Bioinformatics 2019-2020
Date: 15 January 2020
Project report

Bioinformatics@Data Science A.Y. 2019-2020

Precancerous Conditions - Network Medicine project

Omirbanu NURASSILOVA[1], Francesca PORCU[1] and Dusica STEPIC[1]

[1]Group no. 8

Abstract

The goal of the assignment was to perform an analysis of the seed genes collected from DisGeNET dataset *C0032927 - Precancerous Conditions* and collect interaction data from Biogrid and Integrated Interactions Database. Afterward, we have built the interactome networks in three different cases. The first one when we consider seed genes only, the second one when we consider the union of the above-mentioned databases, and in the third one when we consider the intersection of databases. Enrichment analysis was performed to determine overrepresented GO categories and pathways. Different metrics for Union and Intersection networks were calculated in order to determine the general characteristics of the network. MCL and Louvain algorithms were used to identify modules, on which hypergeometric tests were performed to identify putative disease modules. DIAMOnD tool was used in order to compute the putative disease protein list.

Precancerous conditions

Precancerous (premalignant) is a term used to describe a condition or a lesion that may (or is likely to) become cancer. Pathologically, precancerous lesions can range from benign neoplasias, which are tumors which do not invade neighboring normal tissues or spread to distant organs, to dysplasia, which involves collections of abnormal cells which in some cases have an increased risk of progressing to anaplasia and invasive cancer. Premalignant lesions are morphologically atypical tissue which appear abnormal when viewed under the microscope, and which are more likely to progress to cancer than normal tissue. Precancerous conditions and lesions affect a variety of organ systems, including the skin, oral cavity, stomach, colon, and hematological system. Some authorities also refer to hereditary genetic conditions which predispose to developing cancer. Individuals with these conditions have a much higher risk of developing cancer in certain organs.

Seed genes

To get the information about our seed genes, we downloaded the "Precancerous Conditions Curated gene-disease associations data" from DisGeNET DB. DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases. At first, we obtained 107 results. Then, we removed 21 microRNAs involved in the disease because they are not protein coding molecules. Moreover, we checked if the gene symbols are updated and approved on HGNC and UniProt websites. Finally, we stored the data gathered in a table with 86 rows and the following 5 columns (Table 1):

- Gene symbol: approved official gene symbols;
- Uniprot AC: Uniprot alphanumeric 'accession number';
- Protein name: approved protein name (not aliases), taken from HGNC website;
- Entrez Gene ID: NCBI unique identifier of a gene, taken from HGNC website;
- Brief description: description of the protein functions, taken from UniProt website.

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

Gene symbol	Uniprot AC	Protein name	Entrez Gene ID
PTGS2	P35354	prostaglandin-endoperoxide synthase 2	5743
CDKN2A	P42771	cyclin dependent kinase inhibitor 2A	1029
CCND1	P24385	cyclin D1	595
VEGFA	P15692	vascular endothelial growth factor A	7422
HRAS	P01112	HRas proto-oncogene, GTPase	3265
EGF	P01133	epidermal growth factor	1950
MGMT	P16455	O-6-methylguanine-DNA methyltransferase	4255
CDH1	P12830	cadherin 1	999
IGF2	P01344	insulin like growth factor 2	3481
GSTP1	P09211	glutathione S-transferase pi 1	2950

Summary on interaction data

Once we gathered all the information about our seed genes, we collected all binary protein interactions from two different PPI sources:

- Biogrid Human: the Biological General Repository for Interaction Datasets, version 3.5.180;
- IID: the Integrated Interactions Database, version 2018-11.

For each PPI source, we downloaded the entire Homo Sapiens dataset. We filtered the two different datasets using Pandas library in Python and we collected all binary protein interactions that involve at least one seed gene. We also included the non-seed/non-seed interactions, where for non-seed genes we meant those genes that were involved with seed genes in other interactions. Moreover, we considered only human-human (9606-9606) interactions.

