View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publications of the IAS Fellows

A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes

Manoj Bhasin and G P S Raghava*
Institute of Microbial Technology, Sector 39A, Chandigarh 160 036, India
**Corresponding author (Fax, 91-172-690632; Email: raghava@imtech.res.in)*

In the present study, a systematic attempt has been made to develop an accurate method for predicting MHC class I restricted T cell epitopes for a large number of MHC class I alleles. Initially, a quantitative matrix (QM)-based method was developed for 47 MHC class I alleles having at least 15 binders. A secondary artificial neural network (ANN)-based method was developed for 30 out of 47 MHC alleles having a minimum of 40 binders. Combination of these ANN- and QM-based prediction methods for 30 alleles improved the accuracy of prediction by 6% compared to each individual method. Average accuracy of hybrid method for 30 MHC alleles is 92.8%. This method also allows prediction of binders for 20 additional alleles using QM that has been reported in the literature, thus allowing prediction for 67 MHC class I alleles. The performance of the method was evaluated using jack-knife validation test. The performance of the methods was also evaluated on blind or independent data. Comparison of our method with existing MHC binder prediction methods for alleles studied by both methods shows that our method is superior to other existing methods. This method also identifies proteasomal cleavage sites in antigen sequences by implementing the matrices described earlier. Thus, the method that we discover allows the identification of MHC class I binders (peptides binding with many MHC alleles) having proteasomal cleavage site at C-terminus. The user-friendly result display format (HTML-II) can assist in locating the promiscuous MHC binding regions from antigen sequence. The method is available on the web at *www.imtech.res.in/raghava/nhlapred* and its mirror site is available at *http://bioinformatics.uams.edu/mirror/nhlapred/*.

[Bhasin M and Raghava G P S 2006 A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes; *J. Biosci.* **32** 31–42]

1. Introduction

One of the crucial steps in designing subunit vaccine for diseases like cancer involves identification of antigenic peptides that can stimulate cytotoxic T lymphocytes (CTLs). The binding of antigenic peptide to MHC class I molecule is a prerequisite for their recognition by CTLs (Cresswell *et al* 1999). Determination of binding specificity of peptides derived from a protein to MHC class I molecules requires the binding assay of hundreds of multiple overlapping peptides spanning the whole protein. This is a very laborious and time-consuming task. The utility of computational methods to screen the initial hundreds of peptides and to determine the probable candidate peptides for further experiments has been successfully tested in the past (Vordermeier *et al* 2003).

A number of methods have been developed for predicting MHC binders in an antigen sequence on the basis of rules that govern the binding of a peptide to MHC molecule. Most of these methods are knowledge-based where experimentally identified binders are used to derive the rules. These methods can be broadly classified in to four categories; (i) methods based on binding motif; (ii) quantitative matrix based methods; (iii) machine learning techniques and (iv) *ab initio* or structure based methods. In binding motifs-based algorithms, the binding of peptides to

Keywords. Artificial neural network; MHC class I alleles; promiscuous binders; proteasomal cleavage site; quantitative matrices.

Abbreviations used: ANN, Artificial neural network; CTL, cytotoxic T lymphocytes; QM, quantitative matrix; SSE, Sum of Squared Error function.

an allele is examined on the basis of occurrence of specific residues (anchor residues) at specific positions (anchor positions) (Rammensee *et al* 1999). The presence of motifs determines whether a peptide will bind to a specific allele or not. The methods based on motifs have low accuracy of prediction because all the peptides binding with MHC do not have exact motifs (Buus 1999).

In order to overcome the limitations of motif-based approach, quantitative/profile matrix based methods have been developed (Parker *et al* 1994; Singh and Raghava 2001). The quantitative matrix (QM)-based methods consider the contribution of each residue of a peptide rather than just anchor positions/residues. The matrix-based methods predict MHC binding peptides with fair accuracy. One major limitation of the quantitative matrix based methods is its inability to handle the non-linearity in the data. The machine learning techniques can handle the non-linearity in data in a facile way. Such techniques like artificial neural networks (ANNs) have been previously applied for the prediction of MHC binders (Adams and Koziol 1995). The ANN correctly classified the peptides as binders or non-binders in 78% of cases for HLA-A2 and 88% of cases for H-2Kb MHC alleles (Brusic *et al* 1994). ANN was also able to correctly predict 78% of binders for HLA-A*0201 (Adams and Koziol 1995). The ANN-based method predicts fewer false positive binders in comparison to motifs based algorithms (Gulukota *et al* 1997). Thus, It is more reliable in classifying the non-linear data of MHC binders and non-binders in comparison to algorithms based on quantitative matrix and motifs-based. The major constraint in ANN-based methods is that they requires large dataset for training. In case of *ab initio* methods, prediction is based on structural binding of MHC molecule and peptides (Schueler-Furman *et al* 2000; Doytchinova and Flower 2001). There are two major limitations of *ab initio*-based methods: (i) they are very slow and (ii) not possible to develop method for most of the MHC alleles due to unavailability of three-dimensional structural data of these alleles. The high proportion of false positive binders in prediction is another limitation of these methods.

The existing methods have two major drawbacks. First, most of the existing methods have been developed for one or two most common MHC alleles (Brusic *et al* 1994;Adams and Koziol 1995; Honeyman *et al* 1998). These methods are not suitable for identification of potential vaccine candidates, because an ideal candidate should have the capability to bind with many MHC alleles. The peptides binding with variety of MHC alleles are known as promiscuous MHC binders (Hammer *et al* 1993; Sturniolo *et al* 1999; Singh and Raghava 2003). In past, attempts have been made to address this problem by developing methods for prediction for large number of alleles, for example Tepitope and ProPred for MHC class II and ProPred1 for MHC class I (Hammer *et al* 1993; Sturniolo *et al* 1999; Singh and Raghava 2001, 2003). Secondly, these methods have been developed using information from a limited number of binders and non-binders. There is continuous increase in data of MHC binders and non-binders over the years, thus there is a need to develop methods using large and clean dataset (Bhasin *et al* 2003). This is because the performance of knowledge-based methods is directly proportional to size and quality of data used for their development/training.

