAS methods. The frequency-based weighted matrix YKW was third best; HMM, SYFPEITHI, and YK methods were less accurate. The accuracy of the models derived/tested by the expanded data set (Fig. 4C and 4D)

Table 4. Binding motif in HLA-A*201 binding peptides

Peptide Set	Data Set	Peptides	
		Basic Motif*	Extended Motif†
T-cell epitopes	Initial	75 (61%)	114 (92.7%)
	Extended	75 + 28 (56.3%)	114 + 55 (92.3%)
Naturally processed	Initial	13 (39.4%)	33 (100%)
	Extended	13 + 14 (47.4%)	33 + 17 (87.7%)
Poly-Ala	Initial	36 (81.8%)	43 (97.7%)
Other synthetic	Initial	80 (50.3%)	138 (86.8%)
Nonbinders	Initial	23 (2.9%)	243 (30.1%)

*Basic motif defined as x(L,M)xxxxxx(V,L).
†Extended motif defined as x(T,A,V,I,L,M)xxx(T,A,M,J,V,L)

remained unchanged, except for the ANN-based method whose prediction accuracy markedly improved. This finding is consistent with the overall improvement of the ANN-based predictions with the increased size of the peptide set (Fig. 1).

*Prediction of Peptide Binding to HLA-B*3501*

The A_{roc} values for the studied methods (Fig. 5) ranged from 0.65–0.75, indicating poor overall performance across all methods. The A_{roc} values of predictions using experimentally derived matrix