For the two different PPI datasets, we performed the following operations to collect our interaction data. In particular:

- BioGRID: the first step was to consider only *physical interactions* in the "Experimental System Type" column; then, we removed all the possible interactions among homo sapiens and other organisms (e.g. mice, rabbits, etc.). Finally, we considered the interactions that involve at least one seed gene; and, then, we also included the non-seed/non-seed interactions (as explained above).
- IID: the first step was to consider only *experimental data* (exp/exp;ortho/exp;ortho;pred) in "evidence type" column. In the IID dataset, we didn't have to remove interactions among homo sapiens and other organisms because they were not present. Finally, we considered the interactions that involve at least one seed gene; and, then, we also included the non-seed/non-seed interactions (as explained above).

Lastly, to obtain the Summary Table of Interaction Data, we removed from both datasets (Biogrid and IID) all the duplicated interactions (e.g. A-B; A-B) and reversed duplicates (e.g. A-B; B-A).

Table 2. Summary Table of Interaction Data

	Biogrid	IID
n° of seed genes	81	82
n° of interacting proteins	4464	3887
n° of interactions	146083	106044

Interactomes data

In this section, we had to to build and store three different interactome tables (SGI, U and I):

- Seed genes interactome: interactions that involve seed genes only, from all DBs;
- Union interactome: all proteins interacting with at least one seed gene, from all DBs;
- Intersection interactome: all proteins interacting with at least one seed gene confirmed by both DBs.

The format used to store the data of all interactome tables is characterized by the presence of 5 columns: *interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC, interactor B Uniprot AC, database source* (not for I table). In order to obtain these three tables, we used the two datasets mentioned in the previous section and processed them 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.
Enrichment analysis is performed by using the Gene Ontology classes (GO Biological Process, GO Molecular Function, GO Cellular Component), and also using pathway databases (e.g. KEGG 2019 Human). In this step, we had to perform our enrichment analysis on:

- seed genes;
- union interactome genes.

Hence, starting from the SGI and U interactome tables, we extracted the list of the unique genes involved in each one of these two datasets and we uploaded, separately, the two lists of gene symbols on Enrichr. In this way we obtained 8 charts in total (referred to GO categories and 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 (respectively reported from Table 3 to Table 11).

Table 3. GO Biological Process - Seed genes

	Biological Process
1	regulation of cell proliferation
2	regulation of apoptotic process
3	negative regulation of programmed cell death
4	negative regulation of apoptotic process
5	cellular response to cytokine stimulus
6	cytokine-mediated signaling pathway
7	positive regulation of cellular process
8	response to reactive oxygen species
9	positive regulation of cell proliferation
10	positive regulation of macromolecule metabolic process

Table 4. GO Molecular Function - Seed genes

	Molecular Function
1	manganese ion binding
2	protein kinase activity
3	growth factor activity
4	protein kinase binding
5	glutathione transferase activity
6	cyclic nucleotide-dependent protein kinase activity
7	growth factor receptor binding
8	protein serine/threonine kinase activity
9	MAP kinase activity
10	protein serine/threonine/tyrosine kinase activity

Table 5. GO Cellular Component - Seed genes

	Cellular Component
1	secretory granule lumen
2	mitochondrial matrix
3	platelet alpha granule lumen
4	nuclear transcription factor complex
5	azurophil granule lumen
6	platelet alpha granule
7	ficolin-1-rich granule
8	ficolin-1-rich granule lumen
9	actomyosin
10	focal adhesion

Table 7. KEGG 2019 Human - Seed genes

	KEGG Pathways
1	Pathways in cancer
2	FoxO signaling pathway
3	Acute myeloid leukemia
4	AGE-RAGE signaling pathway in diabetic complications
5	Prolactin signaling pathway
6	ErbB signaling pathway
7	Bladder cancer
8	Pancreatic cancer
9	JAK-STAT signaling pathway
10	Human cytomegalovirus infection

Table 8. GO Biological Process - Union interactome genes

	Biological Process
1	regulation of apoptotic process
2	protein phosphorylation
3	cellular protein modification process
4	positive regulation of gene expression
5	phosphorylation
6	positive regulation of transcription, DNA-templated
7	regulation of transcription from RNA polymerase II promoter
8	negative regulation of apoptotic process
9	transmembrane receptor protein tyrosine kinase signaling pathway
10	negative regulation of programmed cell death

