Data Science 10593052 Bioinformatics & Network Medicine 2020-2021

Date: DD Month YYYY

Network Medicine Project report

|  |
| --- |
| Network Medicine@Data Science A.Y. 2020-2021  Wight Gain – Network Medicine project  Alessandro Taglieri1, Yao Appeti1 and Davide Zingaro1  1Group no. 8 Abstract The goal of the assignment was to perform an analysis of the seed genes collected from DisGeNET dataset C0043094 – Wight Gain and collect interaction data from Biogrid Human. Afterward, we have built the interactome networks in two different cases: the first one when we consider seed genes only and the second done when we consider the database mentioned before (Biogrid Human). Enrichment analysis was performed to determine overrepresented GO categories and pathways. Different metrics for seed genes and interactome network were calculated in order to determine the general characteristics of the network. MCL algorithms were used to identify modules, on which hypergeometric test were performed to identify putative disease modules. DIAMOnD tool was used in order to compute the putative disease protein list. |

Weight Gain

Weight gain is an increase in body weight. This can involve an increase in muscle mas, fat deposits, excess fluids such as water or other factors. Weight gain can be symptom of a serious medical condition. Weight gain occurs when more energy (as calories from food and beverage consumption is gained than the energy expended by life activities, including normal physiological processes and physical exercise. If enough weight is gained due to increased body fat deposits, one may become overweight or obese, generally defined as having more body fat (adipose tissue) than is considered good for health. The Body Mass Index(BMI) measures body weight in proportion to height, and defines optimal, insufficient, and excessive weight based on the ratio.

Seed genes

To get the information about our seed genes, we downloaded the “Wight Gain Curated gene-disease associations data” from DisGeNET Databse. This database is a discovery platform containing one of the largest publicly available collecctions of genes and variants associated to human diseases. At first, we obtained 102 results. Moreover, we checked if the gene symbols are updated and approved by HGNC and UniProt websites. Finally, we stored the data gathered in a table with 102 rows an 5 columns that are the following:

* Official Gene symbols: approved and official gene symbols;
* Uniprot AC: Uniprot alphanumeric ‘accession numnber’;
* Protein name: approved protein name taken from HGNC database (not aliases);
* Entrez Gene ID: NCBI unique identifier of the gene, also taken from HGNC database;
* Brief Description: very short description about the protein functions, taken from UniProt website

Table 1. Top-10 rows of the Seed Genes Table (protein description omitted)

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene symbol** | **Uniprot AC** | **Prrotein name** | **Entrez Gene ID** |
| ABCG1 | P45844 | ATP binding cassette subfamily G member 1 | 9619 |
| ACADM | P11310 | acyl-CoA dehydrogenase medium chain | 34 |
| ACE | P12821 | angiotensin I converting enzyme | 1636 |
| ADIPOQ | Q15848 | adiponectin, C1Q and collagen domain containing | 9370 |
| AHR | P35869 | aryl hydrocarbon receptor | 196 |
| AKR1C2 | P52895 | aldo-keto reductase family 1 member C2 | 1646 |
| ANXA2 | P07355 | annexin A2 | 302 |
| ANXA5 | P08758 | annexin A5 | 308 |
| APBB2 | Q92870 | amyloid beta precursor protein binding family B member 2 | 323 |
| APP | P05067 | amyloid beta precursor protein | 351 |

Summary on interaction data

Once we generated all the informations about seed genes involved in our disease, we collected all binary interactions from a PPI sources: Biogrid Human. It is the Biological General Repository for Interaction Datasets, version 4.2.191.

Table 2. Summary Table of Interaction Data

|  |  |
| --- | --- |
|  | Biogrid |
| Number of seed genes collected in DisGenet | 102 |
| Number of seed genes found in Biogrid | 101 |
| Number of interacting proteins | 18910 |
| Number of interactions | 630323 |

# Interactomes data

In this section, we had to build and store two different interactome tables:

* Seed genes interactome (sgi): interactions that involeve seed genes only, from Biogrid DB;
* Disease interactome (di): all proteins interacting with at least one seed gene confimed by Biogrid DB.

We store the data using the same format. All interactome tables are characterized by four columns: interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC and interactor B Uniprot AC. In order to obtain them we’ve pre-processed the Biogrid dataset with Pandas library in Python.

# Enrichment analysis

In this section, we performed an enrichment analysis by Enrichr web service. This method is useful to identify classes of genes or proteins that may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes.

This analysis is performed by using four Gene Ontology classes and also using a pathways databases:

* GO Biological Process;
* GO Molecular Function;
* GO Cellular Component;
* KEGG 2019 Human (pathways databases).

In this step we had to perform our enrichment analysis on disease interactome, that we have performed before.

Hence, starting from disease interactome table, we extracted the list of the unique genes involved in this dataset. After that we uploaded this list of genes on Enrichr website;in this way we obtained five different charts in total ( four charts about GO categories and one for KEGG).

Since we are interested in overrepresented GO categories and overrepresented pathways, we limited our analysis to the firdt 10 results obtained for each main category. The following tables represent these data given from Enrichr website.

