How to generate/replicate the creation of the positive set and negative set:

1. Running the code

To generate the creation of the positive set and negative set, there are two options:

- Option 1: generate positive and negative sets with existing results. Usage: *run_positive_negative_set_option1*.
- Option 2: generate positive and negative sets from original data, involving all required computations. Usage: follow instructions in Option 2 section below.

Here is an overview of the execution of each option. The execution times reported below are measured executing the code in Windows 10 Pro with 256 GB RAM, and AMD Ryzen Threadripper PRO 3995WX 64-Cores CPU with 2.70 GHz. The software environment is MATLAB 2019a. The RStudio version used is 4.3.1.

Option 1: Total execution time ~**8h**

Runs *run_positive_negative_set_option1* to generate both sets with existing results located in data folder. In data folder, there is:

- original_data: contains the raw data downloaded from its source.
- matrix: contains the adjacency matrix of each Yeast DIP network.
- scripts: contains *create_positive_set_Yeast_DIP* and *create_negative_set_Yeast_DIP*.
- table: contains an excel sheet with the Uniprot IDs of mapped proteins in Yeast DIP network.

Option 2: Total execution time ~**9 h**

To generate the positive set and negative set from the original data, please follow the instructions below:

1. In order to download the data, follow the instructions reported in the file *instructions to download the data* located inside the folder “data”.
2. Go to Positive_Negative_set/data_replicated/scripts.
3. Run *create_Scere_DIP_net* to create Yeast DIP adjacency matrix. It returns Yeast_DIP_net.mat, the adjacency matrix of Yeast DIP, in “matrix”; and

Scere_uniprot_ids.csv in “table”, an excel sheet with the Uniprot IDs of mapped proteins in the Yeast DIP network. Total execution time ~**13 s**.

4. Convert the Uniprot IDs in Scere_uniprot_ids.csv to Entrez IDs with DAVID Gene ID conversion tool (https://david.ncifcrf.gov/conversion.jsp). Name the output text file ‘Uniprot_to_EntrezID_Yeast_DIP_net’ and store it in Positive_Negative_set/data_replicated/. Total execution time: **negligible time.**
5. Run in Rstudio *GOSemSIM_DIP.R (we run this function in December 2021, this means that the obtained semantic similarities are associated to the version of the GO annotation database and the library released in this month)* located in “scripts”. The outputs are two text files (Yeast_DIP_BP.txt and Yeast_DIP_CC.txt) saved in Positive_Negative_set/data_replicated/. Total execution time: ~**45 min.**
6. Run the script *run_positive_negative_set_option2* located in “scripts”
    - *Create_positive_set_Yeast_DIP* returns in “data_replicated” a text file with the list of positive protein pairs. Total execution time: ~**8h.**
    - *Create_negative_set_Yeast_DIP* returns in “data_replicated” a text file with the list of negative protein pairs. Total execution time: ~**1 min.**

Note: running option 2 might generate results slightly different in relation to which version of the GO annotation database and library is adopted in the point 4 above.

2. Required library packages for R code

GOSemSim, tidyverse, readxl and data.table.