Table 9. GO Molecular Function - Union interactome genes

	Molecular Function
1	cadherin binding
2	RNA binding
3	protein kinase binding
4	protein kinase activity
5	kinase binding
6	protein serine/threonine kinase activity
7	ubiquitin-like protein ligase binding
8	ubiquitin protein ligase binding
9	purine ribonucleoside triphosphate binding
10	kinase activity

Table 10. GO Cellular Component - Union interactome genes

	Cellular Component
1	focal adhesion
2	nucleolus
3	nuclear body
4	nuclear chromosome part
5	cytoskeleton
6	cytosolic part
7	chromatin
8	nuclear speck
9	mitochondrion
10	nuclear chromatin

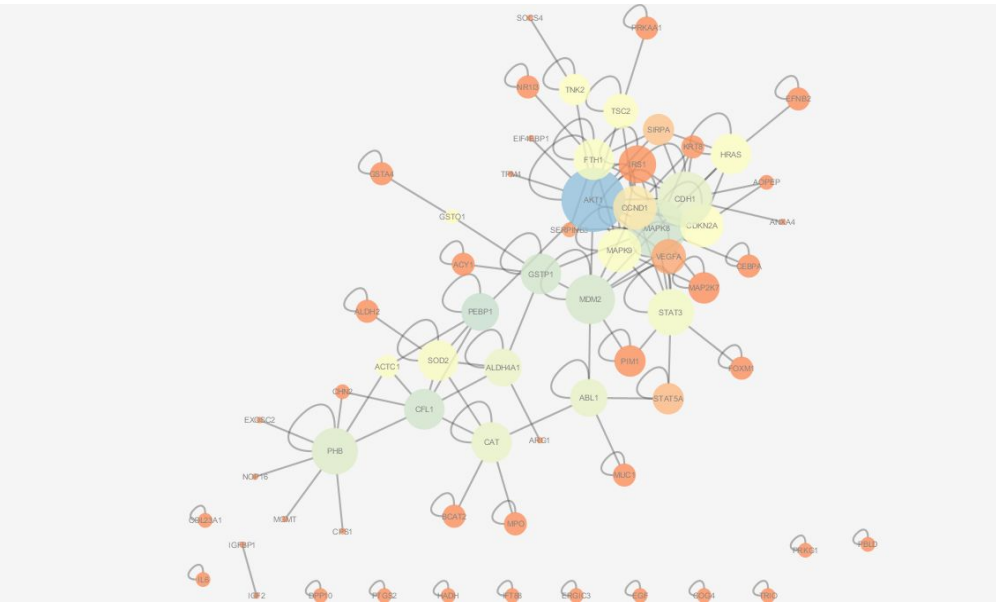
Table 11. KEGG 2019 Human - Union interactome genes

	KEGG Pathways
1	Pathways in cancer
2	MAPK signaling pathway
3	Proteoglycans in cancer
4	Hepatitis B
5	Cellular senescence
6	Neurotrophin signaling pathway
7	FoxO signaling pathway
8	Cell cycle
9	Human T-cell leukemia virus 1 infection
10	PI3K-Akt signaling pathway

Network measures

The main building blocks that were used for executing the following tasks are the seed genes interactome (SGI), the intersection (I), and the union (U) interactomes. They were built in the first part of the project, as described above.