In order to overcome some of the aforesaid limitations and to complement existing methods, in this study, we have developed hybrid method for a large numbers of MHC class I alleles. Firstly, we have developed a QM based method for 47 alleles for which minimum 15 binders were available. The binders were obtained from MHCBN version 1.1 (Bhasin *et al* 2002). Secondly, ANN-based method was developed for 30 alleles out of these 47 alleles had at least 40 binders available in database. Finally, a hybrid method was developed for these 30 MHC class I alleles by combining ANN- and QM-based methods in order to improve the performance. One of the goals of this study is to develop a method for large number of MHC alleles. Therefore, we also developed quantitative matrix based method for 20 more alleles for which quantitative matrices were obtained from literature (Singh and Raghava 2003).

In the recent studies, it has been shown that MHC binding peptides having proteasomal cleavage site at their C terminus have more chances of being recognized by CTL cells (Kessler *et al* 2001; Ayyoub *et al* 2002; Goldberg *et al* 2002). The simultaneous identification of MHC binding and proteasomal cleavage sites leads to prediction of potential CTL epitopes. Thus, we have implemented the matrices described by Toes *et al* (2001) for the identification of MHC binders having proteasome (standard/constitutive proteasome and immunoproteasome) cleavage site at their C terminal (Toes *et al* 2001). The overall structure of prediction method is shown in figure 1.

In summary, our method allows prediction of binding peptides for 67 MHC class I alleles (30 alleles based on hybrid approach, 17 based on new QM and 20 based on old QM). The performance of the prediction method was compared with previously published methods – SYFPEITHI (Rammensee *et al* 1999), BIMAS (Parker *et al* 1994) and SVMHC (Donnes and Elofsson 2002). For alleles like HLA-A*0201 and HLA-A2.1, our method was also compared with RANKPEP (Reche *et al* 2002), PREDEP (Schueler-Furman *et al* 2000) and polynomial (Gulukota *et al* 1997) methods. The results demonstrated that performance of our method is better as compared to existing MHC binders prediction algorithms. This method has been implemented online as nHLAPred at *http://www.imtech.res.in/raghava/nhlapred* to assist in identifying promiscuous MHC class I restricted CTL epitopes.

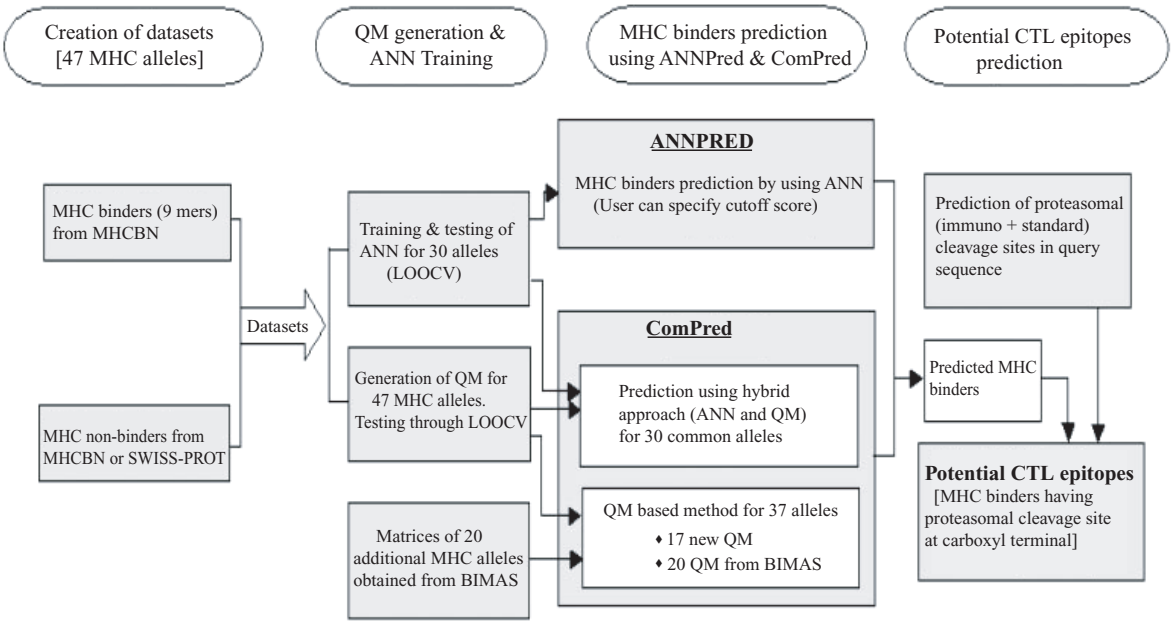

Figure 1. The overall structure of nHLAPred which have four major parts: (i) creation of datasets, (ii) QM generation and ANN training, (iii) MHC binders prediction using ANN and ComPred and (iv) potential CTL epitopes prediction which involve prediction of proteasomal cleavage sites as well as identification of MHC binders having cleavage site at C terminus.

2. Materials and methods

2.1 *Datasets of MHC binders and non-binders*

MHC binders for 47 MHC class I alleles having atleast 15 binders were obtained from MHCBN database (Bhasin *et al* 2003). Equal number of non-binders for these alleles were also obtained from MHCBN. In case equal number of non-binders at MHCBN are not available than remaining were generated from the SWISS-PROT database (Bairoch and Apweiler 2000) by randomly choosing peptides of 9 amino acids. Complete dataset of each MHC allele has nearly equal number of MHC binders and non-binders. The ratio of binders and non-binders was kept 1:1 for developing and evaluating the performance of the method by a single parameter like accuracy at a cutoff score where the sensitivity and specificity are nearly equal.

2.2 *Blind datasets*

The performance of method for few alleles was also evaluated on blind or independent datasets. The blind dataset was generated for each MHC allele. The binders for each allele were obtained from the published literature (for detail visit *http://www.imtech.res.in/raghava/nhalpred/supl/*). Equal number of non-binders for each allele were obtained either from MHCBN database (wherever available) or generated randomly from proteins of SWISS-PROT database. All the binders and non-binders which were used for testing and training of this method were removed from these blind datasets.

