Immunome Research

BioMed Central

Research

Open Access

Prediction of MHC class II binding peptides based on an iterative learning model

Naveen Murugan and Yang Dai*

Address: Department of Bioengineering (MC063), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607, USA

Email: Naveen Murugan - nmurug1@uic.edu; Yang Dai* - yangdai@uic.edu

* Corresponding author

Published: 13 December 2005

Immunome Research 2005, **1**:6 doi:10.1186/1745-7580-1-6

Received: 25 October 2005
Accepted: 13 December 2005

This article is available from: http://www.immunome-research.com/content/1/1/6

Abstract

Background: Prediction of the binding ability of antigen peptides to major histocompatibility complex (MHC) class II molecules is important in vaccine development. The variable length of each binding peptide complicates this prediction. Motivated by a text mining model designed for building a classifier from labeled and unlabeled examples, we have developed an iterative supervised learning model for the prediction of MHC class II binding peptides.

Results: A linear programming (LP) model was employed for the learning task at each iteration, since it is fast and can re-optimize the previous classifier when the training sets are altered. The performance of the new model has been evaluated with benchmark datasets. The outcome demonstrates that the model achieves an accuracy of prediction that is competitive compared to the advanced predictors (the Gibbs sampler and TEPITOPE). The average areas under the ROC curve obtained from one variant of our model are 0.753 and 0.715 for the original and homology reduced benchmark sets, respectively. The corresponding values are respectively 0.744 and 0.673 for the Gibbs sampler and 0.702 and 0.667 for TEPITOPE.

Conclusion: The iterative learning procedure appears to be effective in prediction of MHC class II binders. It offers an alternative approach to this important predictionproblem.

Background

Immune responses are regulated and initiated by major histocompatibility complex (MHC) molecules, which bind to short peptides from antigens and display them on the cell surface for the recognition by T cell receptors. The specificity of this binding can be predicted from the amino acid sequence of a peptide. Such predictions can be used to select epitopes for use in rational vaccine design and to increase the understanding of roles of the immune system in infectious diseases, autoimmune diseases, and cancers.

There are two types of MHC molecules, class I and class II, and both are highly polymorphic. The core binding subsequence of both MHC I and II is approximately 9 amino acids long. However, the MHC I molecules rarely bind peptides much longer than 9 amino acids, while MHC II molecules can accommodate longer peptides of 10–30 residues [1-3]. The presence of the binding core with a uniform length for MHC I molecules makes the prediction of peptide-MHC binding relatively easier. Many different methods have been developed for the prediction of peptide-MHC binding, including simple binding motifs, quantitative matrices, hidden Markov models, and artificial neural networks [4-8]. These methods can be readily applied to MHC I molecules, since the binding motif is well characterized and most of the natural peptides that bind MHC I molecules are of close to equal length.

The prediction of MHC class II binding peptides is a difficult classification problem. MHC class II molecules bind peptides that are 10–30 amino acids long with a core region of 13 amino acids containing a primary and secondary anchor residues [2,9,6,10,11]. Analysis of binding motifs has suggested that a core of only 9 amino acids within a peptide is essential for peptide-MHC binding. Reported binding peptides usually have variable lengths and an undetermined core region for each peptide. Therefore, a search for the binding core region can circumvent the problem of variable lengths.

Efforts have been focused on how to align the peptides such that a block of the peptides can be identified as the binding cores. The alignment of peptides is searched based on evolutionary algorithms [12], the Gibbs sampling method [13], and a recent method motivated by the ant colony search strategy [14]. The former looks for a position scoring matrix with the highest fitness score (predictive power) through the genetic operator of mutation. The latter two methods attempt to find an optimal local alignment by means of Monte Carlo Metropolis sampling in the alignment space or by the collective search strategy of ant colony systems, respectively. The binding cores with same length are identified from the alignment, and a scoring matrix used for prediction is established from these binding cores. In the work of Brusic et al. [12], the alignment of peptides is treated as a pre-processing procedure. Upon the determination of the binding cores, a binary classifier is then learned with artificial neural networks using amino acid sequences presented in the binding core as a positive training set and other non-binding peptides as a negative training set. In Nielsen et al. [13] and Karpenko et al. [14], a position scoring matrix is obtained from the best alignment and used for scoring peptides. Most of these alignment-based predictors have achieved reasonably good performances. However, a common complication involved in these methods is the correct choice of associated parameters. The tuning of the parameters could be complicate. A similar work is by Bhasin et al. who used a pre-processing procedure called MOTs to filter the putative binding core for binding peptide sequences and subsequently trained the classifier based on the support vector machine (SVM) [15] with those binding core sequences and random sequences [16]. Another method using an iterative approach has been developed based on a stepwise discriminant analysis model [17,18]. More recently a model based on Bayesian neural networks has been developed [19].

This work is motivated by a machine learning model designed for a training task with only positive and unlabeled examples in text mining. This type of training set is in evidence in various applications in which the identification of a positive training example is labor intensive and time consuming. The basic idea developed for this learning task is the use of a binary classifier to filter out positive examples from the unlabeled set and include them into the positive set through an iterative procedure [20,21]. A classifier is trained at each iteration by simply assigning positive examples the label 1 and unlabeled examples the label -1 to form normal binary training sets. A classifier can be learned by the use of different binary classification methods such as the Naïve Bayesian or support vector machines.