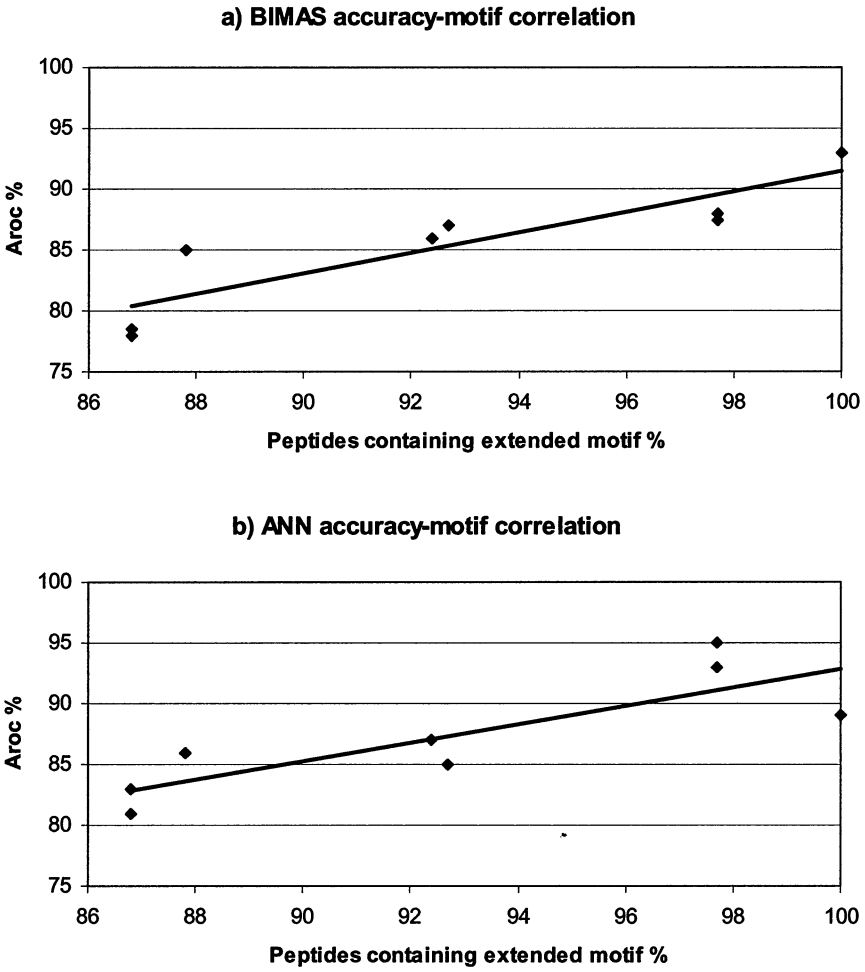

Fig. 3. The accuracy-motif correlation. Representative graphs showing correlation between the accuracy of **(A)** BIMAS and **(B)** ANN predictions and the percentage of peptides containing the extended motif x(T,A,V,I,L,M)xxxxxx(T,A,M,I,V,L). The correlation coefficient values are given in Table 3.

Table 5. Correlation between A_{roc} values and the percentage of peptides containing extended binding motif

Method	YK	YKW	BIMAS	SYFPEITHI	ANN	HMM
Correlation coefficient r	0.89	0.90	0.90	0.88	0.84	0.85
r^2	0.79	0.81	0.81	0.77	0.70	0.72

B35SC, HMM, and YKW0201 were 0.75, 0.75, and 0.72, respectively. Empirically (42), the values $0.7 < A_{roc} < 0.8$ indicate that these methods are potentially useful, but not of high predictive accuracy. ANN, BIMAS, and YK methods had A_{roc} values lower than 0.7, indicating that they are not good predictors with the current data set. Together with the results for HLA-A*0201 predictions, these results show the importance of the number of peptides used to develop predictive models of peptide binding.

Conclusions and Discussion

This comparative analysis demonstrates that no predictive method of peptide–MHC binding consistently outperforms the rest. Rather, the most appropriate predictive model depends on the amount of available data (the number of peptides of known binding affinity for the HLA molecule of interest), the extent of bias in the training data set, and the intended purpose of the prediction. A number of conclusions can be drawn from these results.

Experimentally derived binding matrices, such as BIMAS HLA-A*0201 and B35CS, are useful when relatively small amounts of data are available for the development of predictive models. Over time, as increasing HLA binding peptide data become available, binding matrices will be outperformed by data-driven models such as frequency-based matrices, ANN, or HMM. If a relatively small number of data (approximately 100 binding peptides) are available, the methods of choice are HMM or frequency-based weighted matrices, such as YKW0201 or YKW3501. When larger amounts of data are available, ANN or HMM are the methods of choice. Provided that binding data on a sufficient number of peptides are available, ANN is the most useful method to provide high specificity although at the cost of slightly lower sensitivity. This will be useful where large numbers of proteins are to be scanned to identify at least some high-probability HLA binding peptides. On the other hand, where high sensitivity is required (e.g., screening a single protein for all binding peptides for a particular HLA molecule), HMM may be the most suitable method.

Currently available HLA binding data sets contain significant biases. This is due to selection of peptides for laboratory binding verification based on their similarity to the proposed binding motifs. This results in overrepresentation of such peptides with a relative paucity of random "non-motif-bearing" peptides. Although all of the predictive methods studied here inevitably encode biases present in the data, ANN and HMM appear to be less bias sensitive. ANN and HMM provide better predictions, therefore, of binding peptides that do not fit proposed consensus motifs.

Although quantitative matrices are simple to use they are based on the assumption that each amino acid in a peptide contributes independently to overall binding. The analysis of MHC-peptide crystal structures suggests that binding energy for individual amino acid residues within the peptide is not independent of neighboring residue effects (44). This may account for the failure of matrix-based binding predictions to identify good binders that do not contain proposed binding motifs. By comparison, ANNs have the advantage that they can a) generalize from input data, b) tolerate noise and errors in data, c) deal with nonlinear problems (i.e. beyond simple or extended motifs), and d) are adaptive and self-refine with the addition of new data. The disadvantage of ANNs is that it is difficult to extract the explanation and rules learnt from the data. The poor performance of the ANN for prediction of HLA-B*3501 binding is likely to reflect the insufficient size of the training set of peptides. HMM-based methods for making predictions of peptide binding have not been extensively studied previously. Our results do, however, suggest that HMMs provide similar advantages as ANNs when used for this purpose. An additional advantage of HMMs is that they require a smaller number of training peptides than ANNs to build useful predictive models. The predictions using quantitative matrices can potentially be improved by defining multiple high-specificity matrices for a single HLA molecule.