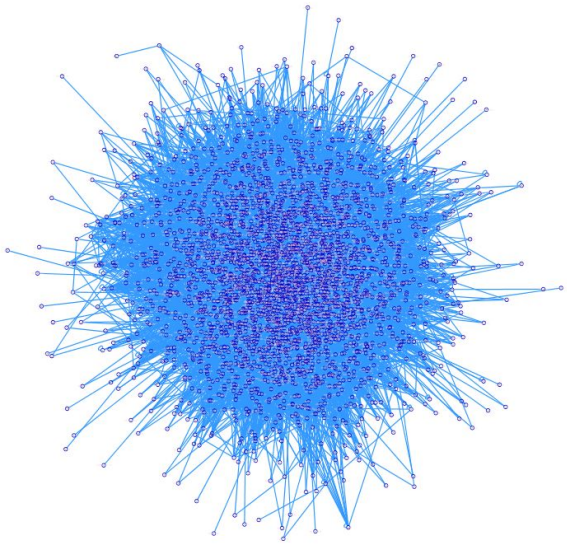
Figure 1. SGI network



Using Cytoscape, we visualized the SGI network, whereby the size of the node is determined by the degree of the node, and the color of the nodes is determined by the value of the betweenness centrality measure. The higher the betweenness centrality, the intense the color of the node is. Since the SGI network is the smallest, clearly it was the best choice to visualize in order to see better every node. As we can see from the image AKT1 is a gene with the highest degree and highest betweenness centrality in the SGI network. The phosphatidylinositol - 3 kinase (PI3K)–AKT pathway is one of the most commonly dysregulated pathways in all of cancer, with somatic mutations, copy number alterations, aberrant epigenetic regulation and increased expression in a number of cancers. (Mundi et al., 2016)

Figure 2 shows the intersection network, and it is also clear from the image that the graph is connected.

Figure 2. Intersection network



The main global network measures are shown in the table below. Moreover, the values in the table were double-checked if they match the output from the Cytoscape NetworkAnalyzer tool.

Table 12. Global network measures

Measures	SGI	U	I
No. of nodes	66	4975	3252
no. of links	135	172052	73318
No. of connected components	14	1	1
No. of isolated nodes	0	0	0
Average path length	2.671	2.42	2.531
Average degree	4.091	69.167	45.091
Average clustering coefficient	0.114	0.182	0.16
Network diameter	7	5	6
Network radius	1	3	4
Centralization	0.17	0.25	0.25

As we can see from the table above, The number of Largest Connected Components is 1 for U and I. This is due to the fact that both of those graphs are connected graphs. Since the LCC components are actually the whole U and I graphs, respectively, therefore, the global measures will be the same for LCC and the whole graphs.
Moreover, local centrality measures for U-LCC and I-LCC were also calculated. The tables below show the centrality measures for the top 20 genes with the highest value of the Betweenness centrality.

Table 13. First 20 genes with the highest Betweenness centrality (U-LCC)

Gene	Node degree	Betweenness centrality	Eigenvector centrality	Closeness centrality	Betweenness/Node degree
ESR2	1280	0.038381	0.088722	0.565935	3.00E-05
TRIM25	1115	0.034437	0.085559	0.555382	3.09E-05
APP	877	0.033625	0.052483	0.539421	3.83E-05
NTRK1	1286	0.029339	0.104017	0.563435	2.28E-05
MYC	1297	0.028788	0.114208	0.565806	2.22E-05
HRAS	649	0.024607	0.01925	0.507655	3.79E-05
ELAVL1	826	0.023993	0.06042	0.534954	2.90E-05
EGFR	884	0.019833	0.068312	0.541771	2.24E-05
CDH1	699	0.01895	0.039017	0.524518	2.71E-05
GRB2	702	0.017097	0.056959	0.530051	2.44E-05
PHB	936	0.015724	0.096152	0.536917	1.68E-05
MDM2	625	0.015292	0.048648	0.518557	2.45E-05
KRAS	640	0.014322	0.037711	0.518233	2.24E-05
HNRNPL	614	0.013984	0.048669	0.519695	2.28E-05
TP53	803	0.012904	0.076538	0.536685	1.61E-05
JUN	851	0.012784	0.087535	0.540065	1.50E-05
XPO1	704	0.011395	0.05825	0.522918	1.62E-05
RNF4	618	0.010069	0.048587	0.51326	1.63E-05
EGLN3	683	0.009877	0.048736	0.517155	1.45E-05
UBC	682	0.00974	0.065417	0.526294	1.43E-05

Table 14. First 20 genes with the highest Betweenness centrality (I-LCC)