The unlabeled and labeled examples in the prediction of peptide-MHC binding can be introduced naturally through the encoding mechanism. A sliding window scheme with a window length of 9 is applied to binding peptides. This procedure breaks a peptide into a set of nonamers of equal length. The binding core, which is unknown, is one of the nonamers. The nonamers from all the binding peptides serve as unlabeled examples in which the positive examples, i.e., nonamers of binding cores, are included. Similarly, all nonamers obtained from the non-binding peptides serve as negative examples. It is noted that the situation in this application is opposite to that of text mining. Here a negative set and an unlabeled set containing potential positive examples are presented. However, the same strategy described previously for text mining can be applied. The approach here is to filter out non-binding nonamers in the unlabeled set iteratively. This iterative learning model enables the use of the nonbinder information for the identification of the binding cores and to generate the predictor simultaneously. This is different from the three alignment based methods mentioned earlier in which the identification of binding cores relied only on binding peptides.

The linear programming (LP) model proposed by Bennett and Mangasarian [22] is used as the learning model for binary classification at each iteration. This model has several advantages over other learning methods such as support vector machines, Naïve Bayesian, and artificial neural networks. First, there are only a few parameters and they are very easy to tune. Second, a linear program can be solved very fast and it embodies favorable properties which allow sensitivity analysis. Therefore, if the subsequent linear program is only different for a small number of constraints, then the corresponding optimal solution can be found through a re-optimization procedure that uses the information of the current optimal solution. This is particularly important for the iterative learning proce-

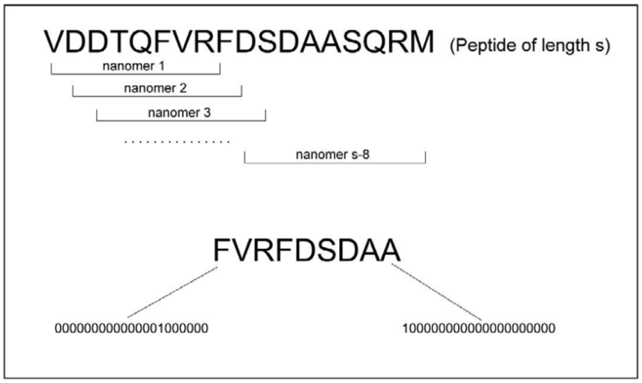


Figure 1
Top: A peptide has been reduced to a set of nonamers. Bottom: A nonamer is encoded as a 180-dimensional vector.

dure as only a small number of nonamers is removed from the positive training set at each iteration.

This model was evaluated with benchmark datasets from MHCBench against other major existing methods. The computational study demonstrates overall that this method can achieve comparable or superior performance in comparison with the competing predictors, such as the Gibbs sampler [13] and TEPITOPE [10]. The average areas under the ROC (Receiver Operating Characteristic) curve [23] obtained from one variant of our model are 0.753 and 0.715 for the original and homology reduced benchmark sets, respectively. The corresponding values are 0.744 and 0.673 for the Gibbs sampler and 0.702 and 0.667 for TEPITOPE.

Methods

LP model for classification

Consider a set of positive examples x_i, $i = 1,...,m_+$ and a set of negative examples x_i, $i = 1,...,m_-$, each of which is a point in an n-dimensional space. The LP model for a binary classification problem in (Bennett et al., 1992) is as follows.

$$\begin{aligned} &\text{Minimize} && \frac{1}{m_+}\sum_{i=1}^{m_+}\xi_i + \frac{1}{m_-}\sum_{i=1}^{m_-}\eta_i \\ &\text{Subject to} && w^T x_i + b \ge 1 - \xi_i, i = 1,\dots,m_+ \quad \text{for } y_i = 1 \\ &&& -w^T x_i - b \ge 1 - \eta_i, i = 1,\dots,m_- \quad \text{for } y_i = -1 \\ &&& \xi_i \ge 0, \eta_i \ge 0, \end{aligned} \qquad (1)$$

where y_i = 1 or -1 is the label assigned to each positive or negative example, respectively.

This model generates a separating hyperplane with the smallest amount of misclassification error. It has been proved that this linear program always returns a non-trivial solution of w, which permits a linear classification function, even in a non-linear separable case [22]. The decision function, denoted by $f(x) = w^T x+b$, assigns a label to an example x by the sign of $f(x)$.

LP model for MHC class II problem

A set of nonamers can be obtained by sliding a window of length 9 along each MHC class II binding peptide as described previously. A peptide of length s will have s - 8 nonamers (see top panel, Figure 1). These nonamers are further reduced to a set of putative nonamers based on the knowledge that the residue in the first position of the nonamer has to be hydrophobic in order for it to bind to an HLA-DR MHC II molecule. This set of putative nonamers is considered as an unlabeled set. Each nonamer in the unlabeled set is assigned the label 1 temporarily.

The negative set of nonamers can be obtained similarly from the non-binding peptides. Each nonamer in this set is assigned the label -1. All redundant nonamers in both sets are removed. The remaining nonamers are subject to further preprocessing steps, which will be described later.

An amino acid at each position of the nonamer can be encoded by a 20-dimensional vector. Each coordinate of the vector is either 1 or 0, representing the presence or the absence of a particular amino acid. Accordingly, each nonamer can then be represented as a 180-dimensional vector (see bottom panel, Figure 1).

