10593052 BIOINFORMATICS & NETWORK MEDICINE 2020-21

Network Medicine project

Part 1 – Data collection

Scope of the project:

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

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

Steps and methods:

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

a) Starting from the disease of interest, explore the DisGeNet dataset "Curated gene-disease associations", and get the list of human genes involved in the disease (hereinafter, "seed genes").

b) For all genes in the seed gene list, **first** check if gene symbols are updated and approved on the HGNC website, reporting any issue/lack of data/potential misinterpretation, and **then** collect the following basic information from the Uniprot:

- **official** (primary) **gene symbol**
- **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

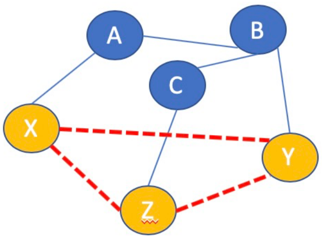
a) For each seed gene, collect all binary protein interactions from the Biogrid Human.

***IMPORTANT**: once you got the list of the non-seed proteins interacting with at least one seed gene, you must also retrieve and include in your interactome the interactions, if any, 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 edgelist:
[interactor 1--interactor2]

A—B
A—X
B—C
B—Y
C—Z
X-?-Y
X-?-Z
Y-?-Z



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

b) Store the data gathered from the Biogrid in a table/matrice 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 collected in Disgenet and no. of seed genes found in Biogrid (some seed genes may be missing in the Biogrid);
b) total no. of interacting genes/proteins found, including seed genes;
c) total no. of interactions found.

1.3) Arrange interaction data

Build and store three tables:

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

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

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

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

1.4) Enrichment analysis

Using the **Enrichr** webservice, find, report in tables and save related charts (4 charts in total) of the overrepresented GO categories (limit to the **first 10 terms** for each main category, BP, MF, CL) and the the overrepresented pathways (KEGG 2019 Human) for the disease interactome.

Part 2 – Data analysis

Scope:

Starting from the disease interactomes built in the first part of the project, compute the main network measures, apply clustering methods for disease modules discovery, carry on an enrichment analysis on the putative disease modules and release a short report.

2.1) Calculate the main network measures for the disease interactome

a) Calculate the following **global** (i.e. concerning the whole network and not the single nodes) measures of the disease interactome:

- 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 the disease interactome 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 the disease interactome 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, if existing, will be the "**putative disease modules**".

Store the putative disease modules results in tables including in each row: *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 (again, always remember to clean the data by purging non human proteins). 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 (i.e. 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 1.2c)
- the charts of the enrichment analysis from Enrichr (point 1.4)
- a table with global measures of the disease interactome LCC
- a figure of the LCC (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 the LCC
- summary table of the putative disease modules found (*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 (i.e. they must be self-consistent); a report template is provided.