Full Length Research Paper

Prediction of MHC Class II binders/non-binders using negative selection algorithm in vaccine designing

S. S. Soam¹*, Feroz Khan², Bharat Bhasker³ and B. N. Mishra⁴

¹Department of Computer Science and Engineering, Institute of Engineering and Technology Gautam Buddha Technical University, Lucknow, India.
²Department of Metabolic and Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India.
³Department of Information Technology and System, Indian Institute of Management, Lucknow, India.
⁴Department of Biotechnology, Institute of Engineering and Technology, Gautam Buddha Technical University, Lucknow, India.

Accepted 5 February, 2013

The identification of major histocompatibility complex (MHC) class-II restricted peptides is an important goal in human immunological research leading to peptide based vaccine designing. These MHC class II peptides are predominantly recognized by CD4+ T-helper cells, which when turned on, have profound immune regulatory effects. Thus, prediction of such MHC class-II binding peptide is very helpful towards epitope based vaccine designing. HLA-DR proteins were found to be associated with autoimmune diseases e.g. HLA-DRB1*0401 with rheumatoid arthritis. It is important for the treatment of autoimmune diseases to determine, which peptides bind to MHC class II molecules. The experimental methods for identification of these peptides are both time consuming and cost intensive. Therefore, computational methods have been found helpful in classifying these peptides as binders or non-binders. We have applied negative selection algorithm, an artificial immune system approach to predict MHC class-II binders and non-binders. For the evaluation of the NSA algorithm, five fold cross validation has been used and six MHC class-II alleles have been taken. The average area under ROC curve for HLA-DRB1*0301, DRB1*0401, DRB1*0701, DRB1*1101, DRB1*1501, DRB1*1301 have been found to be 0.75, 0.77, 0.71, 0.72, 0.69, and 0.84, respectively indicating good predictive performance for the small training set.

Key words: Negative selection algorithm, MHC class-II peptides, artificial immune system, epitope, vaccine designing, human immunology.

INTRODUCTION

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 (Jacques and Steinman, 1998; De Groot et al., 2002). 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 (that is, 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, HLA -A, -B, and -C are the MHC class I type molecules and HLA-DR, -DP and -DQ are the MHC class II type molecules. There are known to be 2DRA, 12DRB, 12DQA, 22DQB, 6DBA and 56 different expressed DPB. HLA-DR proteins were found to be associated with autoimmune diseases e.g. HLA-DRB1*0401 with rheumatoid arthritis. The identification of class II binding peptide epitopes from autoimmune disease-related antigens is an essential step in the development of antigen-specific immune modulation therapy. In the case of type 1 diabetes, two

*Corresponding author. E-mail: sssoam@gmail.com.
DRB1*0401-restricted T cell epitopes from human GAD65, 274-286, and 115-127 are immunogenic in transgenic mice expressing functional DRB1*0401 MHC class II molecules but not in non-transgenic littersmates.

The presentation of these two T-cell epitopes in the islets of DRB1*0401 individuals who are at risk for type 1 diabetes may allow for antigen-specific recruitment of regulatory cells to the islets following peptide immunization. Common allelic variants at the class II HLA-DRB1, DQA1 and DQB1 loci are primarily and jointly associated with the disease (Linda et al., 1996; Christopher and Diane, 2001; Grete et al., 2001). It is important to determine which peptides bind to MHC class II molecules that will help in treatment of the diseases (Sette et al., 2007; Lauemoller et al., 2000; Holden et al., 1998; Emma et al., 2003; Erik et al., 1999). Conventional vaccines comprise live-attenuated microbes, killed inactivated micro-organisms, and purified microbial components, polysaccharide-carrier protein conjugates, or recombinant proteins. In many cases the pathogen to be grown in laboratory conditions, which is both time consuming and costly, and allow for the identification of only the most abundant antigens, which can be purified in quantities for vaccine testing. In case of non-cultivable micro-organisms the conventional vaccine designing approach cannot be applied. The genome sequencing of various microbes and viruses allows the design of vaccines starting from the prediction of all antigens using bioinformatics, independently of their abundance and without the need to grow the pathogen in laboratory. The computational methods are used to identify potential vaccine targets in order to save time and cost (Marirosa et al., 2003; Barbara Capecchi et al., 2004). In our study we have considered six different MHC class II molecules: HLA-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0701, HLA-DRB1*1101, HLA-DRB1*1501, and HLA-DRB1*1301.

