Computational Medicine: 02-518/02-718

Carnegie Mellon University

Homework 5

Version: 1.0; updated 11/7/2020

Due: December 6 by 11:59pm

Hand-in: A single PDF to Gradescope that contains the following items:

1. A cover page that lists your name and Andrew id
    - If you worked on a team, indicate who your teammate(s) is/are by their name(s) and Andrew id(s).
        - Each team should have no more than 3 people.
    - Note: Each team should only hand in one pdf. It does not matter who does the actual upload. We will make sure that the grades are entered appropriately.
2. A PDF export of the Jupyter notebook for question 1

You can combine all three PDFs into one using Adobe Acrobat, or similar tool.

Overview

In this assignment, you will design a subunit vaccine targeting the spike glycoprotein of SARS-CoV-2. Specifically, you will select a set of short peptides that could theoretically be delivered using an RNA Vaccine. Obviously, designing a real vaccine is *much* harder than simply performing the steps in this assignment. These steps may be necessary for certain types of vaccines, but they are certainly not sufficient.

You can download the data for this assignment here: **HW5_Data.zip**. The file contains the training data and other files you will need to complete this assignment.

The goals of the assignment are as follows:

1. To give you the opportunity to practice learning regression models
2. To give you the opportunity to design a vaccine.

Question 1 (100 points)

To complete this question, use the provided jupyter notebook.

Part 1.1 (30 points)

Part 1.1.1 (15 points)

- (10 points) Use the data in the file *MHCI_Binding_Data.csv* to train a regression model for predicting pIC50 values (the final column in the file), given an HLA allele (the second column) and a peptide sequence (the fourth column).
- *Notes*:
    - You only need to be able to make predictions for the 27 HLA alleles in the file *hla_ref_set.class_i.csv*
    - There are several ways to approach this problem and you will have to decide:
        - Whether to build a single regression model for predicting MHC I binding affinities using the allele as an input feature (in addition to the peptide sequence), or build separate regression models for each HLA allele and only take the peptide sequence as an input, or something in-between (ex. a separate model for each locus HLA A, HLA B, …, HLA F)
        - How to encode the peptide sequence as an input feature. The length of the peptides vary from 8 residues to 30. One possibility is to train separate models for each peptide length. Another possibility is to select a single peptide length, and only build models for making predictions for peptides of that length. More advanced approaches might include defining a fixed-length encoding (ex k-mer histograms; or, using a designated value for a 'missing' residue) or *learning* a fixed-length encoding (ex. via auto-encoders).
- (5 points) Create a table where each row corresponds to one of the MHC I binding affinity models above, and the final column is the Coefficient of determination (ie., R^2) of the model using 5-fold cross validation.

Part 1.1.2 (15 points)

- (10 points) Use the data in the file *MHCII_Binding_Data.csv* to train a regression model for predicting pIC50 values, given an HLA allele and a peptide sequence.
- *Notes*:

    - You only need to be able to make predictions for the 27 HLA alleles in the file *hla_ref_set.class_ii.csv*
    - Once again, you will have to address the questions outlined above.
- (5 points) Create a table where each row corresponds to one of the MHC II binding affinity models above, and the final column is the Coefficient of determination (ie., R^2) of the model using 5-fold cross validation.

Part 1.2 (5 points)

- (5 points) The file *SARS-CoV2-Spike.fasta* contains the primary sequence of the spike glycoprotein of SARS-CoV-2. It is a large protein, with 1,273 residues. The models you created in part 1.1 take in peptides. Therefore, you will need to pre-process the spike protein by chopping it up into (overlapping) *k*-mers of the appropriate sizes for the models you created in part 1.1.

Part 1.3 (15 points)

- (10 points) Apply the models you created in part 1.1 to the *k*-mers you created in part 1.2. This will give you a set of labeled (MHC, peptide) pairs, where the label is the predicted pIC50 value.
- (5 points) Complete the following table using the statistics of the *predicted* pIC50 values for the spike protein *k*-mers using your models:

	pIC50s				
MHC Class	***Min***	***Max***	***Mean***	***Median***	***Std. Dev***
I					
II					

Part 1.4 (30 points)

- (10 points) Devise and implement an algorithm for selecting SARS-CoV-2 peptides from the set you created in part 1.3 that maximizes the 'allele coverage' (defined below), subject to the following constraints:

- A maximum number of peptides for use in the design. If the maximum is set to ∞, then there is no limit.
- The *predicted* pIC50 value between each peptide, e, in the design and *some* MHC allele a is at least p_{min}. That is:

$$p_{min} \leq argmin_{e \in Design} argmax_{a \in Alleles} f(e, a)$$

The 'allele coverage' for a design is the percentage of MHC alleles that are predicted to bind to at least one of the peptides in the design with a pIC50 value of at least p_{min}. That is: $Coverage(Design, Alleles) = |\{a: a \in Alleles \text{ and } p_{min} \leq argmax_{e \in Design} f(e, a)\}| / |Alleles|$

- *Note*: You can use any algorithm you see fit. Greedy and/or stochastic algorithms are fine, as are advanced algorithms.
- (10 points) Use your algorithm to compute the values to complete the following table, based on the predictions made using the models for MHC I alleles.

Design limit (# peptides)	p_{min}	**Actual design size**	**Coverage**
∞	3		
∞	6		
∞	9		
10	3		
10	6		
10	9		
20	3		
20	6		
20	9		

- (10 points) Use your algorithm to compute the values to complete the following table, based on the predictions made using the models for MHC II alleles.

Design limit (# peptides)	p_{min}	**Actual design size**	**Coverage**
∞	3		
∞	6		
∞	9		
10	3		
10	6		
10	9		
20	3		
20	6		
20	9		

Part 1.5 (20 points)

- (10 points) Suppose that you are given a design limit of 20 peptides to include in the vaccine. Select 20 peptides for your design and complete the following table. Note that you do not need to have an equal number of peptides 'covering' MHC I and MHC II alleles.

Peptide	Sequence	Target Allele	MHC Class	Predicted pIC50
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

- (5 points) Explain the criteria you used to select peptides.
- (5 points) What is the expected coverage of your design?