Computational binding predictions provide useful data complementary to wet-lab experimentation. Predictions are useful for peptide selection for binding studies, planning of experiments, and better understanding of biological processes. Given the large amount of peptide binding data that are publicly available in respect of human and mouse MHC molecules (2) (data is available for over two dozen human HLA alleles) plus data available privately, ANN or HMM will increasingly be the favored

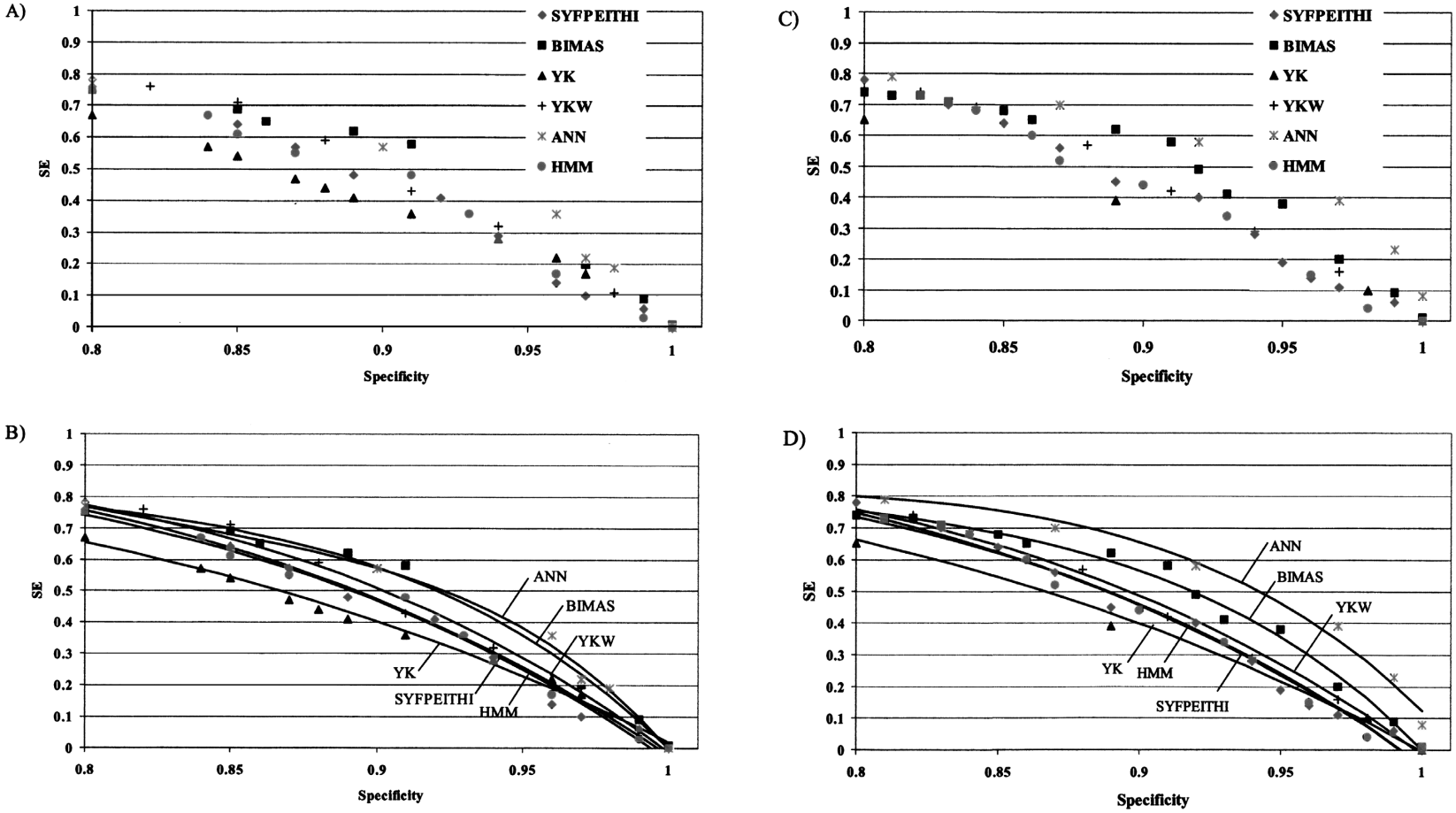

Fig. 4. Values of sensitivity (SE) for various specificity (SP) thresholds. **(A)** Initial HLA-A*0201 peptide set. **(B)** Initial HLA-A*0201 set with fitted curves. **(C)** Extended HLA-A*0201 peptide set. **(D)** Extended HLA-A*0201 peptide set with fitted curves.

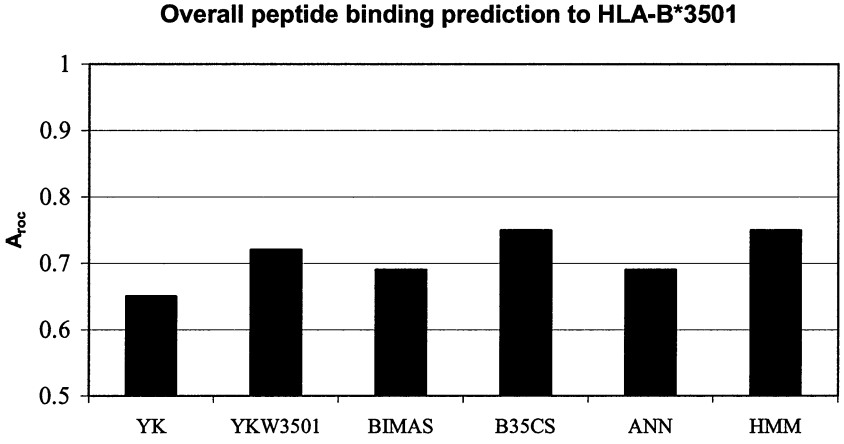

Fig. 5. A_{roc} **values of six models cross-validated for prediction of peptide binding to HLA-B*3501.**

predictive methods. Use of HMM or ANN methods to complement laboratory experiments will increase the efficiency of T-cell epitop e screening. The combination of accurate binding predictions with new experimental methods for identification of T-cell epitopes will allow tracking of antigen-specific CTL responses in clinical studies. These new methods include the use of CTL-independent high performance liquid chromatography mass spectrometry (45) for identification of naturally processed peptides, use of soluble MHC class I-peptide tetramers (46), microarrays (47), and combinatorial peptide libraries (48). The ability of bioinformatic methods to reliably predict MHC binding peptides and thereby potential T-cell epitopes clearly has major implications for clinical immunology, particularly in the area of vaccine design.

Acknowledgments

N.P. is supported by a grant from the Canberra Hospital Salaried Specialists Private Practice Fund.

