Identification of MHC Class II binders/ non-binders using Negative Selection Algorithm

S. S. Soam[1*], Feroz Khan[2], Bharat Bhasker[3], B. N. Mishra[4]

[1]Department of Computer Science & Engg., Institute of Engg. & Tech., G.B. Tech. Univ., Lucknow, India.
[2]Department of Metabolic & Structural Biology, CSIR-Central Institute of Medicinal & Aromatic Plants, Lucknow, India
[3]Department of Information Technology & System, Indian Institute of Management, Lucknow, India.
[4]Department of Biotechnology, Institute of Engg. & Tech., G.B. Tech. Univ., Lucknow, India.

*Corresponding author: sssoam@gmail.com

Abstract

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, and 0.69, and 0.84 respectively indicating good predictive performance for the small training set.

Keywords: Negative selection algorithm, MHC class-II peptides, Artificial immune system, Epitope, Vaccine designing, human immunology.

* Corresponding Author

S.S. Soam
Dept. of Computer Science & Engg.
Institute of Engineering & Technology
Gautam Budh Technical University, Lucknow.
(formerly, U.P. Technical Univ.)
Sitapur Road, Lucknow-226021 (U.P.) India
Phone: +91 9935297376
E-mail:sssoam@gmail.com

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 [1,2]. 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, 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, 126DRB, 12DQA, 22DQB, 6DBA and 56 different expressed DPB. It is important to determine which peptides bind to MHC class II molecules that will help in treatment of the diseases [3, 4]. 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, HLA-DRB1*1301.

The establishment of numerous MHC class-II epitope databases such as SYFPEITHI [5], MHCBN [6], AntiJen [7], EPIMHC [8], and IEDB [9], 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 [10-12], to different scoring matrices based methods [13-16]. The artificial neural network has also been applied for the prediction of HLA-DRB1*0401 binding peptides [17, 18]. Some complex tools for identifying the HLA-DRB1*0401 binding peptides have also been designed *i.e.* an iterative algorithm to optimize MHC class II binding matrix based stepwise discriminant analysis [19]. We have used an artificial immune system based algorithm – the negative selection algorithm to predict MHC Class II binders and non-binders.

Methods and Materials

Negative Selection 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 fundamental 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) [20, 21], 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-idiaotypic activity within the natural immune system, which consequently regulate the population of lymphocytes. Artificial immune network approach is an example of network based model [22].

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 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 Figure1.

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.

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 Є S.
2. Let C is the set of self tolerant artificial lymphocytes to monitor *i.e.* to classify and n_c is the number of elements in the set, and, the element c Є C.
3. S U 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 Є R.
5. Let D is initially an empty set of detectors.
6. While $n_r \neq$ Null
7.   read an element r from set R;
8.   flag = false;
9.   for each self element s Є 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;

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.

Box 2: Algorithm for predicting the elements of set C:

```
While n_r ≠ Null
    read an element c from set C;
    flag = false;
    for each self element d_bn Є D_bn do
        if MATCH (c, d_bn) is less than equal to the affinity threshold t then
        flag = true;
        break;
        end;
    end;
    if flag = true then add 1 to the set R_x
    else add 0 to the set R_x;
    end;
```

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 lymphocyte, s, and randomly generated self tolerant artificial lymphocyte, 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 allele's 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}>=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. This structure includes the extracellular portion of a class II MHC, with a peptide bound. Figure 3 shows crystal structure of the human class II MHC protein HLA-DRB1 complexed with an influenza virus peptide (PDB ID: 1DLH).

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 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. 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 the 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 leaning 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 Table 1 to 5 for various threshold levels. 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 been shown Tables 2-7 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, and 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-binders even for the small data sets. The negative selection algorithm can be used to implement the servers for classification of MHC class – II binders and non-binders and help in designing the epitope based vaccine designing.

Acknowledgements

The authors are thankful to Dr. S.P. Singh, Sr. Lecturer, Biotechnology, Amity University, Lucknow, for their kind help in data preparation.

References

1. Jacques B, Steinman RM (1998) Dendritic cells and the control of immunity. Nature, 392:245-252.
2. De Groot AS, Sbai H, Aubin CS, McMurry J, Martin W (2002) Immuno-informatics: Mining Genomes for Vaccine components. Immunology & Cell Biology, 80: 255-269.