2.3 *Methods*

2.3a *Generation of quantitative matrices:* The quantitative matrices were generated for 47 MHC class I alleles. The coefficient value of each amino acid from position 1 to 9 was obtained by dividing the probability of an amino acid at specific position in binders and in non-binders. It is prerequisite to calculate the cutoff/threshold score for each of the matrices. Due to lack of sufficient data, it was not possible to compute the threshold score using standard methods. So, we have followed the strategy adopted in ProPred, a method for prediction of MHC class I restricted T cell epitopes (Singh and Raghava 2001). In brief, the threshold score of a matrix was determined using following steps. First all overlapping peptides of 9 amino acids were generated from all the proteins of SWISS-PROT database. The score of these peptides were obtained by using quantitative matrix. The peptides were sorted in the descending order and top 1% of the peptides was extracted. Minimum score out of these peptides was considered as threshold score at 1%. Similarly, the threshold score at 2%, 3% etc. were calculated. The quantitative matrices generated above are addition matrices where peptide score is calculated by summing up the scores of each residue at specific position along the peptide sequence.

2.3b *Artificial neural network:* ANN-based method was developed for 30 out of 47 alleles had at least 40 MHC binders. ANN implementation was achieved by using Stuttgart Neural Network Simulator, SNNS 4.2 (Hertz *et al* 1991; Zell A and Mamier G: Stuttgart Neural Network Simulator, version 4.2, University of Stuttgart). A feed-forward neural network with standard backpropagation algorithm having single hidden layer (10 hidden units), 189 (21×9) input units and 1 output unit was used. The input layer consisted of 189 nodes to represent peptide of nine amino acids. Amino acids were represented as binary strings of length 21 (we used X in addition to natural amino acids, to represent termini and unknown amino acids). The linear activation function and random weights were used to initialize ANN training. The training was carried out using standard back propagation with a Sum of Squared Error function (SSE). The magnitude of the SSE on training and testing set was monitored after each cycle. The ultimate number of cycles was determined when the network converges.

2.3c *Hybrid method: Combination of ANN and QM:* A hybrid method was developed by combining QM- and ANN-based methods in order to improve the accuracy of prediction. The ANN predicted binders and non-binders were reexamined using QM-based method. If any of ANN predicted binder had very poor score for QM (lower than cutoff score at 10%) then it was assigned non-binder in final prediction. Similarly, if any ANN predicted non-binder achieved very high score for QM (greater than cutoff score at 1%) then it was considered binder in final prediction. In this manner, hybrid method utilized the positive qualities of both ANN- and QM-based approaches for better prediction.

2.3d *Jack-knife testing:* Jack-knife testing is a well-accepted technique to evaluate the performance of a method (Mardia *et al* 1979; Yuan 1999; Feng 2001; Feng and Zhang 2001; Hua and Sun 2001). Jack-knife validation test was used to evaluate the performance of all the methods developed in this study. In jack-knife validation test, one peptide was used for testing and rest of the peptides were used for training. This procedure was repeated N times (N is total number of peptides in dataset), so that each peptide was used only once for testing. The performance of the method was obtained by averaging the performance over N test sets.

2.3e *Prediction of proteasome cleavage site:* For prediction of proteasomal cleavage sites the proteasomal and immunoproteasomal matrices were obtained from the ProPred I server, derived from the work of Toes *et al* (2001) (Singh and Raghava 2001). The proteasomal cleavage site occurs at the center of 12mer peptides that is six amino acid away from N-terminal. The prediction of proteasomal cleavage site in antigenic proteins was achieved as follows. First overlapping peptides of 12 amino acids were obtained from the antigenic protein. Then score of each peptide was calculated by using proteasomal and/or immunoproteasomal matrices. The peptides with score more than cutoff score at selected threshold were predicted as peptides with proteasomal cleavage sites at their center position, i.e. six positions away from the amino terminal position.

2.3f *Parameters used for assessing the performance:* The performance of the method was evaluated by comparing the prediction results with the experimental findings. The standard parameters (sensitivity, specificity and accuracy) and PPV were used for measuring the performance of the methods. The PPV measures the probability that a predicted binder is in fact a binder. These parameters were computed by following equations:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100, \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100, \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \text{ and} \quad (3)$$

$$\text{PPV} = \frac{TP}{TP+FP}, \quad (4)$$

where *TP* and *TN* are true predicted binders and non-binders respectively. *FP* and *FN* are false predicted binders and non-binders respectively.

3. Results

3.1 *Performance of QM-based method*

The performance of all 47 QM generated in this study was evaluated by using jack-knife validation test. These quantitative matrices are available at *http://www.imtech.res.in/raghava/nhlapred/matrix.html*. The number of binders and non-binders used to derive these quantitative matrices are shown in the table S1 of supplementary material (*http://www.imtech.res.in/raghava/nhlapred/help.html*). Table 1 demonstrates the performance of QM in terms of sensitivity, specificity, PPV and accuracy at default threshold. Performance was measured at default threshold of 5% where sensitivity and specificity were nearly the same for most of the alleles. The accuracy of various alleles ranged from 63% to 100% and average accuracy for 47 alleles was 85.4±8.9%.

Table 1. The performance of QM-based prediction method for 47 alleles. The performance was measured at default threshold of 5%.

