**BT-3172: Special Topics in Bioinformatics: Practical computing for bioinformatics**

**Lab 8: Python advanced programming for bioinformatics.**

**Name: Anushka Udara**

**Index number: s14234**

In this practical, you will be implementing the majority voting algorithm on *Arabidopsis thaliana* DREB2A containing protein network to predict protein candidates for stress tolerance.

After using PyCharm to write your script, **copy the code to the appropriate space below the part III of the question**. Also, submit the Python file separately so they can be tested. Use the following format to name the script: YourIndexNo\_PrimaryQuestion.py.

1. Implementing the majority voting network-based candidate protein prediction algorithm.
   1. Search for the *Arabidopsis thaliana* DREB2A protein in the STRING protein-protein interaction database. Write its STRING ID below. Increase the maximum interactors to 500 (1st shell) and download the interactions in tabular format.

**AT5G05410.1**

* 1. Write the steps of an algorithm to predict the majority voting score of unknown proteins for a given function in a network. Assume that a list of known proteins annotated to the particular function is given as a text file. This should output/print the list of unknown proteins with the predicted majority voting score.

Open the stress proteins file using open function.

Read the lines using a for loop.

Split lines using tab if the lines do not have “#” and empty.

Add uppercase protein names to the stress\_proteins set.

Make the graph object using network package.

Open the string interactions file using open function.

Read the lines using a for loop.

Split lines using tab if the lines do not have “#” and empty.

Add edges to the graph uppercase protein interactions name pairs.

Print the degree of DREB2A protein.

Take all graph nodes using G.nodes() function.

Check every node using a for loop whether they are present in stress proteins set.

If not, add protein name to the unknow protein set.

Print the number of un-known protiens.

Get the intersection of nodes and unknown protein set as nodes.

Get the neighbors of every node in nodes using a for loop.

Check whether every neighbor in neighbors list is present in stress proteins set.

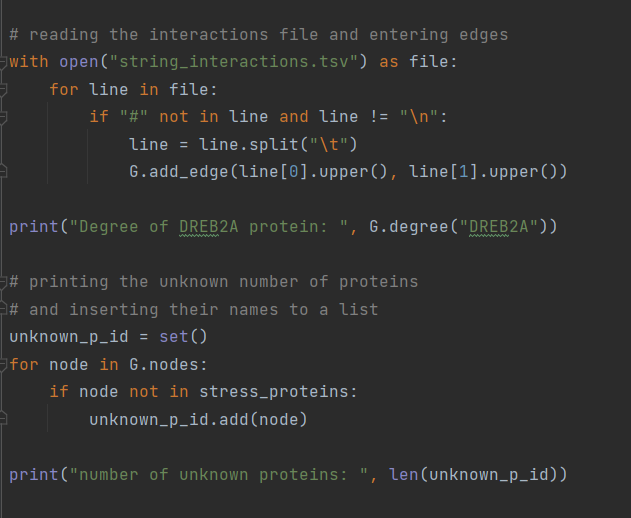
If yes, increment the voting.

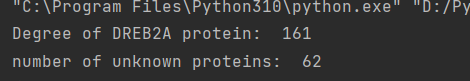
When the processing is done for one node, enter the node as key and voting as value to the majority voting score dictionary.

