Network Medicine Analysis of Novikoff Hepatoma

Laurenti Laura, Mandica Paolo, Testa Lucia
Group no. 1

December 20th, 2020

Data Science 10593052 Bioinformatics Network Medicine 2020-2021

Astract

In this project we analyzed the Novikoff hepatoma desease,with the aim of carrying on a network medicine analysis. We started with the data collection,exploring the related information sources (DisGeNet datasets), collecting the the list of human genes of interest and finally scraping protein-protein interaction data. Proceeding with the data analysis we created the graph, we calculated the main network measures for the disease interactome and we applied the MCL algorithm to find out putative disease modules with hypergeometric test and DIAMOnD tool.

Novikoff Hepatoma

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.

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.
The list of seed genes collected can be found in Table (1).

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" (Table (6)) and the "disease interactome" (Table (7)). 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

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 ((Tables 2, 3, 4, 5) 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.

Notes

In this disease, using MCL we found 62 clusters with 53 of them longer than 10. We didn't find any putative disease module through this algorithm, for this reason it hasn't been possible providing an Enrichment analysis for it.

The code written and used for the analysis of the disease can be found here (github repository).

References

Michael T. White and Krishna K. Tewari (1973) Structural and Functional Changes in Novikoff Hepatoma Mitochondria.

Harris Busch (1974) The Molecular Biology of Cancer.

Table 1: Seed genes of the disease (without description)

geneId	geneSymbol	uniprotAC	proteinName	notes
174	AFP	P02771	Alpha-fetoprotein	-
212	ALAS2	P22557	5-aminolevulinate synthase	-
337	APOA4	P06727	Apolipoprotein A-IV	-
344	APOC2	P02655	Apolipoprotein C-II	-
355	FAS	P25445	Tumor necrosis factor receptor superfamily member 6	-
538	ATP7A	Q04656	Copper-transporting ATPase 1	-
540	ATP7B	P35670	Copper-transporting ATPase 2	-
578	BAK1	Q16611	Bcl-2 homologous antagonist/killer	-
595	CCND1	P24385	G1/S-specific cyclin-D1	-
596	BCL2	P10415	Apoptosis regulator Bcl-2	-
598	BCL2L1	Q07817	Bcl-2-like protein 1	-
967	CD63	P08962	CD63 antigen	-
999	CDH1	P12830	Cadherin-1	-
1317	SLC31A1	O15431	High affinity copper uptake protein 1	-
1356	CP	P00450	Ceruloplasmin	-
1499	CTNNB1	P35222	Catenin beta-1	-
1544	CYP1A2	P05177	Cytochrome P450 1A2	-
1571	CYP2E1	P05181	Cytochrome P450 2E1	-
2026	ENO2	P09104	Gamma-enolase	-
2161	F12	P00748	Coagulation factor XII	-
2305	FOXM1	Q08050	Forkhead box protein M1	-
2705	GJB1	P08034	Gap junction beta-1 protein	-
2752	GLUL	P15104	Glutamine synthetase	-
2922	GRP	P07492	Gastrin-releasing peptide	-
2950	GSTP1	P09211	Glutathione S-transferase P	-
3082	HGF	P14210	Hepatocyte growth factor	-
3265	HRAS	P01112	GTPase HRas	-
3569	IL6	P05231	Interleukin-6	-
3717	JAK2	O60674	Tyrosine-protein kinase JAK2	-
3791	KDR	P35968	Vascular endothelial growth factor receptor 2	-
3845	KRAS	P01116	GTPase KRas	-
4233	MET	P08581	Hepatocyte growth factor receptor	-
4240	MFGE8	Q08431	Lactadherin	-
4283	CXCL9	Q07325	C-X-C motif chemokine 9	-
4313	MMP2	P08253	72 kDa type IV collagenase	-
4609	MYC	P01106	Myc proto-oncogene protein	-
5052	PRDX1	Q06830	Peroxiredoxin-1	-
5371	PML	P29590	Protein PML	-
5471	PPAT	Q06203	Amidophosphoribosyltransferase	-
5594	MAPK1	P28482	Mitogen-activated protein kinase 1	-
5595	MAPK3	P27361	Mitogen-activated protein kinase 3	-
5925	RB1	P06400	Retinoblastoma-associated protein	-
6364	CCL20	P78556	C-C motif chemokine 20	-
6659	SOX4	Q06945	Transcription factor SOX-4	-
6696	SPP1	P10451	Osteopontin	-
6713	SQLE	Q14534	Squalene monooxygenase	-
6774	STAT3	P40763	Signal transducer and activator of transcription 3	-
7010	TEK	Q02763	Angiopoietin-1 receptor	-
7039	TGFA	P01135	Protransforming growth factor alpha [Cleaved into: Transforming growth factor alpha	-

