



Artificial Immune System Algorithm for Peptide-based Vaccine Design

Sudhir Singh Soam¹, Feroz Khan², Bharat Bhasker³, Bhartendu Nath Mishra⁴

¹*Department of Computer Science & Engineering, Institute of Engineering & Technology, A Constituent College of U P Technical University, Lucknow, INDIA.

²Metabolic & Structural Biology Dept., Central Institute of Medicinal & Aromatic Plants, Lucknow, INDIA.
Indian Institute of Management, Lucknow, INDIA.

⁴Department of Biotechnology, Institute of Engineering & Technology, A Constituent College of U P Technical University, Lucknow, INDIA.

Abstract: Peptide based vaccines play an important role in activating the immune response. The small peptides derived from target proteins (epitopes) of an invading pathogen, bacteria, or virus bind with the MHC molecules are recognized by CD4+ T cells or CD8+ T cells. MHC class I molecules are recognized by CD4+ T cells and MHC class – II molecules are recognized by CD8+ T cells. It is very time consuming and expensive to predict those small proteins i.e. peptides, which will bind to MHC molecules for immune response in the laboratory from protein sequence. Therefore, various computational learning algorithms have been used to identify the binders and non-binders. However, the number of known binders / non-binders (BNB) to a specific MHC molecule is limited in many cases, which, still is a computational challenge for binder/ non-binder prediction. In order to improve predictions using a learning algorithm the training data sets has to be sufficient. Here, an artificial immune system approach for small set of known binding and non-binding peptides has been used. In present study, which is an extension to our earlier work for MHC Class-II alleles, we have taken the MHC class I alleles for Sars Corona virus (Tor2 Replicase polyprotein 1ab) for HLA-A and HLA-B type molecules. Five different MHC class I alleles 3 for HLA-A type molecules and 2 for HLA-B type molecules with 5 fold cross validation have been retrieved from IEDB database for BNB. The average area under ROC curve have been found to be 0.70 to 0.81 for various alleles varies small training sets and in some cases it is 0.92.

Keywords: Artificial Immune System, Negative Selection Algorithm, Classification, MHC class I binder/non-binder, epitope prediction, vaccine designing, T-cell immune response.

I. Introduction

Biological Background:

Cytotoxic T cells of the immune system monitors cells for infection by viruses, or intracellular bacteria through scanning their surface for peptides bound to MHC Class – I molecules [1]. The cells, that presents peptides derived from non-self e.g. from viruses or bacteria (after binding to MHC molecule) can trigger a T- cell immune response, which leads to the destruction of such cells [2]. T cells do not recognize soluble native antigen but rather recognize antigens that has been processed into antigenic peptides, which are presented in combination with MHC molecules. It has been observed that peptides of nine amino acid residues (9-mers) bind most strongly; peptides of 8-11 residues also bind but generally with lower affinity than 9-mers [3,4]. Binding of a peptide to a MHC molecule is prerequisite for recognition by T cells and, hence, fundamental to understand the basis of immunity and also for the development of peptide based vaccine designing.

The CD8+ Cytotoxic T cells (CTL) immune response and CD4+ T-helper (Th) immune response is stimulated by binding of peptides to major histocompatibility complex (MHC) Class I and MHC Class II molecules respectively [5]. Intracellular antigens, cut into peptides in the cytosol of the antigen processing cell (APC), bind to MHC Class I molecules and are recognized by CD8+ Cytotoxic T cells (CTLs), which once activated, can directly kill a target cell (i.e. an infected cell). Extra cellular antigens that have entered the endocytic pathway of the APC are processed there. These are generally presented by MHC class II molecules to T-helper cells, which, when turned on, have profound immune regulatory effects. In humans, MHC class I molecules includes HLA-A, -B and -C and MHC class II molecules includes HLA-DR, -DP and -DQ. It is important to determine which peptides bind to MHC class II molecules that will help in treatment of the diseases [6, 7].

Motivation and Review of the work:

The algorithms for prediction of MHC binding peptides are based on two concepts - (i) algorithms based on identifying the patterns in sequences of binding peptides e.g. binding motif, quantitative matrices, and artificial

neural networks, and (ii) algorithms based on three-dimensional structures for modeling peptide/MHC interactions [8-10]. The second approach, i.e. based on structures corresponds to techniques with distinct theoretical lineage and includes the use of homology modeling, docking and 3D threading techniques.

