Project Progress October 2021

SoloDolo
Evan Momen

Project Resources

- All code written in Python 3
- Code is written and executed within VSCode
  - Using Interactive Python integrated into the official Python extension for VSCode
  - ‘# %%’ creates an executable cell

MATLAB and BioGrid PPI data -> csv

- MATLAB function *csvwrite()* converts all PPI data to csv files
- BioGrid text data is parsed in Python
  - For each experimental method, a txt file was written to include all PPIs verified by that particular method (2 protein names per line)
  - Convert each txt file to a matrix and write the matrix out to a csv file

Importing csv data in parallel

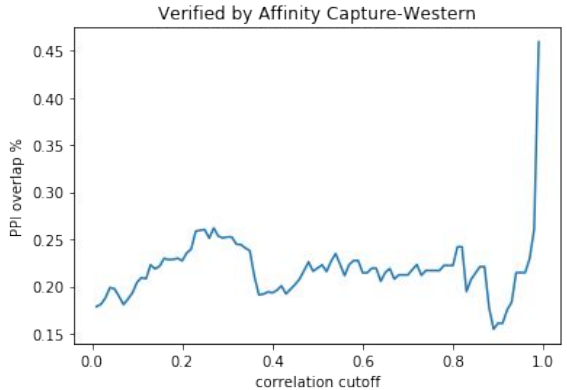
- Goal: speed up the importing of data
- All csv files are imported into Python via a multi-process worker pool
- Pool class in Python’s multiprocessing module

Stage 1: Correlation testing

- Full correlation matrix = matrix of proteins indicating how correlated any pair of proteins is
- Apply correlation cutoff to full correlation matrix to generate a new PPI network
- For each tested cutoff, determine what percentage of PPIs in the new network are verified by the BioGrid methods
    - Involves finding the intersection of the new PPI matrix and the BioGrid matrix
- Also performed in parallel

Visualization of results

- Plots are made with matplotlib
- Percentage of each correlation cutoff's PPI network that is verified by BioGrid is plotted against all cutoffs



Problems

- Verifying data formats are correct
  - Numpy uses float by default
- Verifying assumptions about data
  - BioGrid data contains self-interactions
- Correcting data
  - Main diagonal of full correlation matrix is all 1’s
- Running code in parallel
  - Understand memory space of each process

Currently

- Stage 1 complete
- Stage 2 - generating new PPI networks by degree instead of correlation cutoff
    - Verification by BioGrid data