Predicting Peptides Binding to MHC Class II Molecules Using Boosted Decision Tree

Haneen Tartouri
Supervisors: Hashem Tamimi, Yaqoub Ashhab
Master of Informatics
College of Graduate Studies
Palestine Polytechnic University

Introduction

The Major Histocompatibility Complex (MHC) is a large genomic region or gene family found in most vertebrates that encodes MHC molecules, that plays an important role in the immune system and autoimmunity.

Prediction of peptide-MHC binding represents an important goal in bioinformatics. Prediction of peptides binding to a MHC class II molecule is more difficult than MHC class I due to different length of the binding peptides. Recently, many studied focused on prediction of peptide binding to MHC II depending on different machine learning tools.

This project aims at predicting peptides binding to MHC class II molecules using boosted decision tree.

The experiments results show that the boosted decision tree algorithm can be developed to give good results for MHC II prediction problem.

Project Objectives:

predict peptides binding to MHC class II molecules using boosted decision tree algorithm.

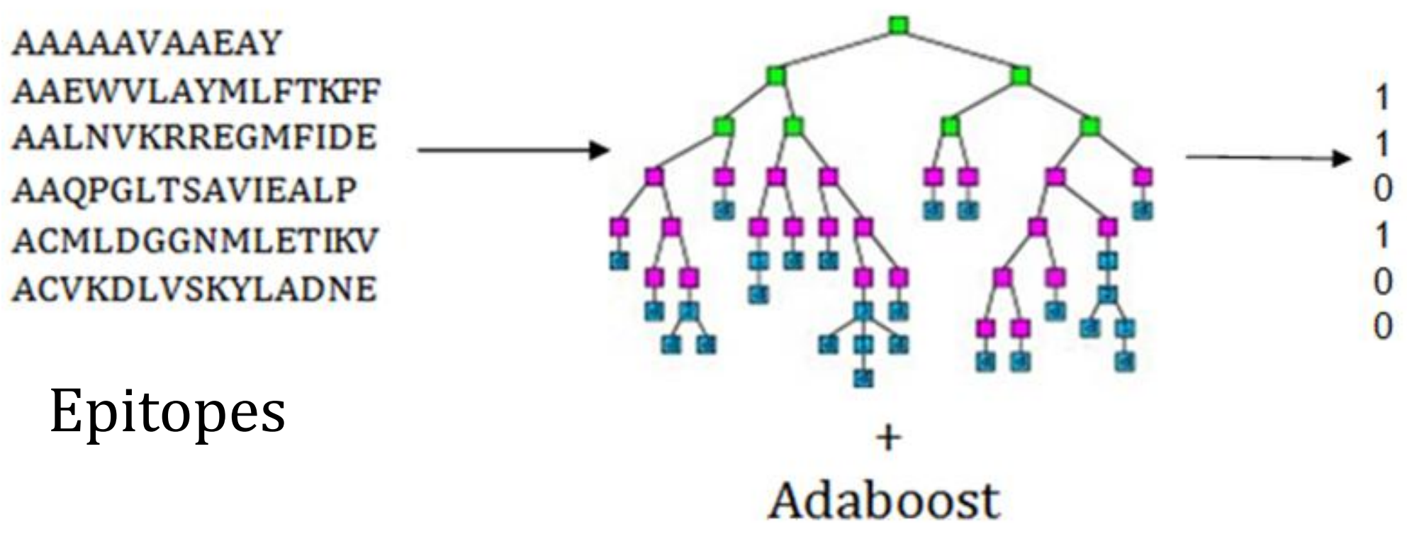

Figure 1: Objective of this project.

Data source:

Peptide datasets used in this study are available from the NetMHCII 2.2 server.

The main characteristics of this data set, that it contains a 5166 epitope sequences, 1656 are non binders, and 3510 are binders.

The second characteristic is that the longest epitope sequence contains 37 amino acid.

Boosted Decision Tree Experiment and Result Using Physicochemical Properties:

In our first experiment we used three physicochemical properties to represent each amino acid; hydrophobic, charged and size.