geneId	geneSymbol	uniprotAC	proteinName	notes
7040	TGFB1	P01137	Transforming growth factor beta-1 proprotein [Cleaved into: Latency-associated peptide	-
7083	TK1	P04183	Thymidine kinase	-
7124	TNF	P01375	Tumor necrosis factor	-
7157	TP53	P04637	Cellular tumor antigen p53	-
8517	IKBKG	Q9Y6K9	NF-kappa-B essential modulator	-
8651	SOCS1	O15524	Suppressor of cytokine signaling 1	-
8795	TNFRSF10B	O14763	Tumor necrosis factor receptor superfamily member 10B	-
8848	TSC22D1	Q15714	TSC22 domain family protein 1	-
9104	RGN	Q15493	Regucalcin	-
9970	NR1I3	Q14994	Nuclear receptor subfamily 1 group I member 3	-
22800	RRAS2	P62070	Ras-related protein R-Ras2	-
23582	CCNDBP1	O95273	Cyclin-D1-binding protein 1	-
27113	BBC3	Q9BXH1	Bcl-2-binding component 3	-
219972	MPEG1	Q2M385	Macrophage-expressed gene 1 protein	-
283120	H19	-	-	Not in Uniprot.
406884	MIRLET7B	-	-	Not in Uniprot.
406886	MIRLET7D	-	-	Not in Uniprot.
406887	MIRLET7E	-	-	Not in Uniprot.
406891	MIRLET7I	-	-	Not in Uniprot.
406902	MIR10A	-	-	Not in Uniprot.
406910	MIR125A	-	-	Not in Uniprot.
406937	MIR145	-	-	Not in Uniprot.
406953	MIR18A	-	-	Not in Uniprot.
406957	MIR181C	-	-	Not in Uniprot.
406959	MIR183	-	-	Not in Uniprot.
406962	MIR186	-	-	Not in Uniprot.
406971	MIR195	-	-	Not in Uniprot.
406984	MIR200B	-	-	Not in Uniprot.
406989	MIR206	-	-	Not in Uniprot.
406994	MIR212	-	-	Not in Uniprot.
407006	MIR221	-	-	Not in Uniprot.
407017	MIR26B	-	-	Not in Uniprot.
407027	MIR301A	-	-	Not in Uniprot.
407029	MIR30A	-	-	Not in Uniprot.
407030	MIR30B	-	-	Not in Uniprot.
407033	MIR30D	-	-	Not in Uniprot.
407034	MIR30E	-	-	Not in Uniprot.
407040	MIR34A	-	-	Not in Uniprot.
407054	MIR98	-	-	Not in Uniprot.
407056	MIR99B	-	-	Not in Uniprot.
442898	MIR324	-	-	Not in Uniprot.
442903	MIR331	-	-	Not in Uniprot.
442904	MIR335	-	-	Not in Uniprot.
442906	MIR338	-	-	Not in Uniprot.
442907	MIR339	-	-	Not in Uniprot.
442908	MIR340	-	-	Not in Uniprot.
442920	MIR196B	-	-	Not in Uniprot.
494332	MIR383	-	-	Not in Uniprot.
554210	MIR429	-	-	Not in Uniprot.
574413	MIR409	-	-	Not in Uniprot.
574457	MIR181D	-	-	Not in Uniprot.
574508	MIR505	-	-	Not in Uniprot.
619552	MIR483	-	-	Not in Uniprot.
664612	MIR539	-	-	Not in Uniprot.
693183	MIR598	-	-	Not in Uniprot.
693235	MIR92B	-	-	Not in Uniprot.
724022	MIR652	-	-	Not in Uniprot.
768213	MIR671	-	-	Not in Uniprot.
100126297	MIR300	-	-	Not in Uniprot.
100126333	MIR708	-	-	Not in Uniprot.
100126348	MIR760	-	-	Not in Uniprot.

