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 meaning that if a gene has been identified as involved in a disease, then same is for the related protein.

---

1) Explore information sources and compile seed gene list:

~~~ NOTE: ***ALL SEED GENES WILL BE PROVIDED*** *~~~*

[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 basic information and arrange it in a table with:

- official gene symbol (check if the symbols are updated and approved on the HGNC website)
- Uniprot AC, 'accession number' (a.k.a. 'Uniprot entry')
- Uniprot ID (a.k.a. 'Uniprot entry name')
- 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

3) Collect interaction data

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

a) Apid Level 2 human interactome (if no interactions are found, switch to Level 1 interactome).
b) Biogrid Human v.3.4.154
c) String (consider 'Experiments', 'Databases', 'Co-expression' interaction sources only)

Note: once you got the list of the proteins interacting with at least one seed gene, you must also retrieve and include the interactions among these proteins, example: seed gene A interacts with protein Z; seed gene B interact with protein X; if there is an interaction between Z and X, then it must be reported.

For each DB, when possible, check if the results are different using gene symbol and Uniprot AC identifiers.

When present, keep data about 'type of interaction'.

3.2 Store the data gathered from the three DBs in three 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 some 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 three tables:

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

protein A gene symbol, protein A Uniprot AC, interaction type, protein B gene symbol, protein B Uniprot AC, database source

4.2-**union interactome**: all proteins interacting with at least one seed gene, from all three DBs above, in the format:

protein A gene symbol, protein A Uniprot AC, interaction type, protein B gene symbol, protein B Uniprot AC, database source

4.3-**intersection interactome**: all proteins interacting with at least one seed gene confirmed by all three PPI sources, in the format:

protein A gene symbol, protein A Uniprot AC, protein B gene symbol, protein B Uniprot AC

5) Enrichment analysis

Find and report in tables the overrepresented GO categories (limit to first ten) and overrepresented pathways (limit to first ten) for:

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

6) Arrange information in a short report including:

- table with seed genes information (point 2)
- summary of interaction data (point 3)
- intersection interactome list (point 4)
- enrichment analysis (point 5)
- notes and comments on the method followed (discrepancies found, lack of data, any other point worth to be mentioned).