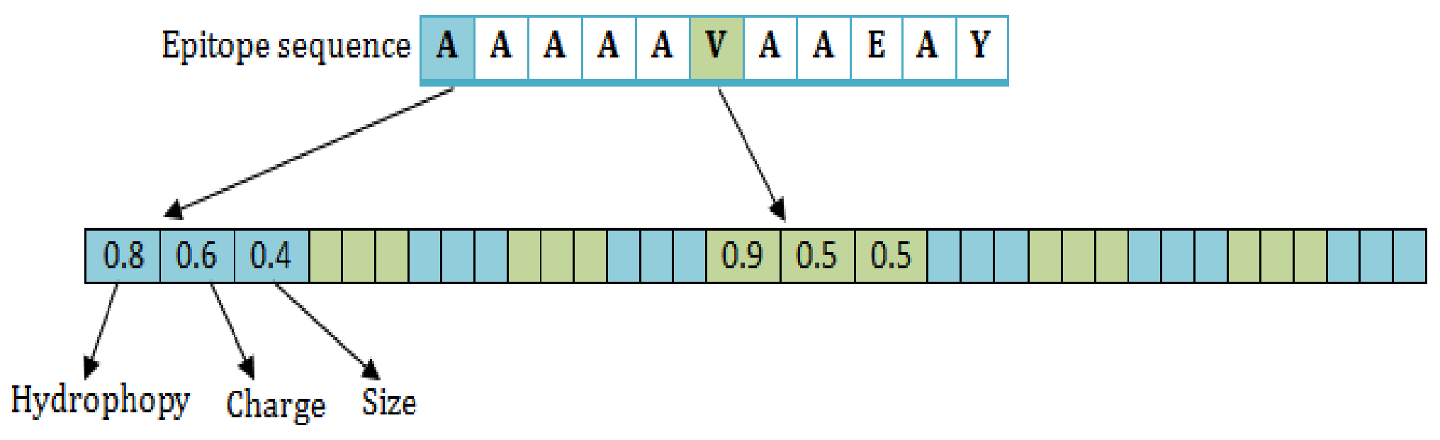

Figure 2: Representing each epitope sequence into physicochemical values.

In this study, we used the boosted recombined weak classifier method to apply boosted decision tree.

We built the decision stumps to apply this method using a suitable threshold for each physicochemical property after normalization.

The boosted recombined weak classifier has one parameter which is r (the level of reuse). The suitable value of r using this method was 5 because it gave good results. Three fold cross validation was used, and we applied the boosting algorithm 10 times, each time we took random sample from binders sample equals non-binders.

Table 1 shows the results for this experiment, and Figure 3 shows the ROC curve.

Table 1: Results for the physicochemical properties experiment for 10 iteration.

Average sensitivity	0.7131
Average specificity	0.71
Average accuracy	0.706522
Average PPV	0.71820
Average NPV	0.697168
Average area under the ROC curve	0.7768