For prediction of T-cell epitope, ANN has been used with genetic algorithms [11, 12] and evolutionary algorithm [13]. Support vector machine has also been used to predict the binding peptides [14, 15]. For improving prediction of MHC class I binding peptides Probability Distribution Functions have also been used [16]. Threading methods and Gibbs motif sampler approach have also been used for prediction of MHC binding peptides [17-19]. In many cases the number of known binders and non-binders to specific MHC alleles are limited which may lead to poor performance. The artificial immune approach based algorithm has been used for various MHC Class – II binding and non-binding peptides [20, 21]. In current study, negative selection algorithm, has been used to improve the predictive performance of the limited number of MHC Class I binding/non-binding peptides for Sars Corona Virus.

II. Materials and Method

Artificial Immune System based Algorithm:

Artificial immune system (AIS), a new computational intelligence paradigm be defined as a system of interconnected components, which emulates a particular subset of aspects originating from the natural immune system in order to accomplish a particular task within a particular environment/domain. The focal point of most AIS research is on how lymphocytes (B-cells and T-cells) mature, adapt, react, and learn in response to a foreign antigen. AIS based models are either population based models or network based models. The algorithms based on population based model are negative selection algorithm (NSA) [22, 23], and clonal selection algorithm (CSA), where the focus is mainly on generating initial population of lymphocytes, and improving and refining that population based on techniques emulated from natural immune system. Network based models are based on anti-idiotype activity within the natural immune system, which consequently regulate the population of lymphocytes. Artificial immune network approach is an example of network based model [24].

The elements found within the thymus are representative of self instead of non-self. As a result, the T-cells containing receptors capable of recognizing these self antigens presented in the thymus are eliminated from the repertoire of T-cells through a process called negative selection. All T-cells that leave the thymus to circulate throughout the body are said to be tolerant to self. The negative selection presents alternative paradigm to perform the pattern recognition/classification by storing information about the complement set (non-self). The main concept behind the negative selection algorithm is to generate a set of detectors. The Negative selection algorithm works as follows:

1. Given the self set (P_b) and non-self set (P_n) i.e. the set of binders and non-binders.
2. Generate a set (M) of random elements using the same representation method.
3. Random element is generated using same representation method.
 4. The randomly generated element is matched with the elements of self set P_b (i.e. the set of binders in this case), if a match occurs, then the random element is rejected; else the element is added to the detector set, D_{bn} .
 5. The randomly generated elements also matched with the elements of the non-self set P_n (i.e. the set of non-binders), if match occurs, then the element is rejected; else the element is added to the detector set D_{bn} created in step 4.

To match the elements of the binder set or non-binder set with the randomly generated element, there are several expressions that can be employed in the determination of the degree of match or affinity between an element of binder or non-binder and an element of randomly generated set which has been created using the same representation method. The simplest way is to calculate the Hamming distance (DH) between these two elements, as given by eq.1.

$$D_H = \sum_{i=1}^L \delta_i, \text{ where } \delta_i = \begin{cases} 1 & \text{if } p_i \neq m_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Here, $p_i \in P_b$ or P_n , and $m_i \in M$

In our algorithm the matching is implemented using the Hamming distance. The Hamming distance between two binary vectors is the number of corresponding bits that differ. The Hamming distance between the self tolerant artificial lymphocyte, p_i and randomly generated self tolerant artificial lymphocyte, m_i is calculated for matching. The p_i and m_i are the binary vector of 180 bits long since these consists of 9 amino acids and an amino acid is represented by 20 bit vector.

The detector set D_{bn} is used to monitor the self set (i.e. binders) and non-self set (i.e. non-binders). Each element of the binder and non-binder prediction set is matched with the element of detector set, if match is found then the predicted element is true positive (TP) or true negative (TN), otherwise the element is false negative (FN) or false positive (FP). The values of TP, TN, FP, FN is used to calculate the sensitivity and specificity of the algorithm.

