

Name: Rafael Wang

USC Student ID: 6189106477

Email: rafaelwa@usc.edu

Program Summary:

In this program (SVM.py), I use numerical descriptions based on biochemical properties alongside ACC as the features for a machine learning algorithm. Based on these features, I then use RFE for feature reduction and then use SVM techniques to be able to predict a new multi-epitope vaccine sequence. I use the NegativeB.txt, NegativeT.txt, PositiveB.txt, and PositiveT.txt as my input files.

Assumptions:

Many key assumptions were made in this program. First, for each individual sequence, I set the max amino acid length as 10, since SVMs must work with constant fixed-length vectors.

Secondly, I randomly imported 175 positive B cells and 175 negative B cells to train the ML model. I randomly imported 150 positive CTL and 150 negative CTL cells to train the ML model. I also did 150 of each for HTL.

I also assumed that sequences of length 11 or less amino acids were CTL cells while those greater than length 11 were HTL cells.

Of the selected sequences, I made 25% of them test sequences.

For predicting a new vaccine, I used the remaining sequences that were not used as part of training each of the models.

The new vaccine takes the top 10 epitopes from each of B cells, HTL, and CTL (that were not used as part of training).

I/O:

The final constructed vaccine is both outputted to the terminal and outfiled to a separated constructed_vaccine_sequence.txt file. The feature selections and results (including accuracies) for each of the B cell, CTL, and HTL are all outputted to the terminal.

Results:

While the performance of the CTL type model was great and the HTL model was acceptable, the B cell model was not as strong, despite me giving it a greater num_to_choose (of 180) as opposed to the 150 for the T cells. I also tried making other modifications within the code, but the tradeoffs in terms of efficiency and runtime was quite large, so I recognize it is an area for future improvement and optimization.