3. Sette A, Peters B (2007) Immune epitope mapping in the post-genomic era: Lessons for vaccines development. Current Opin. Immunol., 19:106-110.
4. Lauemoller SL, Kesmir C, Corbat SL, Fomsgaard A, Holm A, Claesson MH, Brunak S, Buus S. (2000) Identifying Cytotoxic T cell epitopes fromn genomic and proteomic information: The human MHC project. Rev Immunogenet., 2:477-491.
5. Rammensee H, Bachmann J, Emmerich N, Bachor O, Stevanovic S (1999) SYFPEITHI: a database for MHC ligands and peptide motifs. Immunogenetics, 50: 213-219.
6. Bhasin M, Singh H, Raghava GPS (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics, 19: 665-666.
7. Toseland CP, Clayton DJ, Mc Sparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwangama CK, Flower DR (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res., 1: 4.
8. Pedro AR, Zhang H, Paul J, Ellis G, Reinherz L (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics, 21(9):2140-2141.
9. 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.
10. Chicz RM, Urban RG, Gorga JC, Vignali DA, Lane WS, Strominger JL (1993) Specificity and promiscuity among naturally processed peptides bound to HLADR alleles. J Exp Med., 178:27-47.
11. Sette A, Sidney J, Oseroff C, Del Guercio MF, Southwood S, Arrhenius T, Powell MF, Colon SM, Gaeta FC, Grey HM (1993) HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions. J. Immunol., 151: 3163-70.
12. Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B, Sinigaglia F (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. Cell, 74:197-203.
13. Rammensee HG, Friede T, Stevanovic S (1995) MHC ligands and peptide motifs: First listing. Immunogenetics, 41: 178-228.

14. Marshal KW, Wilson KJ, Liang J, Woods A, Zaller D, Rothbard JB (1995) Prediction of peptide affinity to HLA-DRB1*0401. J. Immunol., 154, 5927-5933.
15. Southwood S, Sidney J, Kondo A, Del Guercio MF, Appella E, Hoffman S, Kubo RT, Chesnut RW, Grey HM, Sette A (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. J. Immunol., 160:3363-73.
16. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B (2008) A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. PLoS Computational Biology, 4(4): e1000048.
17. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics, 14:121-30.
18. Honeyman MC, Brusic V, Stone NL, Harrison LC (1998) Neural network based prediction of candidate T-cell epitopes. Nature Biotechnology, 16:966-69.
19. Bhasin M, Raghava GPS (2004) SVM based method for prediction DRB*0401 binding peptides in an antigen sequence. Bioinformatics, 20: 421-423.
20. 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.
21. Igawa K, Ohashi H (2009) A negative selection algorithm fro classification and reduction of the noise effect. Applied Soft Computing, 9:431-438.
22. Hunt JE, Denise EC (1996) Learning using an artificial immune system. Journal of Network and Computer Applications, 19:189-212.

Table 1: Data sets for various MHC class – II alleles

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

Table 2: HLA-DRB1*0301

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.70	0.72	0.71	0.72	0.71	0.73
2	0.68	0.73	0.70	0.72	0.69	0.73
3	0.72	0.72	0.72	0.72	0.72	0.75
4	0.70	0.72	0.71	0.72	0.71	0.75
5	0.72	0.71	0.72	0.72	0.71	0.79
Average	**0.70**	**0.72**	**0.71**	**0.72**	**0.71**	**0.75**

Table 3: HLA-DRB1*0401

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.70	0.72	0.71	0.71	0.70	0.79
2	0.71	0.70	0.71	0.70	0.71	0.76
3	0.69	0.71	0.70	0.71	0.69	0.75
4	0.68	0.71	0.70	0.71	0.69	0.77
5	0.71	0.70	0.70	0.70	0.70	0.76
Average	**0.70**	**0.71**	**0.70**	**0.71**	**0.70**	**0.77**

Table 4: HLA-DRB1*0701

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.64	0.69	0.66	0.68	0.65	0.73
2	0.65	0.67	0.66	0.67	0.65	0.71
3	0.66	0.69	0.68	0.68	0.67	0.70
4	0.63	0.69	0.66	0.69	0.64	0.70
5	0.70	0.68	0.69	0.69	0.70	0.73
Average	**0.66**	**0.68**	**0.67**	**0.68**	**0.66**	**0.71**

Table 5: HLA-DRB1*1101

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.65	0.70	0.67	0.69	0.66	0.70
2	0.66	0.68	0.67	0.68	0.66	0.78
3	0.65	0.69	0.67	0.68	0.65	0.70
4	0.67	0.69	0.68	0.69	0.67	0.72
5	0.68	0.66	0.67	0.66	0.68	0.71
Average	**0.66**	**0.69**	**0.67**	**0.68**	**0.66**	**0.72**

