BIOINFORMATICS PROJECT - Network Biology

Part 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 node as well as node roles according to the method by Guimera & Amaral, 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) measures of SGI, U and I (if n. of nodes >20):

- N. of nodes and links
- N. of connected components
- N. 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 and local (i.e. for each node) measures for both the LCCs only:

a)

- N. of nodes and links
- Average path length
- Average degree
- Average clustering coefficient
- Network diameter & radius
- Centralization

b)

- Node degree
- Betweenness centrality (normalized)
- Eigenvector centrality (normalized)
- Closeness centrality (normalized)
- Node ratio betweenness/degree

Store the results in a suitable matrix format.

2) Apply clustering methods for disease modules discovery

Cluster I-LCC and U-LCC using the following algorithms:

- simulated annealing
- MCL
- Louvain

In each of the clustered partitions, find modules 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, n. of seed genes in the module, total n. of genes in each module, seed gene IDs, all gene IDs in the module, p-value.*

Notes:
- to discover modules using the simulated annealing algorithm you can also directly use the Netcarto tool (point 4), which will also provide the node roles;
- if there are issues in using simulated annealing algorithms implemented in R or python, use the spinglass algorithm, which is similar.

3) Carry on an enrichment analysis on the disease modules

If n. of nodes in the module >20, find overrepresented GO categories (limit to first ten) and overrepresented pathways (limit to first ten) for the putative disease modules.

4) Find roles of I-LCC nodes (according to Guimera & Amaral method)

Using the tool Netcarto on I-LCC, calculate node roles, draw the role cartography map and save results in a table with the columns: node (gene name and/or Uniprot AC), cluster number, z-score, participation coefficient, role (R1, …, R7).

Software and detailed instruction for Netcarto:
https://amaral.northwestern.edu/resources/software/netcarto

Use loose input parameters (such as those for big networks, see instructions) and set 0 randomization iterations. As from the instructions, if you observe a gap of ~seconds between two iterations, the program will likely employs days to complete the task: this is not a problem if you have enough time. If you don't, use a clustered partition of I-LCC (with a clustering algorithm of your choice) and calculate z-scores, participation coefficient and node roles.

5) Find putative disease proteins using the approach from Ghiassian et al.

Using the tool DIAMOnD (or a customized workflow), compute two putative disease protein lists using as reference interactome ("network_file"):

- APID Q1
- latest BioGrid interactome

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).

Compute the intersection list between these two lists.

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 gathered information in the report that must include:

- global measures of SGI, I, U, I-LCC, U-LCC
- a figure of the SGI and of the I networks (obtained with Cytoscape)
- 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- and U-LCCs
- summary table of the putative disease modules found with each of the three 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)
- the role cartography map
- the first 40 genes coming from the DIAMOnD tool, for each of the two reference interactomes, and the intersection of the two putative disease proteins list.
- notes and comments on the method followed, discrepancies, lack of data, any other point worth to be mentioned.