Data Science 10593052 Bioinformatics & Network Medicine 2020-2021

Date: DD Month YYYY

Network Medicine Project report

|  |
| --- |
| Network Medicine@Data Science A.Y. 2020-2021  Manuscript Title  Laurenti Laura, Mandica Paolo, Testa Lucia  Group no. 1 Abstract Max 150 words. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. |

**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.

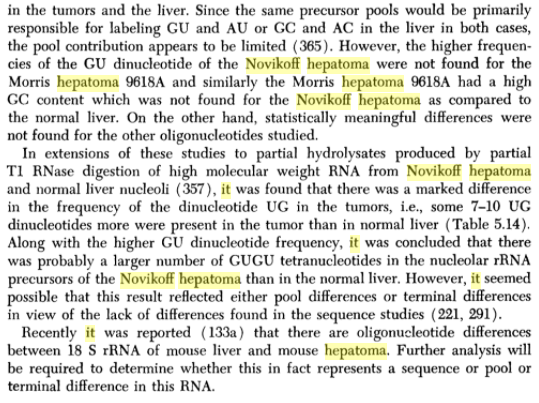
Basic introduction about the disease/process

Max 200 words. Add here few basic information about the disease/process under scrutiny.

DA RIASSUMERE

Novikoff hepatoma mitochondria were studied with regard to their functional and structural organization. The activities of cytochrome oxidase and malate dehydrogenase in mitochondria purified on sucrose gradients were similar in organelles derived from normal liver and hepatoma tissue. However, the activities of reduced nicotinamide adenine dinucleotide oxidase and succinate oxidase in hepatoma mitochondria were reduced to only 10 and 35%, respectively, of the activities in liver mitochondria. Also, monoamine oxidase and rotenone-insensitive reduced nicotinamide adenine dinucleotide-cytochrome c reductase (enzymes localized in the outer mitochondrial membrane) have significantly reduced activities in hepatoma mitochondria. The structural changes in hepatoma mitochondria might be correlated with differences in the banding patterns of liver and hepatoma mitochondria in sucrose gradients. While liver mitochondria banded sharply at a density of 1.187 g/ml, as evidenced by marker enzyme activity and protein assay, hepatoma mitochondria were heterogeneous, banding over a density range of 1.144 to 1.161 g/ml. The addition of 1% bovine serum albumin to the isolation medium increased the density of hepatoma mitochondria in sucrose gradients to 1.171 g/ml and resulted in sharp, homogeneous bands. This increase in density is due in part to protein binding, as evidenced by bovine serum albumin-14C-binding experiments. In addition, hepatoma mitochondria were shown to be heterogeneous by their separation into three density classes on discontinuous bovine serum albumin gradients.

Highly malignant tumor cells have been shown to be well coupled or to possess large numbers of gap junctions. These include Novikoff hepatoma cells. Moreover, other studies have analyzed gap junctions during tumor progression in skin tumors and in hepatocarcinoma and have found that, while gap junctions or coupling may disappear during the transformation of normal cells by cancer-causing agents (such as tumor promoters), their loss seems to follow, rather than lead the changes which transform these normal cells to tumor cells.



Seed genes

To collect the set of seed genes we started by filtering the “Curated gene-disease associations” dataset from DisGeNet in order to find all the genes associated with hepatoma, making sure they were all human genes.

Subsequently, we used the REST API of HGNC to fetch the status (approved or not) of each seed gene. All the 110 genes collected from DisGeNet resulted approved on HGNC, so we parsed the Uniprot dataset and collected the information requested in 1.1.b. We found that only 80 of the 110 seed genes resulted officially reviewed on Uniprot.

All the data collected has been stored in a *.tsv* file.

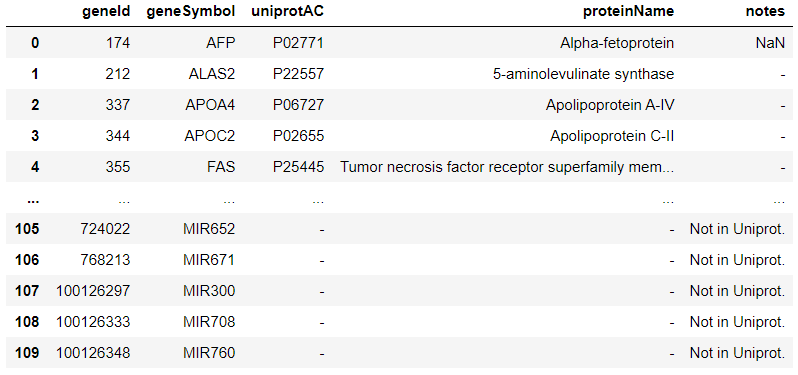


Table 1. Summary table of seed genes without protein description.