The establishment of numerous MHC class-II epitope databases such as SYFPEITHI (Rammensee et al., 1999), MHCBN (Bhasin et al., 2003), AntiJen (Toseland et al., 2005), EPIMHC (Pedro et al., 2005) and IEDB (Peters et al., 2005), has facilitated the development of a large number of prediction algorithms. A number of methods have been developed for the prediction of MHC class-II binding peptides from an antigenic sequence, beginning with, early motif based methods (Chicz et al., 1993; Sette et al., 1993; Hammer et al., 1993), to different scoring matrices based methods (Rammensee et al., 1995; Marshall et al., 1995; Southwood et al., 1998; Wang et al., 2008). The artificial neural network has also been applied for the prediction of HLA-DRB1*0401 binding peptides (Brusic et al., 1998; Honeymam et al., 1998). Some complex tools for identifying the HLA-DRB1*0401 binding peptides have also been designed that is an iterative algorithm to optimize MHC class II binding matrix based stepwise discriminant analysis (Bhasin et al., 2004). We have used an artificial immune system based algorithm; the negative selection algorithm to predict MHC Class II binders and non-binders.

Other computational approaches used for epitope prediction are: genetic algorithm and fuzzy algorithm with artificial neural network, decision tree algorithms, quadratic and linear programming, support vector machine, Gibbs motif sampler, threading methods, structure based methods (Liliana et al., 2003; Soam et al., 2012; Singh and Mishra, 2008; Yael and Hanah, 2004; Ingvar et al., 2004).

METHODS AND MATERIALS

Negative selection algorithm

Artificial immune systems (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. AIS are concerned with computing while the theoretical immune system models focus on understanding the behavior of immune system. The primary function of immune system is to monitor the organisms in search of malfunctioning from their own body or foreign disease causing elements. Thus the immune system is capable of discriminating between self and non-self recognition with certain affinity. The thymus is responsible for the maturation of T-cells; and is protected by a blood barrier capable of efficiently excluding non-self antigens from the thymic environment. Thus, most 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 process is an alternative computational paradigm for pattern recognition by storing information about the complement set (non-self) of the pattern to recognized (self). Therefore, the AIS and NSA are related. Since we have to predict the self/non-self that is, MHC binders and non-binders, an important molecule in activation of immune system, it is better to use AIS as a computational method for prediction. Concept of artificial immune system is based on how lymphocytes (B-cells and T-cells) mature, adapt, react, and learn in response to a foreign antigen. Artificial immune system based models are either population based or network based models. The algorithms on population based model are negative selection algorithm (NSA) (De Castro and Timmis, 2002; Igawa and Ohashi, 2009) and clonal selection algorithm (CSA), focusing mainly on generating initial population of lymphocytes, and improving and refining that population based on techniques emulated from natural immune system. Network 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 (Hunt and Denise, 1996).

Support vector machine is an algorithm for maximizing a particular mathematical function with respect to given collection of data. The separating hyper plane, maximum-margin hyper plane, the soft margin and the kernel function are the main concepts behind SVM (William, 2006). With ANN the adjustments of weights and biases is done during the training, and with SVM the
parameters of a solved model are difficult to interpret. In case of
NSA the appropriate matching function is used to generate a
detector set against which the elements of protected set are
matched for self and non-self elements. The Negative selection
algorithm works as follows:

i. The set of random candidates (generated using any random
number generation algorithm) and the self set is given.

ii. Then each element of the randomly generated set is compared
with the elements of self set. If a match occurs, then that random
element is rejected; else that element is added to the detector set
shown in Figure 1.

