Data Science 10593052 Bioinformatics & Network Medicine 2020-2021
Date: 20 December 2020
Network Medicine Project I report

*Network Medicine@Data Science A.Y. 2020-2021*

# Wight Gain – Network Medicine project

Alessandro Taglieri[1], Yao Appeti[1] and Davide Zingaro[1]

[1]Group no. 8

## Abstract

The goal of the assignment was to perform an analysis of the seed genes collected from DisGeNET dataset C0043094 – Weight 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 mass, 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 “Weight Gain Curated gene-disease associations data” from DisGeNET Databse[1]. This database is a discovery platform containing one of the largest publicly available collections 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[2] and UniProt[3] websites. Finally, we stored the data gathered in a table with 102 rows and 5 columns that are the following:

- Official Gene symbols: approved and official gene symbols;
- Uniprot AC: Uniprot alphanumeric 'accession number';
- 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 web-site

**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 information 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 | 10 |
| Number of interacting proteins | 18909 |
| 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 involves seed genes only, from Biogrid DB[4];
- Disease interactome (di): all proteins interacting with at least one seed gene confirmed 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[5]. 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 first 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

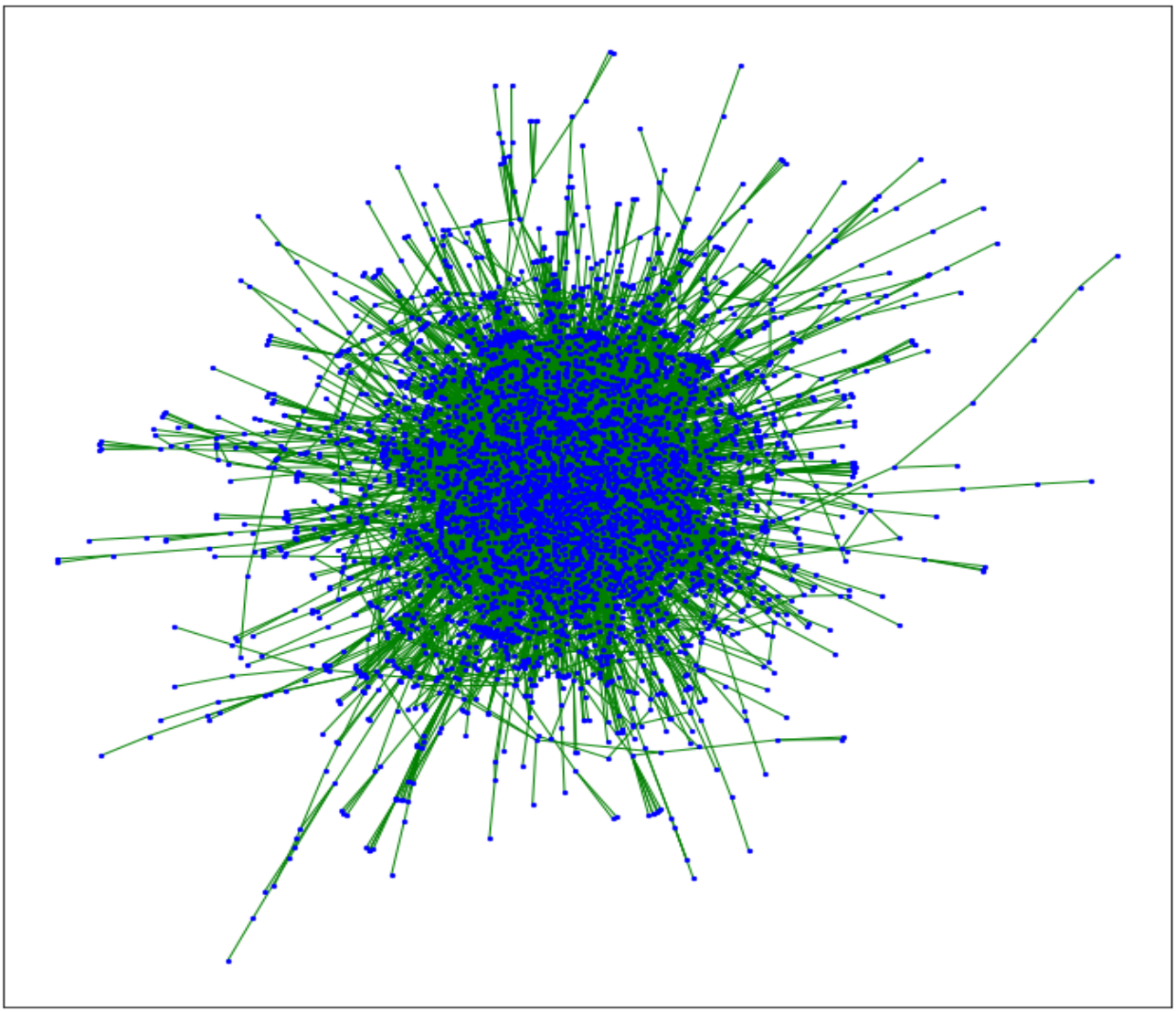
| | **KEGG Pathways** |
|---|---|
| 1 | Pathways in cancer |
| 2 | Cell cycle |
| 3 | Viral carcinogenesis |
| 4 | Human T-cell leukemia virus 1 infection |
| 5 | Cellular senescence |

| | |
|---|---|
| 6 | Hepatitis B |
| 7 | Epstein-Barr virus infection |
| 8 | Endocytosis |
| 9 | Human immunodeficiency 1 infection |
| 10 | Apoptosis |