Summary on interaction data

To collect the interaction data, we started by downloading the full Biogrid dataset and, after that, we wrote a python script to parse the data and extract the interactions. The parsing process was made up of the following steps:

1. Filtering all the interactions which involved only human genes (ID 9606).
2. Filtering the interactions which involved at least one seed gene.
3. Extracting the list of non-seed genes which interacted with seed genes.
4. Collecting all the interactions between the non-seed genes previously extracted.
5. Saving all the interactions (after removing duplicates, if present) in a *tsv* table.

Here are some summary statistics regarding the data collected at this point:

* No. of Disgenet seed genes: 110
* No. of seed genes found in Biogrid: 80
* Total no. of interacting genes: 6319
* Total no. of interactions: 243222

# Interactomes data

The final step into the interactions collection process was to arrange the interactions into two different tables, the “seed genes interactome” and the “disease interactome”.

The first one contains the interactions just between seed genes, while the second one contains all the interactions which include at least one seed gene.

Here are some summary statistics about the interactomes:

* No. of interactions in the seed genes interactome: 139
* No. of interactions in the disease interactome: 13217

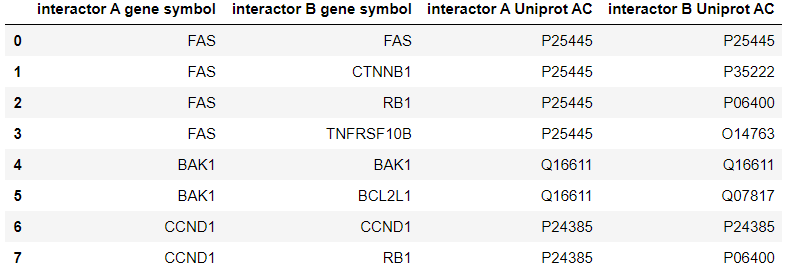


Table 2. First 8 interaction in the seed genes interactome.

# 

Table 3. First 8 interaction in the disease interactome.

# Enrichment analysis

To carry out the enrichment analysis we took advantage of the REST API offered by Enrichr.

Without going too much into the details of the code, what we did was: extracting the set of all the gene symbols present in the disease interactome and then fetching Enrichr to get the charts related to the gene set libraries specified in the homework. After that, we parsed the charts and kept just the first 10 result for each one and we arranged the data into *tsv* tables.

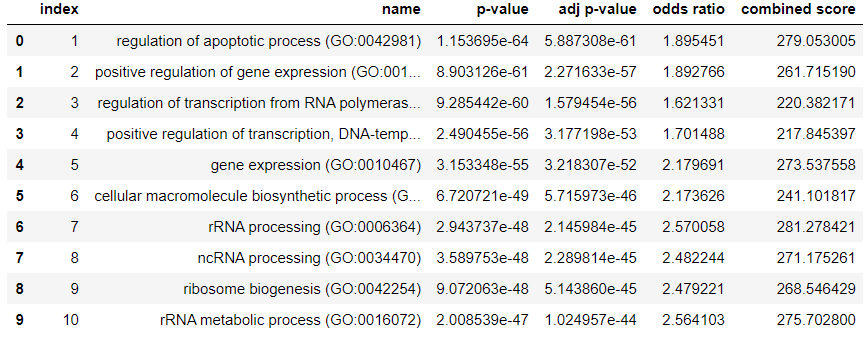


Table 4. GO Biological Process 2018 gene set library.

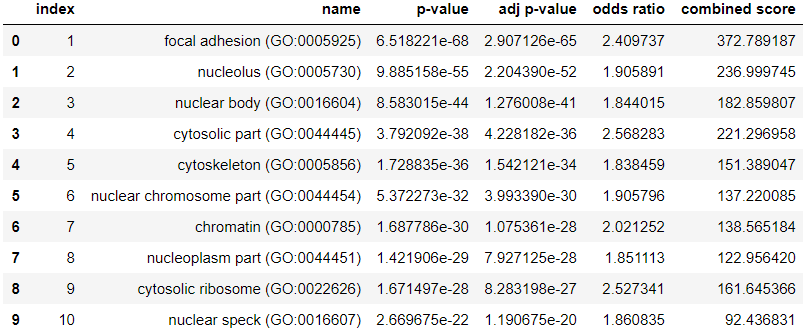


Table 5. GO Cellular Component 2018 gene set library.

