**Prediction of protein candidates for stress tolerance in *Arabidopsis thaliana***

**Author: Hansi Thewarapperuma**

**Date: 24.01.2022**

Majority voting algorithm was implemented on *Arabidopsis thaliana* DREB2A containing protein network to predict protein candidates for stress tolerance.

The initial steps were as follows:

* *Arabidopsis thaliana* DREB2A protein in the STRING protein-protein interaction database was searched. The corresponding STRING ID was [**3702.AT5G05410.1**](https://string-db.org/network/3702.AT5G05410.1)
* The maximum interactions were increased to 500 (1st shell) and the interactions were downloaded in tabular format (“string\_interactions\_short.tsv”)
* We have a text data file containing known *Arabidopsis thaliana* proteins for stress tolerance is provided. (“AT\_stress\_proteins.txt”)

The algorithm to predict the majority voting score of unknown proteins for a given function in a network:

**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 in a text file

**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

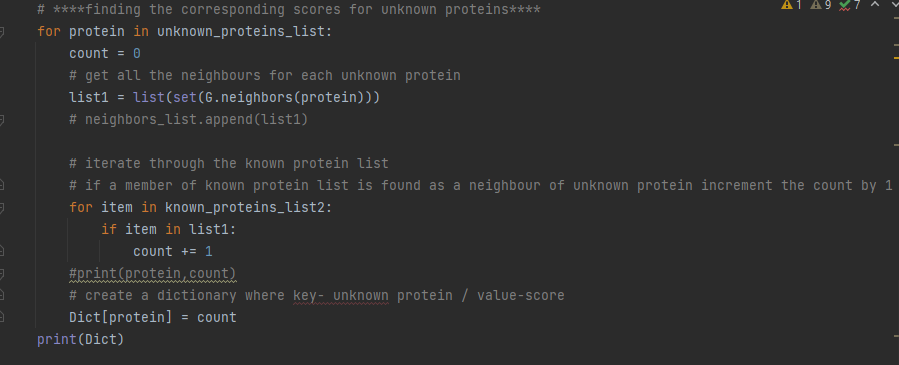
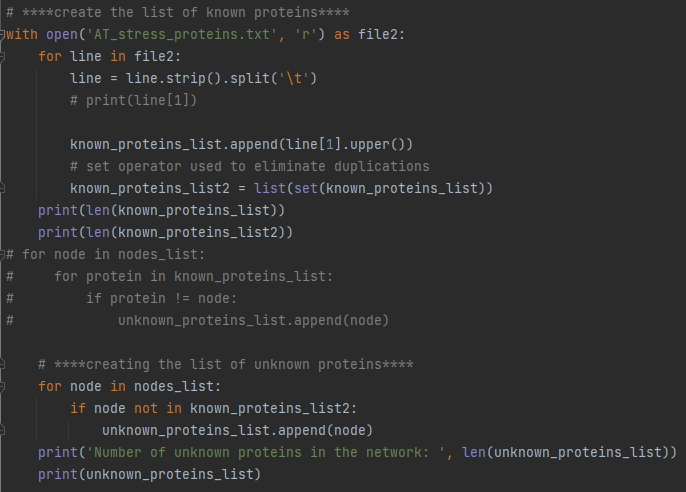
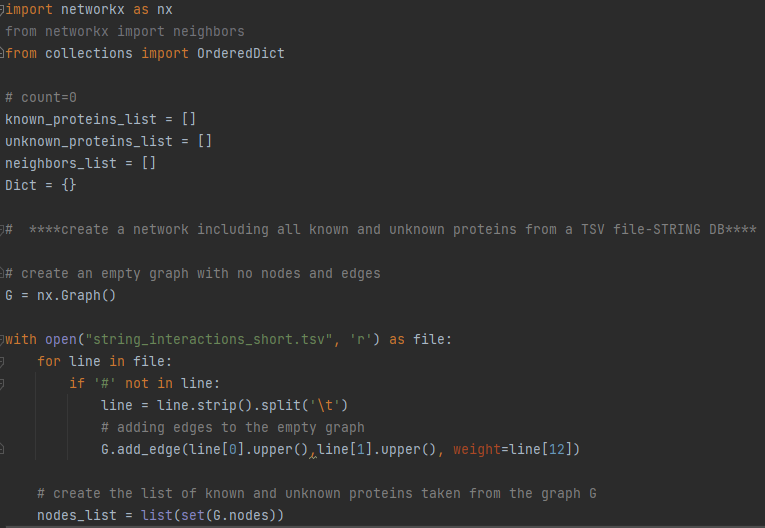
Predicting the majority voting score for the unknown proteins

Assign the count variable to 0

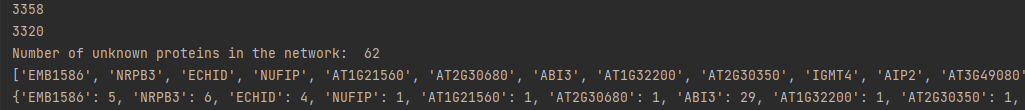
Iterate through the items of unknown\_proteins\_list and get the neighbours for each item and create a list for that

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

**Implementation:**

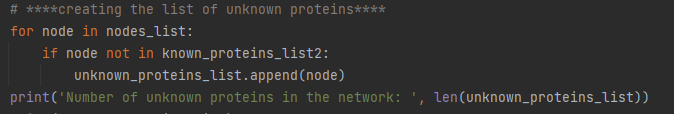
**Result:**

****

**Degree of the ATDREB2A protein and the number of unknown proteins in the network for stress tolerance**

**Code:**

****

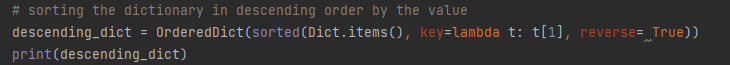
****

**Result:**

****

**After predicting the majority voting scores, the scores were 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.**

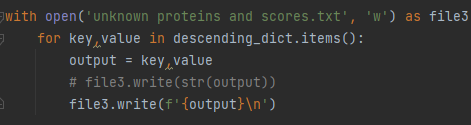
**Code:**

****

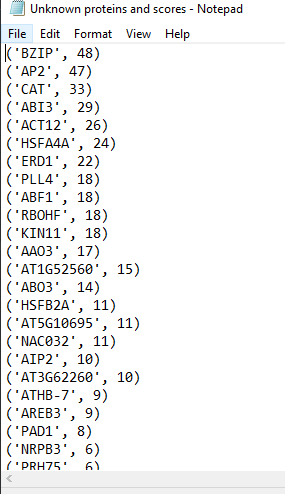
**Result:**

****

**Code:**

****

**Result:**

****

According to the obtained output, BZIP protein is the unknown protein with highest majority vote score (48)

Nevertheless, it is vital 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 methods:

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