BIOINFORMATICS – Network biology project

Part 1 - steps and methods

Scope of the project:

Starting from existing knowledge about a physiological/clinical/pathological condition or process, a) explore the related information sources (experiments and datasets, literature, databases, etc), b) collect the list of human genes of interest ('seed gene list'), c) get protein-protein interaction data, d) carry on a preliminary analysis and e) produce a short report.

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.

----

1) Explore information sources and compile seed gene list:

~~~ *NOTE:* ***ALL SEED GENES WILL BE PROVIDED BY THE INSTRUCTOR*** ~~~

[if the list is provided, proceed to point 2, otherwise:

a) explore existing sources carefully (literature, experiments and datasets, databases, etc) and provide the gene list related to the studied condition/biological process (usually from few to hundreds of genes, the number may vary greatly);
b) justify the inclusion/exclusion of each selected gene in the seed gene list;
c) based on the understanding of the main sources exploration and their scientific grounds, it is possible to discriminate genes involvement/importance in the studied condition across different '*levels of involvement*': in this case, assign seed genes to at least two different classes of importance/involvement, e.g. 1st class genes, considered more directly/strongly involved in the process/disease, 2nd class genes, less directly involved, weaker evidence of involvement; justify the inclusion of each gene in the different classes.]

2) Collect basic information about seed genes

2.1 For all genes in the seed gene list, collect and store the following basic information:

- official gene symbol (check if the symbols are updated and approved on the HGNC website; report any issue/lack of data/potential misinterpretation)
- Uniprot AC, '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

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

3) Collect interaction data

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

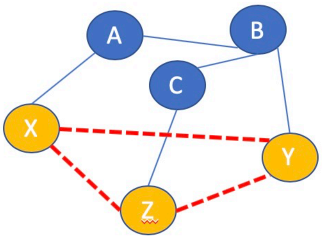
a) Biogrid Human, latest release available
b) IID Integrated Interactions Database (all tissues, unless stated otherwise in further instruction)

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



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.*

For each DB, check if the results are different when carrying out the search by using gene symbols or by using Uniprot AC identifiers. Discrepancies must be reported.

3.2 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).

3.3 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 one of the DBs);
b) total no. of interacting proteins, including seed genes, for each DB;
c) total no. of interactions found in each DB.

4) Arrange interaction data

Build and store three tables/matrices:

4.1 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

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

4.3 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

5) Enrichment analysis

5.1 Using innateDB, find and report in tables the overrepresented GO categories (limit to first ten for each main category, BP, MF, CL) for:

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

5.2 Using the same DB, find the overrepresented pathways (limit to first ten) for the interactomes above.

6) Arrange information in a short ~readable~ report including:

- very short intro about the pathophysiological condition (the seed genes context) and issues with gene IDs, if any
- a table with seed genes information (point 2; omit "protein description")
- a summary table of interaction data (point 3)
- the enrichment analysis (point 5)
- notes and comments on the method followed (discrepancies found, lack of data, any other point worth to be mentioned)

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