Gene	Node degree	Betweenness centrality	Eigenvector centrality	Closeness centrality	Betweenness/Node degree
TRIM25	758	0.058332	0.118517	0.554684	7.70E-05
NTRK1	868	0.052108	0.145426	0.560904	6.00E-05
CDH1	606	0.051084	0.06187	0.526563	8.43E-05
APP	555	0.045811	0.067598	0.528876	8.25E-05
ELAVL1	541	0.037174	0.077257	0.526734	6.87E-05
EGFR	542	0.028131	0.086141	0.531904	5.19E-05
MDM2	446	0.02325	0.073762	0.515868	5.21E-05
XPO1	477	0.021039	0.076838	0.514643	4.41E-05
TP53	530	0.019869	0.105121	0.523932	3.75E-05
UBC	433	0.015993	0.084494	0.518335	3.69E-05
AKT1	313	0.015094	0.043308	0.495126	4.82E-05
BRCA1	532	0.014312	0.115893	0.521244	2.69E-05
STAT3	246	0.014286	0.027895	0.484862	5.81E-05
HSP90AA1	392	0.014201	0.080483	0.517675	3.62E-05
MCM2	458	0.012384	0.104829	0.513343	2.70E-05
CUL3	501	0.012138	0.115223	0.516114	2.42E-05
NXF1	365	0.010958	0.06157	0.502085	3.00E-05
ESR1	411	0.010126	0.091669	0.513748	2.46E-05
GRB2	295	0.01009	0.050495	0.49626	3.42E-05
CDK2	414	0.008982	0.09156	0.501465	2.17E-05

Figure 3. Betweenness (U-LCC)

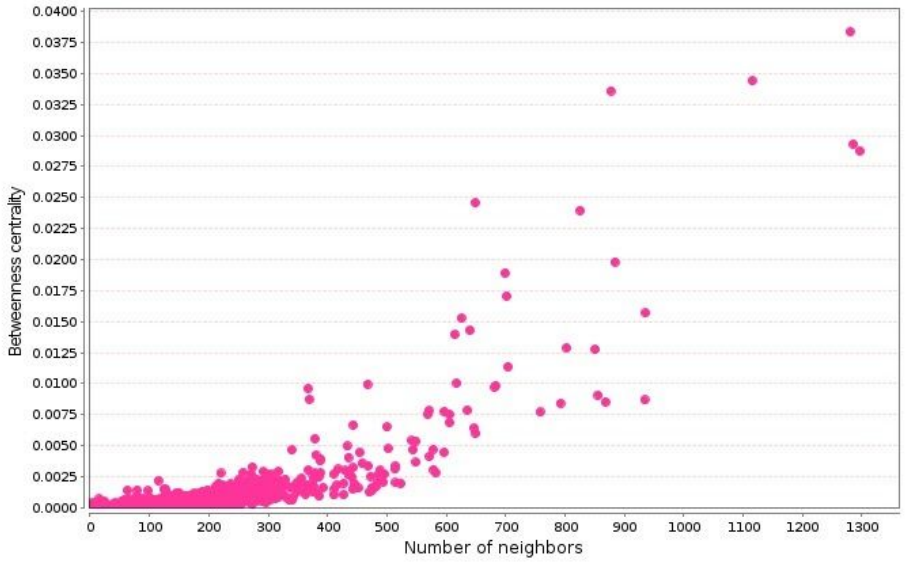
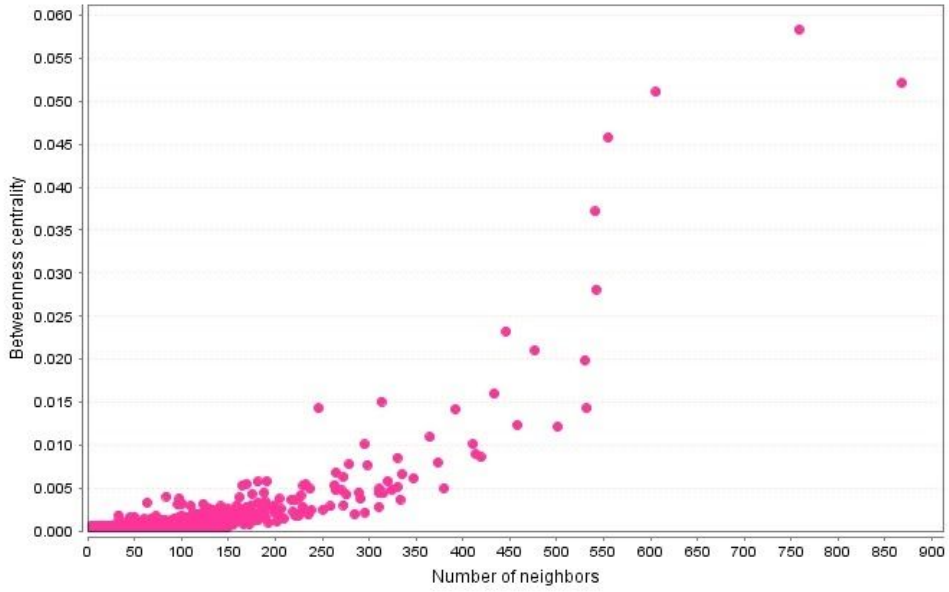