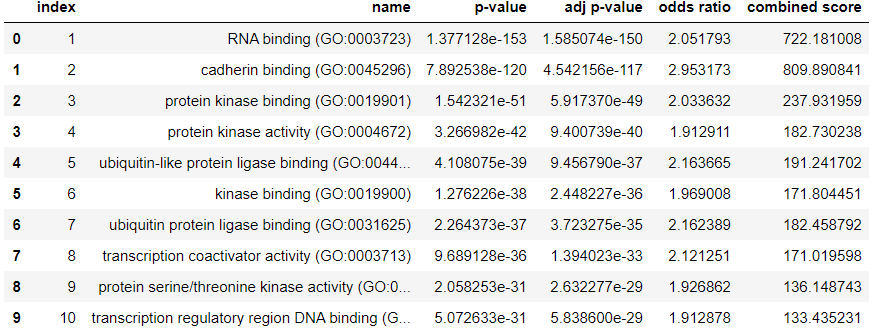


Table 6. GO Molecular Function 2018 gene set library.

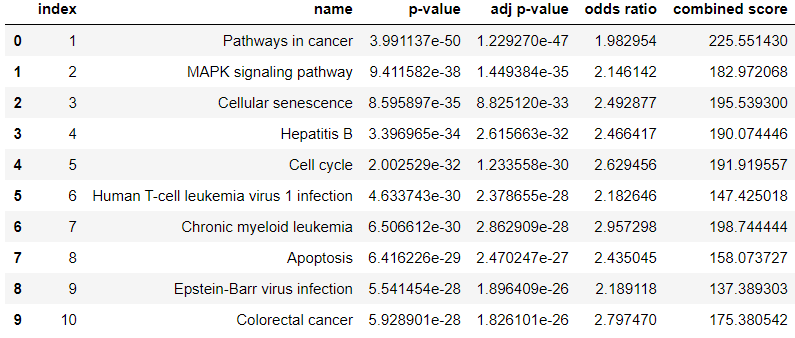
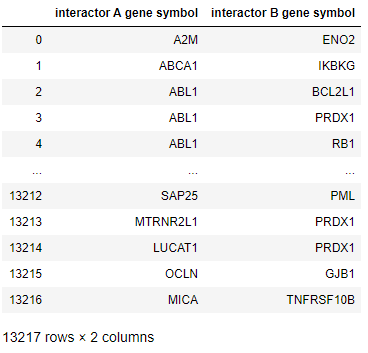
****

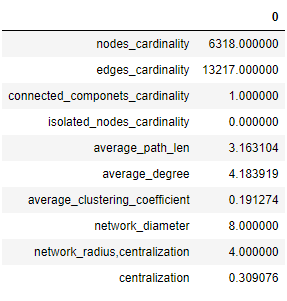
Table 7. KEGG 2019 Human gene set library.

**Table with the first 20 highest ranking genes for betweenness (include in the table**

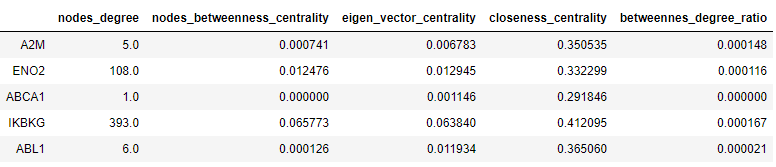
**also all other calculated centrality measures as from 1.2b) for the LCC**

We create the graph using a dataframe with the following informations.



The nodes cardinality is 6318 and the edges cardinality is 13217 (all the rows in the dataframe). No node is isolates, in fact the isolated nodes cardinality is 0 and the connected componets cardinality is 1.   
Since the whole graph is connected, the LCC is the graph itself.   
The other measures are in the table below.  


And here an example:



**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*)**

In this disease, using MCL we found 62 clusters (using the default hyperparameters: expansion=2, inflation=2, loop\_value=1, iterations=100, pruning\_threshold=0.001, pruning\_frequency=1, convergence\_check\_frequency=1), and 53 of them are longer than 10.

In the following table the final results of the method:

AGGIUNGERE TABELLA

# Notes and comments

References (if any, this is the format to be used)

Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design Patterens Applied. Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.