Bioinformatics - Project 2
NETWORK MEDICINE PROJECT

Peikova Kateryna, Sofianos Theodoros, Zavadskaya Katsiaryna
Group 10

January 28, 2020

Abstract

For the purposes of this homework, we analyze the Acute Myeloid Leukemia(AML-M2) pathophysiological condition.We got the genes that are involved in this condition from the DisGeNet database, which we will be calling as seed genes in the rest of this paper. Then for the seed genes we found, we gathered all the binary protein interactions between these seeds but also other proteins. We faced an issue that some seed genes did not have uniprot symbols and/or ids. Furthermore, we encountered various missing or wrong values in the protein interaction datasets. We preprocessed and cleaned these datasets, deciding to only work with the genes that have valid both entrez and uniprot symbols.

1 Basic introduction about the disease/process

Acute myeloid leukemia (AML) is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal cells that build up in the bone marrow and blood and interfere with normal blood cells. Symptoms may include feeling tired, shortness of breath, easy bruising and bleeding, and increased risk of infection. Occasionally, spread may occur to the brain, skin, or gums. As an acute leukemia, AML progresses rapidly and is typically fatal within weeks or months if left untreated. In 2015, AML affected about one million people and resulted in 147,000 deaths globally. It most commonly occurs in older adults. Males are affected more often than females. AML is curable in about 35 percent of people under 60 years old and 10 percent over 60 years old. Older people whose health is too poor for intensive chemotherapy have a typical survival of 5–10 months. It accounts for roughly 1.8 percent of cancer deaths in the United States. [1]

2 Seed genes

In this project, we were analysing genes associated with Acute Myeloid Leukemia (AML-M2). To obtain a list of seed genes we requested DisGeNET service with the given name of the disease and downloaded table of Gene-Disease associations. We merged this data with tables from Uniprot and HGNC by available gene ids. Afterall we succeeded in identifying official gene symbols, UniProt and Entrez ids for almost all seed genes. The result is shown in the Table 1. DLEU2 was missing in UniProt database and three more (H1F0, HIST1H1C, SEPT9) gene symbols weren't found in HGNC dataset.

Table 1: Complete seed genes table

Gene Symbol	UniProtKB	Entrez Id	Name
ADCY7	P51828	113	adenylate cyclase 7

Table 1: Complete seed genes table

Gene Symbol	UniProtKB	Entrez Id	Name
ANXA2	P07355	302	annexin A2
ANXA4	P09525	307	annexin A4
ANXA5	P08758	308	annexin A5
ANXA6	P08133	309	annexin A6
FAS	P25445	355	Fas cell surface death receptor
AQP9	O43315	366	aquaporin 9
ATP1B1	P05026	481	ATPase Na+/K+ transporting subunit beta 1
BCL2	P10415	596	BCL2 apoptosis regulator
CAPG	P40121	822	capping actin protein, gelsolin like
CAPN2	P17655	824	calpain 2
CASP7	P55210	840	caspase 7
RUNX1	Q01196	861	RUNX family transcription factor 1
RUNX1T1	Q06455	862	RUNX1 partner transcriptional co-repressor 1
RUNX3	Q13761	864	RUNX family transcription factor 3
CBFB	Q13951	865	core-binding factor subunit beta
CCND2	P30279	894	cyclin D2
CD9	P21926	928	CD9 molecule
CD33	P20138	945	CD33 molecule
CD44	P16070	960	CD44 molecule (Indian blood group)
CDK6	Q00534	1021	cyclin dependent kinase 6
CEBPA	P49715	1050	CCAAT enhancer binding protein alpha
CEBPD	P49716	1052	CCAAT enhancer binding protein delta
CNR2	P34972	1269	cannabinoid receptor 2
CSF1R	P07333	1436	colony stimulating factor 1 receptor
CSF2	P04141	1437	colony stimulating factor 2
CSF3	P09919	1440	colony stimulating factor 3
CST3	P01034	1471	cystatin C
CTNNA1	P35221	1495	catenin alpha 1
CTSH	P09668	1512	cathepsin H
CTSZ	Q9UBR2	1522	cathepsin Z
DAPK1	P53355	1612	death associated protein kinase 1
DHX15	O43143	1665	DEAH-box helicase 15
DNMT3A	Q9Y6K1	1788	DNA methyltransferase 3 alpha
LPAR1	Q92633	1902	lysophosphatidic acid receptor 1
EIF4EBP1	Q13541	1978	eukaryotic translation initiation factor 4E binding protein 1
ENO2	P09104	2026	enolase 2
ERG	P11308	2078	ETS transcription factor ERG
FHL2	Q14192	2274	four and a half LIM domains 2
FOXO1	Q12778	2308	forkhead box O1
FLT3	P36888	2322	fms related receptor tyrosine kinase 3
GATA2	P23769	2624	GATA binding protein 2
GFI1	Q99684	2672	growth factor independent 1 transcriptional repressor
GTF2I	P78347	2969	general transcription factor IIi
H1F0	P07305		
HIST1H1C	P16403		
HGF	P14210	3082	hepatocyte growth factor
HOXA9	P31269	3205	homeobox A9
HSPB1	P04792	3315	heat shock protein family B (small) member 1
ID2	Q02363	3398	inhibitor of DNA binding 2
IDH1	O75874	3417	isocitrate dehydrogenase (NADP(+)) 1
IDH2	P48735	3418	isocitrate dehydrogenase (NADP(+)) 2
JAK2	O60674	3717	Janus kinase 2
KIT	P10721	3815	KIT proto-oncogene, receptor tyrosine kinase
KRAS	P01116	3845	KRAS proto-oncogene, GTPase
LYL1	P12980	4066	LYL1 basic helix-loop-helix family member
MET	P08581	4233	MET proto-oncogene, receptor tyrosine kinase
KMT2A	Q03164	4297	lysine methyltransferase 2A