Assume that there are m_+ binding and m_- non-binding peptides. Each encoding vector of a nonamer for a peptide i is denoted by x_{i_j}. Assume that each binder i permits i_k putative nonamers. By using the LP model given before, our problem can be formulated as the following linear program:

$$\begin{aligned} &\text{Minimize} && C_1\sum_{i=1}^{m_+}\sum_{j=1}^{i_k}\xi_{i_j} + C_2\sum_{i=1}^{m_-}\sum_{j=1}^{i_k}\eta_{i_j} \\ &\text{Subject to} && w^T x_{i_j} + b \ge 1 - \xi_{i_j}, i = 1,\dots,m_+, j = 1,\dots,i_k \quad \text{for } y_i = 1 \\ &&& -w^T x_{i_j} - b \ge 1 - \eta_{i_j}, i = 1,\dots,m_-, j = 1,\dots,i_k \quad \text{for } y_i = -1 \\ &&& \xi_{i_j} \ge 0, \eta_{i_j} \ge 0, \end{aligned} \qquad (2)$$

where C_1 and C_2 are coefficients that will be determined through cross-validation on the training set. Note that we have extended the LP model (1) by allowing the change of coefficients C_1 and C_2 associated with the error terms in the objective function in LP (1). This extension aims at the control of the weights on error terms so that some non-core nonamers in the positive set are deliberately misclassified. This is a chief characteristic of our learning model.

Iterative procedure

The iterative training process consists of the following major steps. First, a weight vector w and value b are obtained by solving the LP (2) for fixed C_1 and C_2. This

Table 1: Description of HLA-DR4 (B1*0401) benchmark datasets.

Set	Original Dataset			Homology Reduced Dataset		
	Total	Binders	Non Binders	Total	Binders	Non Binders
Set 1	1017	694	323	531	248	283
Set 2	673	381	292	416	161	255
Set 3a	590	373	217	355	151	204
Set 3b	495	279	216	325	128	197
Set 4a	646	323	323	403	120	283
Set 4b	584	292	292	375	120	255
Set 5a	117	70	47	110	65	45
Set 5b	85	48	37	84	47	37
Southwood	105	22	83	99	19	80
Geluk	22	16	6	21	15	6

solution is used to score each nonamer in the positive training set based on the function $f(x) = w^Tx+b$. Nonamers with negative scores from the positive set are moved to the negative set. Subsequently, the LP is solved for the altered training sets. This process is repeated for a number of iterations, which will be determined through cross-validation (CV). The function $f(x)$defined with the final LP solution w, and b is used for the prediction of peptides in the testing set. A peptide that has at least one positively scored nonamer is considered as a binder; and otherwise it is considered a non-binder. If several nonamers from one peptide have a positive score, then the nonamer with the highest score is considered as the binding core for that peptide. Note that there may be no binding core identified for certain binding peptides in the final positive training set.

In addition to the learning model described above, two other variants were considered. In the first variation, the nonamers in the positive set evaluated with a negative score were discarded instead of being appended to the negative set at each iteration, as these nonamers may not necessarily be true non-binders. In the second variation, at most two nonamers with the highest positive scores from each peptide were allowed to remain in the positive set and the remaining was discarded. The approach in this variant of the LP is motivated by the observation that the binding core is likely to occur among the high scoring nonamers for each peptide. (From our preliminary study on peptides from the training set with known binding core regions, it was observed that there was no significant improvement in performance from using the top three or four nonamers over the top two nonamers.)

These variants of the LP method are referred to as LP_append, LP_discard, and LP_top2 in the discussions below. For LP_append, LP_discard, the number of iterations for which the LP process is repeated and the coefficients C_1 and C_2 were determined by a 5-fold CV on the training set. For LP_top2, the CV procedure only determines the coefficients C_1 and C_2, since LP_top2 terminates after the second iteration. The area under the ROC curve was the criterion for the evaluation of predictors. The final predictor for each method was obtained by training the whole training set with the optimal parameters determined from the 5-fold CV. The linear programming package GLPK [24] was used to solve the LP given (2).

Data sets

*Training data sets for HLA-DR4 (B1*0401) allele*

The sequences of peptides binding to the MHC class II molecule HLA-DR4 (B1*0401) from the SYFPEITHI [6] and MHCPEP [12] databases were extracted. Since the SYFPEITHI database has more peptides now than in 1999, peptides sequences added to the database after year 1999 were eliminated to make it comparable to the dataset used in Nielsen et al. [13]. This set consists of 462 unique binding peptide sequences. Non-binders for the MHC class II molecule HLA-DR4 (B1*0401) were extracted from the MHCBN database [25]. This set consists of 177 unique non-binding peptides sequences.

The binding peptides that do not possess a hydrophobic residue (I, L, M, F, W, Y, V) at the first position in putative binding cores were removed [12]. That is, a peptide was removed if no hydrophobic residues are present at the first n-s+1 positions, where n is the peptide length and s is the length of the sliding window. The hydrophobic filter removed 27 peptides. Furthermore, the set was reduced by removing unnatural peptide sequences with an extreme amino acid content of more than 75% alanine. Thus, the pre-processing procedure gives 462 unique binding peptides and 177 unique non-binding peptides. The length distribution in the training set ranges from 9 to 30 residues, with the majority of peptides having a length of 13 amino acids. These peptide sequences were then used to obtain nonamers with the sliding window scheme described earlier. All redundant nonamers and nonamers