Figure 4. Betweenness (I-LCC)



The figures 3 and 4 show how with the increase of the number of neighbors the betweenness centrality is also increasing. Betweenness is a measure of the centrality of a node in a network, and is normally calculated as the fraction of shortest paths between node pairs that pass through the node of interest. Betweenness is, in some sense, a measure of the influence a node has over the spread of information through the network. (Newman, 2005)

Putative Disease Modules Detection

Different clustering techniques can be used to identify modules in a network. We have chosen to perform MCL and Louvain algorithms to determine modules in U-LCC and I-LCC networks. On the identified modules, which had more than 9 nodes, we performed hypergeometric tests and determined the putative disease modules, two for U-LCC and one for I-LCC.

The results are shown in the tables below.

Table 15. Louvain clustering and first 10 genes from the MLC clustering sorted by p-value (U-LCC)

Algorithm	Module ID	no. of seed genes	total no. of genes in each module	p-value
Louvain	6	8	156	**0.004**
Louvain	5	12	383	**0.022**
Louvain	7	1	25	0.341
Louvain	1	18	1022	0.418
Louvain	3	12	733	0.554
Louvain	2	12	859	0.777
Louvain	0	14	1069	0.87
Louvain	4	4	727	0.999
MCL	36	1	10	0.153
MCL	14	1	10	0.153
MCL	23	1	10	0.153
MCL	19	1	12	0.181

MCL	18	1	12	0.181
MCL	17	1	12	0.181
MCL	16	1	12	0.181
MCL	42	1	13	0.195
MCL	41	1	13	0.195
MCL	43	1	14	0.208

Table 16. Louvain clustering and first 10 genes from the MLC clustering sorted by p-value (I-LCC)

Algorithm	Module ID	no. of seed genes	total no. of genes in each module	p-value
Louvain	5	11	255	**0.045477**
Louvain	1	19	545	0.065791
Louvain	6	9	209	0.068537
Louvain	7	6	170	0.238452
Louvain	8	4	104	0.252087
Louvain	4	10	345	0.340303
Louvain	0	11	623	0.922307
Louvain	2	6	514	0.991915
Louvain	3	4	487	0.998826
MCL	9	1	10	0.220754
MCL	30	1	11	0.239983
MCL	6	1	14	0.294902
MCL	14	1	14	0.294902
MCL	24	4	113	0.301948
MCL	27	1	15	0.312322
MCL	4	1	15	0.312322
MCL	11	1	16	0.329318
MCL	26	1	17	0.345898
MCL	10	1	20	0.393249

Table 17. GO Over-Represented Analysis for the ***First*** Putative Disease Module (U-LCC)