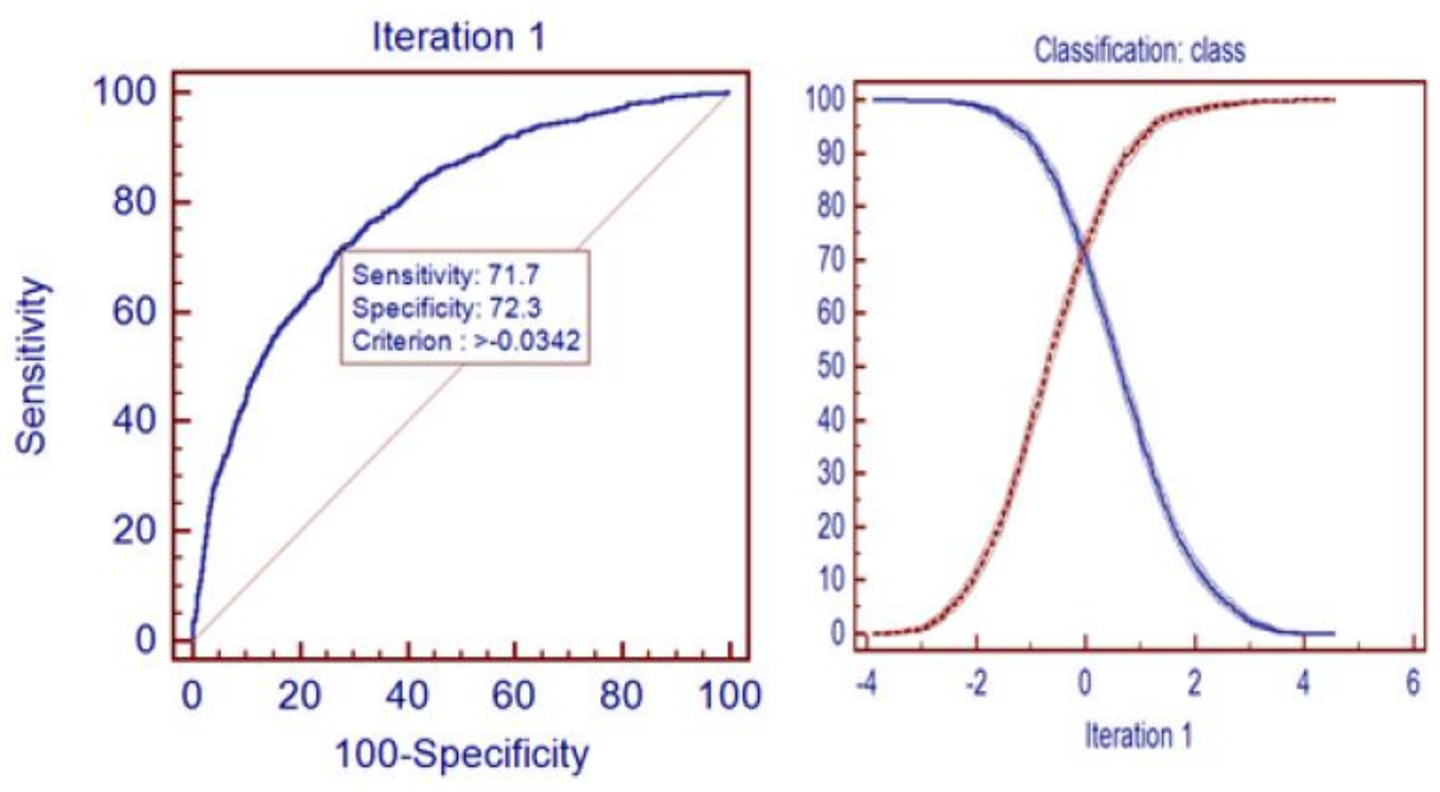

Figure 3: ROC curve for the first iteration, and the cutoff point

Boosted Decision Tree Experiment and Result Using Characters Sequence:

This experiment was done depending on classification of amino acids into five categories using representative properties. These groups are: bulky (WY), small (AG), hydrophobic (IVLFCM), positively charged (RKH), negatively charged (DE), and they excluded (PNQST) amino acids which represent the middle amino acids of the physicochemical properties. We used these groups of amino acids and we completed the sequences into 37 using 0 values. In this experiment the suitable value of (r) that gave a good results was 8.

Table 2 shows the results for this experiment, and Figure 4 shows the ROC curve.

Table 1: Results for the physicochemical characters experiment for 5 iteration.