Table 2: GO Biological Process 2018 gene set library Enrichr results, in descending order of p-value.

index	name	p-value	adj p-value	odds ratio	combined score
1	regulation of apoptotic process (GO:0042981)	1.1536954543308095e-64	5.8873079034501206e-61	1.8954511115383643	279.0530052521251
2	positive regulation of gene expression (GO:0010628)	8.903125971447685e-61	2.271632591614877e-57	1.8927659798102223	261.71519033647223
3	regulation of transcription from RNA polymerase II promoter (GO:0006357)	9.285441957891597e-60	1.5794536770373606e-56	1.6213314965382324	220.3821714889842
4	positive regulation of transcription, DNA-templated (GO:0045893)	2.4904553365523665e-56	3.1771983956066815e-53	1.7014878125989235	217.84539746978362
5	gene expression (GO:0010467)	3.1533481401690182e-55	3.2183071118565e-52	2.1796912848548424	273.5375575972465
6	cellular macromolecule biosynthetic process (GO:0034645)	6.720720930306624e-49	5.715973151225784e-46	2.173626151827787	241.10181696093605
7	rRNA processing (GO:0006364)	2.9437371611871136e-48	2.145984390505422 3e-45	2.570057575558126	281.2784210067212
8	ncRNA processing (GO:0034470)	3.589753229450661e-48	2.2898138412358404e-45	2.4822442835169216	271.1752614840283
9	ribosome biogenesis (GO:0042254)	9.072063103298842e-48	5.143859779570444e-45	2.479220768306314	268.5464286990561
10	rRNA metabolic process (GO:0016072)	2.0085387411338678e-47	1.0249573196006128e-44	2.5641025641025643	275.7027997628168

Table 3: GO Cellular Component 2018 gene set library Enrichr results, in descending order of p-value.

index	name	p-value	adj p-value	odds ratio	combined score
1	focal adhesion (GO:0005925)	6.518220647501117e-68	2.907126408785498e-65	2.4097371163538455	372.7891871775142
2	nucleolus (GO:0005730)	9.885157659329896e-55	2.2043901580305667e-52	1.9058911235111105	236.99974480488933
3	nuclear body (GO:0016604)	8.583014796459521e-44	1.2760081997403156e-41	1.8440147889986076	182.85980668285177
4	cytosolic part (GO:0044445)	3.792091715979627 6e-38	4.2281822633172846e-36	2.5682834907153564	221.29695794631144
5	cytoskeleton (GO:0005856)	1.728834863525763e-36	1.5421206982649806e-34	1.8384591034163684	151.38904672821178
6	nuclear chromosome part (GO:0044454)	5.372273337480802e-32	3.993389847527396e-30	1.9057955565892073	137.22008505588101
7	chromatin (GO:0000785)	1.6877862873034785e-30	1.0753609773390734e-28	2.021252021252021	138.56518395154035
8	nucleoplasm part (GO:0044451)	1.4219058753615052e-29	7.927125255140391e-28	1.8511129622240732	122.95641994075733
9	cytosolic ribosome (GO:0022626)	1.6714974391967449e-28	8.283198420908314e-27	2.527341237018656	161.64536572901162
10	nuclear speck (GO:0016607)	2.6696752095386683e-22	1.1906751434542461e-20	1.8608351941685275	92.43683051843422

Table 4: GO Molecular Function 2018 gene set library Enrichr results, in descending order of p-value.

index	name	p-value	adj p-value	odds ratio	combined score
1	RNA binding (GO:0003723)	1.3771279461465885e-153	1.5850742660147234e-150	2.051793287874358	722.181007577404
2	cadherin binding (GO:0045296)	7.892537957342989e-120	4.54215559445089e-117	2.9531729922216257	809.8908405650458
3	protein kinase binding (GO:0019901)	1.5423205567483836e-51	5.917369869391298e-49	2.0336316632612927	237.9319591004644
4	protein kinase activity (GO:0004672)	3.2669815024257683e-42	9.400739273230149e-40	1.9129107281587245	182.73023765681094
5	ubiquitin-like protein ligase binding (GO:0044389)	4.10807549651439e-39	9.456789792976125e-37	2.1636647151050443	191.24170248714836
6	kinase binding (GO:0019900)	1.276225965072945e-38	2.448226809664933e-36	1.9690078169610332	171.80445050094374
7	ubiquitin protein ligase binding (GO:0031625)	2.2643725500870413e-37	3.7232754359288354e-35	2.1623887038597522	182.4587924906022
8	transcription coactivator activity (GO:0003713)	9.689127527512636e-36	1.3940232230208805e-33	2.121250689406474	171.01959815955425
9	protein serine/threonine kinase activity (GO:0004674)	2.0582534404375994e-31	2.632277455492974e-29	1.9268618302434728	136.14874292321335
10	transcription regulatory region DNA binding (GO:0044212)	5.072632712430293e-31	5.838600252007267e-29	1.9128777298711936	133.43523056587406