Evaluation Parameters

The predictive performance for the model has been evaluated using receiver operating characteristics (ROC) analysis. The area under the ROC curve (AROC) provides a measure of overall prediction accuracy, $AROC < 70\%$ for poor, $AROC > 80\%$ for good, and $AROC > 90\%$ for excellent prediction. The ROC curve is generated by plotting sensitivity (SN) as a function of 1-specificity (SP). The sensitivity, $SN = TP / (TP + FN)$ and $SP = TN / (TN + FP)$, gives percentage of correctly predicted binders and non-binders respectively. The $PPV = ((TP) / (TP + FP)) * 100$ and $NPV = ((TN) / (FN + TN)) * 100$ gives the positive probability value i.e. the probability that a predicted binder will actually be a binder, and negative probability value i.e. the probability that a predicted non-binder will actually be a non-binder. 10-fold cross validation has been used for training and prediction. The terms TP, FP, TN & FN related to threshold T are true positive, false positive, true negative and false negative respectively. A web based tool has been used to calculate the area under the ROC curve available at [www.rad.jhmi.edu/jeng/javarad/roc /JROCFITi.html].

Data Resources

The data sets used for training and testing of (BNB) have been obtained from IEDB [25] Beta 2.0 database [http://www.immuneepitope.org] for HLA-A*0201, HLA-A*1101, HLA-A*6801, and HLA-B*4002, HLA-B*4501 MHC Class I alleles. The strong binders have been retrieved for $IC_{50} < 500$. All 9-mers have been filtered after removing the duplicates. For strong non-binders, the records with $IC_{50} > 5000$ have been retrieved. The duplicates from binders and non-binders sets have been removed. Further, to keep the ratio of binders and non-binders nearly 1:1, so as to reduce the biasness in learning, the additional 9-mer non-binders were retrieved through EBI-Expasy protein database available at [http://www.expasy.ch]. Final sets of binders and non-binders for various alleles have been shown in Table 1. Data used in each set of various SARS Corona viruses MHC alleles for training and prediction using five fold Cross Validation has been given in Table 2.

Table 1: The number of Binders and Non-binders for various alleles for SARS Corona virus:

MHC Alleles	Binders			Non Binders		
	Retrieved	9-mers	Final*	Retrieved	9-mers	Final*
A*0201	136	103	103	0	0	104
A*1101	202	128	128	16	8	129
A*6801	148	85	85	1	0	85
B*4002	91	36	36	37	9	36
B*4501	75	37	37	41	16	37

Table 2: Data used in each set of various SARS Corona viruses MHC alleles for training and prediction using Fivefold Cross Validation:

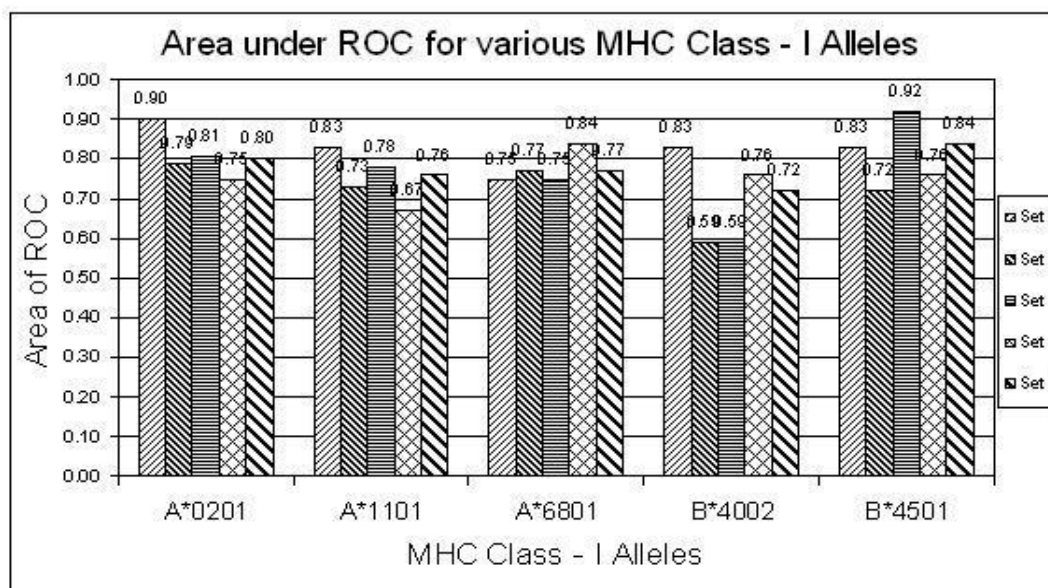
MHC Alleles	Binders			Non Binders		
	Total	Training Set	Prediction Set	Total	Training Set	Prediction Set
A*0201	103	83	20	104	84	20
A*1101	128	103	25	129	104	25
A*6801	85	68	17	85	68	17
B*4002	36	29	7	36	29	7
B*4501	37	30	7	37	30	7

III. Results and Discussion

The 5-fold cross validation has been used to validate the results. The negative selection algorithm has been trained 5 times for each MHC allele using adaptive learning, each time leaving one of the subsets out of 5, and

using the left out subset for prediction. The average area under ROC curve for various MHC Class – I alleles has been shown in figure1.

