HW4: Sequence Alignment, Sequence Profiles, Machine Learning, Network Analysis, and Proteomics

Due: Monday, December 6th, 2021 at 5pm

Submission Guidelines

Please read the instructions carefully! Name all files using the following convention:

- lastName_firstName_shortDescriptionOfFile.extension.

You must submit **one** pdf file (with all solutions to questions) and **one** file containing your code. If using R this can either mean including:

- one pdf file and one .R file **or**
- one pdf file and one .Rmd (R markdown) file.

1 Semi-global Sequence Alignment

1. Suppose you want to perform a pairwise alignment of two sequences, S and T, but you do not want to penalize gaps at the beginning of S and at the end of T (for example, the sequences may only share similarity in an "overlap" region consisting of the end of the first sequence and the beginning of the second sequence and you wish to capture this). How would you modify the dynamic programming algorithm we discussed in class? Note that you will not have to modify the recurrence itself.

2. Using the modified algorithm from above, complete the alignment matrix for sequences S and T given below. Do NOT penalize gaps at the beginning of S and at the end of T. Use Match: 2, mismatch: -1, and gap: -2 as the similarity scoring function. (You may attach a picture of handwritten matrix for your response)

   S: GCAAGT
   T: ATGCTG

		G	C	A	A	G	T
A							
T							
G							
C							
T							
G							

3. Perform traceback on the alignment matrix from above, and report the final alignment of S and T (refer to the example below for format) and their alignment score. Clearly indicate the traceback using arrows.

   $s : \_\_CDEF$
   $t : ABCD\_\_$

2 Sequence Profiles

Suppose you are given the following alignment of a protein domain and you would like to build a HMM profile to model it and then use this profile to search for additional sequences:

```
1234567
MKEVNVR
M-KVQIE
M-DANIR
NRKA-VE
M-RVNID
```

You decide that your profile-HMM will directly model the 1st, 3rd, 4th, 5th and 6th, and 7th columns.

1. For your model, compute estimates for the probability of observing each of the 20 amino acids for the 5th column. Make sure to correct each estimate by adding a pseudocount ($\frac{1}{20}$) to each observation. Why is this correction necessary?
2. For your model, you will need to estimate the probability of having insertions and deletions after each modeled column. Give an estimate for having an insertion, deletion or neither after the 1st column. Make sure to to correct each estimate by adding a pseudocount ($\frac{1}{3}$) to each.
3. You notice that the 3rd and 7th columns of the alignment are correlated. Whenever there is a positively charged amino acid (K or R) in one, there is a negatively charged amino acid (D or E) in the other. Do HMM-profiles effectively capture these correlations? Why or why not?
4. What is the advantage of profile-HMMs over regular profiles/PSSMs?

3 Machine Learning to Classify Cancer Type

You are interested in developing a a biomarker panel for detecting the presence of breast cancer from information that can be easily collected during a blood test. You will use data from a proof-of-concept study published in 2018 by Patricio et al., “Using Resistin, glucose, age and BMI to predict the presence of breast cancer.” It showed how new cases of breast cancer could be effectively screened for by using simply obtained metabolic markers. Here, you will reproduce part of their analysis by training and testing a machine learning model.

There are two files containing 81 training samples (**training.csv**) and 35 testing samples (**testing.csv**), randomly split 70-30 from the data used in the paper. These files contain 9 columns of features for each individual including age, BMI, and various metabolite measures, along with their classification status as a healthy control (breast cancer negative, denoted 1) or a patient (breast cancer positive, denoted 2). Please see the Machine Learning Repository (where the data was accessed) or the original paper for more details.

1. Import the two data files **training.csv** and **testing.csv** as data frames into your working environment. If you are using R, install and import the packages below:

   ```
   install.packages("caret")

   library(caret)
   ```

   Note that the **Classification** column constitutes the labels (breast cancer negative (1)/positive (2)) for these data. The caret package requires that labels be given as factors. After importing your data, use the as.factor() function to change the **Classification** column in both data frames to a factor.