Average sensitivity	0.7414
Average specificity	0.7828
Average accuracy	0.7601
Average PPV	0.7701
Average NPV	0.7522
Average area under the ROC curve	0.838

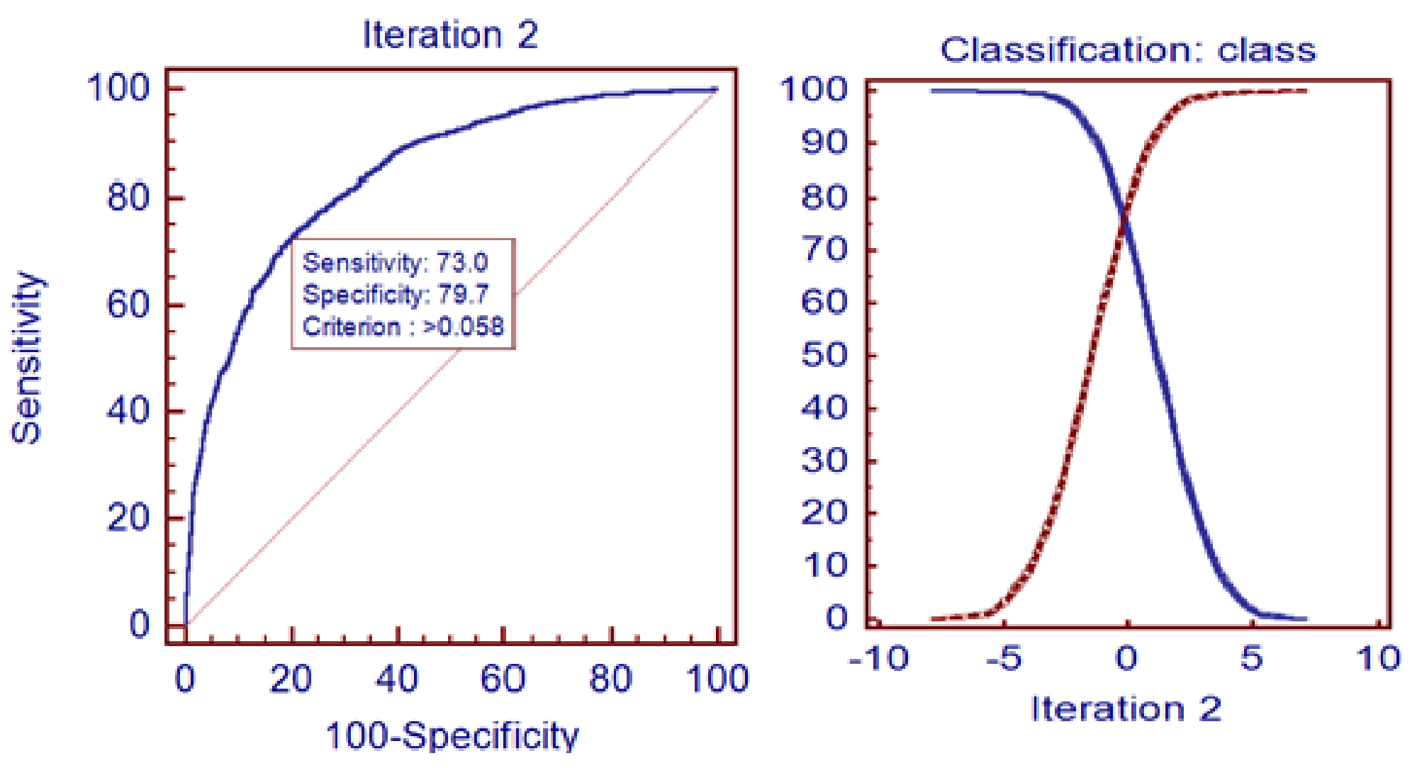

Figure 4: ROC curve for the second iteration, and the cutoff point

The Comparison of AUC Values on Allele DRB1-0101

Boosted decision tree based on physicochemical properties	0.78
Boosted decision tree based on characters sequence	0.84
An artificial neural network-based alignment algorithm including data redundancy step-size rescaling and P1-PSSM encoding (NN-W-P1) (Nielsen, et al., 2009)	0.88
An artificial neural network-based alignment algorithm including data redundancy step-size rescaling (NN-W) (Nielsen, et al., 2009)	0.87
SVRMHC (support vector regression) (Wang, et al., 2008)	0.69
MHC2Pred (support vector machine) (Wang, et al., 2008)	0.67