Table 1: Complete seed genes table

Gene Symbol	UniProtKB	Entrez Id	Name
MN1	Q10571	4330	MN1 proto-oncogene, transcriptional regulator
MX1	P20591	4599	MX dynamin like GTPase 1
MYC	P01106	4609	MYC proto-oncogene, bHLH transcription factor
MYH11	P35749	4629	myosin heavy chain 11
NF1	P21359	4763	neurofibromin 1
NPM1	P06748	4869	nucleophosmin 1
NRAS	P01111	4893	NRAS proto-oncogene, GTPase
NUP98	P52948	4928	nucleoporin 98 and 96 precursor
PDE4B	Q07343	5142	phosphodiesterase 4B
POU4F1	Q01851	5457	POU class 4 homeobox 1
PTPN11	Q06124	5781	protein tyrosine phosphatase non-receptor type 11
RGS2	P41220	5997	regulator of G protein signaling 2
S100A8	P05109	6279	S100 calcium binding protein A8
S100A10	P60903	6281	S100 calcium binding protein A10
SGK1	O00141	6446	serum/glucocorticoid regulated kinase 1
SPARC	P09486	6678	secreted protein acidic and cysteine rich
SPI1	P17947	6688	Spi-1 proto-oncogene
STAT3	P40763	6774	signal transducer and activator of transcription 3
SVIL	O95425	6840	supervillin
TCEA2	Q15560	6919	transcription elongation factor A2
TRH	P20396	7200	thyrotropin releasing hormone
TRIO	O75962	7204	trio Rho guanine nucleotide exchange factor
TSC2	P49815	7249	TSC complex subunit 2
TUBB2A	Q13885	7280	tubulin beta 2A class IIa
WT1	P19544	7490	WT1 transcription factor
PXDN	Q92626	7837	peroxidasin
ASMTL	O95671	8623	acetylserotonin O-methyltransferase like
TNFSF10	P50591	8743	TNF superfamily member 10
DLEU2			
SYNGR1	O43759	9145	synaptogyrin 1
RASGRP1	O95267	10125	RAS guanyl releasing protein 1
IFI30	P13284	10437	IFI30 lysosomal thiol reductase
GAS2L1	Q99501	10634	growth arrest specific 2 like 1
SEPT9	Q9UHD8		
EHMT2	Q96KQ7	10919	euchromatic histone lysine methyltransferase 2
PIM2	Q9P1W9	11040	Pim-2 proto-oncogene, serine/threonine kinase
PSIP1	O75475	11168	PC4 and SFRS1 interacting protein 1
VSIG4	Q9Y279	11326	V-set and immunoglobulin domain containing 4
EHD3	Q9NZN3	30845	EH domain containing 3
ZBTB7A	O95365	51341	zinc finger and BTB domain containing 7A
CHMP5	Q9NZZ3	51510	charged multivesicular body protein 5
FXYD6	Q9H0Q3	53826	FXYD domain containing ion transport regulator 6
ASXL2	Q76L83	55252	ASXL transcriptional regulator 2
ENAH	Q8N8S7	55740	ENAH actin regulator
KMT2C	Q8NEZ4	58508	lysine methyltransferase 2C
BACH2	Q9BYV9	60468	BTB domain and CNC homolog 2
BAALC	Q8WXS3	79870	BAALC binder of MAP3K1 and KLF4
VOPP1	Q96AW1	81552	VOPP1 WW domain binding protein
SPRY4	Q9C004	81848	sprouty RTK signaling antagonist 4
AGRN	O00468	375790	agrin

3 Summary on interaction data

(1) From the Biogrid Dataset, we downloaded all the interactions and then filtered them,by setting as criteria that both organism interactors must have id 9606, which is for homo sapiens, and also the Experimental System Type must be a physical interaction. Then, we removed

the rows that have duplicates or non existing values, for the columns involving UniprotAC Symbols A and B.