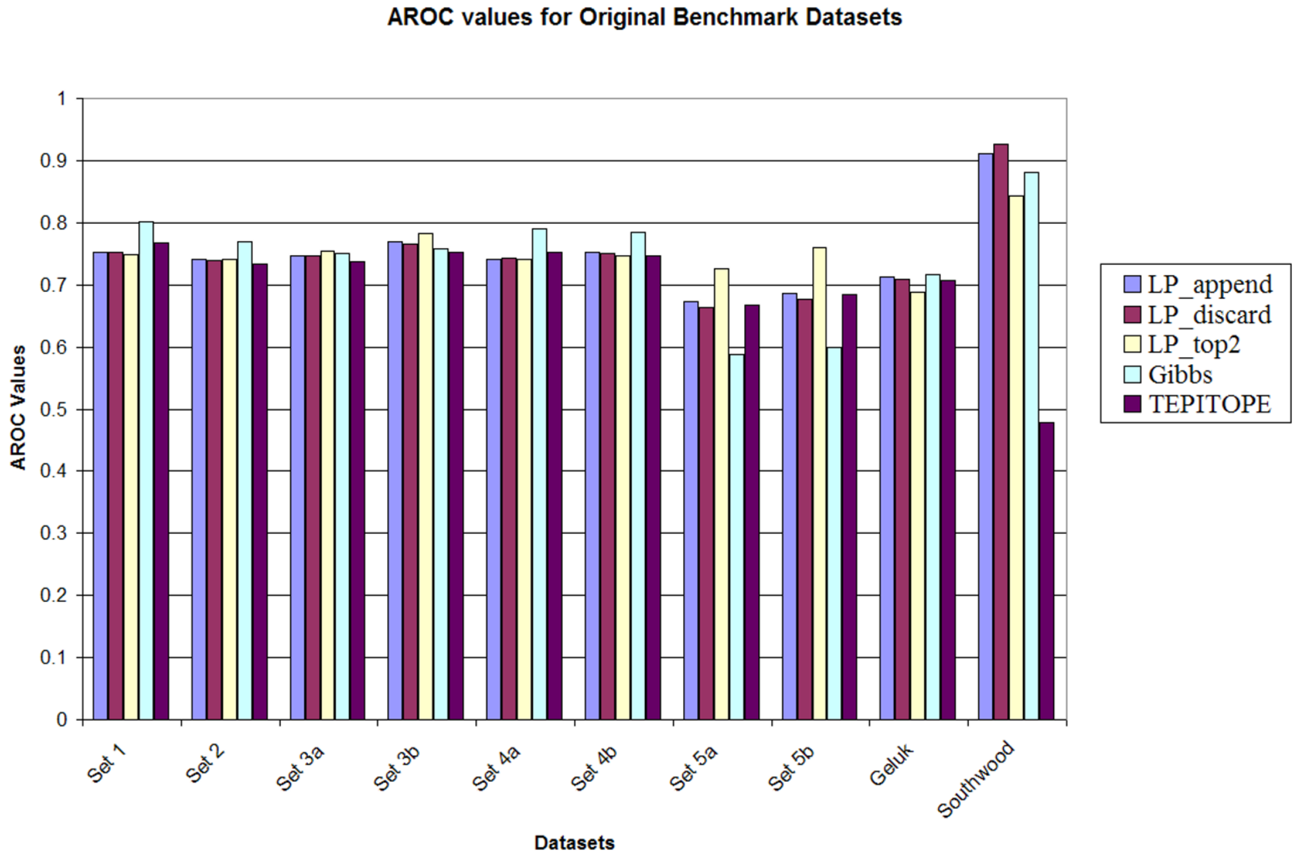


Figure 2
Prediction accuracy of the various methods on the original benchmark datasets.

that do not have a hydrophobic residue at position 1 were removed. The final numbers of nonamers obtained from the binding and non-binding peptides are 796 and 903, respectively.

*Testing data sets for HLA-DR4 (B1*0401) allele*

Ten benchmark datasets used in Nielsen [13] were considered in our study. These 10 datasets consist of the 8 datasets described in MHCBench [26] and 2 datasets described by Southwood [27] and Geluk [28]. The same procedure presented in Nielsen et al for the determination of binders and non-binders was followed in our study. More specifically, for the 8 MHCBench datasets, peptides with an associated binding value of zero were considered be non-binders, and all other peptides were binders. For the datasets of Southwood and Geluk, an affinity of 1000 nM was taken as the threshold for peptide binding [27]. In order to reduce the chance of over-prediction, the benchmarking was also performed on the homology-reduced datasets. The homology reduction was carried out so that no peptide in the evaluation sets had a match in the training set with sequence identity >90% over an alignment length of at least nine amino acids. Table 1 shows a summary of the original and the homology-reduced benchmark datasets, respectively. Note that there is small discrepancy in the numbers in some of the reduced sets compared with the ones reported in Nielsen [13] (From the email communication with Dr. Nielsen, there was an error in reporting the numbers in the table in their paper; however, the results on prediction presented there were based on the numbers shown in Table 1).

*Data sets of HLA-DRB1*0101 and HLA-DRB1*0301 for cross-validation test*

Two other datasets for the MHC class II molecules HLA-DRB1*0101 and HLA-DRB1*0301 were obtained from the MHCBN database [25]. The dataset for HLA-DRB1*0101 consists of 475 binder and 105 non-binder peptides. The dataset for HLA-DRB1*0301 consists of 219 binder and 150 non-binder peptides. The same pre-processing procedure described earlier was applied to these two sets.