Figure 1: Average area under ROC for various MHC Class-I alleles.



Predictions of MHC binding and non-binding peptides have the potential to provide important information for rational research and development of vaccines and immunotherapy. To screen out the binding and non-binding peptides although the experimental methods can be used, but this approach is time consuming as well as costly. Computational approaches can be applied to predict the binders and non-binders. For a useful prediction, using any machine learning approach, the data in the training set should be sufficient. In case of small the training data set the prediction will not be useful. In many cases the numbers of known binding and non-binding peptides for MHC Class - I alleles is not sufficient for prediction using the machine learning approaches. Here, in the present study the application of artificial immune algorithm i.e. the negative selection approach has been applied for the prediction of MHC Class – I binding and non-binding peptides which has shown useful predictions in case of small data sets also.

Negative selection algorithm is preferred over the other two artificial immune algorithms because it is theoretical simple and also allows any matching function to be employed. Different matching functions have different detecting regions and thus have direct influence on the performance of the algorithm. We have taken a simple matching function based on Hamming distance rule. The function given in eq. (1) has been used as matching function used in negative selection algorithm. The function calculates the Hamming distance between the self tolerant lymphocyte, p , and randomly generated self tolerant artificial lymphocyte, m . Hamming distance 0 indicates that the two strings are perfectly matched with each other. The maximum score is 18 that indicate the strings are fully mismatched. The value of affinity threshold can be between 0, 2, 4, 6, 8, 10, 12, 14, 16 and 18. In our study the values of thresholds 4, 6, 8, 10, 12 have been taken. The results for various evaluation parameters viz. sensitivity, specificity, positive predictive value (PPV; $PPV = TP / (TP + FP)$), negative predictive value (NPV; $NPV = TN / (TN + FN)$), accuracy and area under ROC curve for five sets has been calculated. A general rule of thumb is that an AROC value > 0.7 indicates a useful prediction performance and a value > 0.85 indicates a good prediction. The summary of the average area under receiver operating characteristics curve for HLA-A*0201, HLA-A*1101, HLA-A*6801, and HLA-B*4002, HLA-B*4501 MHC Class I alleles of Sars Corona Virus have been shown in figure 1 using fivefold cross validation.

The average area under ROC for MHC Class – I alleles i.e. HLA-A*0201, HLA-A*1101, HLA-A*6801, and HLA-B*4002, HLA-B*4501 for various sets have been found to be 0.81, 0.75, 0.78, 0.70 and, 0.81 respectively indicating good predictive performance for small set of binding and non binding peptides.

The above study shows that the artificial immune system algorithm gives useful predictive performance for MHC class – I binding and non-binding peptides small training sets. The artificial immune system algorithm approach i.e. the negative selection algorithm can be used to implement the servers for classification of MHC Class – I binding and non-binding peptides and help in designing the peptide based vaccine designing. The C programming language has been used for classification and training modules.

The study shows that the quality of the prediction of MHC Class – I binding and non-binding peptides can be substantially improved using the artificial immune systems based algorithm for small data sets. The approach can also be used for various other applications too, where the datasets for training are limited. Further analysis can be done for different matching functions in negative selection algorithm.

IV. Acknowledgement

The authors acknowledge the support of Dr. S P Singh Department of Biotechnology, Amity University, Lucknow, in reading the entire manuscript and providing the useful suggestions.