After generating the detector set, the system is monitored for non-
self element. The protected set is compared with the elements of
detector set. If match occurs then the non-self is detected otherwise
it continue to match as shown in Figure 2. The binding process of
MHC class I or MHC class-II molecules with antigenic peptides
within the natural immune systems is basically simulated by affinity
threshold functions. For a given lymphocyte, x, and an antigen, y, a
number of matching rules can be defined to determine whether x
and y match. Some of the commonly used affinity functions are as
follows: Hamming distance rule, r-Contiguous bits rule, r-chunks
rule. Hamming distance rule have been used to simulate the affinity
threshold function in present study.

The detector set for binders (D_b) and non-binders (D_n) generated
using the above algorithm. Monitoring the elements of the set C_x (x
is replaced by either b or n depending upon the protected set for
binders and non-binders) to test the resultant population of artificial
lymphocytes against detector set D_bn (D_bn is union of D_b and D_n). In
case of match, value 1 is stored; otherwise value 0 is stored in the
set R_x. The values in sets R_b and R_n are used to obtain the values
of evaluation parameters FP, FN, TP, and TN. The algorithm for
generation of detector set is given in Box 1 and algorithm for
predicting the element of protected set is given in Box 2.

The MATCH() function has been implemented based on the
concept of Hamming distance. The Hamming distance between two
binary vectors is the number of corresponding bits that differ. For
example, if A = (1, 0, 0, 1) and B = (1, 1, 0, 1) then the Hamming
distance between A and B is 1. Here, the MATCH() function
calculates the Hamming distance between the self tolerant artificial
lym-
**Box 1. Algorithm for generation of detector set.**

1. Let $S$ is the set of self tolerant artificial lymphocytes to train and $n_S$ is the numbers of elements in the set, and, the element $s \in S$.
2. Let $C$ is the set of self tolerant artificial lymphocytes to monitor that is, to classify and $n_C$ is the number of elements in the set, and, the element $c \in C$.
3. $S \cup C$ is the set of total number of self tolerant artificial lymphocytes.
4. Let $R$ is the set of all randomly generated self tolerant artificial lymphocytes and $n_R$ is the number of elements in the set and the element $r \in R$.
5. Let $D$ is initially an empty set of detectors.
6. While $n_R \neq \text{Null}$
   7. read an element $r$ from set $R$;
   8. flag = false;
   9. for each self element $s \in S$ do
      10. if MATCH $(s, r)$ is greater than the affinity threshold $t$ then
          11. flag = true;
          12. break;
          13. end;
      14. end;
      15. if flag = false
          16. add $r$ to $D$;
          17. end;
   18. end.

**Box 2. Algorithm for predicting the elements of set $C$.**

1. While $n_R \neq \text{Null}$
   2. read an element $c$ from set $C$;
   3. flag = false;
   4. for each self element $dbn \in Dbn$ do
      5. if MATCH $(c, dbn)$ is less than equal to the affinity threshold $t$ then
          6. flag = true;
          7. break;
          8. end;
      9. end;
     10. if flag = true then add 1 to the set $Rx$
     11. else add 0 to the set $Rx$;
     12. end.

phocyte, $r$. The $r$ and $s$ is the binary vector of 180 bits long since these consists of 9 amino acids and an amino acid is represented by 20 bit vector.

**Training and validation dataset**

We have assembled dataset of peptide binding and nonbinding affinities for six MHC class II alleles’ molecules 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]. The binding affinities ($IC_{50}$) of these peptides, quantitatively measured by immunological experiments have been used for binders and non-binders. The $IC_{50}$ values have been scaled to binding scores ranging from 0 to 100 using linear transformation, where score $IC_{50} \geq 33$ are taken as binders $IC_{50} < 33$ as non-binders. The data sets have been shown in Table 1 after removing the duplication. In order to reduce biasness in prediction, the ratio of binder and non-binders has been kept 1:1 by adding randomly generated non-binders to the non-binders set. The number epitopes in training sets as well as in the prediction set has also shown in the Table 1. Five fold cross validation have been used for prediction. To measure the generalization ability of a computational model i.e. the quality of its inductive bias, the test data has to be outside the training data set. For this the data has to be divided into two parts that is, in training set and validation set. In k-fold cross validation the dataset is divided randomly into k equal parts. To create training and validation sets, one of the k parts is kept as validation set and remaining k-1 parts forms the training set. This is done k times, each time leaving out another one of the k
Table 1. Data sets for various MHC class–II alleles.