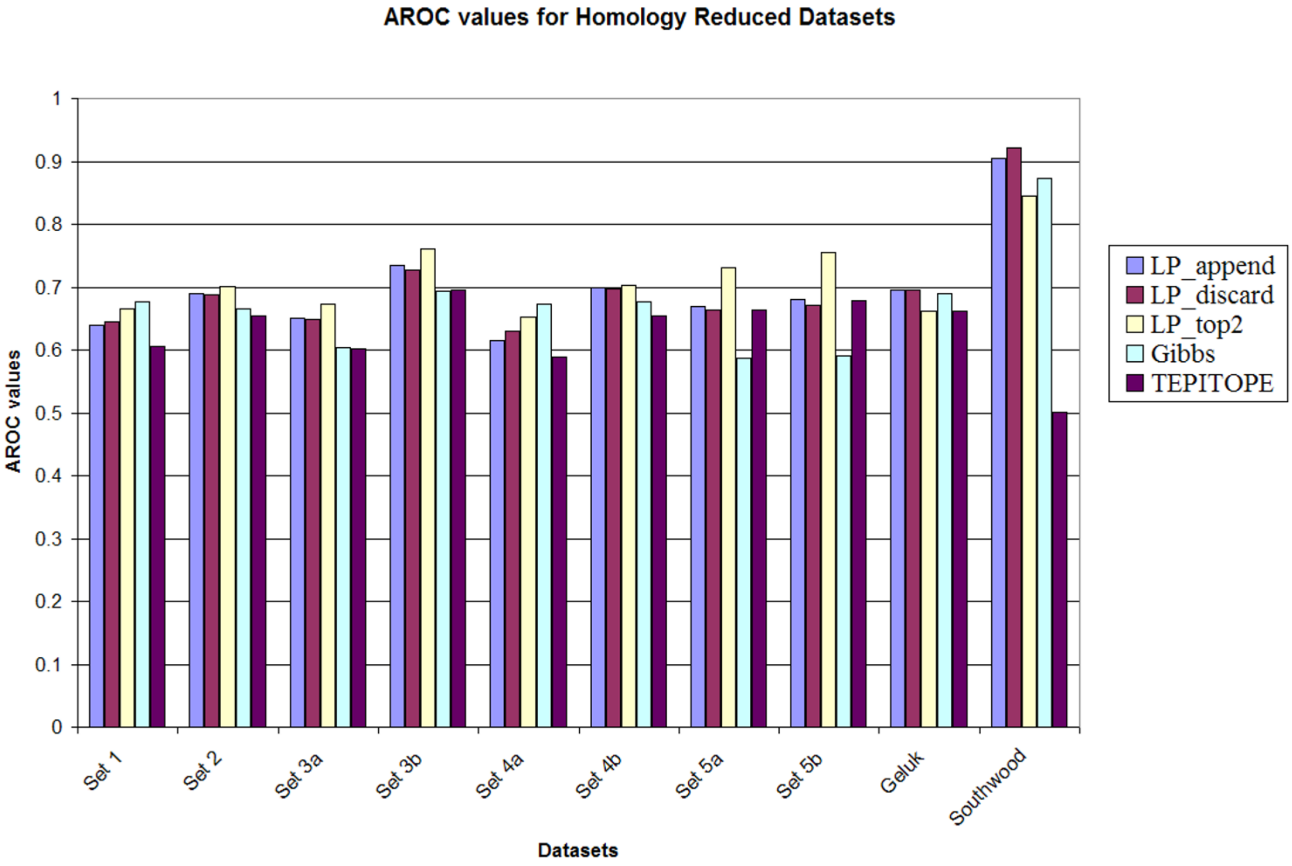


Figure 3
Prediction accuracy of the various methods on the homology reduced datasets.

Results

*Testing of the benchmark data for HLA-DR4(B1*0401)*

The results of the three methods on the benchmark datasets are compared with those obtained from the Gibbs sampling technique [13] and TEPITOPE [10]. The results of the Gibbs sampler were calculated with the scoring matrix provided by Dr. Nielsen; and the results of TEPITOPE were obtained with the use of the scoring matrix from ProPred [29], which is based on the one from TEPITOPE. The performance, evaluated by the area under the ROC curve (Aroc), of each method on the 10 benchmark datasets is presented in Figure 2 and Figure 3. Table 2 gives the performance of the methods averaged over the 10 benchmark datasets. It is observed that among the three proposed methods, LP_top2 has a slightly higher average Aroc value than those obtained from the other two variants. It is also observed that all the three LP variants have higher Aroc values compared to the Gibbs sampler and TEPITOPE.

A notable observation is that the performance of the Gibbs sampler appears to deteriorate for set 5A (0.588) and set 5B (0.600), whereas the LP methods maintain the performance for those two datasets, e.g., LP_top2 has Aroc values of 0.725 and 0.760 for the original benchmark sets 5a and 5b respectively. These two datasets have higher cysteine content as compared to the training set. However, since the LP methods use both binders and non-binders to train the classifier (unlike most of the other methods in which only binders are used for training), the LP methods are more robust in performance. In addition, upon testing the method by substituting all occurrences of cysteine in all the sets by alanine [13], it was observed that the LP_top2 method obtained Aroc values of 0.815 and 0.859 for the original benchmark sets 5a and 5b, respectively; while the Gibbs sampler obtained Aroc values of 0.621 and 0.661, respectively. The details of the results are provided in Tables S1, S2, S3 and S4 in the supplementary document (see Additional files add1.doc – add4.doc). Also TEPITOPE had a poor performance for the South-

Table 2: The average Aroc values for different methods.

