Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vaccine 22 (2004) 3195–3204
Vaccine
www.elsevier.com/locate/vaccine

Prediction of CTL epitopes using QM, SVM and ANN techniques☆

Manoj Bhasin, G.P.S. Raghava*

Institute of Microbial Technology, Sector 39A, Chandigarh, India

Received 16 August 2003; received in revised form 15 December 2003

Available online 5 March 2004

Abstract

Cytotoxic T lymphocyte (CTL) epitopes are potential candidates for subunit vaccine design for various diseases. Most of the existing T cell epitope prediction methods are indirect methods that predict MHC class I binders instead of CTL epitopes. In this study, a systematic attempt has been made to develop a direct method for predicting CTL epitopes from an antigenic sequence. This method is based on quantitative matrix (QM) and machine learning techniques such as Support Vector Machine (SVM) and Artificial Neural Network (ANN). This method has been trained and tested on non-redundant dataset of T cell epitopes and non-epitopes that includes 1137 experimentally proven MHC class I restricted T cell epitopes. The accuracy of QM-, ANN- and SVM-based methods was 70.0, 72.2 and 75.2%, respectively. The performance of these methods has been evaluated through Leave One Out Cross-Validation (LOOCV) at a cutoff score where sensitivity and specificity was nearly equal. Finally, both machine-learning methods were used for consensus and combined prediction of CTL epitopes. The performances of these methods were evaluated on blind dataset where machine learning-based methods perform better than QM-based method. We also demonstrated through subgroup analysis that our methods can discriminate between T-cell epitopes and MHC binders (non-epitopes). In brief this method allows prediction of CTL epitopes using QM, SVM, ANN approaches. The method also facilitates prediction of MHC restriction in predicted T cell epitopes. The method is available at http://www.imtech.res.in/raghava/ctlpred/.

Keywords: Artificial Neural Network; Support Vector Machine; LOOCV; CTL epitopes

1. Introduction

T cells are a vital component of the machinery of protective immunity, both directly by recognizing and eliminating the self-altered cells and indirectly by controlling the production of antibodies by the cells of B lineage [1]. The former function is controlled by cytotoxic T lymphocytes (CTL) [2]. The CTL cells recognize proteolysed fragments of the protein in combination with MHC class I molecules [3,4]. They recognize short peptides of 8–10 amino acids. The interaction of T cell receptor (TCR) with MHC peptide complex can be highly flexible, so that a single TCR can recognize large number of peptides in the context of single MHC molecule [5]. Hence, identification of CTL epitopes is crucial in understanding the rules of T cell activation and designing of synthetic vaccines [6]. The identification of CTL epitopes have paved a way towards cancer immunotherapy and many other infectious diseases.

In the past, a number of methods have been developed for prediction of T cell epitopes from protein sequences. These methods can be classified as direct and indirect methods. In 1980s, direct prediction methods based on structural and sequential analysis of T cell epitopes were developed [7–10]. DeLisi and Berzofsky [7] proposed that the critical requirement of T cell epitopes is its ability to form stable amphipathic structure. Based on this hypothesis, a program AMPHI was developed [8,9]. Another algorithm SOHHA was developed based on the assumption that T cell epitopes consist of a helix of 3–5 helical turns with a narrow strip of hydrophobic residues on one side. These approaches were superseded after analysis of MHC peptide complex by X-ray crystallography, which demonstrated that peptide bound in MHC groove have extended conformation [12].

Sequential models for T cell epitope prediction were also developed, which relies on the occurrence of motifs in the primary sequence rather than considering the secondary structure [13–15]. In 1988, Rothbard and Taylor collected

Abbreviations: MHC, major histocompatibility complex; CTL, cytotoxic T-lymphocytes; ANN, Artificial Neural Network; SVM, Support Vector Machine; QM, Quantitative Matrix; LOOCV, Leave One Out Cross-Validation

☆ Supplementary data associated with this article can be found at doi: 10.1016/j.vaccine.2004.02.005.

* Corresponding author. Tel.: +91-172-690557/695225; fax: +91-172-690632/690585.

E-mail address: raghava@imtech.res.in (G.P.S. Raghava).

URL: http://imtech.res.in/raghava/.

0264-410X/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.vaccine.2004.02.005

nearly 57 T cell epitopes and based on the patterns, they published a list of motifs [14]. The proposed motifs are 3–4 residues consisting of glycine followed by hydrophobic residues. Further, an algorithm was developed based on association of cysteine containing T cell epitopes and certain other residues. The algorithm searches for triplets including CAK, CLV, CKL and CGS in the peptide sequence [13]. In 1995, two computational T cell epitope prediction tools EpiMer and OptiMer were developed based on knowledge of MHC binding motifs [11]. OptiMer predicts amphipathic segments of protein with high motif density and EpiMer locates the segments of protein with high motif density. These direct prediction methods based on structural or sequential models have low accuracy [16]. The main cause of low accuracy may be insufficient data and less specificity of T cell receptors (TCRs).

In the last decade, a number of indirect methods have been developed that predict MHC binders instead of T cell epitopes. The currently available indirect methods are based on structure, binding motifs, matrices or Artificial Neural networks (ANNs) [17–24]. Due to more specific interaction of MHC and peptides, performance of these methods are better in comparison to direct T cell epitope prediction methods. The major limitation of these methods is that they cannot discriminate between T cell epitopes and non-epitope MHC binders. These methods only predict the MHC binders from antigenic sequences.

In this study, an attempt has been made to develop a direct method for prediction of CTL epitopes. The data of CTL epitopes and non-epitopes was obtained from MHCBN version 1.1, a comprehensive database of MHC binders and non-binders [25]. The methods based on QM, SVM and ANN have been developed to discriminate CTL epitope and non-epitopes.

