BIOINFORMATICS PROJECT - ***Network Biology***

Part 1.2 - steps and methods

Scope of the project:

Starting from the seed genes interactome (SGI), the intersection (I) and the union (U) interactomes built in the first part of the project, compute the main network measures for I, U and their nodes, apply clustering methods for disease modules discovery, carry on an enrichment analysis on the putative disease modules and produce a short report.

----

1) Calculate the main network measures for SGI, I and U

1.1 Calculate the following **global** (i.e. concerning the whole network and not the single nodes) measures of SGI, U and I (only if no. of nodes >20):

- No. of nodes and no. of links
- No. of connected components
- No. of isolated nodes
- Average path length
- Average degree
- Average clustering coefficient
- Network diameter & radius
- Centralization

1.2 Isolate the largest connected component (LCC) of I and U and calculate the following **global** (a) and **local** (b, i.e. for each node) measures:

a)

- N. of nodes and no. of links
- Average path length
- Average degree
- Average clustering coefficient
- Network diameter & radius
- Centralization

b)

- Node degree
- Betweenness centrality
- Eigenvector centrality
- Closeness centrality
- ratio Betweenness/Node degree

Store the results in a suitable matrix format of your choice.

2) Apply clustering methods for disease modules discovery

Cluster I-LCC and U-LCC using the following algorithms to get the modules:

- MCL
- Louvain

Once you have clustered the networks, find modules with no. of nodes >= 10 in which seed genes are statistically overrepresented ($p<0.05$) by applying a hypergeometric test: such modules will be the “putative disease modules”.

Store the results for both U-LCC and I-LCC in tables including in each row: *clustering algorithm used, module ID, no. of seed genes in the module, total no. of genes in each module, seed gene IDs, all gene IDs in the module, p-value.*

3) Carry on an enrichment analysis on the disease modules

Find overrepresented GO categories (limit to first ten) and overrepresented pathways (limit to first ten) for the putative disease modules.

4) Find putative disease proteins using the DIAMOnD tool

Using the tool DIAMOnD, compute the putative disease protein list using as reference interactome (“network_file”) the latest BioGrid interactome already used to collect PPIs.

Software and instruction for DIAMOnD:
https://github.com/barabasilab/DIAMOnD

As “seed_file” use your seed gene list, limit the number of putative disease proteins (“n”) to 200, and omit the “alpha” parameter (it will be set by default to 1).

Find overrepresented GO categories (limit to first ten) and overrepresented pathways (limit to first ten) for the intersection list joined with your seed genes list.

6) Summarize the following information in the report:

- global measures of SGI, I, U, I-LCC, U-LCC

- a figure of the SGI and of the I-LCC networks (do not forget figure captions)

- a table with the first 20 highest ranking genes for betweenness (include in the table all other calculated centrality measures as from 1.2b) for I-LCC and U-LCC

- summary table of the putative disease modules found with each of the two clustering algorithms (*clustering algorithm used, n. of modules, n. of seed genes in each module, total n. of genes in each module, ratio n. seed genes/total genes in the module, p-value of the enrichment using the hypergeometric test*)

- the first 40 genes coming from the DIAMOnD tool

- notes and comments on the method followed, discrepancies, lack of data, any other point worth to be mentioned.