Method	Average Aroc values for the benchmark datasets	
	Original	Homology Reduced
LP_append	0.749	0.698
LP_discard	0.748	0.699
LP_top2	**0.753**	**0.715**
Gibbs method	0.744	0.673
TEPITOPE (Propred)	0.702	0.667

wood dataset (Aroc values 0.703 and 0.630 for the original and homology datasets) due to a biased amino acid composition at position P1 and that if a modified TEPITOPE matrix at the P1 position was used, TEPITOPE could increase Aroc values to 0.786 and 0.794 for the original and homology reduced Southwood datasets, respectively [13]. For the other benchmark datasets, the performance of the modified TEPITOPE is similar to the original TEPITOPE matrix.

In order to investigate the statistical significance of the results, 1000 datasets were generated by random sampling N data points with replacement for each dataset. Here, *N* is the number of data points in the original dataset. The performance of the different methods was evaluated for each of the original and homology reduced datasets. It was observed that among the LP variants, LP_top2 had a slightly improved performance when compared to LP_append and LP_discard methods. However there was no significant difference observed in their performance. The overall average performance of the methods for the sampled datasets was also not very different from that for the original and homology reduced datasets. The details are provided in Tables S1 and S2 in the supplementary document (see Additional files add5.doc – add6.doc)

For comparison with the Gibbs sampler, the *p*-value for the hypothesis that the Gibbs method performs better than the LP method is estimated as the fraction of experiments where the Gibbs sampler has a better performance. LP_top2 was selected in this comparison. It was observed that for the original benchmark datasets, for 7 of the 10 datasets (sets 1, 2, 3a, 3b, 4a, 5a and 5b), LP_top2 performed better than the Gibbs sampling method ($p < 0.05$). For the remaining 3 datasets, there was no significant difference in performance ($0.05 < p < 0.95$). In case of the homology reduced datasets, for 8 of the 10 datasets (sets 1, 2, 3a, 3b, 4a, 4b, 5a and 5b), LP_top2 performed better than the Gibbs sampling method ($p < 0.05$). For the remaining 2 datasets there was no significant difference in performance ($0.05 < p < 0.95$). The same comparison was made between LP_top2 and TEPITOPE. It was observed that for the original benchmark datasets, for 2 of the 10 datasets (sets 5b and Southwood), LP_top2 performed better than the TEPITOPE sampling method ($p < 0.05$). For the remaining 8 datasets, there was no significant difference in performance ($0.05 < p < 0.95$). In case of the homology reduced datasets, for 7 of the 10 datasets (sets 1, 2, 3a, 3b, 4a, 5a and Southwood), LP_top2 performed better than TEPITOPE ($p < 0.05$). For the remaining 3 datasets there was no significant difference in performance ($0.05 < p < 0.95$). Details are given in Table S3 in the supplementary document (see Additional files add7.doc).

Results for cross-validation

The LP method (LP_top2) was also evaluated using a 5-fold cross-validation for the datasets of HLA-DRB1*0101 and HLA-DRB1*0301. The results were compared against those obtained from TEPITOPE (see Table 3). The TEPITOPE matrix was downloaded from ProPred [29], and was used on the testing folds. The LP method produced Aroc values 0.779 for HLA-DRB1*0101 data set and 0.721 for HLA-DRB1*0301 dataset. The corresponding values generated from TEPITOPE are 0.842 and 0.585, respectively. The LP method appears to be more consistent in performance over different alleles.

Prediction of binding core

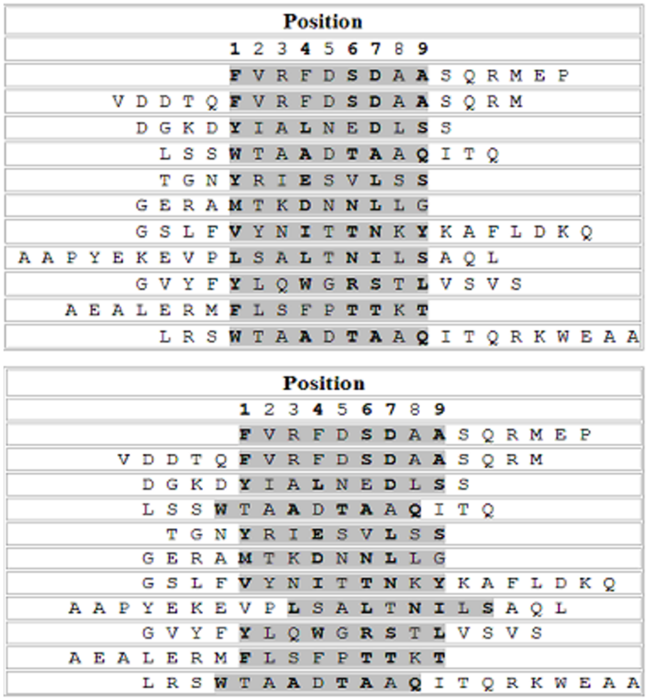
The predictive ability of the LP method (LP_append) for the identification of binding cores in binding peptides was assessed for the HLA-DR4 (B1*0401) allele. The 68 peptide sequences which have information on experimentally determined binding cores, contained in the SYFPEITHI database were used for the verification. Nonamers in the initial set of putative binding cores for the HLA-DR4(B1*0401) allele that are identical to any binding cores in the 68 binding peptides were removed. It resulted in a new training set of 755 binding nonamers. The same negative nonamer set for the HLA-DR4(B1*0401) allele was used. The classifier was trained with the use of the previously described procedure. Among the 68 binding peptides, Fifty one binders which produce distinct binding cores were selected from the 68 binders. However, 6 of those had cores with a length less than 9 amino acids. After the removal of these exceptions, 45 peptides were left for the testing.