The methods based on QM, ANN and SVM achieved an accuracy of 70.0, 72.2 and 75.5%, respectively, when evaluated through Leave One Out Cross-Validation (LOOCV). The results clearly illustrate that machine-learning techniques are better in comparison to quantitative matrices. The performance of machine learning techniques was further enhanced by devising consensus and combined approaches based on SVM and ANN. The combined prediction approach achieved a sensitivity of 79.4%, which is better as compared to any individual methods. The specificity of consensus approach is 88.4%, which is better as compared to any other individual methods.

The methods developed in this study were also evaluated on a blind dataset that does not contain any pattern used in training or testing. The performance of these methods were evaluated on two subgroups: (i) one subgroup having CTL epitopes and non-epitopes MHC binders, (ii) second subgroup having CTL epitopes and MHC non-binders. The performance of all methods was fairly good on both subgroups as shown in Table 6. This demonstrates that methods developed in this study are able to discriminate between CTL epitopes and non-epitopes MHC binders, which is not possible through MHC binder prediction methods.

Finally, MHC restriction of predicted CTL epitopes were examined using quantitative matrices-based MHC binder prediction method [23]. The quantitative matrices-based method will determine MHC binding specificity of T cell epitopes. A schematic view of prediction method has been shown in Fig. 1. In summary, this comprehensive method will speed up the process of vaccine development for various dreadful diseases like cancer and AIDS.

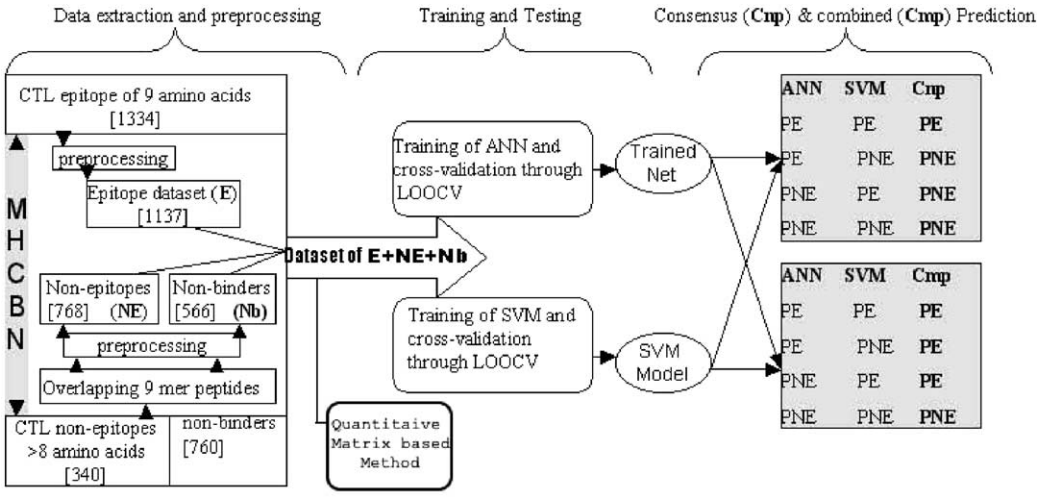

Fig. 1. The overall architecture of CTLPred showing ANN-, SVM- and QM-based methods. The method is divided in three parts: (1) data extraction and preprocessing, (2) training and testing of method, (3) consensus and combined prediction approaches. Where E: epitopes, NE: non-epitopes, Nb: non-binders, LOOCV: Leave One Out Cross-Validation, ANN: Artificial Neural Network, SVM: Support Vector Machine, PE: means predicted epitopes, PNE: predicted non-epitopes, Cnp: consensus prediction, and Cmp: combined prediction.

2. Material and methods

2.1. Datasets

All peptide sequences of the CTL epitopes and non-epitopes were drawn from MHCBN version 1.1 [25]. Initially, 1334 CTL epitopes of 9 amino acids with varying T cell activity were obtained from the database. All duplicate epitopes and epitopes having unnatural amino acids were removed. The final dataset consisted of 1137 CTL epitopes interacting with nearly 170 MHC class I molecules. A total of 340 CTL non-epitopes of 9 or more amino acids were extracted from MHCBN. They were chopped to obtain overlapping nonamer peptides. All duplicate non-epitopes and non-epitopes having unnatural amino acids were removed. The dataset finally consisted of 786 non-epitopes of 9 amino acids. To equalize the number of epitopes and non-epitopes we added 348 MHC non-binders to the dataset of non-epitopes. The final dataset consisted of 1137 CTL epitopes and 1134 non-epitopes. The final ratio of CTL epitopes and non-epitopes was kept nearly 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. This is important as unequal ratio of epitope and non-epitopes can mislead the user.

2.2. Blind dataset

The methods developed in this study were evaluated on a blind dataset to obtain unbiased performance of these methods. The blind dataset was divided in two subgroups to analyze whether the prediction method is better at separating the T cell epitopes from MHC binders (non-epitopes) or non-binding peptides. Sixty-three CTL epitopes of various HIV proteins were collected from HIV database [26]. First subgroup consists of 63 CTL epitopes and equal number of non-epitope MHC binders randomly extracted from MHCBN database. Second, subgroup having 63 CTL epitopes and 63 MHC non-binders were randomly chosen from MHCBN. The blind dataset have unique experimentally proven CTL epitopes, non-epitope MHC class I binders and MHC non-binders. The MHC class I binders (non-epitopes) and non-binders were obtained from the MHCBN database. The peptides of blind dataset had no similarity with T cell epitopes and non-epitopes used in development of various prediction methods of this study. The list of CTL epitopes of blind dataset has been shown in table S1 of supplementary material.