**Network measures**

Starting from the data that we build in the first part of the project, we compute the main network measures. First we have built a graph from disease interactome. From this graph we noticed that it was composed by a several connected components and we take only the LCC (large connected component) and we start to work on it. The following figure represents the graph that we draw with networkx[6] (python library).

**Figure 1.** Interactome LCC plot



From this graph we computed the following measures:

- Number of nodes;
- Number of links;
- Average path length;
- Average degree;
- Average clustering coefficient;
- Network diameter;
- Network radius;
- Network centralization.

The Table 7 shows all data about these measures.

**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 |

In the following table it's represented a list of first 20 genes in LCC ranked by their betweenness centrality.

**Table 8.** 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 |

### Putative Disease Module Detection

Different clustering techniques can be used to identify modules in a network. We have chosen to perform MCL algorithm to determine modules in interactome LCC network. On the identified modules, which had more than 9 nodes, we performed hypergeometric test and determinated the putative disease modules. The results are shown in the following table and we mark the only one putative disease module that has p-value less than 0.05.

**Table 9.** Summary table of the module found from MCL clustering

| Module ID | No. of seed genes in the module | Total no. of genes in the module | Ratio no. of sed genes/total genes in the module | P-value |
|---|---|---|---|---|
| MOD_0 | 1 | 54 | 0.0187 | 0.19358 |
| MOD_1 | 1 | 50 | 0.02 | 0.18252 |
| MOD_2 | 1 | 26 | 0.038 | 0.10581 |
| MOD_3 | 1 | 14 | 0.071 | 0.06015 |
| MOD_4 | 1 | 51 | 0.0196 | 0.18533 |
| MOD_5 | 1 | 30 | 0.033 | 0.11989 |
| MOD_6 | 1 | 10 | 0.1 | 0.04375 |

From the previous table we can see that there is only one putative disease modules, i.e. a module with a p-value less than 0.05. This is the MOD_0. For all genes involved in this module we did an Enrichment analysis to get overrepresented GO categories and overrepresented pathway (both limit to ten).

**Table 10.** GO over-represented Analysis for Putative Disease Module (interactome LLC)

| GO Biological Process | Cellular Component | Molecular Function | KEGG Pathways |
|---|---|---|---|
| RNA splicing, via transesterification reactions with bulged adenosine as nucleophile | U2-type prespliceosome | RNA binding | Spliceosome |
| mRNA splicing, via spliceosome | Prespliceosome | Poly(U) RNA binding | Ferroptosis |
| mRNA processing | Nuclear spck | RNA polymerase binding | Cardiac muscle contraction |
| mRNA 3'-splice site recognition | Spliceosomal snRNP complex | Poly-pyrimidine tract binding | Hypertrophic cardiomyopathy |
| mRNA splice site selection | U2-type prespliceosome complex | RNA polymerase II transcription corepressor activity | Dilated cardiomyopathy |
| RNA processing | Spliceosomal complex | RNA polymerase II repressing transcription factor binding | Adrenergic signaling cardiomyocytes |
| Nucleic acid metabolic process | Nuclear body | Transcriptional repressor activity, RNA polymerase II transcription factor binding | |
| mRNA metabolic process | U2 snRNP | Repressing transcription factor binding | |
| RNA metabolic process | U1 snRNP | PRNA polymerase II transcription cofactor activity | |
| Negative regulation of endoplasmic reticulum unfolded protein response | Contractile actin filament bundle | Single-stranded DNA binding | |

**Putative Disease Proteins Detection (DIAMOnD tool)**

A tool named DIAMonD[7] was used to perform putative modules detection. This tool allow to use DIAMOnD algorithm that is a DIseAse MOdulee Detection (DIAMOnD) Algorithm based on a systematic analysis of connectivity patterns of disease proteins in the Human Interactome. It was originally implemented in python 2 and with some changes this algorithm is converted to python 3. It takes as input several parameters:

- Path of txt file that contains the seed genes list involved in our disease;
- Path of txt file containing protein-protein interaction network. In our case these informations comes from BioGrid interactome. This file contains every interaction for every line of the txt file;
- Iteration number, in our case it's 200;
- Alpha (seed weight). It is set to default value 1;
- Path where file containing results will be stored.

