Supplementary Material

Pcleavage: A SVM based Method for Prediction of Consitutive Proteasome and Immunoproteasome Cleavage Sites in Antigenic Sequences

Manoj Bhasin[§] and G. P. S. Raghava[§*]

Material AND Methods:

MHC class I ligand data

For the prediction of cleavage sites with proteasome and immunoproteasome specificity, classifiers were trained on data of MHC class I ligands obtained from MHCBN database (15). The MHCBN have 1288 HLA-A and HLA-B restricted ligands that are either natural T cell epitopes or natural peptides eluted from MHC molecules. All ligands used by saxova et al., 2003 for evaluation of existing methods were discarded from the dataset. We have also discarded the peptides <8 and >12 amino acids. All duplicate ligands and ligands with unknown source protein are also removed from the dataset. This dataset have 855 unique ligands interacting with 100 different HLA alleles. To prevent the bias toward specific HLA binding motifs, we made sure that final data set do not have more than 5% ligands binding with a given allele. The peptides whose flanking regions on C and N terminal is not able to reconstruct uniquely by 10 amino acids. To classify amino acids within a protein sequence into cleavage and non-cleavage sites, one need examples of both negative and positive types of sites. We considered that C terminal of all ligands have proteasomal cleavage sites where as the position within the ligands are non-cleavage

sites.For example a ligand of 9 amino acids will have a cleavage site at C terminal and 8 non-cleavage sites between N and C terminal (excluding C terminal residue). After processing, a final dataset of 506 ligands of more than 250 proteins was derived this dataset as shown in Figure 1.

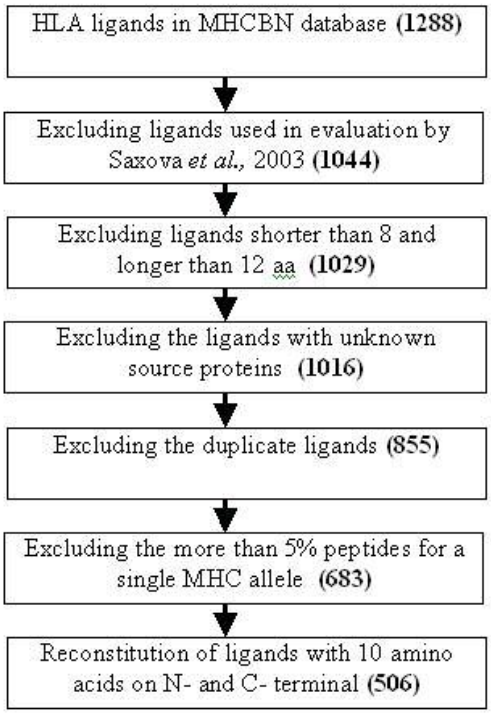

Figure 1: Diagram summarizing the compilation of MHC class I ligand dataset used in the development of this method.

Performance Measures:

On cross-validation these parameters can be derived from following four scalar quantities: TP is number of true positives (experimentally verified cleavage sites which are also predicted cleavage sites), TN is the number of true negative (experimentally verified non-cleavage sites which are also predicted non-cleavage sites), FP is the number of false positives (experimentally verified non-cleavage sites which are predicted as cleavage sites), FN is the number of false negatives (experimentally verified cleavage sites which are predicted as non-cleavage sites). During testing on independent dataset the assignment of these scalar quantities is slightly modified as described by Saxova et al., 2003.

The six parameters can be derived from these four quantities are: (1) Sensitivity or percent coverage is the percentage that cleavage sites that are predicted correctly; (2) Specificity is the percentage of non-cleavage sites that are predicted correctly; (3) PPV is the probability that predicted cleavage sites are actually cleavage site; (4) NPV is the probability that predicted non-cleavage site is actually non-cleavage site; (5) Accuracy is the percentage of correctly predicted cleavage and non-cleavage sites and (6) Matthew's Correlation Coefficient (MCC) is a measure to estimate how well method perform on positive and negative examples. The MCC took in account both over and under prediction so it is the best parameter to measure the performance of a method when the dataset is unbalance. The parameters can be calculated by following equation 1-6.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \quad \text{.....1}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100 \quad \text{.....2}$$

$$PPV = \frac{TP}{TP + FP} \times 100 \quad \text{.....3}$$

$$NPV = \frac{TN}{TN + FP} \times 100 \quad \text{.....4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad \text{.....5}$$

$$\text{MCC} = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \quad \text{.....6}$$

Results:

Comparison of performance of Quantitative matrices with machine leaning techniques:

To compare the performance of machine learning techniques with linear statistical methods, we have generated a quantitative matrix from the dataset used for training of above discussed SVM based method. The quantitative matrix is based on the probability of amino acids occurring at specific position in peptides with cleavage and non-cleavage sites. We have considered the optimum window length of machine learning [19] techniques to generate quantitative matrix. The generation of quantitative matrices require positive as well as negative examples. The peptides possessing C terminal cleavage sites are considered as positive examples and peptides without C terminal cleavage sites are considered as negative examples. The quantitative matrix generated using this approach is shown in Table S1. The performance of quantitative matrix in recognizing the cleavage and non-cleavage sites was measured on independent dataset at "0" cutoff score. Surprisingly, quantitative matrix achieved MCC of 0.23, which is better in comparison existing methods. The quantitative matrix is able to recognize ~73% of cleavage sites, which is nearly similar to the performence of ANN, based method " NetChop" on this dataset. Surprisingly, quantitative matrix has outperformed all the existing methods as well as methods developed in this study in recognizing the non-cleavage sites. The quantitative matrix is able to recognize 49.8% of non-cleavage sites correctly. The better performance of quantitative matrix based methods may due to the fact that it was generated using MHC ligands data. These results demonstrate that linear methods are also able to capture the cleavage and non-cleavage sites as accurately as machine-learning techniques based methods.

