Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Prediction of MHC binding peptide using Gibbs motif sampler, weight matrix and artificial neural network

Satarudra Prakash Singh[1, 2] and Bhartendu Nath Mishra[2, *]

[1]Amity Institute of Biotechnology, Amity University Uttar Pradesh, Gomti Nagar, Lucknow-226010, India; [2]Department of Biotechnology, Institute of Engineering and Technology, U.P. Technical University, Sitapur Road, Lucknow-226021, India;
Bhartendu Nath Mishra* - Email: profbnmishra@gmail.com; * Corresponding author

received September 21, 2008; accepted November 05, 2008; published December 06, 2008

Abstract:
The identification of MHC restricted epitopes is an important goal in peptide based vaccine and diagnostic development. As wet lab experiments for identification of MHC binding peptide are expensive and time consuming, *in silico* tools have been developed as fast alternatives, however with low performance. In the present study, we used IEDB training and blind validation datasets for the prediction of peptide binding to fourteen human MHC class I and II molecules using Gibbs motif sampler, weight matrix and artificial neural network methods. As compare to MHC class I predictor based on sequence weighting (Aroc=0.95 and CC=0.56) and artificial neural network (Aroc=0.73 and CC=0.25), MHC class II predictor based on Gibbs sampler did not perform well (Aroc=0.62 and CC=0.19). The predictive accuracy of Gibbs motif sampler in identifying the 9-mer cores of a binding peptide to DRB1 alleles are also limited (40%), however above the random prediction (14%). Therefore, the size of dataset (training and validation) and the correct identification of the binding core are the two main factors limiting the performance of MHC class-II binding peptide prediction. Overall, these data suggest that there is substantial room to improve the quality of the core predictions using novel approaches that capture distinct features of MHC-peptide interactions than the current approaches.

Keywords: MHC; weight matrix; ANN; Gibbs sampler; motif; epitope

Abbreviations: ANN-artificial neural network; MHC-major histocompatibility complex; Aroc-area under receiver operating characteristic; CC-correlation coefficient; IEDB- immune epitope database

Background:
A major task of the immune system is to identify cells that have been infected by pathogens and discriminate them from healthy cells. This is realized by the MHC class-I and II antigen processing and presentation pathway and the duty is assigned to helper T-lymphocytes (HTL) and cytotoxic T-lymphocytes (CTL). The activation of CD8+ cytotoxic T-cells in the immune system requires presentation of endogenous antigenic peptides by MHC class-I molecules **[1]**. The activation of CD4+ helper T-cells is also essential for the development of adaptive immunity against pathogens. A critical step in CD4+ T cell activation is the recognition of exogeneous peptides presented by MHC class-II molecules **[2]**. The peptides bound to MHC molecules that trigger an immune response are referred as T-cell epitopes. Identifying T-cells epitopes is of high importance to immunologists, because it allows the development of diagnostics, peptide based vaccine and immunotherapy **[3]**. Therefore, the computational prediction of MHC class-I and II binding epitopes is of immense importance as their experimental identification is costly and time consuming **[4, 5]**.

A number of prediction methods for MHC binding peptides have been developed using peptide binding data from different databases such as SYFPEITHI **[6]**, MHCBN **[7]**, AntiJen **[8]** and IEDB **[9]**. The first method was based on the identification of allele-specific anchor residues [10]. This simple motif-based method was later replaced by various weight matrix-based methods **[11, 12]**. Similarly, other methods were based on scoring matrices derived from multiple peptide alignments such as RANKPEP **[13]** and the contribution of different residues in a peptide binding based on quantitative binding data such as ARB **[14]** and SMM-align **[15]**. The accumulation of more epitope data resulted into the development of different types of machine learning algorithms for prediction, including support vector machines **[16]** and artificial neural networks **[17]**. The other methods are also available based on structural template information for the prediction of MHC binding peptides **[18, 19]**. In order to assess the current state of the MHC class-I and II binding peptide predictions, a number of research groups have established a systematic and quantitative benchmarks **[20, 21]**.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)