In order to run the tool, we have prepared the two input files using the 'gene symbol' column from DisGeNet dataset for the first txt file and using the columns named 'Official Symbol Interactor A' and 'Official Symbol Interactor A' (splitted in gene pairs) from the result Biogrid interactome for the second txt file. Then we run the following command from the command line:

Python3 ./DIAMOnD.py ppi.txt seed_genes.txt 200

As a result, we obtain a text file containing a list of putative disease proteins. The first 30 elements of the result are shown in the following table

**Table 11.** First 30 elements of the result from DIAMOnD algorithm

| rank | DIAMOnD node |
|---|---|
| 1 | MARK4 |
| 2 | PPM1B |
| 3 | MLNR |
| 4 | EDNRB |
| 5 | MAP2 |
| 6 | TTF1 |
| 7 | PPIL3 |
| 8 | CEACAM1 |
| 9 | ITGB5 |
| 10 | APBA2 |
| 11 | CLSTN1 |
| 12 | DGKZ |
| 13 | ANXA1 |
| 14 | MPDZ |
| 15 | SCP2 |
| 16 | AQP3 |
| 17 | PLD2 |
| 18 | PLCG1 |
| 19 | EEF1A1 |
| 20 | PPP1R12C |
| 21 | SELE |
| 22 | SERPING1 |
| 23 | SELP |
| 24 | SNX17 |
| 25 | IKBKB |
| 26 | LTA4H |
| 27 | RASD2 |
| 28 | MYH9 |
| 29 | PLCD4 |
| 30 | VIL1 |

From the result obtained with DIAMOnD algorithm we can do and Enrichment analysis. In this way we can:

- Fin overrepresented GO categories (limit to first ten);
- Find overrepresented pathways (limit to first ten).

We used all result (200 node) that we obtain for the previous algorithm. Then we show the result obtained from Enrichr in different table ranked by the p-value.

**Table 12.** Top ten of overrepresented GO categories

| GO Biological Process | Cellular Component | Molecular Function |
|---|---|---|
| transmembrane receptor protein tyrosine kinase signaling pathway | Focal adhesion | Cadherin binding |
| ERBB signaling pathway | Catenin complex | Phosphotyrosine resiude binding |
| Enzyme linked receptor protein signaling pathway | Actin cytoskeleton | Protein tyrosine kinase activity |
| Epidermal growth factor receptor signaling pathway | Actomyosin | Protein phosphorylated amino acid binding |
| Fc-gamma receptor signaling pathway involved in phagocytosis | Cytoskeleton | Phosphatidylinositol 3-kinase activity |
| Fc-gamma receptor signaling pathway | Contractile actin filament bundle | Phosphatidylinositol-4,5-bisphosphate 3-kinase activity |
| Fc receptor mediated stimulatory signaling pathway | Stress fiber | Phosphatidylinositol bisphosphate kinase activity |
| Peptidyl-tyrosine phosphorylation | Cortical actin cytoskeleton | Non-membrane spanning protein tyrosine kinase activity |
| Adherens junction organization | Membrane raft | Protein kinase binding |
| Peptidyl-tyrosine autophosphorylation | Cortical cytoskeleton | Ephrin receptor |

**Table 13.** Top ten of overrepresented pathways

| Term | P-value | Adjusted P-Value |
|---|---|---|
| Regulation of actin cytoskeleton | 8.182e-26 | 2.520e-23 |
| Focal adhesion | 3.485e-24 | 5.366e-22 |
| Bacterial invasion of epithelial cells | 5.547e-22 | 4.271e-20 |
| ErbB signaling pathway | 1.002e-20 | 6.172e-19 |
| Adherens junction | 1.187e-20 | 6.094e-19 |
| Chronic myeloid leukemia | 1.144e-18 | 5.034e-17 |
| Endometrial cancer | 1.199e-17 | 4.104e-16 |
| Gastric cancer | 3.904e-17 | 1.202e-15 |
| Leukocyte transendothlial | 1.176e-15 | 3.293e-14 |
| Non-small cell lung cancer | 8.676e-14 | 2.055e-12 |

**Notes and comments:**

All the code and files are stored in the following github repository:

https://github.com/AlessandroTaglieri/Project_Bioinformatics

In dataset folder you can find all files that we use to do this project and all output files. It's organized in several sub-folder, one for every module. One of input files, about biogrid data, is not present in the github repository (for memory storage problem).
As it's written in the code, you can find this file at the following drive folder:
https://drive.google.com/drive/folders/1SW9HTjI_RtX_eigLe5OozuN279C_A7EB?usp=sharing
It's necessary to put this file in dataset/1.2/ .

**References**

1- https://www.disgenet.org
2- https://www.genenames.org
3- https://www.uniprot.org
4- https://thebiogrid.org
5- https://maayanlab.cloud/Enrichr/
6- https://networkx.org
7- https://github.com/barabasilab/DIAMOnD