Furthermore, We have also tested the quantitative matrix used for the prediction of immunoproteasome cleavage sites in Propred1 server (Singh and Raghava, 2003). This quantitative matrix was originally derived from the work of Toes et al., 2001. Matrix are able to recognize the ~43% of cleavage sites at 10% threshold, which is very less stringent condition for prediction. The most probable reason for poor performance is the limited size of dataset used derives this matrix.

Table S1: A quantitative matrix derived from the MHC ligands data. The P1----P19 specify the position of the peptides where the actual cleavage site occurs between P9 and P10.

AA	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19
A	0.19	-0.02	-0.04	-0.20	-0.35	-0.18	-0.21	-0.02	-0.07	-0.52	0.26	0.14	0.16	0.11	-0.19	-0.11	0.04	-0.03	0.01
C	-0.22	-0.82	-0.54	-0.50	-0.64	-0.33	-0.67	-0.78	-0.33	-0.75	0.36	0.00	0.00	-0.16	0.04	-0.33	-0.08	-0.18	-0.45
D	0.07	0.13	-0.25	0.41	0.31	-0.08	-0.27	-0.33	-0.16	-0.91	-0.14	-0.23	-0.05	0.26	0.43	0.13	0.08	-0.37	-0.18
E	0.12	0.28	-0.08	0.02	0.31	0.13	-0.06	0.10	0.35	-0.85	0.44	0.42	0.33	0.32	0.35	0.27	0.15	0.15	0.03
F	-0.17	0.18	0.18	0.18	0.13	0.22	0.24	-0.04	-0.27	0.24	-0.50	-0.17	-0.16	-0.22	-0.27	-0.40	-0.18	-0.25	-0.23
G	0.21	0.10	-0.11	0.30	0.32	0.22	0.14	-0.09	-0.19	-0.94	0.06	0.11	-0.23	-0.05	0.10	0.28	0.16	0.08	0.00
H	0.07	-0.24	0.09	-0.15	-0.01	-0.27	0.26	0.12	0.13	-0.71	0.10	-0.18	0.10	0.30	0.00	0.33	0.23	0.05	0.10
I	-0.02	0.16	-0.04	0.06	0.06	0.14	0.29	0.06	-0.21	0.17	-0.09	0.12	-0.36	-0.15	-0.03	0.11	0.15	0.15	0.04
K	0.24	0.15	-0.50	0.33	0.29	0.50	0.33	0.27	0.51	0.59	0.23	0.00	-0.04	-0.13	-0.13	-0.07	-0.06	-0.01	0.00
L	-0.40	-0.48	-0.08	-0.52	-0.34	-0.49	-0.29	-0.18	-0.23	0.34	-0.28	-0.09	-0.01	-0.12	-0.10	0.02	-0.17	-0.12	-0.15
M	0.18	0.27	0.00	-0.26	-0.33	-0.56	-0.10	0.08	0.00	0.06	-0.20	-0.23	-0.23	0.33	0.00	0.06	-0.08	-0.25	0.52
N	0.00	0.14	-0.50	0.20	0.33	0.25	0.08	0.46	0.10	-0.80	0.50	0.38	0.54	0.70	0.70	0.44	0.52	0.40	0.52
P	-0.25	-0.44	0.43	0.37	0.29	0.21	0.17	-0.04	-0.55	-0.89	-0.40	-0.11	-0.21	0.17	0.05	0.17	0.32	0.32	0.11
Q	0.11	-0.15	-0.31	-0.14	0.14	0.27	0.18	0.27	0.33	-0.56	0.25	0.17	0.11	0.21	0.19	0.10	0.08	0.14	0.06
R	0.25	0.38	0.26	0.09	-0.09	-0.16	-0.33	-0.28	0.02	0.12	0.32	0.19	0.21	-0.02	0.25	0.06	0.03	0.00	0.10
S	0.07	-0.10	-0.17	-0.02	-0.13	-0.17	-0.03	-0.08	0.13	-0.90	0.23	0.02	0.10	-0.06	-0.04	0.01	0.08	0.11	-0.05
T	0.02	-0.11	-0.14	-0.40	-0.21	-0.08	0.03	-0.03	0.30	-0.76	-0.05	-0.03	0.01	0.06	-0.13	-0.13	-0.23	0.18	2.00
V	-0.22	-0.20	0.00	-0.23	-0.22	0.20	0.16	0.31	0.01	0.16	-0.32	-0.24	-0.09	-0.30	-0.20	-0.33	-0.20	-0.34	-0.09
W	-1.00	-0.50	-0.20	0.29	0.37	0.23	-0.22	0.04	-0.06	0.44	-0.50	-0.50	-0.26	-0.29	-0.43	-0.50	-0.40	0.23	0.44
Y	0.03	0.36	0.45	0.28	-0.33	-0.20	-0.32	-0.33	-0.14	0.50	-0.26	-0.31	0.09	-0.36	-0.36	-0.25	-0.10	0.00	0.04