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

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

**Hansi Thewarapperuma – 14716**

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.

[**3702.AT5G05410.1**](https://string-db.org/network/3702.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.

**Input**: The text file containing known proteins annotated to a particular function(AT\_stress\_proteins.txt) / TSV file of interactions between all the proteins obtained from STRING DB (string\_interaction\_short.tsv)

**Output**: The list of unknown proteins with the predicted majority voting score

**Procedure**:

Create a network using all known and unknown proteins

Import networkx to generate a graph

Create an empty graph without no nodes and edges

Open the string\_interaction\_short.tsv file and do strip and split to relevant lines

Add edges to the empty graph from line[o] node to line[1] node meanwhile nodes are updated with .upper()

Create a list of all the nodes (here includes both known and unknown proteins) in the generated graph ( nodes\_list)

Create the list of known proteins

Create an empty list, namely known\_proteins\_list

Open the AT\_stress\_proteins.txt file and do strip and split to relevant lines

Append 1st index of every line (line[1].upper())to the above mentioned list

Use the set operator to eliminate duplications

Create the list of unknown proteins

Create an empty list, namely unknown\_proteins\_list

Iterate through the nodes\_list and check the nodes that are absent in known\_proteins\_list

Append those nodes to the above mentioned list

Predicting the majority voting score for the unknown proteins

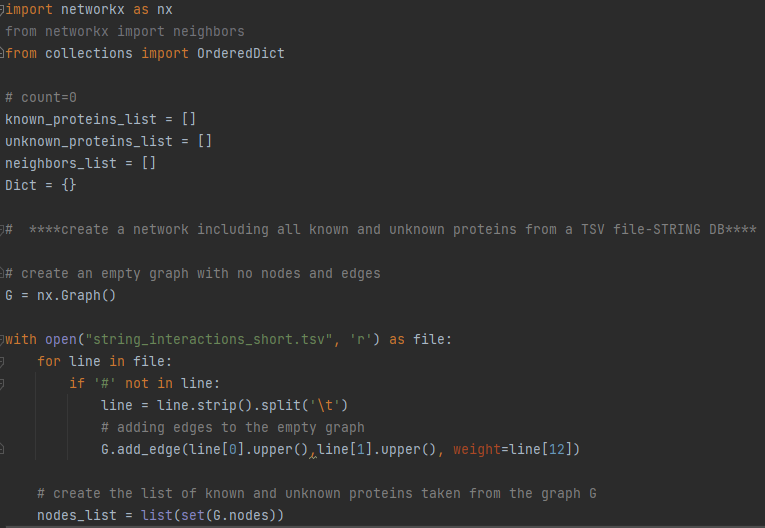
Assign the count variable to 0

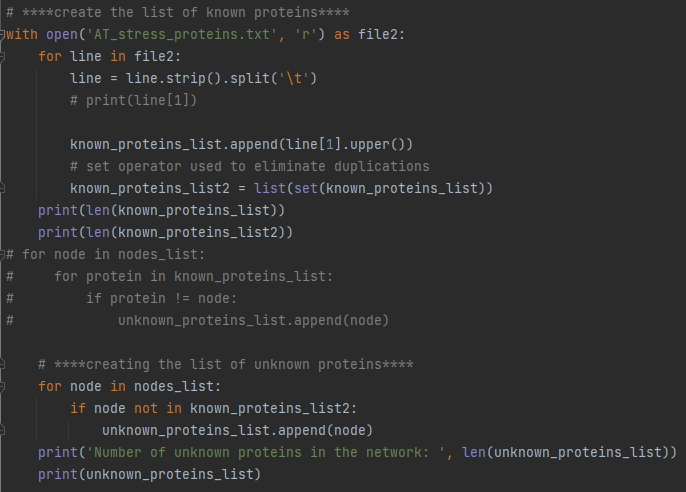
Iterate through the items of unknown\_proteins\_list and get the neighbours for each item and create a list for that

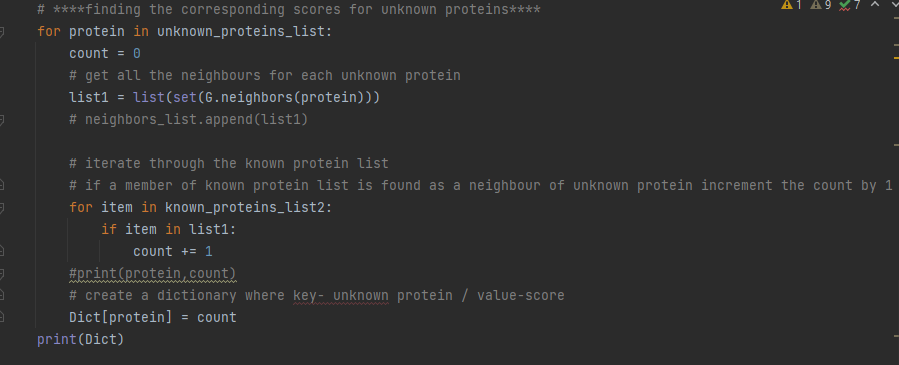
Iterate through the known protein list and if a member if known protein list is found as an element in neighbours list;

Count is incremented by 1

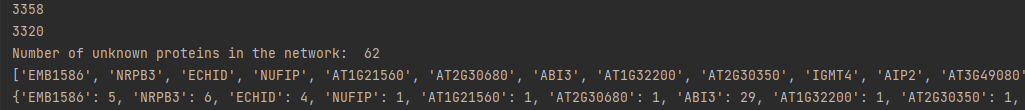
* 1. Implement the above algorithm in Python to predict the majority voting scores for all the unknown protein members for the stress tolerance biological process of the ATDREB2A network you downloaded in question (I). A data file containing known *Arabidopsis thaliana* proteins for stress tolerance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







**RESULT:**

****

* + 1. Print and write the degree of the ATDREB2A protein and the number of unknown proteins in the network for stress tolerance below.

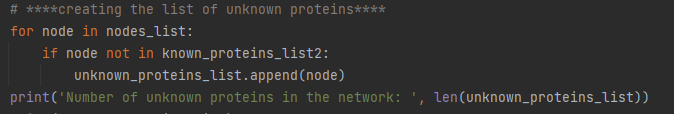
**CODE:**

****

**RESULT:**

****

**CODE:**

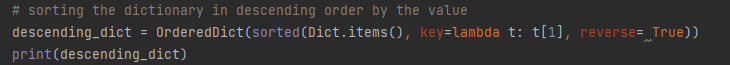
****

**RESULT:**

****

* + 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.

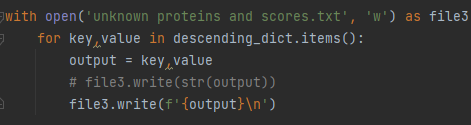
**CODE:**

****

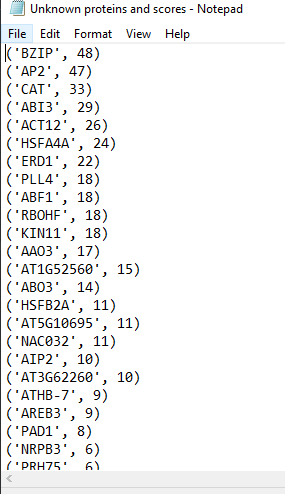
**RESULT:**

****

**CODE:**

****

**RESULT:**

****

*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.
    2. 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.
    3. 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.

Wet lab :

By conducting gene knockouts/ knockdowns as well as investigating on overexpression and ectopic expression

Computational :

By conducting Sequence based methods /Network based methods or using Phylogenetic profiles