2.3. Generation of quantitative matrices

To classify the data of CTL epitopes and non-epitopes a quantitative matrix was generated from the above compiled dataset. Following equation was used to geneate the quantitative matrix:

$$\boldsymbol{Q}_{(i,r)} = \boldsymbol{P}_{(i,r)} - \boldsymbol{N}_{(i,r)} \quad (1)$$

$$\boldsymbol{P}_{i,r} = \frac{E_{i,r}}{NA_{i,r}} \quad (2)$$

$$\boldsymbol{N}_{i,r} = \frac{A_{i,r}}{NA_{i,r}} \quad (3)$$

where $\boldsymbol{Q}_{(i,r)}$ is the weight of residue r at position i in the matrix. r can be any natural amino acid and the value of i can vary from 1 to 9, $\boldsymbol{P}_{(i,r)}$ and $\boldsymbol{N}_{(i,r)}$ is the probability of residue r at position i in CTL epitopes and non-epitopes, respectively, $E_{i,r}$ and $A_{i,r}$ is number residue r at position i in epitopes and non-epitopes, respectively, and $NA_{i,r}$ is number of epitopes and non-epitopes having residue r at position i.

The quantitative matrix geneated by using Eq. (1) has been shown in Table 1. This matrix is an addition matrix where the score of a peptide is calculated by summing up the scores of each residue at specific position along peptide sequence as

$$\text{Score} = \sum_{i=1}^{l} \boldsymbol{Q}_{i,r}$$

where l is the length of the peptide.

For example, the score of peptide "ILKEPVHGV" is calculated as follows

$$\text{Score} = \text{I}_1 + \text{L}_2 + \text{K}_3 + \text{E}_4 + \text{P}_5 + \text{V}_6 + \text{H}_7 + \text{G}_8 + \text{V}_9 \quad (4)$$

The peptides achieving score more than cutoff score were considered as CTL epitopes.

2.4. Artificial Neural Network (ANN)

The ANN consists of nodes that receive signals through interconnecting arcs [27]. ANN was trained by implementing Stuttgart Neural Network Simulator, SNNS version 4.2 [28]. The main feature of this package is that it allows incorporation of resulting networks in ANSI C functions for use in stand-alone code. The number of hidden nodes in the hidden layer and other learning parameters were optimized after spending hundreds of hours of computational power. A feed-forward backpropagation type of ANN with a single hidden layer (20 nodes), 180 (20×9) input units and 1 output unit was been used in this study. The input layer consisted of 180 nodes to represent the peptide of nine amino acids. Amino acids were represented as binary string of length 20 where 19 "0" and a unique position set to "1" for each amino acid. The output unit consisted of single binary number 0 or 1, which meant true or false. A linear activation function and random weights were used for initializing the net. The training was carried out using error back propagation with Sum of Squared Error function (SSE). The magnitude of SSE on training was monitored after each cycle. The ultimate number of cycles was determined where the network converges, means the value of error is minimum. The value of the linear parameter was set to 0.01. The training was carried out for

Table 1
Example of quantitative matrix for classifying CTL epitopes and non-epitopes

Amino acids	Amino acid positions within the peptide								
	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	0.28	0.32	−0.13	−0.22	−0.04	0.01	0.03	0.21	−0.23
C	−0.08	−0.47	−0.20	0.06	−0.09	0.13	0.04	−0.07	−0.23
D	−0.06	−0.64	0.01	0.25	−0.29	−0.33	−0.10	−0.18	−0.90
E	−0.07	−0.06	−0.29	−0.05	−0.23	−0.18	−0.21	0.33	−0.95
F	0.42	−0.16	−0.07	−0.14	0.29	0.18	0.13	−0.08	0.20
G	−0.14	−0.39	−0.15	0.09	−0.14	−0.33	−0.56	−0.33	−0.75
H	0.22	−0.74	0.22	0.43	−0.38	−0.33	0.16	−0.02	−0.68
I	0.02	0.04	−0.05	0.10	0.06	0.20	0.08	0.09	0.05
K	0.03	−0.52	0.01	−0.22	0.15	−0.43	−0.26	−0.17	0.19
L	−0.40	0.04	0.00	−0.16	−0.19	−0.04	−0.11	−0.05	0.35
M	−0.04	−0.01	−0.09	−0.29	−0.13	0.30	0.31	0.09	0.25
N	−0.03	−0.20	0.21	−0.07	0.34	−0.14	0.14	0.14	−0.85
P	−0.66	−0.09	0.09	0.10	−0.20	−0.09	0.07	−0.52	−0.92
Q	−0.24	−0.22	−0.33	0.31	−0.21	−0.18	−0.02	−0.15	−0.85
R	0.36	0.41	0.16	0.00	0.26	0.04	0.22	0.09	0.28
S	0.17	−0.12	−0.17	−0.16	0.00	0.16	−0.15	−0.06	−0.85
T	0.09	−0.09	−0.10	−0.29	−0.10	−0.18	0.15	0.24	−0.64
V	−0.22	0.15	−0.08	−0.30	0.07	0.17	0.11	0.00	0.06
W	0.11	−0.69	−0.08	0.30	−0.09	0.12	0.32	−0.03	0.24
Y	0.03	0.18	0.39	−0.11	−0.07	0.48	0.22	0.14	−0.07

Each residue at each position in a 9 mer is assigned a weight which is used to calculate the score. The matrix is able to discriminate between CTL epitopes and non-epitopes with 70% accuracy.

400 epochs and the learning was terminated when the error reached a stable value. The stable error meant very small decrease in the error in the subsequent cycles of learning.

2.5. Support Vector Machine

The SVM implementation is achieved by using the package SVM_LIGHT [29,30]. The SVM map the inputs into a higher dimensional feature space that separates a given set of binary training data with an optimal hyperplane. The optimal hyperplane found by SVM is one at which the maximum separation between the CTL epitope and non-epitope data is obtained.

A training set consists of N samples or input vectors $\{x_1, x_2, x_3, \ldots, x_i, \ldots, x_N\}$ with known class labels $\{y_1, y_2, y_3, \ldots, y_i \cdot y_N\}$, $y_i \in \{+1, -1\}$. The x corresponds to the amino acid sequence of CTL epitopes and non-epitopes and y_i represents epitope or non-epitope. A new value $\boldsymbol{x}$ is assigned to each example by the SVM.

$$\int(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i, \quad k(x_i, x) + b\right)$$

where k is kernel function that define the feature space; b is the bias value, α_i is the number obtained by solving the a quadratic programming (QP) problem that gives the maximum margin hyper plane.The aim is to maximize α_i

$$0 \leq \alpha_i \leq C$$

where C is controlling the trade off between the margin and training error. This is a kernel function that determine the feature space which means that different kernels represent the input vectors in different ways [31].