Allele Name	Sensitivity	Specificity	PPV	Accuracy
HLA-A1	84.09	97.7	97.3	90.9
HLA-A2	69.5	98.4	97.8	83.9
HLA-A*0201	79.74	92.75	91.68	86.24
HLA-A*0202	77.3	97.33	96.6	87.3
HLA-A*0203	68.8	98.3	97.6	83.6
HLA-A*0204	55.2	100	100	77.3
HLA-A*0205	60.8	90.9	87.5	75.5
HLA-A*0206	67.95	92.31	89.83	80.13
HLA-A2.1	81.8	82.2	81.8	82
HLA-A3	64.6	94.7	92.4	79.7
HLA-A*0301	48.8	96.7	94.1	72.7
HLA-A11	76.4	97.6	97	87.2
HLA-A*1101	76.6	98.7	98.3	87.6
HLA-A24	57.89	94.44	91.67	75.68
HLA-A*2402	87.18	98.29	98.08	92.74
HLA-A31	71.1	93.3	91.4	82.2
HLA-A*3301	35.2	93.5	85.7	63.08
HLA-A*6801	81.9	96.7	96.1	89.3
HLA-A*6802	32.5	96.3	93.3	57.1
HLA-B7	90.7	89.3	89.6	90
HLA-B*0702	91.6	94.4	94.2	93
HLA-B8	83.33	93.3	92.5	88.3
HLA-B14	87.2	91.4	91.1	89.3
HLA-B27	96.8	94	92.3	95.2
HLA-B*2703	84.3	100	100	92.1
HLA-B*2704	76.4	100	100	87.5
HLA-B*2705	92.5	96.3	96.1	94.4
HLA-B*2706	39.13	100	100	69.5
HLA-B*2902	90	100	100	94.8
HLA-B35	50	82.8	75	66.2
HLA-B*3501	90.3	89	89.1	89.6
HLA-B44	83.3	100	100	91.3
HLA-B51	64.3	96.5	94.9	80.4
HLA-B*5101	87.2	97.8	97.6	92.5
HLA-B*5102	81.2	95.8	96.3	87.5
HLA-B*5103	73.3	95.8	95.6	83.3
HLA-B*5301	84.2	98.2	97.9	91.2
HLA-B*5401	92.98	96.49	96.36	94.74
H-2Kb	80.43	96.74	96.1	88.59
H-2Kd	64.21	95.79	93.85	80
H-2Db	91.1	84	85.1	87.6
H-2Ld	96.1	92.3	91.3	94
H-2Dd	78.12	96.88	96.15	87.5
H-2Qa	100	100	100	100
HLA-Cw*0401	94.12	100	100	96.7
Mamu-A*04	82.14	92.59	92	87.27
HLA-G	88.8	88.8	88.8	88.8
Mean±STDEV	76.4±16.4	95±4.45	94.0±5.25	85.4±8.9

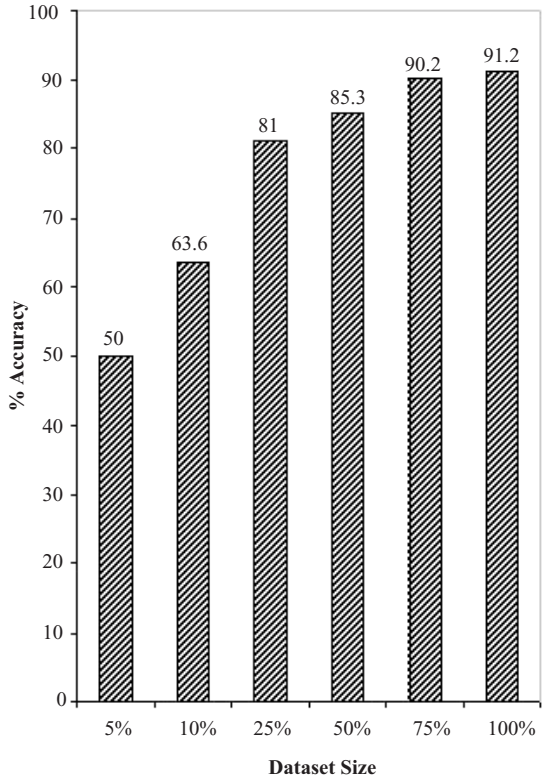

Figure 2. The correlation between the performances of ANN-based method and size of datasets used for training. The results shown are obtained by using various fractions of data (5%, 10%, 25%, 50%, 75% and 100%). The original dataset have 117 binder and equal number of non-binders.

3.2 *Pformance of ANN method*

The ANN learning rate and number of epochs were optimized by spending hours of computational power. Learning parameter was set to 0.01. The training of ANN was optimized to 2000 epochs when there was no further reduction in the error compared to previous cycles. We observed that the performance of ANN-based methods depends on the size of dataset used for training. In order to demonstrate the correlation between size and performance, we computed the accuracy of ANN method trained on various fractions (5%, 10%, 25%, 50%, 75%, 100%) of original dataset. The original dataset is consist of 117 binders and equal number of non-binders for HLA-A*2402 allele. As sown in figure 2, the performance of methods is directly proportional to the dataset used for development. The detailed performance of methods trained on various fractions of datasets is shown in table S2 of supplementary material (*http://www.imtech.res.in/raghava/nhlapred/supl/index.html*). A good correlation (0.85) was observed between the prediction accuracy and number of peptides used for training/testing. Thus, we developed the ANN-based method only for those 30 MHC alleles having more than 40 MHC binders. The performance of the methods was assessed using jack-knife testing. The performance of ANN for 30 alleles at a default cut off score 0.5 (where sensitivity and specificity are nearly equal) is shown in table 2. The accuracy of the prediction varied from 67 % to 96% for various MHC alleles with average accuracy of 87.3±5.9. These results support the fact that ANN is superior in classifying the MHC binders and non-binders.

3.3 *Combination of QM and ANN*

As demonstrated in table 3, the accuracy of ANN-based prediction is higher than QM based prediction for a number of alleles and reverse is also true for other alleles (means accuracy of QM is better than ANN). We developed hybrid method by combining ANN and QM based methods for alleles, which are common in both the methods. How to combine two methods is an important issue in developing hybrid method, in order to utilize their full potential. It is possible that hybrid method may perform poorer than individual method. In this study ANN is used as base method at default cutoff 0.5 for prediction and QM was used for re-examining these predicted binders and non-binders. The prediction was changed in case where QM predicted them as different from ANN with very high confidence (see § 2). The prediction results for 30 alleles using hybrid approach are shown in the table 3. These results clearly demonstrate that hybrid approach based method classified data ~7% and ~5% more accurately in comparison with individual QM- and ANN-based methods respectively.

3.3a *Performance on blind/independent datasets:* The performance of method for few alleles was also evaluated on blind datasets. The performance of the prediction method on blind dataset will be unbiased because these datasets do not contain any binder or non-binder used in the development of method. The performance of the hybrid method of different alleles on blind is shown in table 4. The performance was measured at default threshold (0.5). The performance of the most of alleles on blind dataset is similar to performance observed during jack-knife validation test. The accuracies of the alleles on bind data vary from 76% to 95%. These results demonstrate that method developed in this study is generalized, not biased to dataset used in training. So, it is worth to use this method for prediction of MHC class I restricted binders.