Table 3. GO Biological Process – Disease interactome genes

|  |  |
| --- | --- |
|  | **GO Biological Process** |
| 1 | Positive regulation of gene expression |
| 2 | Positive regulation of transcription, DNA-templated |
| 3 | Regulation of transcription from RNA polymerase II promoter |
| 4 | Transcription from RNA polymerase II promoter |
| 5 | Regulation of transcription, DNA-templated |
| 6 | Regulation of apoptotic process |
| 7 | mRNA processing |
| 8 | Positive regulation of nucleic acid-templated transcription |
| 9 | Positive regulation of transcription from RNA polymerase II promoter |
| 10 | mRNA splicing, via spliceosome |

Table 4. GO Molecular Function – Disease interactome genes

|  |  |
| --- | --- |
|  | **GO Molecular Function** |
| 1 | RNA binding |
| 2 | Transcription coactivator activity |
| 3 | Kinase binding |
| 4 | Protein kinase binding |
| 5 | Cadherin binding |
| 6 | Protein kinase activity |
| 7 | Transcription regulatory region DNA binding |
| 8 | DNA binding |
| 9 | Protein serine/threonine kinase activity |
| 10 | Ubiquitin-like protein ligase binding |

Table 5. GO Cellular Component – Disease interactome genes

|  |  |
| --- | --- |
|  | **GO Cellular Component** |
| 1 | Nuclear body |
| 2 | Focal adhesion |
| 3 | Nuclear chromosome part |
| 4 | RNA polymerase II transcription factor complex |
| 5 | Nucleoplasm part |
| 6 | Chromatin |
| 7 | Nuclear speck |
| 8 | nucleolus |
| 9 | Nuclear chromatin |
| 10 | cytoskeleton |

Table 6. KEGG Pathways – Disease interactome genes

|  |  |
| --- | --- |
|  | **KEGG Pathways** |
| 1 | Pathways in cancer |
| 2 | Cell cycle |
| 3 | Viral carcinogenesis |
| 4 | Human T-cell leukemia virus 1 infection |
| 5 | Cellular senescence |
| 6 | Hepatitis B |
| 7 | Epstein-Barr virus infection |
| 8 | Endocytosis |
| 9 | Human immunodeficiency 1 infection |
| 10 | Apoptosis |

# Network measures

Figure of interactome network

Table 7. Global measures of the disease interactome LCC

|  |  |
| --- | --- |
| **Measures** | **Interactome network** |
| **Number of nodes** | 4391 |
| **Number of links** | 9619 |
| **Average path length** | 5.2675 |
| **Average degree** | 18.7061 |
| **Average clustering coefficient** | 0.0987 |
| **Network diameter** | 15 |
| **Network radius** | 8 |
| **Centralization** | 0.02839 |

Table 7. First 20 genes with the higher Betweenness centrality (LCC)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Node degree** | **Betweenness centrality** | **Eigenvector centrality** | **Closeness centrality** | **Betweenness / Node degree** |
| **EWSR1** | 100 | 0.075241 | 0.013749 | 0.278943 | 0.000752 |
| **CCDC85B** | 129 | 0.075241 | 0.040199 | 0.281374 | 0.000566 |
| **AR** | 71 | 0.066159 | 0.021188 | 0.282606 | 0.000932 |
| **BRCA1** | 95 | 0.065715 | 0.073733 | 0.279475 | 0.000692 |
| **SFN** | 102 | 0.060212 | 0.009421 | 0.271574 | 0.000590 |
| **TRAF2** | 87 | 0.049687 | 0.020990 | 0.273367 | 0.000571 |
| **MDFI** | 98 | 0.035781 | 0.016117 | 0.259594 | 0.000365 |
| **MAGEA11** | 61 | 0.035190 | 0.030719 | 0.271625 | 0.000577 |
| **FXR2** | 52 | 0.033764 | 0.024577 | 0.273095 | 0.000649 |
| **CTNNB1** | 45 | 0.030914 | 0.005073 | 0.263664 | 0.000687 |
| **TP53** | 57 | 0.030907 | 0.019865 | 0.268617 | 0.000542 |
| **PLSCR1** | 66 | 0.030335 | 0.020364 | 0.266367 | 0.000460 |
| **MYC** | 45 | 0.029432 | 0.010809 | 0.265819 | 0.000654 |
| **VHL** | 53 | 0.028636 | 0.002145 | 0.251850 | 0.000540 |
| **RNPS1** | 58 | 0.025673 | 0.003891 | 0.256425 | 0.000443 |
| **HDAC1** | 60 | 0.024104 | 0.023803 | 0.263759 | 0.000402 |
| **UBE2I** | 42 | 0.023160 | 0.003558 | 0.247533 | 0.000551 |
| **KRTAP4-12** | 75 | 0.020981 | 0.017257 | 0.256276 | 0.000280 |
| **LNX1** | 43 | 0.020624 | 0.016834 | 0.268485 | 0.000480 |
| **YWHAQ** | 55 | 0.01946 | 0.003126 | 0.248909 | 0.000354 |

**TO CONTINUE..**

Sample numbered list, if necessary.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

1. 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.

* Sample bullet list, if necessary.
* 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. 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.

**Table 1.**Sample table. This should be the table format, add/remove columns and rows according to the data to be shown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| c | Predicted cost | Timing | Predicted speed | Speed |
| 1 | S219.20(100%) | 68m43s | 1.00 | 1.00 |
| 2 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 4 | 219.20(100%) | 68m43s | 1.00 | 1.00 |
| 10 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 20 | 219.20(100%) | 68m43s | 1.00 | 9.5 |

This is table foot note sample text This is table foot note sample text This is table foot note sample text

# 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.