The choice of the proper kernel function is an important issue for SVM training because the power of SVM comes from the kernel representation that allows the non-linear mapping of input space to a higher dimensional feature space. The use of appropriate decision function can give better classification. The experiments were conducted by using every type of kernel dot, RBF and polynomial to achieve better results. The results were evaluated in terms of accuracy at a cutoff score where sensitivity and specificity were nearly equal. The best results were obtained by using the polynomial kernel where the sensitivity, specificity and accuracy obtained were much better as compared to the other kernels. The value of the kernel function d was optimized to 2.0. After choosing a particular type of kernel, the value of the regularization parameter C needs to be tuned. The value of the C parameter was optimized to 0.01. The value of the different functions were optimized by looking at the accuracy of the prediction method at the cutoff score where the sensitivity and specificity were nearly equal.

2.5.1. Input for SVM

Same data were used for training and testing of the SVM as used in case of ANN. Each amino acid of 9 mer peptide was represented by a 20-dimensional vector. The CTL epitope was represented by the +1 and CTL non-epitope by −1.

2.6. Combined and consensus prediction

The machine learning-based methods were used to perform consensus and combine prediction of CTL epitopes. In *consensus prediction*, epitopes predicted by both methods were considered as epitopes, otherwise they were considered as non-epitopes. In *combined prediction*, epitope predicted by either of methods were considered as epitopes. We investigated a variety of techniques and generated various models for consensus and combined prediction using SVM and ANN. In the first model, SVM was used as a base method (at default cutoff) and ANN was used at various cutoff scores. In the second model, ANN was used as the base method whereas SVM was used at various cutoff scores. The performance of combined and consensus prediction were computed for both models.

2.7. Quantitative matrices for MHC restriction prediction

The quantitative matrices for determining the MHC restriction of T cell epitopes were obtained from ProPred1 server [23]. These matrices were originally obtained from BIMAS server and literature [20]. These quantitative matrices for 46 MHC alleles are available at http://www.imtech.res.in/raghava/propred1/matrix.html. The matrices are either multiplication or addition matrices. In multiplication matrices, score of the peptide is obtained by multiplying scores of each amino acid. In addition matrices, score of the peptide is obtained by summing up the scores of individual residue. The prediction method for testing the MHC restriction of T cell epitope was developed by implementing these matrices. The prediction for MHC restriction of CTL epitope has been performed at default threshold.

2.8. Evaluation of methods

The LOOCV procedure was employed to estimate the performance of the prediction methods. The LOOCV procedure involves removing one peptide from the training data; training is done on the basis of remaining data and then testing was done on this removed peptide. In this manner, if the training data consisted of 100 peptides, then 100 networks were produced by using each of the peptide as test set while using the other peptides as the training data. This is the most extreme test of the cross validation. It is the most accurate way to estimate the performance of method when the training data is small. The performance of methods was computed using following measures,

$$\text{Senstivity or recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (5)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (6)$$

$$\text{Precision or PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (8)$$

where TP and TN are correctly predicted CTL epitopes and non-epitopes, respectively. FP and FN are wrongly predicted epitopes and non-epitopes, respectively.

3. Results

3.1. Quantitative matrices

In case of QM, the contribution of each residue for each position of peptide in T cell activity was quantified. A matrix with weights for each amino acid residue in every position of peptide was generated using Eq. (1). The QM is shown in Table 1. The effect of each residue on T cell activity of peptide could be easily estimated. The QM-based method was able to classify the data with 70.0% accuracy at default threshold where sensitivity and specificity of prediction was nearly equal. The performance of the QM at different thresholds is shown in Table 2. At default threshold (0.0), the sensitivity, specificity and accuracy of prediction are 65.2, 74.9 and 70.0%, respectively. As shown in Table 2, the sensitivity of method is directly proportional and specificity is inverse proportional to the threshold. The stringency of prediction varies with the thresholds so the selection of threshold is very crucial. The performance of quantitative matrix-based method was evaluated using LOOCV test.

3.2. Machine learning approaches

The elegant machine learning approaches, SVM and ANN were applied for CTL epitope prediction. These approaches could handle the non-linearity of data. The performance of these methods was evaluated through LOOCV test.

3.3. ANN

Artificial Neural Network was trained with single sequence encoded in the binary bits with a window size of

Table 2
The performance of QM-based method at various cutoff scores or thresholds

Threshold	Sensitivity	Specificity	Precision	Accuracy
−2.5	99.9	1.06	50.3	50.6
−2.0	99.8	4.3	51.1	52.1
−1.5	99.4	12.3	53.2	55.9
−1.0	96.0	28.1	57.2	62.0
−0.5	85.9	51.4	63.9	68.7
0.0	**65.2**	**74.9**	**72.2**	**70.0**
0.5	39.0	89.2	78.4	64.1
1.0	25.2	95.8	85.7	60.5
1.5	14.0	97.7	86.0	55.8
2.0	3.7	99.2	82.4	51.4

Bold values show sensitivity, specificity, precision and accuracy at default cutoff score.

Table 3
Summary of variation in predictive measures of ANN-based prediction method

Cutoff score	Sensitivity	Specificity	Precision	Accuracy
0.01	98.2	17.1	54.3	57.7
0.11	92.4	43.0	61.9	67.8
0.21	89.2	52.0	65.1	70.6
0.31	83.9	58.9	67.2	71.4
0.41	79.1	65.7	69.8	72.4
0.51	**73.2**	**71.2**	**71.8**	**72.2**
0.61	67.2	75.7	73.5	71.4
0.81	48.2	87.1	79.0	67.6
0.96	17.9	97.8	89.1	57.8

Bold values show sensitivity, specificity, precision and accuracy at default cutoff score.