Biological processes	Cellular Component	Molecular function	KEGG Pathways
secretory granule lumen	secretory granule lumen	insulin-like growth factor II binding	Fluid shear stress and atherosclerosis
platelet alpha granule lumen	platelet alpha granule lumen	serine-type peptidase activity	Complement and coagulation cascades
platelet alpha granule	platelet alpha granule	endopeptidase inhibitor activity	Renin secretion
endoplasmic reticulum lumen	endoplasmic reticulum lumen	insulin-like growth factor I binding	Ferroptosis
vacuolar lumen	vacuolar lumen	serine-type endopeptidase activity	HIF-1 signaling pathway
lysosomal lumen	lysosomal lumen	insulin-like growth factor binding	Phagosome
tertiary granule lumen	tertiary granule lumen	endopeptidase activity	Lysosome
azurophil granule lumen	azurophil granule lumen	protease binding	Proteoglycans in cancer
lysosome	lysosome	peptidase activity, acting on L-amino acid peptides	Salivary secretion
azurophil granule lumen	azurophil granule	serine-type endopeptidase inhibitor activity	Rap1 signaling pathway

Table 18. GO Over-Represented Analysis for the ***Second*** Putative Disease Module (U-LCC)

Biological processes	Cellular Component	Molecular function	KEGG Pathways
secretory granule lumen	secretory granule lumen	insulin-like growth factor II binding	Fluid shear stress and atherosclerosis
platelet alpha granule lumen	platelet alpha granule lumen	serine-type peptidase activity	Complement and coagulation cascades
platelet alpha granule	platelet alpha granule	endopeptidase inhibitor activity	Renin secretion
endoplasmic reticulum lumen	endoplasmic reticulum lumen	insulin-like growth factor I binding	Ferroptosis
vacuolar lumen	vacuolar lumen	serine-type endopeptidase activity	HIF-1 signaling pathway
lysosomal lumen	lysosomal lumen	insulin-like growth factor binding	Phagosome
tertiary granule lumen	tertiary granule lumen	endopeptidase activity	Lysosome
azurophil granule lumen	azurophil granule lumen	protease binding	Proteoglycans in cancer
lysosome	lysosome	peptidase activity, acting on L-amino acid peptides	Salivary secretion
azurophil granule	azurophil granule	serine-type endopeptidase inhibitor activity	Rap1 signaling pathway

Table 19. GO Over-Represented Analysis for Putative Disease Module (I-LCC)

Biological processes	Cellular Component	Molecular function	KEGG Pathways
hexose biosynthetic process	mitochondrial matrix	cadherin binding	Glycolysis / Gluconeogenesis
glucose metabolic process	cytoplasmic vesicle lumen	aldehyde dehydrogenase (NAD) activity	Pyruvate metabolism
glucose catabolic process to pyruvate	mitochondrion	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	Valine, leucine and isoleucine degradation
canonical glycolysis	ficolin-1-rich granule lumen	glutathione transferase activity	Cysteine and methionine metabolism
glycolytic process through glucose-6-phosphate	secretory granule lumen	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	Glutathione metabolism
gluconeogenesis	microbody	RNA binding	Propanoate metabolism
pyruvate metabolic process	peroxisome	protein homodimerization activity	Peroxisome
glutathione metabolic process	ficolin-1-rich granule	hydro-lyase activity	Alanine, aspartate and glutamate metabolism
glutathione derivative metabolic process	focal adhesion	metal ion binding	Drug metabolism
glutathione derivative biosynthetic process	peroxisomal part	transition metal ion binding	Tryptophan metabolism

Putative Disease Proteins Detection (DIAMOnD tool)

A tool with name DIAMOnD was used in order to perform putative modules detection. In this tool, the DIAMOnD algorithm was implemented. A DIseAse MOdule Detection (DIAMOnD) Algorithm was based on a systematic analysis of connectivity patterns of disease proteins in the Human Interactome. An algorithm was implemented in python 2 and it takes as input several parameters:

- seed genes list in txt format;
- txt file containing protein-protein interaction network in our case information that comes from BioGrid interactome;
- iteration number, in our case 200;
- alpha (seed weight), this is optional parameter, by default is 1;
- name of the file where results will be stored, this parameter is also optional.