Despite, the large number of available computational methods, prediction of MHC class-I and II restricted epitopes remains a challenging problem even today. An essential step in developing accurate prediction tool is to gather a set of experimentally consistent training and validation dataset. In present study, we compiled a large IEDB dataset for training and blind datasets for validation to the fourteen MHC class- I and II molecules (seven for each class) that were experimentally determined under uniform conditions collected from Dana-Farber Repository (http://bio.dfci.harvard.edu/DFRMLI/). The computational methods like, Gibbs motif sampler **[22]**, sequence weighting schemes **[23, 24]** and feed-forward backpropagation ANN **[25]** were used to predict the peptide binding to MHC molecules. The Gibbs sampler and weight matrix approaches are well suited to describe sequence motifs of fixed length. For MHC class I and II, the peptide binding motif is in most situations assumed to be of a fixed length of 9 amino acids. The weight-matrix approach is only suitable for prediction of a binding event in situations where the binding specificity can be represented independently at each position in the motif and this assumption can only be considered to be an approximation. In the binding of a peptide to the MHC molecule the amino acids might for instance compete for the space available in the binding grove. The neural networks with a hidden layer are designed to describe sequence patterns with such a higher order correlations. The superiority of these sequence based approaches to the structure are believed to be the consequence of two main features, i.e. the flexibility in optimizing the Gibbs motif sampling parameter and sequence weighting schemes and also in optimizing the ANN training parameter according to the dataset. Finally, the developed prediction models would predict the HTL and CTL epitopes, which shall provide better insight into further research of peptide based vaccine and diagnostics against diseases ranging from malaria to cancer.

Methodology:

Data collection

We assembled dataset of peptide binding and nonbinding affinities for fourteen MHC class-I and II molecules (seven for each class) from DRFMLI repository (http://bio.dfci.harvard.edu/DFRMLI/). These dataset of high quality MHC binding and nonbinding peptides were taken from IEDB database [9] (Table 1 and 2, see supplementary material). The binding affinities (IC_{50}) of these peptides were quantitatively measured by immunological experiments. They were then scaled to binding scores ranging from 0 to 100 using linear transformation **[21]**, where score >=99 are strong binders, 90-98 are moderate binders, 33-89 are border cases and <33 are non-binders. These dataset were used as the training data to develop computational models based on Gibbs sampler, sequence weighting and ANN to predict MHC binding peptides.

Three sets of validation data were used to evaluate the prediction performance of MHC class-I binding peptides. First, referred as survivin dataset derived from a full overlapping study of 134 nonamer peptides spanning the full length of the tumor antigen survivin, second, CMV dataset contains 42 peptides spanning a 50 amino acids long construct containing cytomegalovirus (CMV) internal matrix protein pp65 peptides and third, combination of survivin and CMV dataset referred as combined-I dataset contains 176 peptides for each seven human MHC class-I molecules. One hundred three binding and nonbinding peptides were derived from four protein antigens, i.e. bee venom allergen, LAGE-1, dog allergen Can f1 and Nef protein for each seven MHC class-II molecules, referred as combined-II dataset for the validation of predictions. The original binding scores were measured by iTopia™ Epitope Discovery System and then scaled to scores ranging from 0 to100 (Table 1 and 2, see supplementary material). In an attempt to check the ability of Gibbs sampler method to predict the 9-mer peptide cores revealed in crystal structures of MHC-peptide complexes, a total of 10 structures were compiled from Protein Data Bank for DRB1 alleles (Table 3 in supplementary material).

Algorithms used for the prediction of MHC binding peptides

Gibbs motif sampler

MHC class-II binding peptides have a broad length distribution complicating the development of prediction methods. Identifying the correct alignment of a set of peptides known to bind the MHC class-II molecule using Gibbs motif sampler is a crucial part of the algorithm to identify the core of an MHC class-II binding peptide **[22]**. Here, we used the default Gibbs sampling parameters to find the 9-mer motif in a set of MHC class-II binding peptide data using the web-server EasyGibbs available at http://www.cbs.dtu.dk/biotools/EasyGibbs/.

Sequence weighting

Three different sequence weighting methods i.e. Henikoff and Henikoff 1/nr **[23]**, clustering at 62% identity **[24]** and no clustering are available, which can be used to weight 9-mer peptide sequences. The Henikoff method is fast as the computation time only increases linearly with the number of sequences, whereas in the Hobohm clustering algorithm, computation time increases as the square of the number of sequences. Here, we used the web-server EasyPred available at http://www.cbs.dtu.dk/biotools/EasyPred/ to generated the weight matrix for the prediction of MHC binding peptides by applying all three sequence weighting schemes with weight on pseudo counts is 200.