Table 4
Summary of predictive measures for SVM-based prediction method at various cutoff scores

Cutoff score	Sensitivity	Specificity	Precision	Accuracy
0.01	84.5	60.4	68.2	72.5
0.11	81.9	65.9	70.6	73.9
0.21	78.6	70.6	72.9	74.6
0.36	**73.8**	**77.0**	**76.3**	**75.4**
0.51	65.6	82.5	79.0	74.1
0.61	60.7	85.5	85.5	73.1
0.81	48.9	89.8	82.7	69.3
0.96	37.7	92.2	82.8	64.9

Bold values show sensitivity, specificity, precision and accuracy at default cutoff score.

nine. The performance of ANN-based method was measured at different thresholds (Table 3). As shown in Table 3 prediction accuracy varied from 57.7 to 72.4%. The best performance of the method was obtained at threshold 0.51 (default threshold) where sensitivity and specificity were nearly equal. At default threshold, sensitivity, specificity, precision and accuracy of method was 73.2, 71.2, 71.8 and 72.2%, respectively. The user can choose low threshold value to increase the sensitivity (percent coverage). For example, at threshold "0.11" one may achieve 92.4, 43.0 and 68.7% sensitivity, specificity and accuracy, respectively (see Table 3). Similarly one may select higher threshold to achieve higher specificity at the cost of sensitivity. The performance of the method is slightly better than QM-based method.

3.4. SVM

Support Vector Machine was also applied on the same training data to develop a method for discriminating CTL epitopes and non-epitopes. The best results were obtained by using the polynomial kernel where the sensitivity, specificity, accuracy obtained was much better as compared to the other kernels. In case of the polynomial kernel, the value of kernel parameter (d) was optimized to 2 and value of the regularization parameter C optimized to 0.01. The values of the different functions optimized to achieve high prediction accuracy at default cutoff score. The sensitivity, specificity, precision and accuracy of the optimized SVM model at different cutoff scores are shown in Table 4. The SVM was able to achieve an accuracy of 75.2% at a default cutoff score whereas the sensitivity, specificity and precision were 73.8, 77.0 and 76.3%, respectively. The SVM-based method was able to classify the data with ∼3 and ∼5% more accuracy as compared to ANN- and QM-based methods. These results indicate that SVM is better than ANN and QM in discriminating CTL epitopes and non-epitopes.

3.5. Combined and consensus prediction

In order to improve the accuracy of prediction, the positive qualities of both techniques SVM and ANN were utilized. This is important to study both methods in depth otherwise combination of two may be counter predictive to each other. In this study, two models were investigated. In the first model, SVM was kept as the base method (at default cutoff score) and ANN was utilized at various cutoff scores. In the second model, ANN was kept as base method (at default cutoff score) and SVM was applied at different cutoff scores. The consensus and combined prediction were performed for both type of models. The values of sensitivity and specificity in combined and consensus prediction using first model are shown in plots A and B of Fig. 2. The specificity of consensus prediction approach is nearly 7% higher as compared to the specificity of ANN and SVM at a cutoff score where sensitivity and specificity are nearly equal, as shown in plot A of Fig. 2. In consensus prediction, the accuracy of prediction was 77.6%, which is nearly 4% more than ANN and ∼2% higher than SVM. In case of combined prediction the sensitivity (79.7%) of the prediction method was much higher as compared to individual prediction methods. In case of the combined method, the accuracy of prediction is nearly ∼3% higher than ANN prediction and marginally greater than SVM-based prediction. The plots C and D of Fig. 2 depicts the sensitivity and specificity of consensus and combined prediction using second model. The results indicate that first model is better in comparison to second model. The best results of the consensus and combined prediction are shown in Table 5.

3.6. Performance of methods on blind dataset

In the past, it has been observed that cross-validation is not an unbiased test to evaluate the performance of method [24]. Thus, it is important to evaluate newly developed

Table 5
The best values of sensitivity, specificity, precision and accuracy obtained through consensus and combined prediction based on SVM and ANN

Prediction approach	Sensitivity	Specificity	Precision	Accuracy
Consensus prediction	66.9	88.4	85.2	77.6
Combined prediction	79.7	71.9	74.0	75.8

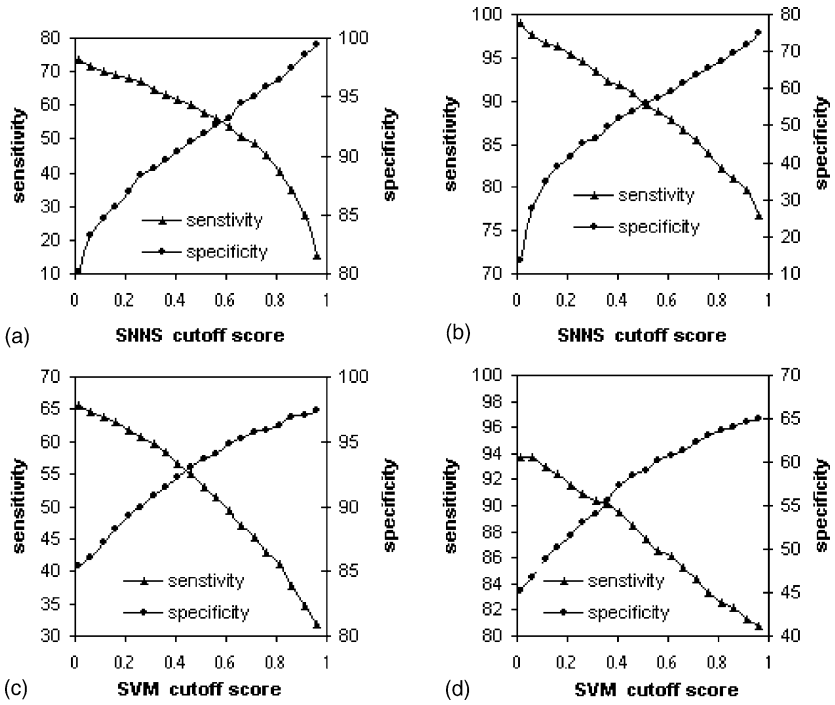

Fig. 2. The consensus and combined prediction on ANN- and SVM-based methods using various models. The plots A and B depict consensus and combined prediction using the first model. In first model SVM at default cutoff score [base method] is utilized with ANN-based methods at different cutoff scores. The plots depict the variation in sensitivity and specificity at various cutoff scores of ANN-based methods. The plots C and D represent the combined and consensus prediction second model. The second model was developed method developed by using ANN at default cutoff score with various cutoff scores of SVM. The specificity line is marked with (●) and sensitivity line is marked by (▲).