In order to run the tool, we needed to use the command line and write all parameters consequently. For the first parameter, we used the ‘gene symbol’ column from the DisGeNet dataset. And as the second parameter, we used text file which contains columns called 'Official Symbol Interactor A' and 'Official Symbol Interactor B' from the BioGrid interactome file. As a result, we got a text file containing a list of putative disease proteins. The first 30 elements of the result are shown in the table below.

Table 20. First 30 elements of the result

rank	DIAMOnD_node
1	HSP90AA1
2	HSPA4
3	ESR1

4	HSPE1
5	SP1
6	HSPA9
7	SIRT1
8	JUN
9	AR
10	MAPK1
11	GSK3B
12	EP300
13	ATF2
14	RB1
15	CEBPB
16	FOXO1
17	SMAD3
18	JUND
19	FOS
20	ATF3
21	PPARG
22	SMARCA4
23	HDAC1
24	HIF1A
25	CREBBP
26	KAT5
27	MYOD1
28	CREB1
29	TP53
30	TP73

With the result obtained after running the tool we can go to the next step:

- finding overrepresented GO categories (limit to first ten)
- finding overrepresented pathways(limit to first ten)

In order to perform this tasks we used Enrichr. By using it we can perform both of the searches. We sorted obtained results according to adjusted p-values and took first 10 rows. They are shown on the tables below.

Table 21. Top-10 of overrepresented GO categories

Biological Process	Cellular Component	Molecular Function
regulation of transcription from RNA polymerase II promoter	nuclear chromatin	transcription regulatory region DNA binding
regulation of transcription DNA-templated	nuclear chromosome part	RNA polymerase II regulatory region sequence-specific DNA binding
positive regulation of transcription DNA-templated	chromatin	DNA binding
negative regulation of transcription DNA-templated	RNA polymerase II transcription factor complex	transcription coactivator activity

positive regulation of transcription from RNA polymerase II promoter	BAF-type complex	core promoter proximal region sequence-specific DNA binding
positive regulation of gene expression	SWI/SNF complex	transcription regulatory region sequence-specific DNA binding
negative regulation of transcription from RNA polymerase II promoter	histone methyltransferase complex	regulatory region DNA binding
positive regulation of nucleic acid-templated transcription	nuclear transcription factor complex	RNA polymerase II core promoter proximal region sequence-specific DNA binding
negative regulation of gene expression	npBAF complex	RNA polymerase II distal enhancer sequence-specific DNA binding
negative regulation of nucleic acid-templated transcription	CHD-type complex	enhancer sequence-specific DNA binding

Table 22. Top-10 overrepresented pathways

Term	P-value	Adjusted P-value
Thyroid hormone signaling pathway	4.02E-31	1.24E-28
Transcriptional misregulation in cancer	4.76E-29	7.33E-27
Pathways in cancer	7.14E-27	7.33E-25
Hepatocellular carcinoma	1.30E-23	1.00E-21
Human T-cell leukemia virus 1 infection	8.63E-22	5.32E-20
Viral carcinogenesis	2.46E-20	1.26E-18
Hepatitis B	8.54E-19	3.76E-17
Th17 cell differentiation	2.45E-17	9.45E-16
Kaposi sarcoma-associated herpesvirus infection	3.15E-15	1.08E-13
Chronic myeloid leukemia	2.52E-14	7.77E-13

Notes and comments

All the code and files are stored in the following github repository:
https://github.com/dusicastepic/bioinformatics-and-digital-epidemiology/tree/master/BI_HW2

References

Newman, M. E. (2005). A measure of betweenness centrality based on random walks. Social networks, 27(1), 39-54.
Mundi, P. S., Sachdev, J., McCourt, C., & Kalinsky, K. (2016). AKT in cancer: new molecular insights and advances in drug development. *British journal of clinical pharmacology*, *82*(4), 943–956. doi:10.1111/bcp.13021

https://cytoscape.org/
https://systems.crump.ucla.edu/hypergeometric/index.php
https://en.wikipedia.org/wiki/Precancerous_condition
https://github.com/barabasilab/DIAMOnD
http://amp.pharm.mssm.edu/Enrichr/
https://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/index.html