3.4 *Comparison with existing methods*

In order to evaluate comparative performance of the new method, we compare its performance with performance of

Table 2. The performance of ANN-based method for 30 MHC class I alleles at default cutoff score of 0.5.

MHC Allele	Sensitivity	Specificity	PPV	Accuracy
HLA-A1	95.4	95.4	95.4	95.4
HLA-A2	83.0	88.5	64.9	87.5
HLA-A*0201	83.6	86.1	79.9	85.3
HLA-A*0202	87.3	84.5	84.9	85.9
HLA-A*0203	81.9	81.9	81.9	81.9
HLA-A*0206	84.9	85.8	85.7	85.2
HLA-A2.1	77.2	77.7	77.2	77.5
HLA-A3	83.4	81.2	81.6	82.3
HLA-A*0301	86.3	89.3	89.0	87.7
HLA-A11	88.3	88.3	88.3	88.3
HLA-A*1101	88.3	87.0	87.1	87.6
HLA-A*2402	90.6	91.9	91.8	91.2
HLA-A31	84.4	84.4	84.4	84.4
HLA-A*6801	88.5	85.2	85.7	86.8
HLA-A*6802	67.9	66.6	76.6	67.1
HLA-B7	88.1	86.6	87.0	87.4
HLA-B*0702	97.2	95.2	95.8	96.5
HLA-B8	90.0	88.3	89.3	89.1
HLA-B14	89.3	91.4	91.3	90.4
HLA-B27	96.2	96.8	96.7	96.5
HLA-B*2705	88.8	89.8	89.7	89.3
HLA-B*3501	86.7	86.1	86.2	86.4
HLA-B51	88.5	86.2	86.5	87.3
HLA-B*5101	87.2	85.1	85.4	86.1
HLA-B*5301	94.8	94.8	94.8	94.8
HLA-B*5401	94.7	94.7	94.7	94.7
H-2Kb	82.6	82.6	82.6	82.6
H-2Kd	89.4	90.5	90.4	90
H-2Db	84.9	84.9	84.9	84.9
H-2Ld	91.0	92.3	91.0	91.7
Mean ±STDEV	87.3±5.8	87.3±6.0	86.6±6.6	87.3±5.9

existing methods like (i) SYFPEITHI (Rammensee *et al* 1999), (ii) BIMAS (Parker *et al* 1994), and (iii) SVMHC (Donnes and Elofsson 2002). The performance of BIMAS matrices for 20 MHC alleles (present in both BIMAS and nHLAPred) was tested on the dataset used in this study. The table 5a clearly illustrates that the average accuracy of BIMAS matrices (79.6%) was lower than ANN (88.3%), QM (88.1%) and the hybrid approach (93.6%) developed in this study. To analyse the predictive power of hybrid approach, we have also developed a hybrid method by combining the BIMAS matrices and ANN methods for these 20 MHC alleles. It was interesting to note that even combination of old matrices and ANN outperformed any individual method. This demonstrated that QM and ANN based methods complement each other. The performance of nHLAPred was also compared with SYFPEITHI for 10 MHC alleles (common in nHLAPred and SYFPEITHI). The accuracy of SYFPEITHI matrices was considered at a cutoff score where sensitivity and specificity were nearly equal. As shown in table 5b the average accuracy of nHLAPred (93.9%) was nearly 8% higher than accuracy of SYFPEITHI (85.9%). The nHLAPred was also compared with SVMHC,

Table 3. The performance of hybrid approach for different MHC alleles. The performance is shown at the cutoff score (0.5) where sensitivity and specificity are nearly equal.

MHC Allele	Sensitivity	Specificity	PPV	Accuracy
HLA-A1	98.8	96.5	96.6	97.7
HLA-A2	82.2	92.7	79.3	90.8
HLA-A*0201	83.1	92.9	88.6	89.0
HLA-A*0202	87.3	97.1	96.8	92.2
HLA-A*0203	83.6	100	100	91.8
HLA-A*0206	83.3	96.1	95.5	89.7
HLA-A2.1	97.2	91.1	91.4	94.3
HLA-A3	83.4	93.9	93.2	88.7
HLA-A*0301	74.2	98.4	98.0	86.3
HLA-A11	91.1	96.5	96.3	94.1
HLA-A*1101	89.6	98.7	98.7	94.1
HLA-A*2402	94.8	98.2	98.2	96.5
HLA-A31	88.8	100	100	94.4
HLA-A*6801	95.0	96.7	96.6	95.9
HLA-A*6802	55.8	100	100	72.8
HLA-B7	96.0	89.3	90.1	92.7
HLA-B*0702	98.6	98.6	98.6	98.6
HLA-B8	88.3	95	94.6	91.6
HLA-B14	91.4	93.6	93.4	92.5
HLA-B27	98.4	96.8	95.8	97.4
HLA-B*2705	94.4	97.2	97.1	95.8
HLA-B*3501	95.4	90.3	90.7	92.9
HLA-B51	88.5	98.8	98.7	93.6
HLA-B*5101	91.4	97.8	97.7	94.6
HLA-B*5301	98.2	93.1	93.4	95.6
HLA-B*5401	98.2	98.2	98.2	98.2
H-2Kb	84.7	96.7	96.2	90.7
H-2Kd	91.5	100	100	95.7
H-2Db	92.0	88.4	88.8	90.2
H-2Ld	94.8	96.7	96.1	95.8
Mean±SD	91.8±5.4	94.9±3.45	94.0±4.9	93.6±2.92

a support vector machine (SVM) based method (Donnes and Elofsson 2002), for 19 alleles in terms of Matthews Correlation Coefficient (MCC). The figure 3a clearly demonstrates that the performance of our method was better than SVMHC for most of the MHC alleles.