methods on blind or independent datasets that are not used in the method for training or testing. Firstly, all the methods were tested on a dataset of 63 CTL epitopes at default cutoff score. The percent coverage of different methods was measured. Nonetheless, the percent coverage is a useful measure to evaluate the ability of methods for identification of CTL epitopes, but it does not provide any information specificity or accuracy of prediction methods. This cannot be a rigorous test to estimate the performance of newly developed method because the dataset do not consist of non-epitopes. In order to evaluate methods rigorously, two subgroups of blind dataset were created; first subgroup having CTL epitopes and non-epitope MHC binders and second subgroup having CTL epitopes and MHC non-binders.

The performance of the three methods is evaluated on first subgroup to assess the ability of these methods in separating T cell epitopes from non-epitope MHC binders. The prediction accuracy is 64.3, 69.0 and 62.0% for SVM, ANN and QM methods, respectively (Table 6). As shown in Table 6, the sensitivity and specificity varies significantly with respect to threshold of these methods. The analysis of results illustrated that methods developed in this study are able to discriminate between CTL epitopes and non-epitopes with fair accuracy. This is the unique feature of these prediction methods, which is not possible with indirect method of T cell epitope prediction (MHC binder prediction methods).

Similarly, we examined the performance of methods on second subgroup in order to access the ability of methods in separating CTL epitopes from MHC non-binders (Table 6). This analysis is important because antigenic sequences consist of large number of MHC non-binders in comparison to MHC binders. The accuracies of SVM-, ANN- and QM-based methods on second subgroup are 69.1, 65.1 and 60.3%, respectively. This clearly indicates that these methods had the capability to predict the CTL epitopes in antigenic sequences. The overall results of analysis demonstrated that the methods developed in this study are able to separate T cell epitopes from MHC binders as well as T cell epitopes and non-binders.

3.7. MHC restriction of predicted T cell epitopes

Due to MHC polymorphism, the predicted T cell epitopes may be functional in one population and non-functional in another population. The functionality of T cell epitopes is related to MHC molecule, because binding of peptides to MHC molecules is the bottleneck for functioning as a T cell epitope. To explore the MHC restriction of particular T cell epitope, we developed quantitative matrices-based method. The quantitative matrices were obtained from ProPred1 server [23]. The cutoff score used for MHC binder prediction was same as suggested in ProPred1 server. The method will predict whether predicted CTL epitope bind

Table 6
The subgroup analysis of SVM, ANN and QM on the blind dataset

QM				SVM				ANN			
Thrs	SEN	SPEC	ACC	Thrs	SEN	SPEC	ACC	Thrs	SEN	SPEC	ACC
T cell epitopes and MHC binders (non-epitopes)											
−1.6	98.4	0.0	49.2	−0.6	100.0	3.2	51.6	0.05	95.2	9.5	52.4
−0.6	93.7	19.0	56.4	−0.3	93.7	9.5	51.6	0.2	92.1	19.1	55.6
−0.5	90.5	22.2	56.4	−0.2	93.7	17.5	55.6	0.3	87.3	25.4	56.4
−0.4	88.9	25.4	57.1	0.1	87.3	30.2	58.7	0.5	82.5	42.9	62.7
−0.1	82.5	34.9	58.7	0.2	84.1	39.7	61.9	0.55	79.4	47.6	63.5
0.1	74.6	44.4	59.5	0.3	79.4	47.6	63.5	0.65	74.6	57.1	65.9
0.3	63.5	58.7	61.1	**0.4**	**74.6**	**54.0**	**64.3**	**0.7**	**73.0**	**65.1**	**69.1**
0.4	**57.1**	**66.7**	**62.0**	0.5	60.3	60.3	60.3	0.75	68.3	68.3	68.3
0.7	38.1	88.9	63.5	0.8	41.3	69.8	55.6	0.8	58.7	76.2	67.5
0.8	30.2	90.5	60.3	0.9	25.4	74.6	50.0	0.85	42.9	77.8	60.3
0.9	17.5	90.5	54.0	1.0	17.5	82.5	50.0	0.9	33.3	82.5	57.9
1.1	11.1	95.2	53.2	1.1	6.4	87.3	46.8	0.95	22.2	88.9	55.6
1.3	4.8	98.4	51.6	1.3	3.2	93.7	48.4	1.0	0.0	100.0	50.0
T cell epitopes and MHC non-binders											
−1.0	96.8	4.8	50.8	−0.6	100.0	4.8	52.4	0.05	95.2	9.5	52.4
−0.9	96.8	7.9	52.4	−0.3	93.7	25.4	59.5	0.2	92.1	23.8	57.9
−0.5	90.5	15.9	53.2	−0.2	93.7	27.0	60.3	0.25	87.3	27.0	57.1
−0.4	88.9	23.8	56.4	0.1	87.3	47.6	67.5	0.3	87.3	31.8	59.5
−0.3	85.7	25.4	55.6	0.2	84.1	52.4	68.3	0.4	85.7	38.1	61.9
−0.1	82.5	33.3	57.9	0.3	79.4	57.1	68.3	0.5	82.5	49.2	65.9
0.1	74.6	46.0	60.3	**0.4**	**74.6**	**63.5**	**69.1**	0.55	79.4	50.8	65.0
0.3	**63.5**	**57.1**	**60.3**	0.5	60.3	68.3	64.3	**0.7**	**73.0**	**57.1**	**65.0**
0.4	57.1	61.9	59.5	0.8	41.3	76.2	58.7	0.75	68.3	60.3	64.3
0.5	49.2	68.3	58.7	0.9	25.4	77.8	51.6	0.8	58.7	68.3	63.5
0.8	30.2	81.0	55.6	1	17.5	77.8	47.6	0.85	42.9	71.4	57.1
0.9	17.5	85.7	51.6	1.1	6.4	84.1	45.2	0.9	33.3	77.8	55.6
1.3	4.8	95.2	50.0	1.3	3.2	95.2	49.2	0.95	22.2	87.3	54.8