The predicted binding core is considered as the nonamer with the highest score. The numbers of identified binding cores that were within two positions from the exact binding core by the LP method, TEPITOPE, and the Gibbs sampler are respectively 41, 43, and 42. That is, each identified binding core shares at least 7 consecutive residues with the reported cores. The reason for verifying the predicted core with a shift of a few positions of the reported binding core is because that the binding affinity is not completely determined by the binding core and the flanking amino acids on both sides of the real core may contribute to the

Table 3: The average Aroc values from 5-fold cross validations.

Method	HLA-DRB1*0101	HLA-DRB1*0301
LP_top2	0.779	0.721
TEPITOPE (Propred)	0.842	0.585

binding affinity and stability [19,30,31]. It should be noted that the Nielsen matrix used was obtained from the original training set, which includes those 68 binders. It appears that the three methods performed almost the same. The core alignment of 11 peptides out of the 45 testing peptides obtained from the LP method and the original core alignment from the SYFPEITHI database are presented in Figure 4.



```
                 Position
            1 2 3 4 5 6 7 8 9
            F V R F D S D A A S Q R M E P
  V D D T Q F V R F D S D A A S Q R M
    D G K D Y I A L N E D L S S
      L S S W T A A D T A A Q I T Q
      T G N Y R I E S V L S S
    G E R A M T K D N N L L G
    G S L F V Y N I T T N K Y K A F L D K Q
A A P Y E K E V P L S A L T N I L S A Q L
    G V Y F Y L Q W G R S T L V S V S
  A E A L E R M F L S F P T T K T
      L R S W T A A D T A A Q I T Q R K W E A A
```

```
                 Position
            1 2 3 4 5 6 7 8 9
            F V R F D S D A A S Q R M E P
  V D D T Q F V R F D S D A A S Q R M
    D G K D Y I A L N E D L S S
      L S S W T A A D T A A Q I T Q
      T G N Y R I E S V L S S
    G E R A M T K D N N L L G
    G S L F V Y N I T T N K Y K A F L D K Q
A A P Y E K E V P L S A L T N I L S A Q L
    G V Y F Y L Q W G R S T L V S V S
  A E A L E R M F L S F P T T K T
      L R S W T A A D T A A Q I T Q R K W E A A
```

Figure 4
Top: The alignment of actual binding cores (shadowed) from SYFPEITHI database. Bottom: The alignment of the predicted binding cores by the LP method.

Discussion

It is important to note that the Gibbs sampler involves a set of parameters that need to be optimized using a complicated procedure before the training, whereas the LP method is very simple and the only parameters that need to be determined are the coefficients for the misclassification errors and the number of iterations. Both these parameters are easily and very quickly determined through cross-validation. This process involves no modification when applied to peptide sequences respect to MHC alleles.

A similar iterative approach for predicting HLA DR1 alleles using a stepwise discriminant analysis (SDA) has been reported [17,18]. This approach trains a linear discriminant function at each iteration and uses it to evaluate nonamers obtained from the original binding peptide sequences. Those nonamers passed the prediction threshold forms the positive training set in the next iteration. Therefore, the positive training set is dynamically changing over iteration to iteration. The negative training set remains unchanged. In this sense, Mallios' method is similar to our LP_discard or LP_top2. The discriminative features are selected based on the F-statistic from a one-way analysis of variance. The Mallios model is essentially a multiple linear regression which minimizes the sum of squared errors, while our model minimizes a weighted sum of errors.

In a recent work, a Bayesian neural network [19] was used for the prediction of MHC class II peptide binding. They concluded that their method outperforms the neural network model [12] and the SVM model [16]. Since their datasets were not available, a direct comparison could not be performed.

Conclusion

An iterative supervised learning model has been developed for the prediction of peptide binding to MHC class II molecules. This approach was motivated by a model for building a classifier with the positive and unlabeled training sets in text mining. The major feature of this method is its iterative extracting of binding core nonamers. The iterative training process functions like an 'adaptive loop', feeding back useful information by validating against the training data. The results indicate that the performance of the new method for HLA-DR4 (B1*0401) allele is competitive to other methods. Furthermore, the method can incorporate new peptides into the training data easily. This feature makes the method far more adaptive. It is expected that the predictive accuracy will be improved, if the information on other key anchor positions is incorporated [13] and a support vector machine learning model is adapted.