The performance of RANKPEP and PREDEP for the most commonly used MHC alleles, HLA-A*0201 allele was also evaluated on our dataset. Similarly, we also evaluated performance of polynomial matrix described by Gulukota *et al* (1997) for HLA-A2.1 allele on our dataset. The performance of these algorithms is illustrated in figure 3. The figures 3b,c illustrate that our method performs better in comparison to above discussed methods.

It has to be noted that all the methods compared above (except nHLAPred) were evaluated on our dataset without cross validation. It is possible that these methods have used some of the peptides in our dataset, for their training. Their performance will decrease if tested using standard cross validation procedure. Authors were unable to check some of the previously developed methods, as these are not available

Table 4. The performance of different alleles on blind or independent dataset. The binders of blind dataset are obtained from literature. The list of binders of blind dataset along with their bibliographic information is provided in supplementary information at *http://www.imtech.res.in/raghava/nhlapred/supl/index.html*.

Allele	Dataset (B+N)	Sensitivity	Specificity	PPV	Accuracy
HLA-A*0201	216	86.1	95.4	94.9	90.7
HLA-A2	204	63.7	95.1	92.1	79.4
HLA-A3	98	79.6	93.9	92.9	86.7
H-2Db	48	95.8	95.8	95.8	95.8
HLA-B7	102	78.4	94.1	93.0	86.3
HLA-A*0203	40	80	85	84.2	82.5
HLA-A*0301	46	56.5	95.6	92.8	76.0

Table 5a. The performance individual methods (QM, ANN and BIMAS QM) and combination of ANN with QM and BIMAS QM-based methods for 20 common alleles.

Name	ANN	QM (at default threshold)	ANN + QM	Only BIMAS QM	ANN +*BIMAS* QM
Sensitivity	88.2±4.9	82.6±10.0	91.8±5.4	60.7±16.0	91.3±5.0
Specificity	88.2±4.8	93.7±4.6	94.9±3.4	96.9±2.7	88.2±6.0
PPV	86.7±7.3	93.2±4.4	94.0±4.9	95.2±4.4	87.4±7.2
NPV	89.1±4.9	85.0±7.1	93.2±4.1	72.3±8.7	92.0±5.1
Accuracy	88.3±4.8	88.1±4.8	93.6±2.9	79.6±8.4	90.0±4.6

Table 5b. Comparison of accuracy of the hybrid approach with already existing SYFPEITHI prediction methods. The accuracy is compared at a cutoff score where the sensitivity and specificity are nearly equal.

MHC alleles	Accuracy	
	nHLAPred	SYFPEITHI
H2.Db	90.2	87.6
H2-Kb	90.7	87.6
H2-Kd	95.7	85.2
H2-Ld	95.8	86.9
HLA-A1	97.7	94.8
HLA-A*03	88.7	86.3
HLA-B*0702	98.6	63.8
HLA-B*08	91.6	82.5
HLA-B*2705	95.8	90.2
HLA-B*5101	94.6	93.6
Mean± STDEV	93.9±3.4	85.9±8.6

to public as software/web-server (Brusic *et al* 1994; Adams and Koziol 1995). However, the accuracy reported by these methods was lower than our method. In conclusion, a hybrid approach based method developed in this study, is a highly accurate method for prediction of MHC class I binding peptides.

3.5 *Prediction of potential CTL epitopes*

We combined the prediction of MHC class I binders with the information on proteasomal cleavage sites, for efficient prediction of potential T cell epitopes. The prediction of proteasomal and immunoproteasome cleavage sites was achieved by implementing the matrices of Toes *et al* (2001). Customizing the threshold values could vary the stringency of proteasomal prediction. The lower the value of threshold the more stringent will be the prediction. The performance of this hybrid approach based method in prediction of MHC binders and potential T cell epitopes is shown through the following case study.

3.5a *MHC binders:* The method was tested on recently published data. The binders of HLA-A*0201 allele in a tumour associated antigenic protein PRAME, were predicted using hybrid method. These predicted binders were compared with experimentally determined binders for same allele. The hybrid method classified correctly 73% high affinity and 81% intermediate affinity binders as shown in table 3S (*http://www.imtech.res.in/raghava/nhlapred/supl/*) of supplementary material. This has further affirmed