References

1. Rammensee HG, Friede T, Stevanoviic S. (1995) MHC ligands and peptide motifs: 1st listing. *Immunogenetics* **41:** 178–228.
2. Brusic V, Rudy G, Harrison LC. (1998) MHCPEP—a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.* **26:** 368–371.
3. Wang RF, Rosenberg SA. (1999) Human tumor antigens for cancer vaccine development. *Immunol. Rev*. **170:** 85–100.
4. Wang R, Doolan DL, Le TP, et al. (1998) Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine. *Science* **282:** 467–480.
5. Berzofsky JA, Ahlers JD, Derby MA, et al. (1999) Approaches to improve engineered vaccines for human immunodeficiency virus and other viruses that cause chronic infections. *Immunol. Rev*. **170:** 151–172.
6. Brusic V, Zeleznikow J. (1999) Computational binding assays of antigenic peptides. *Letters in Peptide Science* **6:** 313–324.
7. Prilliman KR, Jackson KW, Lindsey M, et al. (1999) HLA-B15 peptide ligands are preferentially anchored at their C termini. *J. Immunol*. **162:** 7277–7284.
8. Parker KC, Bednarek MA, Coligan JE. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide sidechains. *J. Immunol*. **152:** 163–175.
9. Hammer J, Bono E, Gallazzi F, et al. (1994) Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J. Exp. Med*. **180:** 2353–2358.
10. Schönbach C, Ibe M, Shiga H, et al. (1995) Fine tuning of peptide binding to HLA-B*3501 molecules by nonanchor residues. *J. Immunol.* **154:** 5951–5958.
11. Mallios RR. (1999) Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* **15:** 432–439.
12. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50:** 213–219. Available at URL: www.uni-tuebingen.de/uni/kxi/.
13. Brusic V, Rudy G, Harrison LC. (1994) Prediction of MHC binding peptides using artificial neural networks. In Stonier RJ, Yu XS (eds). *Complex Systems: Mechanism of Adaptation*, Amsterdam/OHMSHA Tokyo: IOS Press; 253–260. Also published in *Complexity International* **2:** 1995.
14. Adams HP, Koziol JA. (1995) Prediction of binding to MHC class I molecules. *J. Immunol. Methods* **185:** 181–190.
15. Gulukota K, Sidney J, Sette A, Delisi C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* **267:** 1258–1267.
16. Brusic V, Rudy G, Honeyman M, Hammer J, Harrison LC. (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* **14:** 121–130.
17. Mamitsuka H. (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* **33:** 460–474.
18. Lim JS, Kim S, Lee HG, et al. (1996) Selection of peptides that bind to the HLA-A2.1 molecule by molecular modeling. *Mol. Immunol*. **33:** 221–230.
19. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V. (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem*. **42:** 4650–4658.
20. Doytchinova IA, Flower DR. (2001) Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J. Med. Chem*. **44:** 3572–3581.
21. Zhang, C., Anderson, A. and DeLisi, C. (1998) Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol*. **281,** 929–947.
22. Davenport MP, Ho Shon IA, Hill AV. (1995) An empirical method for the prediction of T-cell epitopes. *Immunogenetics* **42:** 392–397.
23. Godkin AJ, Davenport MP, Willis A, et al. (1998) Use of complete eluted peptide sequence data from HLA-DR and -DQ