Artificial neural network

Here, we used a conventional feed-forward neural network **[25]** with an input layer (180 neurons), one hidden layer (2 neurons) and a single neuron output layer using the web server Easypred available at http://www.cbs.dtu.dk/biotools/EasyPred/. The default setting parameters (one bin for balanced training, running upto 300 training epochs and top 80% of the training set) were used to train the neural network.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)

Evaluation parameters

Based on these datasets and algorithms, we have developed computational models which could predict the binding affinity between MHC molecules and peptides. The efficiency of algorithms was determined by discrimination between binders and nonbinders. A predicted peptide belongs to one of the four categories, i.e. True Positive (TP); an experimentally binding peptide predicted as a binder, False Positive (FP); an experimentally nonbinding peptide predicted as a binder, True Negative (TN); an experimentally nonbinding peptide predicted as a nonbinder and False Negative (FN); an experimentally binding peptide predicted as nonbinder. Here, we used non-parametric performance measures, area under receiver operator characteristic (Aroc) curve and Pearson correlation coefficient (CC) to evaluate the predictive performance of the applied algorithms. The ROC curve is a plot of the true positive rate TP/(TP+FN) on the vertical axis *vs* false positive rate FP/(TN+FP) on the horizontal axis for the complete range of the decision thresholds and the Pearson correlation coefficient (CC) is used to measure the association between pairs of values i.e. predicted and experimental **[26]**.

Discussion:

We assembled a dataset of peptide binding and nonbinding affinities for fourteen MHC class-I and II molecules (seven for each class) from DRFMLI repository (http://bio.dfci.harvard.edu/DFRMLI/). The Table 1 and 2 (shown under supplementary material) gives an overview of the training and validation dataset, encompassing a total of 16,771 peptide training data determined experimentally including 10,303 MHC class-I and 6,468 MHC class-II binding affinities **[21]**. Compared to the training datasets publicly available on the IEDB database **[9]**, our evaluation dataset expands the number of measured peptide-MHC interactions, 1,232 for MHC class-I from Survivin and CMV whereas, 712 for MHC class-II molecules from four protein antigens bee venom allergen, LAGE-1, dog allergen Can f1 and Nef protein. As the validation dataset not included in IEDB database, it is equivalent to a blind test. From the experimental data, peptides were classified into binders (IC_{50}<1000 nM) and nonbinders (IC_{50}≥1000 nM) based on measured affinities. From these dataset, the performance of the prediction methods were then measured by area under ROC curves (Aroc) and Pearson correlation coefficient (CC). The calculation of Aroc provides a highly useful measure of prediction quality, which is 0.5 for random predictions and 1.0 for perfect predictions and correlation coefficient value of one corresponds to a perfect correlation, a value of zero corresponds to a random prediction and a value of minus one to a perfect anti-correlation.

The prediction performances of the weight matrix are better than the non-linear predictor (ANN) for the MHC class-I molecules using all the validation datasets measured in terms of Aroc and CC (Figure 1). The weight matrix performance in term of Aroc and CC is maximum (1.0, 0.69) using CMV validation dataset and the minimum (0.85, 0.51) using combined-I validation dataset for allele HLA-A*1101 (Figure 1). The prediction performance of the Gibbs motif sampler for the DRB1*1301(MHC class-II) molecule is maximum (Aroc=0.71, CC=0.32) using Henikoff and Henikoff 1/nr weighting scheme (Figure 2) which is lower than the minimum performance of MHC class I binding prediction (Figure 1). The average performance of MHC class-II predictor based on Gibbs sampler (Aroc=0.62, CC=0.19) is also lower than the MHC class-I predictor based on sequence weighting (Aroc=0.95, CC=0.56) and artificial neural network (Aroc=0.73, CC=0.25).

From the above results it is clear that the size of training dataset may be an important factor contributing to better performance of sequence weighting and artificial neural network methods for the prediction of MHC class-I binding peptides. A key difference between MHC class-I and MHC class-II molecule is that the binding groove of class-II molecules is open at both ends. As a result, the length of peptide binding to class-II molecules can vary considerably, typically ranges 13-25 amino acids long. Therefore, a requisite for all MHC class-II binding prediction approaches is the capacity to identify the correct 9-mer core residues within longer peptide sequences that mediate the binding interaction. For Gibbs sampler method, we compared the predicted cores with the true cores extracted from crystal structures (Table 3 under supplementary material). Gibbs sampler methods had limited success (40%) as shown in the Table 4 (see supplementary material), although they still perform above random prediction (the probability to randomly guess the right core for a 15-mer peptide is 1 out of 7 or 14%). Thus the Gibbs sampler and weight matrix approaches are only suited to describe sequence motifs of fixed length e.g. 9-mer amino acids and suitable for prediction of a binding event in situations, where the binding specificity can be represented independently at each position in the motif This assumption can only be considered to be an approximation and in the binding of a peptide to the MHC molecule the amino acids might for instance compete for the space available in the binding grove where artificial neural networks with a hidden layer are generally used to describe sequence patterns. Overall, these data suggest that there is a substantial room to improve the quality of the core predictions using novel approaches that capture distinct features of MHC-peptide interactions.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