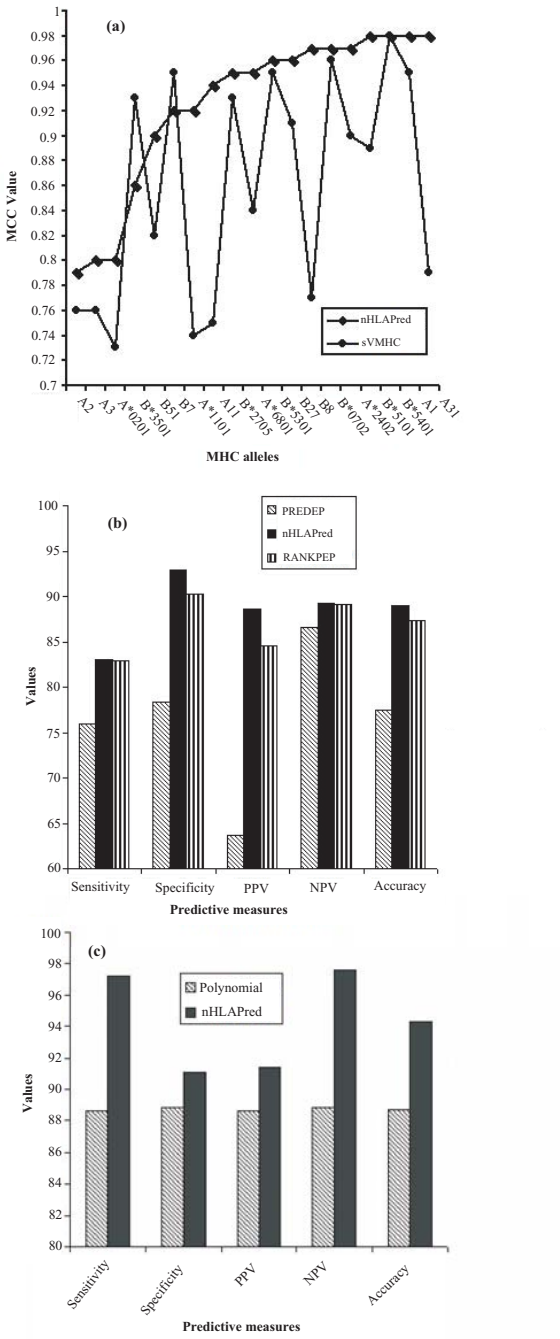

Figure 3. **(a)** The plot represents the performance of an SVM-based method SVMHC with hybrid approach-based method nHLAPred for 19 MHC alleles, which are present in both methods. The comparison was done in term of Matthews Correlation Coefficient (MCC) value. The nHLAPred line is marked with (■) and SVMHC line is marked with (●). Almost for all MHC alleles our hybrid approach based method performed better as compared to SVMHC. **(b)** The plot illustrate the comparison of PREDEP and RANKPEP methods with our hybrid approach based method for HLA-A*0201 allele. **(c)** The plot demonstrates the compassion of our hybrid approach with polynomial method for HLA-A2.1 allele.

the fact that it is worthwhile to use prediction methods for identifying MHC binding peptides.

3.5b *CTL epitopes:* T cell epitopes generally have proteasomal cleavage sites at their carboxyl terminal (Ayyoub *et al* 2002; Goldberg *et al* 2002). The experimentally identified four regions of the tumour-associated antigen PRAME were used to estimate the performance of the hybrid method in identification of CTL epitopes (Kessler *et al* 2001). The HLA-A*0201 binding regions were identified by using combined approach at default cutoff score of 0.5. Thereafter, all proteasomal cleavage sites were predicted at various thresholds (1–10%). The MHC binders having the C terminal position coinciding with proteasomal cleavage site were predicted as potential CTL epitopes. The correctly identified regions are shown in table 6. It is observed that at 7% threshold of the proteasomal filters (standard or immunoproteasome) the hybrid method was able to correctly identify the 75% regions of experimentally proven antigens by Kessler *et al* (2001).

3.6 *Implementation of web server*

We have developed a web server nHLAPred based on this hybrid approach, running on SUN server 420R under Solaris environment. The server is partitioned into two major parts, ComPred and ANNPred. ComPred allows to prediction of binders for 67 MHC alleles. The prediction for 30 MHC alleles is based on combined approach (ANN and QM) and prediction for the remaining 37 alleles is based on QM (17 were generated in this study and 20 were obtained from literature). ANNPred allows the prediction of binders for only 30 alleles purely based on the artificial neural network. Both parts of the server use proteasomal matrices to predict the MHC binders possessing proteasomal cleavage site at C terminal. The server can read input protein sequence in any of the standard formats as it uses the ReadSeq (developed by Dr Don Gilbert).

4. Discussion

The prime objective of this study is to assist biologists searching for potential vaccine candidates which can bind to a large number of MHC class I alleles and can also stimulate CTL cells. Presently available MHC class I binders prediction methods are based on motifs, quantitative matrices (Parker *et al* 1994 Rammensee *et al* 1999) neural networks (Brusic *et al* 1994; Adams and Koziol 1995) and structural information (Schueler-Furman *et al* 2000). In this study an attempt has been made to improve the accuracy of MHC class I restricted CTL epitopes prediction. Towards that, we have developed a quantitative matrix based method

Table 6. The analysis of nHLAPred on the four regions of PRAME protein (A:90-116; B:133-159; C:290-316; D:415-441) identified as T cell epitopes by Kessler *et al* (2001). The MHC binders are predicted by using a cutoff score 0.5 and then predicted MHC binders are filtered by Proteasomal filters at varying thresholds.

Name of filter	Correctly predicted T cell epitopes in the PRAME protein at different thresholds (out of 4)				
	3%	5%	7%	9%	10%
Standard Proteasome (SP)	1(A)	2(A)	2(A,D)	3(A,C,D)	3(A,C,D)
Immunoproteasome (IP)	2(A,D)	2(A,D)	3(A,C,D)	3(A,C,D)	3(A,C,D)
BOTH (SP or IP)	2(A,D)	2(A,D)	3(A,C,D)	3(A,C,D)	3(A,C,D)

by generating new matrices for large number of MHC class I alleles. Further, an artificial neural network based method has been developed for alleles having sufficient, number of experimentally proven binders. Finally, ANN- and QM-based methods were combined to develop a hybrid method, which integrated the merits of both the methods while minimizing their weaknesses. The performance of the method was evaluated using the jack-knife validation test. The performances of few alleles are also tested on an independent datasets wherever the independent data is available in literature. Both analysis demonstrated that method could predict the MHC binders with high accuracy.

The comparison of performance of different existing methods was very difficult as different criteria were used to assess the different algorithms and in most of the cases, cross validation was not used. The predictive accuracy of hybrid approach was compared with the algorithms of most of available algorithms for MHC class I prediction. Our method performed better than all algorithms as shown in the results section, despite the fact that accuracy of earlier methods was not evaluated through jack-knife testing whereas the predictive accuracy of our method was obtained through jack-knife testing. The summary of performance of different methods when compared with our method is illustrated in table 5 and figure 3. This demonstrates that our method not only provides prediction for large number of alleles but also give more accurate prediction than the methods reported till date. The prime reason of improvement in the accuracy of prediction was due to larger dataset and combination of the two prediction methods. The output display formats of the server assists the users in locating the promiscuous MHC binding peptides in their query antigen sequence.

It is a well established fact that binding of a peptide to MHC molecule is a prerequisite to be recognized by T cell epitopes, but it is not necessary that all peptides bound to MHC molecule, will activate the T cells. Thus it is not necessary that all MHC binders are T cell epitopes. Recent studies have demonstrated that MHC binders having proteasome cleavage site at C terminal have higher potential to be T cell epitopes (Kessler *et al* 2001; Singh and Raghava 2003). In order to implement this observation, the prediction of proteasome cleavage site has been integrated in this method. The server allows the prediction of MHC binders possessing proteasomal cleavage site at C terminus, which are known as potential T cell epitopes. The methods for proteasome cleavage prediction are less accurate due to these three factors: (i) broad specificity of proteasome as compared to MHC-peptide binding specificity, (ii) availability of very limited amount of proteasome digested data and (iii) cleavage specificity is not only dependent on the residues occurring at cleavage site but also on neighbouring residues. Thus, the prediction of only proteasome cleavage prediction can have more false positive results. Only a small fraction of the peptides produced by proteasome are passed through TAP transporter, binds to MHC and are recognized by T cell receptors. Therefore combing of proteasome cleavage with MHC binders and TAP binders prediction can lead to reduction in false positive results. The method is able to predict not only MHC binders but also CTL epitopes with high accuracy as shown in table 6. Thus, this strategy not only allows the users to predict the MHC binders but will also allow identification of T cell epitopes, which are suitable candidates for subunit vaccine design.