molecules to predict T cell epitopes, and the influence of the nonbinding terminal regions of ligands in epitope selection. *J. Immunol*. **161:** 850–858.
24. Manici S, Sturniolo T, Imro MA, et al. (1999) Melanoma cells present a MAGE-3 epitope to CD4(+) cytotoxic T cells in association with histocompatibility leukocyte antigen DR11. *J. Exp. Med*. **189:** 871–876.
25. Vissers JL, De Vires JJ, Schreurs MW, et al. (1999) The renal cell carcinoma-associated antigen G250 encodes a human leukocyte antigen (HLA)-A2.1-restricted epitopes recognized by cytotoxic T lymphocytes. *Cancer Res*. **59:** 5554–5559.
26. Zarour HM, Kirkwood JM, Kierstead LS, et al. (2000) Melan-A/MART-1(51–73) represents an immunogenic HLA-DR4-restricted epitope recognized by melanoma-reactive CD4(+) T cells. *Proc. Natl. Acad. Sci. U.S.A.* **97:** 400–405.
27. Honeyman MC, Brusic V, Stone NL, Harrison LC. (1998) Neural network-based prediction peptides binding major histocompatibility complex molecules. *Nat. Biotechnol.* **16:** 966–969.
28. Khanna R, Burrows SR, Nicholls J, Poulsen LM. (1998) Identification of cytotoxic T cell epitopes within Epstein-Barr virus (EBV) oncogene latent membrane protein 1 (LMP1): evidence for HLA A2 supertype-restricted immune recognition of EBV-infected cells by LMP1-specific cytotoxic T lymphocytes. *Eur. J. Immunol.* **28:** 451–458.
29. Jin X, Roberts CG, Nixon DF, et al. (2000) Identification of subdominant cytotoxic T lymphocyte epitopes encoded by autologous HIV type 1 sequences, using dendritic cell stimulation and computer-driven algorithm. *AIDS Res. Hum. Retroviruses.* **16:** 67–76.
30. De Lalla C, Sturniolo T, Abbruzzese L, et al. (1999) Cutting edge: identification of novel T cell epitopes in Lol p5a by computational prediction. *J. Immunol*. **163:** 1725–1729.
31. Deavin AJ, Auton TR, Greaney PJ. (1996) Statistical comparison of established T-cell epitope predictors against a large database of human and murine antigens. *Mol. Immunol*. **33:** 145–155.
32. Borrás-Cuesta F, Golvano J, Garcia-Granero M, et al. (2000) Specific and general HLA-DR binding motifs: comparison of algorithms. *Hum. Immunol*. **61:** 266–278.
33. Andersen MH, Tan L, Sondergaard I, et al. (2000) Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules. *Tissue Antigens*, **55:** 519–531.
34. Weiss SM, Kulikowski CA. (1990) *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufman Publishers.
35. Beale R, Jackson T. (1990) *Neural Computing: An Introduction*. Bristol, UK: Adam Hilger.
36. Miyata Y. (1991) *A User's Guide to Planet Version 5.6*. Boulder, CO: Computer Science Department, University of Colorado.
37. Rumelhart DE, Hinton E, Williams J. (1986) Learning internal representation by error propagation. In Rumelhart D, McClelland J, and the PDP Research Group (eds). *Parallel Distributed Processing, Vol. 1*. Cambridge, MA: MIT Press; 318–362.
38. Hughey R, Krogh A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci*. **12:** 95–107.
39. Krogh A, Brown M, Mian IS, et al. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.
40. Eddy SR. (1998) HMMer user's guide (version 2.1.1). *Profile hidden Markov models for biological sequence analysis.* Available at URL: http://hmmer.wustl.edu/hmmer-html/.
41. Brusic V, Bucci K, Schonbach C, et al. (2001) Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding. *J. Mol. Graph. Model*. **19:** 405–411, 467.
42. Swets JA. (1988) Measuring the accuracy of diagnostic systems. *Science* **240:** 1285–1293.
43. Kast WM, Brandt RM, Sidney J, et al. (1994) Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J. Immunol.* **152:** 3904–3912.
44. Madden DR, Garboczi DN, and Wiley DC. (1993) The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **75:** 693–708.
45. Schirle M, Keilholz W, Weber B, et al. (2000) Identification of tumor-associated MHC class I ligands by a novel T cell-independent approach. *Eur. J. Immunol*. **30:** 2216–2225.
46. Altman JD, Moss PA, Goulder PJ, et al. (1996) Phenotypic analysis of antigen-specific T lymphocytes. *Science* **274:** 94–96.
47. Mathiassen S, Lauemoller SL, Ruhwald M, et al. (2001) Tumor-associated antigens identified by mRNA expression profiling induce protective anti-tumor immunity. *Eur. J. Immunol*. **31:** 1239–1246.
48. Linnemann T, Tumenjargal S, Gellrich S, et al. (2001) Mimotopes for tumor-specific T lymphocytes in human cancer determined with combinatorial peptide libraries. *Eur. J. Immunol*. **31:** 156–165.