   How many breast cancer negative controls are in the training set?

2. Train a classifier on the training data to distinguish breast cancer presence using the Support Vector Machine (SVM) method as discussed in lecture. If using R, the **train()** function from the **caret** package will be useful here - you can directly specify the machine learning method ("svmLinear" in this case, we will use a linear kernel) with the **method** parameter. Be sure to specify that the **Classification** column is the labels and preprocess the data using the options ’center’ and ’scale’.

3. Now, test your trained classifier on the testing dataset by predicting the hidden cancer status of the testing data (**predict()** function if using R). Generate a confusion matrix and summary statistics (accuracy, precision, and recall) for your classifier (**confusionMatrix()** function if using R). Print out your confusion matrix.

4. Do you believe the classifier you trained is a good predictor of breast cancer status? Explain why or why not, citing the summary statistics from the previous problem. Does this conflict with the author’s assessment? Name at least one difference between our analysis and the authors’ analysis.

5. The authors use a process called cross-validation to build 95% confidence intervals for their summary statistics. Briefly describe what cross-validation is and why it is important in machine learning.

6. SVM is one type of method developed for use in machine learning. Another category of machine learning methods that is extremely popular in recent years is deep learning (neural network) methods. Neural networks have been proven to be highly effective and have enabled great strides in many fields, such as image recognition and natural language proccessing. However, deep learning is not without its limitations. In a few sentences, explain why neural networks might or might not be a good choice for the dataset and learning task we chose for this problem.

4 Network Analysis

1. Use *www.thebiogrid.org* to download the protein-protein interaction network for Stukalov, 2020 "Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV":

   - https://thebiogrid.org/222410/publication/multi-level-proteomics-reveals-host-perturbation-strategies-of-sars-cov-2-and-sars-cov.html
   - Only examine interactions with SARS-CoV-2 for this problem, not SARS-COV

   If using tidygraph, the following code with get you started converting the table to a tidygraph object. *Hint: A tidygraph object can be manipulated with dplyr verbs, i.e mutate, left_join, arrange, etc.*

```
library(tidygraph)
library(tidyverse)
library(ggraph)

biogrid <- read_tsv("BIOGRID-PUBLICATION-222410-4.4.203.DOWNLOADS.zip") %>%
  filter(`Experimental System` == "Affinity Capture-MS") %>%
  select(
    node_1=`Official Symbol Interactor A`,
    node_2 =`Official Symbol Interactor B`,
    spec_1 = `Organism Name Interactor A`,
    spec_2 =`Organism Name Interactor B`) %>%
  filter(spec_1 != "Severe acute respiratory syndrome-related coronavirus"
         & spec_2 != "Severe acute respiratory syndrome-related coronavirus") %>%
  filter(node_1 != node_2) # Remove self edges

edges <- biogrid %>% select(node_1, node_2)
g <- as_tbl_graph(edges, directed = TRUE)
```

2. Briefly describe the experimental approach in this paper to create the interaction network (2-3 sentences)

3. How many nodes and edges are present in the graph?

4. Plot the degree distribution of the coronavirus proteins. Which two proteins have the highest degree? What, if anything, is known about role of each of these two proteins during infection (1-2 sentences).