The first subgroup has T cell epitopes and MHC binders (which are not T cell epitopes) and seconds subgroup has T cell epitopes and MHC non-binders. Thrs, SEN, SPEC, ACC means threshold, sensitivity, specificity, accuracy, respectively. The bold values show the performance of different prediction methods at cutoff score where sensitivity and specificity are nearly equal.

to MHC molecule or not at default threshold. The MHC restriction study of T cell epitopes will help in choosing promiscuous CTL epitopes, which binds many MHC alleles. The MHC restriction will also help in selecting heterogeneous CTL epitopes (having diverse MHC specificity) for inclusion in synthetic vaccines so that an effective immune response can be obtained in large population.

3.8. Comparison with other methods

The methods developed in this study were compared with direct CTL epitope prediction methods. The methods were not compared with MHC binder prediction methods because they are allele specific and our method predicts the CTL epitope irrespective of MHC restriction. We implemented two popular T cell epitope prediction AMPHI and EpiMer on our dataset [11,15]. The prediction accuracy of AMPHI method was 53% at default threshold. The prime cause of low accuracy may be a significant increase in the number of T cell epitopes since the advent of the AMPHI. The EpiMer located the regions of protein having high MHC motif density. The EpiMer was also able to classify the data with 62% of accuracy at default threshold. The restricted length of peptides may be responsible for low accuracy of method on our dataset. The dataset consisted of peptides of nine amino acids. This clearly demonstrated that our prediction method was superior to existing methods. So our method will complement all existing T cell epitope prediction methods. This high accuracy of our method may be due to (i) large dataset of epitopes and non-epitopes, as accuracy of knowledge-based methods is directly proportional to the quality and quantity of data; (ii) use of elegant artificial intelligence techniques which can classify the data more accurately by handling the non-linearity in data as compared to motif or quantitative matrix-based methods.

3.9. Description of CTLPred server

The method was developed and implemented on Sun Microsystems SPARC 420R under the Solaris environment. The method CTLPred is available for public from web site http://www.imtech.res.in/raghava/ctlpred/ and http://bioinformatics.uams.edu/mirror/ctlpred/. The server can read input protein sequence in any of the standard formats as it uses the ReadSeq (developed by Dr. Don Gilbert). The server allows the prediction by using ANN, QM or

SVM. The server also provides a number of options including selection of threshold for QM, ANN and SVM and various output formats for displaying results. The server further allows consensus and combined prediction using SVM- and ANN-based methods.

4. Discussion and conclusions

It was observed in mid 1990s that the performance of all the previously published T cell epitope prediction methods was quite poor [16]. The performance of these methods were not even significantly better than random prediction. The lack of sufficient amount of data about T cell epitopes may be the prime cause of poor performance [16]. The success of a prediction method depends on the quality and quantity of data. To predict T cell epitopes with fair accuracy, a large number of MHC binders prediction methods were developed during last decade [23,24,32,33]. These methods were successful in predicting MHC binders due to more specific binding of MHC and peptides [19]. These methods could help the immunologist in searching potential T cell epitopes because T cell epitopes formed a subset of MHC binders. These methods were specifically developed for individual MHC allele having sufficient amount of data. These methods could predict the MHC binders with fair accuracy but it was not necessary that all the MHC binding peptides would stimulate T cells [34]. To the authors knowledge, no direct T cell epitopes prediction method has been developed during last one decade.

In the last 10 years, the data on MHC binders/non-binders and T cell epitopes has increased tremendously [25]. Thus, in this report, a systematic attempt has been made to develop direct methods for predicting CTL epitopes. We restrict our study on CTL epitopes (MHC class I restricted T cell epitopes) as they are very important for cancer therapy. Ideally, one should develop method for predicting MHC allele restricted T cell epitopes, which is not practically possible due to lack of sufficient number of CTL epitopes corresponding to different MHC alleles. Thus, we have developed a common method for predicting CTL epitopes irrespective of MHC alleles.

In the past, a number of pattern recognition techniques were used to develop methods for predicting MHC binding peptides. These included motif search, quantitative matrix and machine learning techniques. QM provides a very detailed model in which contribution of each amino acid at each position is quantified. In the past, a number of prediction methods were developed using machine learning techniques like ANN and SVM. Though machine-learning techniques can handle the non-linearity of data, they require a large amount of data for training. Thus, both machine learning- and QM-based techniques have their own merits and demerits.

Here, we have developed methods based on QM, SVM and ANN, for prediction of CTL epitopes. The performance of all methods was evaluated through LOOCV. As shown in Tables 2–4, all methods classified the data with accuracy more than 70.0% at default cutoff score. The performance of these methods was better than any direct CTL epitope prediction methods reported in literature. The consensus and combined prediction of machine-learning techniques further improve the performance of method and allow user to predict with high sensitivity and/or specificity. The combination of two methods also resulted in improvement of accuracy as compare to individual method.