VI. References

- [1] Harriet L Robinson *et. al.* "T cell vaccines for microbial infections", *Nature Medicine*, vol. 11, no. 4, pp. S25-S32, 2005.
- [2] Anne S De Groot *et. al.* "Genome-derived vaccines", *Expert Rev. Vaccines*, vol. 3 No. 1, pp. 59-76, 2004.
- [3] Anne S De Groot *et. al.* "Immuno-informatics: Mining Genomes for Vaccine components". *Immunology and Cell Biology*, vol. 80, pp. 255-269, 2002.
- [4] Rino Rappuoli, "Reverse vaccinology, a genome based approach to vaccine development", *Vaccine* vol. 19, pp. 2688-2691, 2001.
- [5] Jacques B, Steinman R M: Dendritic cells and the control of immunity. *Nature*, 392:245-252, 1998.
- [6] Sette A, Peters B: Immune epitope mapping in the post-genomic era: Lessons for vaccine development. *Current Opin. Immunol.*, 19:106-110, 2007.
- [7] Laumoller SL, Kesmir C, Corbat SL, Fomsgaard A, Holm A, Claesson MH, Brunak S, Buus S. (2000) Identifying Cytotoxic T cell epitopes from genomic and proteomic information: The human MHC project. *Rev Immunogenet.*, 2:477-491, 2000.
- [8] Joo Chuan Tong *et. al.* "Methods and protocols for prediction of immuno-genic epitopes", *Briefings in Bioinformatics*, vol. 8, no.2, pp. 96-108, 2006.
- [9] Manoj Bhasin *et. al.*, "Prediction of CTL epitopes using QM, SVM, ANN techniques", *Vaccine*, vol. 22, pp. 3195-3204, 2004.
- [10] Vladimir Brusic *et. al.*, "Computational methods for prediction of T-cell epitopes-a framework for modeling, testing, and application" *Methods* vol. 34 pp.436-443, 2004.
- [11] Giuliano Armano *et. al.* "A Hybrid Genetic-Neural System for predicting Protein Secondary Structure", *BMC Bioinformatics*, vol. 6, No 4, 2005.
- [12] Yeon-Jin Choi *et. al.*, "Prediction Rule Generation of MHC Class I Binding Peptides Using ANN and GA" *L. ICNC 2005, LNCS 3610*, pp. 1009 – 1016, 2005.
- [13] Vladimir B, G Rudy, M Honyman, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, *Bioinformatics*, vol. 14, no.2, pp. 121-130, 1998.
- [14] Henning Riedesel *et. al.* "Peptide Binding at Class I Major Histo-compatibility Complex Scored with Linear Functions and Support Vector Machines." *Genome Informatics* vol. 15, No.1, pp.198-212 2004.
- [15] Manoj Bhasin, G P S Raghava, "A Hybrid approach for predicting promiscuous MHC class I restricted T Cell epitopes, *J. Biosci.*, vol. 32(1), pp. 31-42, 2007.
- [16] S S Soam *et. al.*, "Prediction of MHC class I Binding Peptides using Probability Distribution Functions" *Bioinformation*. Vol. 3, Issue 9, pp. 403-408, 2009.
- [17] Vladimir Brusic *et. al.* "Data Learning: understanding Biological Data", *American Association for Artificial Intelligence*, 1998.
- [18] S P Singh *et. al.* "Evaluation of threading based method for prediction of peptides binding to MHC class I alleles" *International Journal of Integrative Biology*, vol. 4, no.1, pp. 16-20, 2008.
- [19] S P Singh *et. al.* "Prediction of MHC binding peptides using Gibbs motif sampler, weight matrix and artificial neural network" *Bioinformation* vol. 3, No.4, pp. 150-155, 2008.
- [20] Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B: A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. *PLoS Computational Biology*, 4(4): e1000048, 2008.
- [21] S S soam *et. al.*, "Prediction of MHC Class II binders/ non-binders using negative selection algorithm in vaccine designing" published in *Journal of Bioinformatics and Sequence Analysis* Vol. 5(2), pp. 16-24, February, 2013.
- [22] De Castro LN, Timmis J (2002) Artificial Immune Systems: A Novel Paradigm to Pattern Recognition. *Artificial Neural Networks in Pattern Recognition* (J.M. Corchado, L. Alonso, and C. Fyfe (eds.) SOCO-2002, University of Paisley, UK):67-84, 2002.
- [23] Igawa K, Ohashi H: A negative selection algorithm for classification and reduction of the noise effect. *Applied Soft Computing*, 9:431-438, 2009.
- [24] Hunt J E, Denise E C: Learning using an artificial immune system. *Journal of Network and Computer Applications*, 19:189-212, 1996.
- [25] Peters B, Sidney J, Bourne P, Bui HH, Buus S (2005) The immune epitope database and analysis resource: From vision to blueprint. *PLoS Biol.*, 91, 2005.