5. Which human proteins are found with more than two different SARS-COV-2 proteins?

6. Create a network visualization of this network, sizing nodes by their degree, and coloring by species.

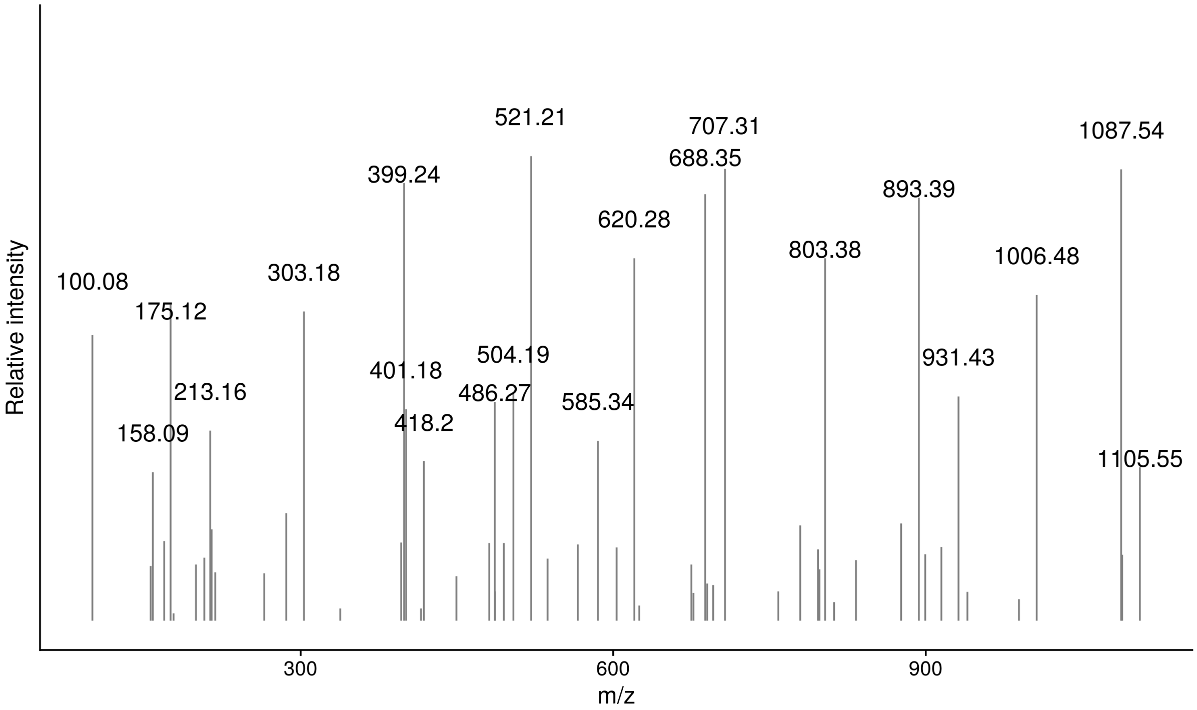
7. What is one way you could use this interaction network to learn about the biology of SARS-COVID-2 infection in humans?

5 Proteomics

1. Determine the amino acid sequence of the following spectrum using the provided monoisotopic mass table on the next page. The spectrum contains the full series of B and Y ions and a partial series of A and Z ions.

2. Describe briefly how you figured out the sequence (2-3 sentences)

3. How would a Post-Translation Modification of a peptide be identified from a mass spectrum?



3-letter code	**1-letter code**	**Monoisotopic mass**	**Chemical formula**
Gly	G	57.021	C2H3ON
Ala	A	71.037	C3H5ON
Ser	S	87.032	C3H5O2N
Pro	P	97.053	C5H7ON
Val	V	99.068	C5H9ON
Thr	T	101.048	C4H7O2N
Cys	C	103.009	C3H5ONS
Ile	I	113.084	C6H11ON
Leu	L	113.084	C6H11ON
Asn	N	114.043	C4H6O2N2
Asp	D	115.027	C4H5O3N
Gln	Q	128.059	C5H8O2N2
Lys	K	128.095	C6H12ON2
Glu	E	129.043	C5H7O3N
Met	M	131.04	C5H9ONS
His	H	137.059	C6H7ON3
Phe	F	147.068	C9H9ON
Arg	R	156.101	C6H12ON4
Tyr	Y	163.063	C9H9O2N
Trp	W	186.079	C11H10ON2

6 Project Update

1. Submit a separate R file containing all new code that has been written by you (as an individual) for your final project since the project update that was due November 19th. Please submit only code that we have not seen before, and people in the same group should not be submitting the same code; we are checking that each person is actively contributing to progress being made on your projects. Good luck!