Though the prediction accuracy of MHC (class I and class II) binder prediction methods has increased over the years, but still it is far away from perfection. Therefore, it is essential to experimentally validate the predicted MHC ligands before considering these as potential sub-unit vaccine candidates for peptide based vaccines. It would be advisable that one should test these predicted ligands using various *in vitro* screening methods (e.g. ELISPOT assay, T cell proliferation assay or MHC binding assay) to prove whether these contain naturally processed T cell epitopes or not.

Acknowledgements

The authors are thankful to Dr Balvinder Singh and Dr Naresh Kumar for useful suggestions in the preparation of the manuscript. We are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), New Delhi,for financial assistance.

MB is a recipient of a fellowship from CSIR. This report has IMTECH communication No: 056 /2002.

References

Adams H P and Koziol J A 1995 Prediction of binding to MHC class I molecules; *J. Immunol. Methods* **185** 181–190

Ayyoub M, Stevanovic S, Sahin U, Guillaume P, Servis C, Rimoldi D, Valmori D, Romero P, Cerottini J C, Rammensee H G, Pfreundschuh M, Speiser D and Levy F 2002 Proteasome-assisted identification of a SSX-2-derived epitope recognized by tumor-reactive CTL infiltrating metastatic melanoma; *J. Immunol.* **168** 1717–1722

Bairoch A and Apweiler R 2000 The SWISS-PROT protein sequences database and its supplyment TrEMBL in 2000; *Nucleic Acids Res.* **28** 45–48

Bhasin M, Singh H, Raghava G P S 2003 MHCBN A comprehensive database of MHC binding and non-binding peptides; *Bioinformatics* **19** 665–666

Brusic V, Rudy G and Harrison L C 1994 Prediction of MHC binding peptides by using artificial neural networks; in *Complex mechanism of adaptation* (Amsterdam: IOS Press) pp 253–258

Buus S 1999 Description and prediction of peptide-MHC binding: the 'human MHC project'; *Curr. Opin. Immunol.* **11** 209–213

Cresswell P, Bangia N, Dick T and Diedrich G 1999 The nature of the MHC class I peptide loading complex; *Immunol. Rev.* **172** 21–28

Donnes P and Elofsson A 2002 Prediction of MHC class I binding peptides, using SVMHC; *BMC Bioinformatics* **3** 25

Doytchinova I A and Flower D R 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

Feng Z P 2001 Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition; *Biopolymers* **58** 491–499

Feng Z P and Zhang C T 2001 Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids; *Int. J. Biol. Macromol.* **28** 255–261

Goldberg A, Cascio P, Saric T and Rock K 2002 The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides; *Mol. Immunol.* **39** 147–164

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, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B and Sinigaglia F 1993 Promiscuous and allele-specific anchors in HLA-DR-binding peptides; *Cell* **74** 197–203

Hertz J A, Palmer R G and Krogh A S 1991 *Introduction to theory of neural computation* (Redwood City: Addison-wesley)

Honeyman M C, Brusic V, Stone N L and Harrison L C 1998 Neural network-based prediction of candidate T-cell epitopes; *Nat. Biotechnol.* **16** 966–999

Hua S and Sun Z 2001 Support vector machine approach for protein subcellular localization prediction; *Bioinformatics* **17** 721–728

Kessler J H, Beekman N J, Bres-Vloemans S A, Verdijk P, vanVeelen P A, Kloosterman-Joosten A M, Vissers D C J, ten Bosch G J A, Kester M G D, Sijts A, Drijfhout J W, Ossendrop F, Offringa R and Melief C J M 2001 Efficient identification of novel HLA-A*0201-presented cytotoxic T lymphocyte epitopes in the widely expressed tumor antigen PRAME by proteasome-mediated digestion analysis; *J. Exp. Med.* **193** 73–88

Mardia K V, Kent J T and Bibby J M 1979 *Multivariate analysis* (London*:* Academic Press) pp 322–381

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

Rammensee H G, Bachmann J, Emmerich N P N, Bachor O A and Stevanovi S 1999 SYFPEITHI: database for MHC ligands and peptide motifs; *Immunogenetics* **50** 213–219

Reche P, Glutting J and Reinherz E 2002 Prediction of MHC class I binding peptides using profile motifs; *Hum. Immunol.* **63** 701–709

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** 1838–1876

Singh H and Raghava G P S 2003 ProPred1 Prediction of Promiscuous MHC class I binding sites; *Bioinformatics* **19** 1009–1014

Singh H and Raghava G P S 2001 ProPred: prediction of HLA-DR binding sites; *Bioinformatics* **17** 1236–1237

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti M P, Sinigaglia F and Hammer J 1999 Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices; *Nat. Biotechnol.* **17** 555–561

Toes R E, Nussbaum A K, Degermann S, Schirle M, Emmerich N P N, Kraft M, Laplace C, Zwinderman A, Dick T P, Muller J, Schonfisch B, Schmid C, Fehling H J, Stevanovic S, Rammensee H G and Schild H 2001 Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products; *J. Exp. Med.* **194** 1–12

Vordermeier M, Whelan A O and Hewinson R G 2003 Recognition of Mycobacterial Epitopes by T Cells across Mammalian Species and Use of a Program That Predicts Human HLA-DR Binding Peptides To Predict Bovine Epitopes; *Infect. Immun.* **71** 1980–1987

Yuan Z 1999 Prediction of protein subcellular locations using Markov chain models; *FEBS Lett.* **451** 23–26

*e*Publication: 15 September 2006