Table 5: KEGG 2019 gene set library Enrichr results, in descending order of p-value.

index	name	p-value	adj p-value	odds ratio	combined score
1	Pathways in cancer	3.991137479344575e-50	1.229270343638129e-47	1.9829537649244147	225.55142970444584
2	MAPK signaling pathway	9.411581557631911e-38	1.4493835598753145e-35	2.1461415058401876	182.97206801613268
3	Cellular senescence	8.59589650232882e-35	8.825120409057588e-33	2.4928774928774926	195.53929957766306
4	Hepatitis B	3.396964902469259e-34	2.6156629749013296e-32	2.4664169176779946	190.07444640948324
5	Cell cycle	2.0025292045904078e-32	1.233557990027691e-30	2.6294560344739555	191.91955741314376
6	Human T-cell leukemia virus 1 infection	4.633742676388832e-30	2.3786545738796007e-28	2.1826455109052776	147.42501849640024
7	Chronic myeloid leukemia	6.506611632484118e-30	2.8629091182930117e-28	2.957298278935706	198.74444406341306
8	Apoptosis	6.416226372855718e-29	2.4702471535494513e-27	2.435045170087905	158.0737271880907
9	Epstein-Barr virus infection	5.541454023086531e-28	1.896408710122946e-26	2.189117722561614	137.3893033929319
10	Colorectal cancer	5.9289007456297025e-28	1.8261014296539484e-26	2.7974704977288956	175.38054165139712

Table 6: First 9 interactions in the seed genes interactome.

interactor A gene symbol	interactor B gene symbol	interactor A Uniprot AC	interactor B Uniprot AC
FAS	FAS	P25445	P25445
FAS	CTNNB1	P25445	P35222
FAS	RB1	P25445	P06400
FAS	TNFRSF10B	P25445	O14763
BAK1	BAK1	Q16611	Q16611
BAK1	BCL2L1	Q16611	Q07817
CCND1	CCND1	P24385	P24385
CCND1	RB1	P24385	P06400
CCND1	STAT3	P24385	P40763

Table 7: First 9 interactions in the disease interactome.

interactor A gene symbol	interactor B gene symbol	interactor A Uniprot AC	interactor B Uniprot AC
A2M	ENO2	P01023	P09104
ABCA1	IKBKG	O95477	Q9Y6K9
ABL1	BCL2L1	P00519	Q07817
ABL1	PRDX1	P00519	Q06830
ABL1	RB1	P00519	P06400
ABL1	TP53	P00519	P04637
ACAA1	PRDX1	P09110	Q06830
ACACA	BCL2	Q13085	P10415
ACACA	PRDX1	Q13085	Q06830

Table 8: Global measures of the disease interactome LCC.

Measure	Value
average_degree	4.183918961696739
connected_componets_cardinality	1.0
centralization	0.3090763749235911
nodes_cardinality	6318.0
network_radius,centralization	4.0
network_diameter	8.0
edges_cardinality	13217.0
average_path_len	3.1631040976721945
average_clustering_coefficient	0.19127438839462943
isolated_nodes_cardinality	0.0

Figure 1: Disease interactome graph.

