BIOINFORMATICS – Network Medicine project

Part 1 – Data collection

Scope of the project:

Starting from existing knowledge about a pathological condition: a) explore the related information sources (DisGeNet datasets), b) collect the list of human genes of interest (hereinafter 'seed gene list'), c) get protein-protein interaction data, d) carry on a preliminary analysis.

Note: in this project we will often use the terms 'gene' and 'protein' as synonyms, even if they are not, from the purely biological point of view.

Steps and methods:

1.1) Explore information sources and compile the seed gene list:

a) Explore the DisGeNet dataset, find the disease of interest and get the list of human genes involved.

b) For all genes in the seed gene list, collect the following basic information from the Uniprot:

- official (primary) **gene symbol** (check if the symbols are updated and approved on the HGNC website; report any issue/lack of data/potential misinterpretation)
- **Uniprot AC**, alphanumeric 'accession number' (a.k.a. 'Uniprot entry')
- **protein name** (the main one only, do not report the aliases)
- **Entrez Gene ID** (a.k.a. 'GeneID')
- very **brief description** of its function (keep it very short, i.e. max 20 words)
- notes related to the above information, if any and if relevant

c) Store the data gathered in a table in an easily accessible format of your choice (csv, tab, excel, etc).

1.2) Collect interaction data

a) For each seed gene, collect all binary protein interactions from two different PPI sources:

i) Biogrid Human, latest release available
ii) IID Integrated Interactions Database (experimental data only, all tissues, unless stated otherwise in further instruction)

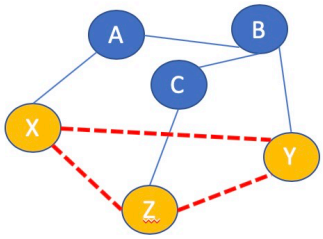
***Note**: once you got the list of the proteins interacting with at least one seed gene, you must also retrieve and include in your interactome the interactions among these non-seed proteins, as from this example:*

A, B and C are seed genes;

*X, Y, Z are **not** seed genes, but they interact with at least one seed gene (blue lines in the figure below):*

interaction table:
[interactor 1--interactor2]

A—B
A—X
B—C
B—Y
C—Z
X-?-Y
X-?-Z
Y-?-Z



*<u>if there are interactions among X, Y, or Z</u> (red dotted lines in the figure) then **<u>these interactions must be reported</u>**, even if they do not involve any seed gene.*

b) Store the data gathered from the two DBs in two different tables/matrices in an easily accessible format of your choice (csv, tab, excel, etc).

c) Summarize the main results in a table reporting:

a) no. of seed genes found in each different DBs (some seed genes may be missing in the DBs);
b) total no. of interacting proteins, including seed genes, for each DB;
c) total no. of interactions found in each DB.

1.3) Arrange interaction data

Build and store three tables:

a) seed genes interactome: interactions that **involve seed genes only**, from all DBs, in the format:

interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC, interactor B Uniprot AC, database source

b) union interactome: all proteins interacting with at least one seed gene, from all DBs, same format as above.

c) intersection interactome: all proteins interacting with at least one seed gene confirmed by both DBs, in the format:

interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC, interactor B Uniprot AC

Always check that interactors are both human (i.e. organism ID is always **9606,** Homo Sapiens)

1.4) Enrichment analysis

a) Using the service **Enrichr**, find, report in tables and save related charts (8 charts in total) of the overrepresented GO categories (limit to the **first 10** for each main category, BP, MF, CL) and the the overrepresented pathways (KEGG 2019 Human) for:

a) the seed genes,
b) the union interactome genes

Part 2 – Data analysis

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.

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

a) 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

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

i)

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

ii)

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

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

2.2) Apply clustering methods for disease modules discovery

Cluster I-LCC and U-LCC using the **MCL** algorithm to get the modules.

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.*

2.3) 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 genes belonging to each putative disease module.

2.4) Find putative disease genes using the DIAMOnD tool

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

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

Find overrepresented GO categories (limit to first ten) and overrepresented pathways (limit to first ten) of such 200 newly found genes.

Part 3 – Reporting

3.1) Summarize the following information in a short report which includes:

- very short intro (10 lines max) about the pathophysiological condition (the seed genes context) and, if any, issues with gene IDs
- a table with seed genes information (point 1b; omit "protein description")
- a summary table of interaction data (point 2)

- the 8 charts of the enrichment analysis from Enrichr (point 4)
- a table with 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 also 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 (*for each module: no. of seed genes in each module, total no. of genes in each module, ratio no. seed genes/total genes in the module, p-value of the enrichment using the hypergeometric test*)
- the list of the first 30 genes identified by the DIAMOnD tool and charts from Enrichr.
- notes and comments on the method followed, discrepancies, lack of data, any other point worth to be mentioned.

Notes: all tables and figures must have a caption (they must be self-consistent); a report template is provided.