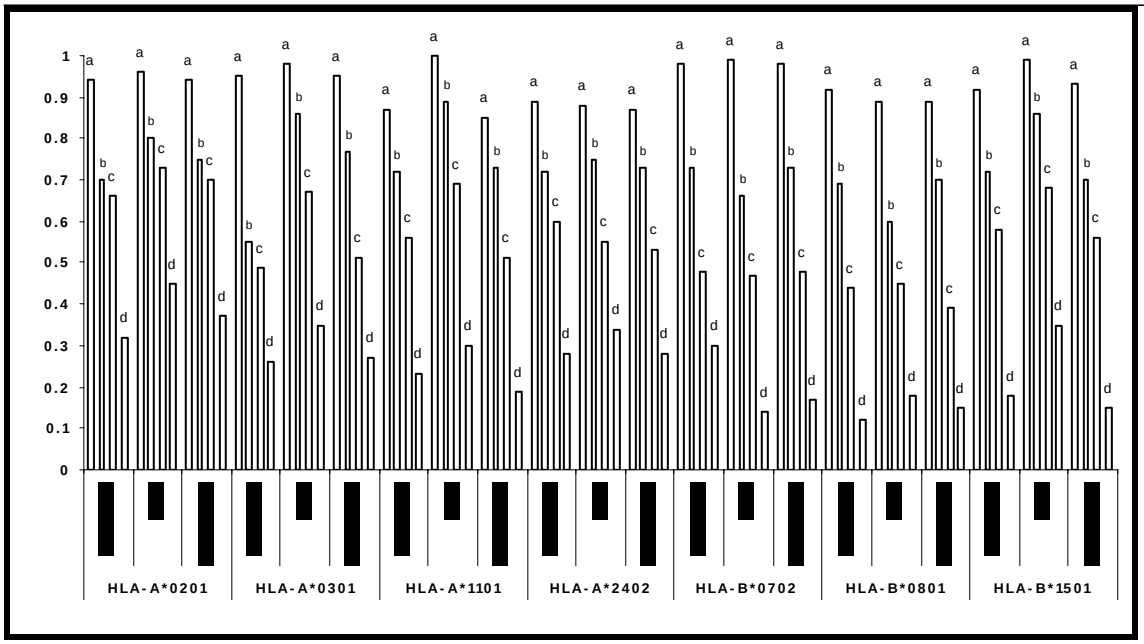

Figure 1: Performance of the MHC class I binding peptide predictor based on sequence weighting schemes and ANN using the validation datasets Survivin, CMV and Combined-I. Vertical axes shows the values of Aroc (a for weight marix, b for ANN) and CC (c for weight matrix, d for ANN) while horizontal axes shows HLA class I alleles and validation datasets.