Table 6: HLA-DRB1*1301

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.65	0.67	0.66	0.67	0.65	0.78
2	0.63	0.65	0.64	0.65	0.63	0.94
3	0.61	0.71	0.66	0.69	0.64	0.88
4	0.69	0.69	0.69	0.68	0.70	0.78
5	0.67	0.71	0.69	0.70	0.68	0.82
Average	**0.65**	**0.68**	**0.67**	**0.68**	**0.66**	**0.84**

Table 7: HLA-DRB1*1501

Set #	Sensitivity	Specificity	Accuracy	PPV	NPV	Area ROC
1	0.65	0.66	0.66	0.66	0.66	0.69
2	0.65	0.68	0.67	0.67	0.66	0.60
3	0.68	0.66	0.67	0.67	0.68	0.72
4	0.67	0.69	0.68	0.69	0.68	0.71
5	0.67	0.68	0.67	0.68	0.67	0.71
Average	**0.67**	**0.67**	**0.67**	**0.67**	**0.67**	**0.69**

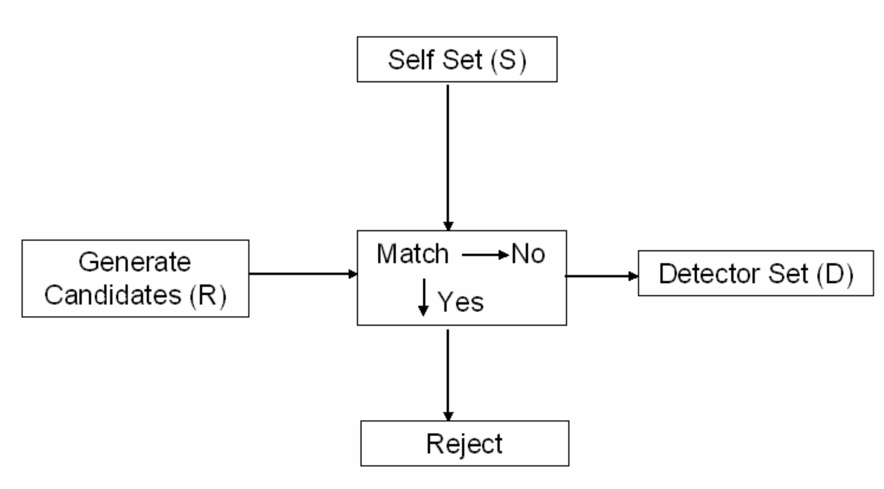

Figure 1: Generating the set of detectors.

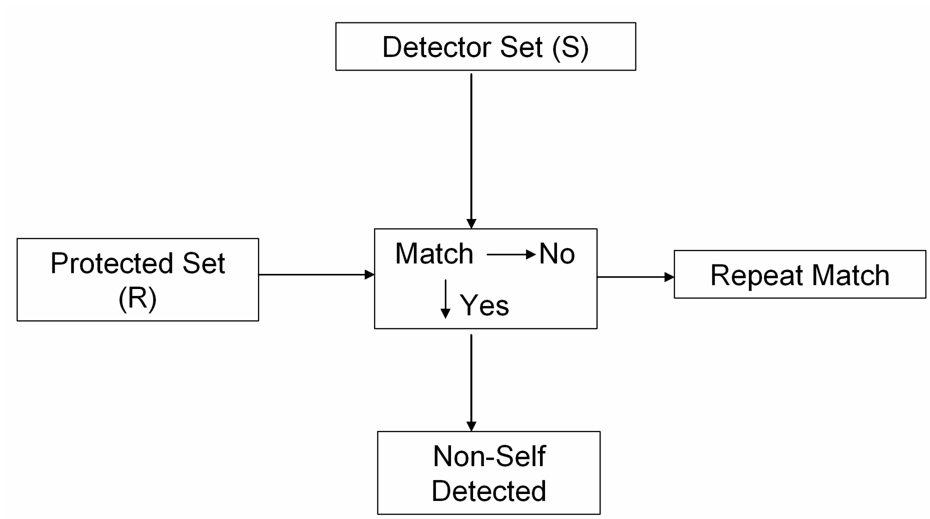

Figure 2: Detecting non-self elements.