<table>
<thead>
<tr>
<th>Allele Name</th>
<th>Total</th>
<th>Bind&gt;=33</th>
<th>NBind&lt;33</th>
<th>Binders</th>
<th>N_Binders</th>
<th>Final N_B</th>
<th>Train_Set</th>
<th>Pred_Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1-0301</td>
<td>605</td>
<td>430</td>
<td>175</td>
<td>396</td>
<td>156</td>
<td>396</td>
<td>317</td>
<td>79</td>
</tr>
<tr>
<td>DRB1-0401</td>
<td>615</td>
<td>450</td>
<td>165</td>
<td>408</td>
<td>143</td>
<td>408</td>
<td>327</td>
<td>81</td>
</tr>
<tr>
<td>DRB1-0701</td>
<td>608</td>
<td>468</td>
<td>140</td>
<td>430</td>
<td>120</td>
<td>430</td>
<td>344</td>
<td>86</td>
</tr>
<tr>
<td>DRB1-1101</td>
<td>623</td>
<td>494</td>
<td>129</td>
<td>444</td>
<td>114</td>
<td>444</td>
<td>356</td>
<td>88</td>
</tr>
<tr>
<td>DRB1-1301</td>
<td>133</td>
<td>55</td>
<td>78</td>
<td>40</td>
<td>57</td>
<td>40</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>DRB1-1501</td>
<td>623</td>
<td>415</td>
<td>208</td>
<td>380</td>
<td>180</td>
<td>380</td>
<td>304</td>
<td>76</td>
</tr>
</tbody>
</table>

Figure 3. Molecular structure of Class II Histocompatibility antigen (HLA-DR1) (PDB ID: 1DLH) revealing binding domain in beta sheets representation of secondary structure sequence.

Evaluation parameters

The prediction accuracy of the algorithm for generation of detector set (Box 1) and for predicting the elements of set (Box 2) have been determined using discrimination between binders and non-binders. In order to classify peptides into binders (positive data) and non-binders (negative data), a threshold value between 0, 2, 4, 6, 8, 10, 12, 14, 16 and 18 based on the Hamming distance between the binary vectors r and s may be taken. Here, in our study the threshold values 4, 6, 8, 10, 12 have been used. 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 non-binder and False Negative (FN); an experimentally binding peptide predicted as non-binder. A non-parametric performance measure, area under receiver operating characteristic (AROC) curve has been used to evaluate the prediction 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.

**RESULTS AND DISCUSSION**

Predictions of T-cell epitopes have the potential to provide important information for rational research and development of vaccines and immunotherapy. To screen out the binders and non-binders, 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. Various computational methods viz. ANN, SVM etc. have been used for predictions. For a useful prediction, using any machine learning approach, the data in the training set should be sufficient. In case of small training data set the prediction will not be useful. In many cases the numbers of known binders and non-binders for MHC class-II alleles is not sufficient for prediction using the machine learning approaches. Further, the available HLA-II servers do not match prediction capabilities of HLA-I servers. Currently available HLA-II prediction server offer only limited prediction accuracy and the development of improved predictors is needed for large-scale studies, such as proteome-wide epitope mapping and for the cases where the small data sets are available. Here, in the present study the application of negative selection algorithm (an artificial immune system paradigm) has been applied for the prediction of MHC class-II T-cell epitopes, 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. MATCH () function calculates the Hamming distance between the self tolerant artificial lymphocyte, s, and randomly generated self tolerant artificial lymphocyte, r. 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 are 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 are shown in Tables 1 to 5 for various threshold levels. For a predictive performance of an algorithm, the number of training data should be sufficient. Here, in case of negative selection algorithm, a large complementary set of self has to be generated using any random algorithm against the known data set. This leads to better predictive performance. We have used the Hamming distance for matching function. 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-DRB1*0301, HLA-DRB1*0401, HLA-DRB1*0701, HLA-DRB1*1101, HLA-DRB1*1501, HLA-DRB1*1301 have