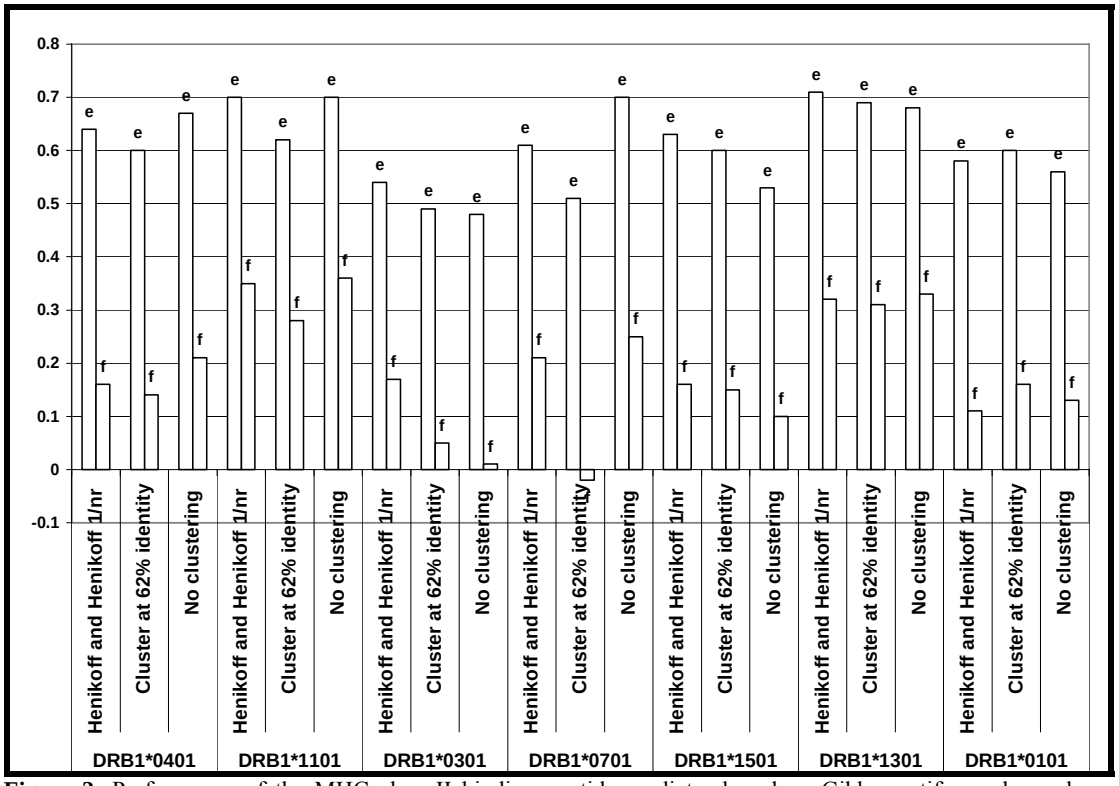

Figure 2: Performance of the MHC class II binding peptide predictor based on Gibbs motif sampler and sequence weighting schemes using the validation dataset combined-II. Vertical axes shows the values of Aroc (e) and CC (f) while horizontal axes shows HLA class II alleles and sequence weighting schemes.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)

Conclusion:

As, the identification of MHC class-I and II restricted epitopes using wet lab experiments are expensive and time consuming, the computational methods can be used as a fast alternatives. Although the prediction of peptide that bind to MHC class-II did not perform well as the MHC class-I molecules, however, it is able to identify the 9-mer cores of a binding peptide with limited accuracy (40%) above the random prediction (14%). Therefore, the size of dataset (training and validation) and the correct identification of the binding core are the two main factors limiting the performance of MHC binding prediction and thus, there is a substantial room to improve the quality of the core predictions. Finally, we hope that novel approaches that capture distinct features of MHC class-I and II peptide interactions could lead to more useful predictions than the current approaches.

Acknowledgment:

We are grateful to Mr. Akhilesh Singh, Amity Institute of Biotechnology, Amity University, Lucknow campus for their critical reading of the manuscript and valuable suggestions.We are also thankful to U.P Technical University, Lucknow and Amity University Uttar Pradesh, Lucknow for their laboratory support to research work.

References:

[01] B. Lankat-Buttgereit and R.Tampe, *Physiol Rev*., 82: 187 (2002) [PMID: 11773612]

[02] M. G. Rudolph, *et al., Annu Rev Immunol*., 24: 419 (2006) [PMID:16551255]

[03] L. F. Wang and M.Yu, *Curr Drug Targets*, 5: 1 (2004) [PMID: 14738215]

[04] C. Sylvester-Hvid, *et al.,Tissue Antigens*, 63: 395 (2004) [PMID:15104671]

[05] S. P. Singh, *et al*., *Online Journal of Bioinformatics*, 7: 69 (2006)

[06] H. Rammensee, *et al., Immunogenetics,* 50: 213 (1999) [PMID: 10602881]

[07] M. Bhasin, *et al., Bioinformatics,* 19: 665 (2003) [PMID: 12651731]

[08] C. P. Toseland, *et al.*, *Immunome research,* 6: 4 (2005) [PMID: 16305757]

[09] Q. Zhang, *et al*., *Nucleic Acids Res*.36: W513 (2008) [PMID: 18515843]

[10] H. G. Rammensee, *et al., Immunogenetics,* 41: 178 (1995) [PMID: 7890324]

[11] K. C. Parker, *et al., J Immunol,* 152: 163 (1994) [PMID: 8254189]

[12] H. Singh and G.P.S. Raghava, *Bioinformatics,* 17: 1236 (2001) [PMID: 11751237]

[13] P. A. Reche, *et al., Immunogenetics,* 56: 405 (2004) [PMID: 15349703]

[14] H. H. Bui, *et al., Immunogenetics,* 57: 304 (2005) [PMID: 15868141]