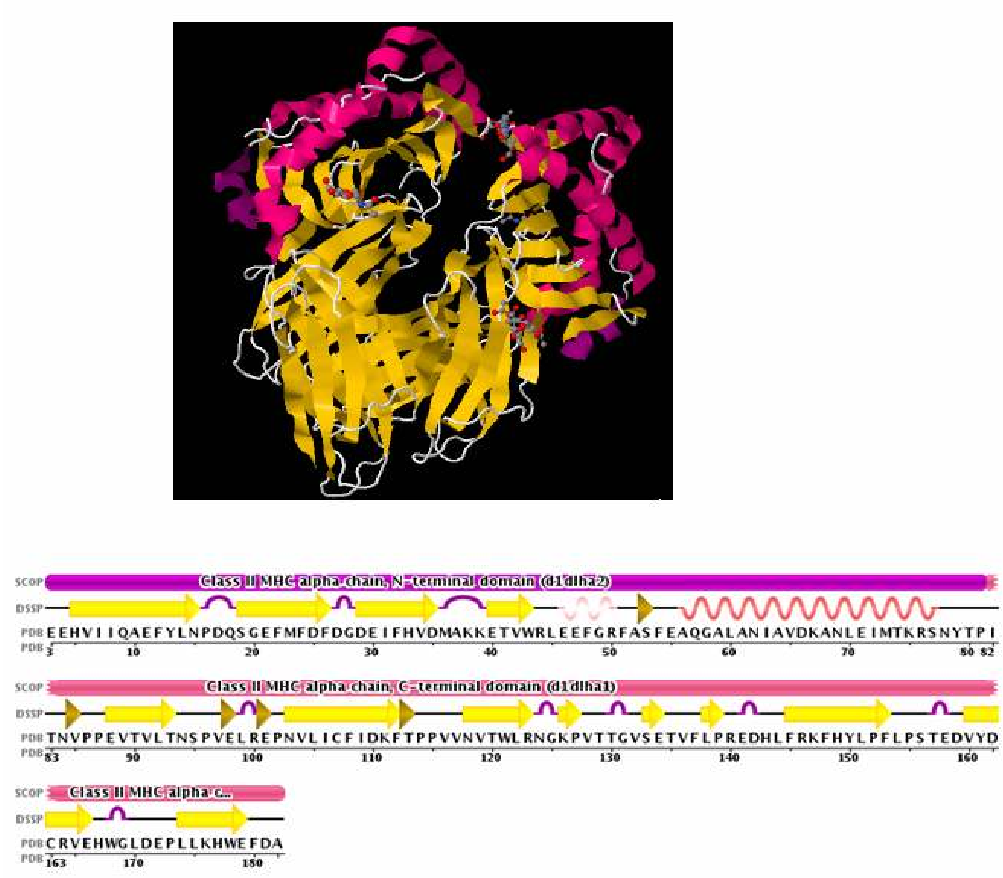

Figure 3: Molecular structure of Class II Histocompatibility antigen (HLA-DR1) (PDB ID: 1DLH) revealing binding domain in beta sheets representation of secondary structure sequence.

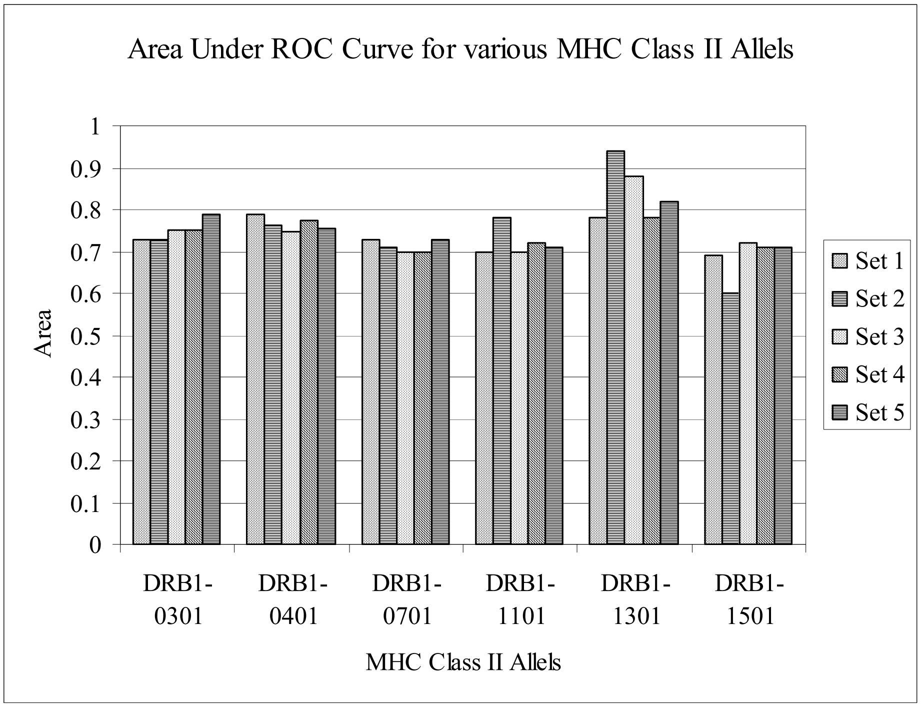

Figure 4: Performance comparison of various MHC Class – II alleles for different sets.