Additional material

Additional File 1

This file includes Table S1 – The average of Aroc values and standard deviation for the 1000 random sampling datasets on the original benchmark datasets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S1.doc]

Additional File 2

This file includes Table S2 – The average of Aroc value and standard deviations for the 1000 random sampling datasets on the homology reduced benchmark datasets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S2.doc]

Additional File 3

This file includes Table S3 – P values for the statistical tests.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S3.doc]

Additional File 4

This file includes Table S4 – The Aroc values for the original benchmark datasets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S4.doc]

Additional File 5

This file includes Table S5 – The Aroc values for the reduced benchmark datasets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S5.doc]

Additional File 6

This file includes Table S6 – The Aroc values for the original benchmark datasets (Cysteine substituted).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S6.doc]

Additional File 7

This file includes Table S7 – The Aroc values for the reduced benchmark datasets (Cysteine substituted).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-7580-1-6-S7.doc]

Acknowledgements

The authors are grateful to Lei Huang for helpful discussion and Deepa Vijayaraghavan for the assistance with the computing environment. The authors also thank Dr. Morten Nielsen for sharing of the homology reduced dataset and a scoring matrix. This research is supported in part by National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016).

References

1. Sette A, Buus S, Appella E, Smith JA, Chesnut R, Miles C, Colon SM, Grey HM: **Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis.** *Proc Natl Acad Sci USA* 1989, **86:**3296-3300.
2. Max H, Halder T, Kropshofer H, Kalbus M, Muller CA, Kalbacher H: **Characterization of peptides bound to extracellular and intracellular HLA-DR1 molecules.** *Hum Immunol* 1993, **38:**193-200.
3. Castellino F, Zhong G, Germain RN: **Antigen presentation by MHC class II molecules: invariant chain function protein trafficking and the molecular basis of diverse determinant capture.** *Hum Immunol* 1997, **54:**159-169.
4. 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:**163-175.
5. Brusic V, Rudy G, Harrison LC: **Prediction of MHC binding peptides using artificial neural networks.** In *Complex Systems: Mechanism of Adaptation* Edited by: Stonier RJ, Xu XH. Amsterdam, Holland: IOS Press; 1994:253-260.
6. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanoviæ S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50:**213-219.
7. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S: **Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach.** *Tissue Antigens* 2003, **62:**378-384.
8. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O: **Reliable prediction of T-cell epitopes using neural networks with novel sequence representations.** *Protein Sci* 2003, **12:**1007-1017.
9. Marshall KW, Wilson KJ, Liang J, Woods A, Zaller D, Rothbard JB: **Prediction of peptide affinity to HLA DRB1*0401.** *J Immunol* 1995, **154:**5927-5933.
10. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J: **Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices.** *Nat Biotechnol* 1999, **17:**555-561.
11. Borras-Cuesta F, Golvano J, Garcia-Granero M, Sarobe P, Riezu-Boj J, Huarte E, Lasarte J: **Specific and general HLA-DR binding motifs: comparison of algorithms.** *Hum Immunol* 2000, **61:**266-278.
12. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L: **Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network.** *Bioinformatics* 1998, **141:**121-130.
13. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: **Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach.** *Bioinformatics* 2004, **20:**1388-1397.
14. Karpenko O, Shi J, Dai Y: **Prediction of MHC class II binders using the ant colony search strategy.** *Artificial Intelligence in Medicine* 2005, **35:**147-156.
15. Vapnik VN: *The Nature of Statistical Learning Theory* Springer; 1995.
16. Bhasin M, Raghava GPS: **SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence.** *Bioinformatics* 2004, **20:**421-423.
17. Mallios RR: **Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 1999, **15:**432-439.
18. Mallios RR: **Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 2001, **17:**942-948.
19. Burden FR, Winkler DA: **Predictive Bayesian neural network models of MHC class II peptide binding.** *J Mol Graph Model* 2005, **23:**481-489.
20. Lee WS, Liu B: **Learning with positive and unlabeled examples using weighted logistic regression.** *Proceedings of International Conference on Machine Learning (ICML)* 2003:448-455.
21. Liu B, Dai Y, Li X, Lee WS, Yu P: **Building text classifiers using positive and unlabeled examples.** In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)* Melbourne Florida; 2003:179-188.
22. Bennett KP, Mangasarian OL: **Robust Linear Programming Discrimination of Two Linearly Inseparable Sets.** *Optimization Methods and Software* 1992, **1:**23-34.
23. Swets JA: **Measuring the accuracy of diagnostic systems.** *Science* 1988, **240:**1285-1293.
24. **GLPK** [http://www.gnu.org/software/glpk/glpk.html]
25. Bhasin M, Singh H, Raghava GPS: **MHCBN: a comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics* 2003, **19:**665-666.
26. **MHCBench** [http://www.imtech.res.in/raghava/mhcbench]

27. Southwood S, Sidney J, Kondo A, del Guercio MF, Appella E, Hoffman S, Kubo RT, Chesnut RW, Grey HM, Sette A: **Several common HLA-DR types share largely overlapping peptide binding repertoires.** *J Immunol* 1998, **160:**3363-3373.
28. Geluk A, van Meijgaarden KE, Schloot NC, Drijfhout JW, Ottenhoff TH, Roep BO: **HLA-DR binding analysis of peptides from islet antigens in IDDM.** *Diabetes* 1998, **47:**1594-1601.
29. Singh H, Raghava GPS: **ProPred: prediction of HLA-DR binding sites.** *Bioinformatics* 2001, **17:**1236-1237.
30. Gulukota K, Sidney J, Sette A, DeLisi C: **Two complementary methods for predicting peptides binding major histocompatibility complex molecules.** *Journal of Molecular Biology* 1997, **267:**1258-1267.
31. Godkin AJ, Smith KJ, Willis A, Tejada-Simon MV, Zhang J, Elliott T, Hill AVS: **Naturally processed HLA Class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions.** *J Immunol* 2001, **166:**6720-6727.