Project 6 - Bioinformatics Data Analysis

Data

Data originate from The Saccharomyces Genome Database (SGD) that provides comprehensive integrated biological information for the budding yeast Saccharomyces cerevisiae. Data encompass 1799 genes and their observed expressions under 740 different conditions from SGD.

Aim of the project is to evaluate the discovery of functional relationships between genes.

Tasks

1. Calculate Pearson correlations between genes. Present pairwise matrix as heatmap.
2. Calculate correlations of randomly selected genes and permuted data (permute values across rows), repeat procedure 10000 times. Present results as histogram along with histogram of correlation values obtained in the first step.
3. Create weighted graph from original correlation matrix, use 99^{th} percentile of correlations obtained on random data as a threshold: keep only edges with weights higher than this threshold.
4. Calculate global graph properties (diameter, radius, average clustering coefficient) for obtained graph. Visualize graph or parts of the graph.
5. Create histograms of degree distribution and shortest paths. Present histograms.
6. Analyse local properties of graphs (degrees, centrality measures, page ranks, hubs, clustering coefficient…)
7. Rank genes according to local properties of graph and for the top ranked genes and its ego graphs discover what are their functions (for this step use Orange package Bioinformatics add-on)
8. Propose your own analysis step that could be interesting for this project.

Project Delivery

Prepare project presentation (~15 slides) and python code (.py files or jupyter notebook).

Project presentation should include:

- Domain description (motivation)
- Data description
- Results

- Implementation information (list of Python packages, functions)
- Conclusions
- Ideas for future work