The performance of the above-developed methods was evaluated by applying them on the independent blind dataset. The testing on blind dataset provides the unbiased performance of the prediction method. All prediction methods (QM, ANN and SVM) developed in this study performed significantly on blind dataset. The sensitivity or percent coverage of methods was more than 75.0% when applied on 63 CTL epitopes at default cutoff scores (data not shown). Further analysis on two subgroups of blind dataset showed that the method can discriminate between T cell epitopes and non-epitope MHC binders. The subgroup analysis also illustrated that methods developed in this study were able to discriminate between T cell epitopes and non-binders with good accuracy. It was noticed that performance of methods on blind dataset was lower than performance during cross-validation. It may be due to small size of blind dataset that consisted only 63 epitopes and 63 non-epitopes. The size of dataset was too small to fairly evaluate any prediction method. It is also possible that methods got over trained during development. The authors suggest that users should use all prediction methods developed in this study to predict CTL epitopes. Hence, it is of worth to use these prediction methods in speeding up the process of subunit vaccine development.

This method will help the researchers in finding tumors antigens to eradicate dreadful diseases like cancer. The approach has a larger potential for improvement of prediction accuracy, especially in view of highly growing superior quality CTL epitopes data. The prediction accuracy may be improved by adding more features to the prediction.

Acknowledgements

The authors are thankful to Sanjoy Paul and Amrita Lama for carefully reading the manuscript. The authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India for financial assistance. Manoj Bhasin is a recipient of a fellowship from CSIR. This report has IMTECH communication No. 016/2003.

References

[1] De Groot AS, Sbai H, Aubin CS, McMurry J, Martin W. Immuno-informatics: mining genomes for vaccine components. Immunol Cell Biol 2002;80:255.

[2] Long EO, Jacobson S. Pathways of viral antigen processing and presentation to CTL: defined by the mode of virus entry? Immunol Today 1989;10:45.
[3] Hammerling GJ, Vogt AB, Kropshofer H. Antigen processing and presentation—towards the millennium. Immunol Rev 1999;172:5.
[4] Watts C, Powis S. Pathways of antigen processing and presentation. Rev Immunogenet 1999;1(60):74.
[5] Buus S. Description and prediction of peptide-MHC binding: the 'human MHC project'. Curr Opin Immunol 1999;11:209.
[6] Brunak S, Buus S. Identifying cytotoxic T cell epitopes from genomic and proteomic information: "The human MHC project". Rev Immunogenet 2000;2:477.
[7] DeLisi C, Berzofsky JA. T-cell antigenic sites tend to be amphipathic structures. Proc Natl Acad Sci USA 1985;82:7048.
[8] Cornette JL, Margalit H, DeLisi C, Berzofsky JA. The amphipathic helix as a structural feature involved in T cell recognition. In: Epand RM, editor. The amphipathic helix. Boca Raton: CRC Press; 1993.
[9] Spouge JL, Guy HR, Cornette JL, Margalit H, Cease K, Berzofsky JA, et al. Strong conformational propensities enhance T cell antigenicity. J Immunol 1987;138:204.
[10] Stille CJ, Thomas LJ, Reyes VE, Humphreys RE. Hydrophobic strip of helix algorithm for selection of T cell-presented peptides. Mol Immunol 1987;24:1021.
[11] Meister GE, Roberts CG, Berzofsky JA, De Groot AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. Vaccine 1995;13:581.
[12] Stern LJ, Brown JH, Jardefzky TS, Gogra JC, Urban RG, Strominger JL, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 1994;368:215.
[13] Mouritsen S, Meldal M, Ruud-Hansen J, Werdelin O. T-helper-cell determinants in protein antigens are preferentially located in cysteine-rich antigen segments resistant to proteolytic cleavage by cathepsin BL, D Scand. J Immunol 1991;34:421.
[14] Rothbard JB, Taylor WR. A sequence pattern common to T cell epitopes. EMBO J 1988;7:93.
[15] Margalit H, Spouge JL, Cornette JL, Cease KB, DeLisi C, Berzofsky JA. Prediction of immunodominant helper T cell antigenic sites from the primary sequence. J Immunol 1987;138:2213.
[16] Deavin AJ, Authon TR, Greaney PJ. Statistical comparison of established T cell epitope predictors against a large database of human and murine antigens. Mol Immunol 1996;33:145.
[17] Brusic V, Rudy G, Harrison LC. Prediction of MHC binding peptides by using artificial neural networks. In: Complex mechanism of adaptation. Amsterdam: IOS Press; 1994. p. 253–60.
[18] Gulukota K, Sidney J, Sette A, DeLisi C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 1997;267:1258.
[19] Rammensee HG, Friede T, Stevanoviic S. MHC ligands and peptide motifs: first listing. Immunogenetics 41;1995:178 [Review].
[20] Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 1994;152:163.
[21] Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. J Immunol Methods 1995;185:181.
[22] Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinformatics 2001;17:1236.
[23] Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC class I binding sites. Bioinformatics 2003;19:1009.
[24] Bhasin M, Raghava GPS. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics 2003;20:421.
[25] Bhasin M, Singh H, Raghava GPS. MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics 2003;19:666.
[26] Korber B, Brander C, Haynes B, Koup R, Kuiken C, Moore J, et al. HIV monoclonal antibodies. In: HIV molecular immunology 2001. Los Alamus, New Mexico, USA: Theoretical Biology and Biophysics Group T-10, Mail Stop K710, Los Alamos National Laboratory; 2001. IV-B-1–278.
[27] Hertz JA, Palmer RG, Krogh AS. Introduction to theory of neural computation. Redwood City: Addison-Wesley; 1991.
[28] Zell A, Mamier G. Stuttgart Neural Network Simulator version 4.2. University of Stuttgart; 1997.
[29] Joachims T. Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smola A, editors. Advances in kernel methods—support vector learning. Cambridge, MA: MIIT Press; 1999.
[30] Cristianini N, Shawe-Taylor J. Support vector machines and other kernel-based learning methods. Cambridge, England: Cambridge University Press, The Edinburg Building; 2000.
[31] Vapnik VN. The nature of statistical learning theory. New York: Wiley; 1998.
[32] Doytchinova IA, Flower DR. 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 2001;44:3572.
[33] Reche PA, Glutting J, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. Human Immunol 2002;63: 701.
[34] Schönbach C, Yu K, Brusic V. Large-scale computational identification of HIV T-cell epitopes. Immunol Cell Biol 2002;80:300.