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.70</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>0.72</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>3</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>0.70</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>Average</td>
<td>0.70</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.70</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>3</td>
<td>0.69</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
<td>0.69</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>0.68</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>Average</td>
<td>0.70</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
<td>0.70</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Table 4. HLA-DRB1*0701.

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.64</td>
<td>0.69</td>
<td>0.66</td>
<td>0.68</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
<td>0.67</td>
<td>0.65</td>
<td>0.71</td>
</tr>
<tr>
<td>3</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>0.63</td>
<td>0.69</td>
<td>0.66</td>
<td>0.69</td>
<td>0.64</td>
<td>0.70</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>0.68</td>
<td>0.69</td>
<td>0.69</td>
<td>0.70</td>
<td>0.73</td>
</tr>
<tr>
<td>Average</td>
<td>0.66</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.66</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 5. HLA-DRB1*1101.

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.65</td>
<td>0.70</td>
<td>0.67</td>
<td>0.69</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>0.66</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.66</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>0.65</td>
<td>0.69</td>
<td>0.67</td>
<td>0.68</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>5</td>
<td>0.68</td>
<td>0.66</td>
<td>0.67</td>
<td>0.66</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>Average</td>
<td>0.66</td>
<td>0.69</td>
<td>0.67</td>
<td>0.68</td>
<td>0.66</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 6. HLA-DRB1*1301.

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
<td>0.67</td>
<td>0.65</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>0.63</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td>0.63</td>
<td>0.94</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.71</td>
<td>0.66</td>
<td>0.69</td>
<td>0.64</td>
<td>0.88</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.71</td>
<td>0.69</td>
<td>0.70</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>Average</td>
<td>0.65</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.66</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 7. HLA-DRB1*1501.

<table>
<thead>
<tr>
<th>Set #</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>PPV</th>
<th>NPV</th>
<th>Area ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.65</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>0.68</td>
<td>0.66</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>Average</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.69</td>
</tr>
</tbody>
</table>

been shown Tables 2 to 7, respectively. The value of AROC for HLA-DRB1*1501 is 0.84 which has small training set size of 32.

The comparison of AROC for various MHC class–II alleles for different sets has been shown in Figure 4. The average area under ROC curve for HLA-DRB1*0301, DRB1*0401, DRB1*0701, DRB1*1101, DRB1*1501, DRB1*1301 have been found to be 0.75, 0.77, 0.71, 0.72, 0.69, and 0.84, respectively indicating good predictive performance. The above study shows that the negative selection algorithm gives useful predictive performance for MHC class - II binders and non-binders even for small training sets. The above method can be applied for the classification of MHC class – II binders and non-bind ers even for the small data sets. The negative selection algorithm can be used to implement the servers for
Area Under ROC Curve for various MHC Class II Allels

![Area Under ROC Curve for various MHC Class II Allels](image)

**Figure 4.** Performance comparison of various MHC Class – II alleles for different sets.

classification of MHC class-II binders and non-binders and help in designing the epitope based vaccine designing.

**ACKNOWLEDGEMENT**

The authors are thankful to Dr. Satarudra Pratap Singh, Sr. Lecturer, Biotechnology, Amity University, Lucknow, for their kind help in data preparation and analysis.

**REFERENCES**


Liliana Florea, Bjarni Halldorsson, Oliver Kohlbacher, Russell Schwartz, Stephen Hoffman, Sorin Istrail (2003). Epitope prediction algorithms