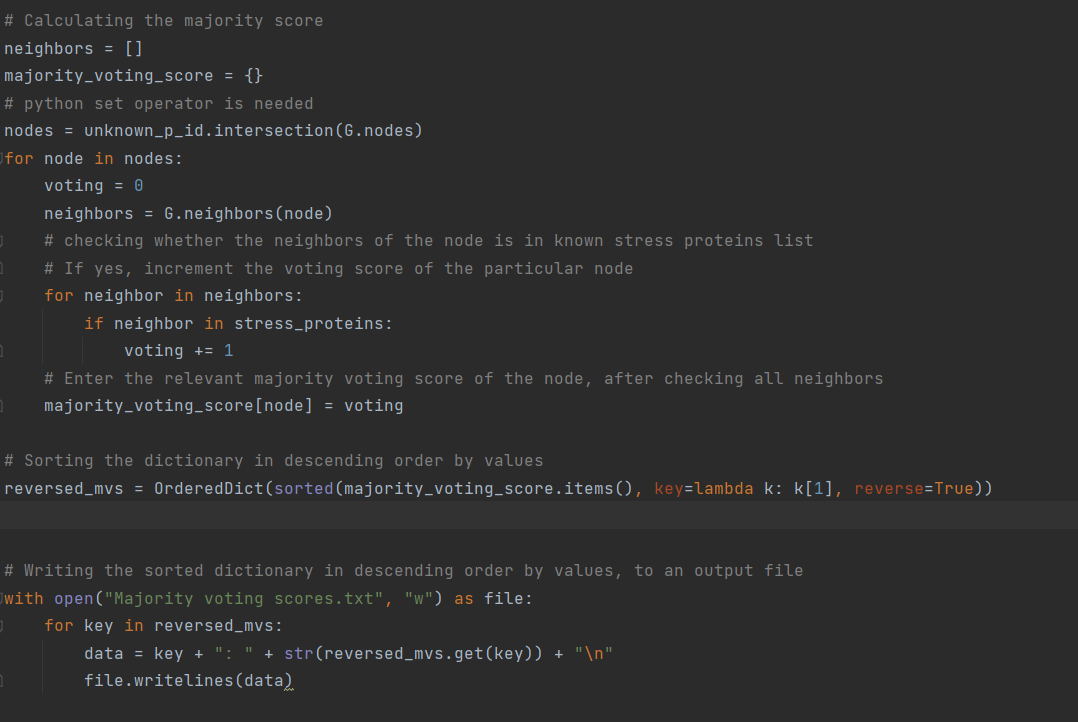
Sort the dictionary in descending order using values.

Write the key and value pairs of sorted-dictionary to a text file.

* 1. Implement the above algorithm in Python to predict the majority voting scores for all the unknown protein members for the stress tolerance biological process of the ATDREB2A network you downloaded in question (I). A data file containing known *Arabidopsis thaliana* proteins for stress tolerance is provided. (“AT\_stress\_proteins.txt). Please make sure you complete the following tasks.
     1. Print and write the degree of the ATDREB2A protein and the number of unknown proteins in the network for stress tolerance below.

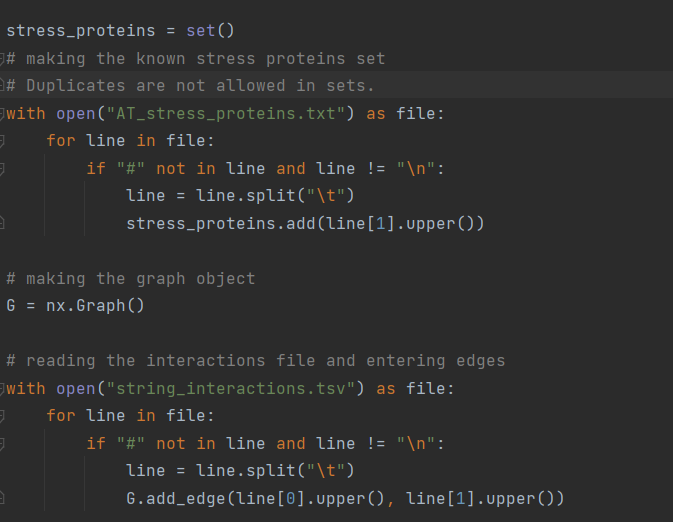




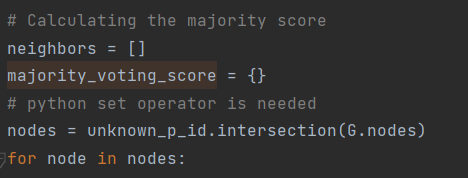
* + 1. After predicting the majority voting scores, you should sort them in descending order based on the scores, with proteins with high scores at the top. Then, write the ordered list to an output file.

*Hint: you can use OrderedDict submodule from the Collections Python package for sorting a dictionary based on values.*

* + 1. Pay a special attention to the names of the proteins. During the counting step, you have to match the protein names from the network and the input list. Pay a special attention to the sentence case of the protein names.



* + 1. You can use Python set operators to perform set matching, difference and removal of duplicates from a list. Please refer to a Python tutorial for more information.



* + 1. What is the unknown protein with highest majority vote score?

BZIP: 48

* 1. Briefly explain how different methods (computational and wet-lab) can be used to biologically validate your protein function predictions for abiotic stress.

First, computational methods can be used to predict the function of unknown proteins using the existing Protein-Protein interaction networks and known genes for abiotic stress. Then, wet-lab methods such as gene knockout, gene knock down, overexpression and ectopic expression are used to check the effect of deactivation or expression of a particular gene on the plant and validate the protein function predictions.