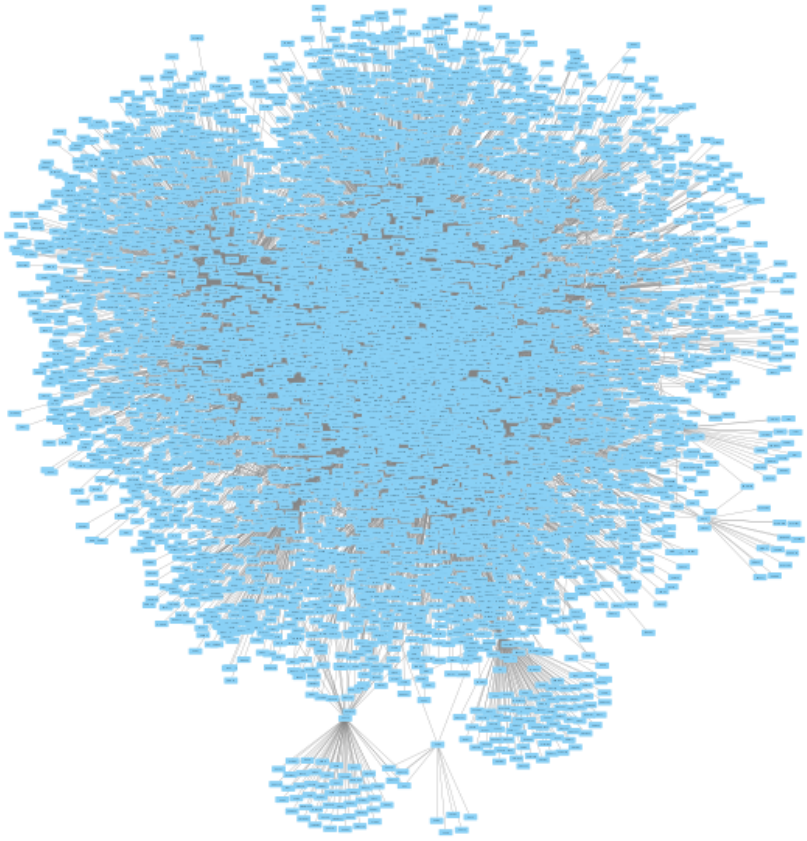

Table 9: First 20 highest ranking genes for betweenness in the LCC.

gene_symbol	nodes_degree	nodes_betweenness_centrality	eigen_vector_centrality	closeness_centrality	betweennes_degree_ratio
MYC	1956	0.35205	0.47437	0.49275	0.00018
KRAS	1593	0.27662	0.30863	0.44779	0.00017
TP53	1328	0.22744	0.28933	0.47593	0.00017
CTNNB1	708	0.08584	0.18391	0.45121	0.00012
CDH1	657	0.08236	0.13068	0.41339	0.00013
HRAS	633	0.07918	0.11843	0.42877	0.00013
IKBKG	393	0.06577	0.06384	0.41209	0.00017
MAPK1	330	0.04663	0.07341	0.44735	0.00014
STAT3	294	0.03999	0.05674	0.39772	0.00014
PML	276	0.03591	0.07554	0.44891	0.00013
TK1	204	0.03518	0.03541	0.3988	0.00017
TNF	301	0.03042	0.05335	0.35493	0.0001
RB1	259	0.02751	0.06408	0.411	0.00011
MAPK3	243	0.02748	0.05554	0.41812	0.00011
CCNDBP1	131	0.0271	0.00782	0.33523	0.00021
PRDX1	217	0.02303	0.04798	0.39141	0.00011
TGFB1	258	0.02283	0.05543	0.35419	9e-05
CCND1	224	0.01943	0.05819	0.38142	9e-05
GSTP1	147	0.01909	0.02936	0.38287	0.00013
BCL2L1	181	0.0188	0.04599	0.40634	0.0001

Table 10: First 30 putative disease genes found by DIAMOnD.

GeneSymbol	Degree	No. of neighbors	p-value
EGFR	20	20	8.630468140178172e-40
HNRNPK	19	19	1.1643122221052507e-37
TRIM25	19	19	1.5154540033751182e-37
HSP90AA1	19	19	1.96535441062697e-37
HNRNPA1	19	19	2.5398426229640867e-37
HNRNPL	18	18	3.075725293082426e-35
HNRNPM	18	18	3.8898878706630207e-35
FUS	18	18	4.904641228227464e-35
SYNCRIP	18	18	6.165834686914531e-35
DDX3X	18	18	7.729004044160159e-35
LARP7	18	18	9.661255055200051e-35
HNRNPH1	17	17	1.025486245339492e-32
HNRNPA2B1	17	17	1.257929794283084e-32
SFPQ	17	17	1.539308827214786e-32
YBX1	17	17	1.8791562306257936e-32
LRPPRC	17	17	2.288715921916043e-32
DDX1	17	17	2.781224411442583e-32
HNRNPF	16	16	2.6236817829744176e-30
C1QBP	16	16	3.135619691847283e-30
YBX3	16	16	3.7400764999145166e-30
NONO	16	16	4.452472023707753e-30
DARS	16	16	5.2905844046407935e-30
SF3B2	15	15	4.54604177658059e-28
RARS	15	15	5.32093526122495e-28
CDK2	15	15	6.217722103004547e-28
IARS	15	15	7.254009120171834e-28
EIF2AK2	15	15	8.44972490921051e-28
G3BP2	15	15	9.827397448755915e-28
QARS	15	15	1.141246155339441e-27
HNRNPR	15	15	1.3233599035320023e-27