[15] M. Nielsen, *et al., BMC Bioinformatics*, 8: 238 (2007) [PMID: 17608956]

[16] P. Donnes and A. Elofsson, *BMC Bioinformatics,* 3: 25 (2002) [PMID: 12225620]

[17] M. C. Honeyman, *et al., Nat Biotechnol,* 16: 966 (1998) [PMID: 9788355]

[18] S. P Singh and B. N. Mishra, *Bioinformation*, 3: 72 (2008)

[19] O. Schueler-Furman *et al., Protein Sci.,* 9: 1838 (2000) [PMID: 11045629]

[20] B. Peters, *et al., PLoS Comput Biol*, 2: e65 (2006) [PMID: 16789818]

[21] P. Wang, *et al., PLoS Comput Biol*, 4: e1000048 (2008) [PMID: 18389056]

[22] M. Nielsen, *et al., Bioinformatics*, 20: 1388 (2004) [PMID: 14962912]

[23] U. Hobohm, *et al., Protein Sci*., 1: 409 (1992) [PMID: 1304348]

[24] S. S. Henikoff and J.G. Henikoff, *J. Mol. Biol*., 243: 574 (1994) [PMID: 7966282]

[25] M. Nielsen, *et al., Protein Sci*., 12: 1007 (2003) [PMID: 12717023]

[26] J. A. Swets, *Science,* 240: 1285 (1988) [PMID: 3287615]

Edited by P. Kangueane

Citation: Singh & Mishra, Bioinformation 3(4): 150-155 (2008)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Tables:

S.No	MHC allele	Total no. of IEDB training dataset	Total no. of positive IEDB training set #	Total no. of validation dataset		
				Survivin	CMV	Combined-I
1	HLA-A*0201	3087	1641	134	42	176
2	HLA-A*0301	2093	838	134	42	176
3	HLA-A*1101	1983	953	134	42	176
4	HLA-A*2402	195	138	134	42	176
5	HLA-B*0702	1261	319	134	42	176
6	HLA-B*0801	708	92	134	42	176
7	HLA-B*1501	976	299	134	42	176

Table 1: MHC class-I binding peptides used in the study. # Number of binding peptides scaled score>33.

S.No	MHC allele	Total no. of IEDB training dataset	Total no. of positive IEDB training set**	Total no. of validation dataset (combined-II)
1	DRB1*0401	512	423	103
2	DRB1*1101	520	440	103
3	DRB1*0301	502	400	103
4	DRB1*0701	502	400	103
5	DRB1*1501	520	396	103
6	DRB1*1301	30	30	103
7	DRB1*0101	1974 #	1974	103

Table 2: MHC class-II binding peptides used in the study. # Number of strong binders, i.e. scaled binding score<99; **Number of peptides binding scaled score>33.

9-mer Core	Binding Peptide	Chain	PDB ID	MHC class-II allele
FKGEQGPKG	AGFKGEQGPKGEPG	E	2FSE	DRB1*0101
IGILNAAKV	GELIGILNAAKVPAD	C	1KLG	DRB1*0101
VIPMFSALS	PEVIPMFSALSEGATP	C	1SJE	DRB1*0101
WRFLRGYHQ	GSDWRFLRGYHQYA	C	1AQD	DRB1*0101
YSDQATPLL	AAYSDQATPLLLSPR	C	1T5W	DRB1*0101
YVKQNTLKL	PKYVKQNTLKLAT	C	2G9H	DRB1*0101
MRMATPLLM	PVSKMRMATPLLMQA	C	1A6A	DRB1*0301
MRADAAAGG	AYMRADAAAGGA	E	2SEB	DRB1*0401
YVKQNTLKL	PKYVKQNTLKLAT	C	1J8H	DRB1*0401
VHFFKNIVT	ENPVVHFFKNIVTPR	C	1BX2	DRB1*1501

Table 3: MHC class II-peptide complex structures used to evaluate the performance of Gibbs sampler method.

S.No.	MHC class-II allele	PDB ID	Known core	Number of core regions identified correctly using Gibbs sampling
1	DRB1*0101	2FSE,1T5W,2G9H,KLG, 1SJE, 1AQD	6	2
2	DRB1*0301	1A6A	1	0
3	DRB1*0401	2SEB,1J8H	2	1
4	DRB1*1501	1BX2	1	1
Accuracy (Correct/Total)		Total=10		4/10=0.40

Table 4: Accuracy of the Gibbs sampler for identifying 9-mer epitope core region in a binding peptide.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 150-155 (2008)