(2) From the IID Dataset, we downloaded all the human annotated PPIs. We filtered the evidence type, by t the criteria that it must include the word exp ie experimental in it. We then removed yet again the rows that involve duplicates or non existing values for UniprotAC Symbols A and B, and we addressed an issue with some Symbols having either multiple or invalid values.

(3)Some further details about the data cleaning and preprocessing for our two datasets can be found in the Notes section of this paper.

Total Number of	**Biogrid**	**IID**
Seed Genes involved:	101	107
Interacting proteins:	5037	4745
Interactions:	308003	330116

Table 2: Summary of protein interactions for each dataset

4 Interactomes data

For creating the intersection interactome, we created a graph, with its edges being the interactions between proteins A and B, for both Biogrid and IID datasets. Then, for all the proteins-nodes that belong to the seed genes, we get their neighbours-interactions for each dataset. The proteins confirmed by both datasets is the intersection between neighbouring proteins of seed genes, belonging to each of the two datasets.After finding the proteins confirmed by both Biogrid and IID datasets, we keep only the interactions in which either UniprotAC Symbol A or UniprotAC Symbol B belongs to the confirmed proteins, for both datasets, and we concatenate them in one file. Table 3 shows the global measures for the Seed Genes, Union and Intersection interactomes

Name	**Nodes and edges**	**Connected Components**	**Isolated Nodes**	**Avg path length**	**Avg Degree**	**Avg. Clustering Coefficient**	**diameter and radius**	**Centralization**
SGI	5951 – 12501.	1	0	3.3457	100.0343	0.1684	6 – 3	0.3554
I	16449 – 304749.	1	0	2.9234	225.089	0.1547	6 – 3	0.1459
U	16820 – 343100.	1	0	2.9148	227.0400	0.1406	6– 3	0.14247

Table 3: global measures of networks
(since we have 1 connected component,LCC-U=U , LCC-I=I)

Table 4 shows the local measures for largest connected component Intersection interactome. Since Intersection Interactome is a connected graph with one connected component,its connected component is practically the whole network.

Uniprot Name	**Node Degree**	**Betweenness centrality**	**Eigenvector centrality**	**Closeness centrality**	**Ratio betweenness/Node Degree**
P05067	2210	0.0723	0.0579	0.4886	3.27E-05
Q14258	2184	0.0517	0.0904	0.4984	2.37E-05
Q15717	1824	0.0463	0.0654	0.4812	2.54E-05
Q92731	2437	0.0463	0.1018	0.4989	1.90E-05
P14866	1517	0.0413	0.0520	0.4771	2.72E-05
P01106	2119	0.0315	0.1207	0.5000	1.49E-05
P00533	1436	0.0244	0.0690	0.4806	1.70E-05
P04629	1976	0.0230	0.1079	0.4891	1.16E-05
P62993	1092	0.0176	0.0588	0.4715	1.61E-05
P02545	826	0.0165	0.0489	0.4594	2.00E-05
P78317	1208	0.0161	0.0513	0.4579	1.33E-05
P0CG48	1116	0.0157	0.0650	0.4726	1.41E-05
P05412	1601	0.0153	0.1087	0.4881	9.58E-06
O14980	1263	0.0149	0.0623	0.4660	1.18E-05
P04637	1170	0.0140	0.0771	0.4771	1.20E-05
Q9UBU9	1144	0.0136	0.0524	0.45371	1.19E-05
Q13618	1215	0.0113	0.0873	0.4741	9 .31E-06
Q9HCE1	1030	0.0104	0.0467	0.4475	1.01E-05
P01116	850	0.0100	0.0362	0.4531	1.17E-05
P07900	919	0.0097	0.0611	0.4714	1.06E-05

Table 4: local measures of top 20 LCC I network nodes,sorted by betweenness centrality

Table 5 shows the local measures for largest connected component Union interactome. Since Union Interactome is a connected graph with one connected component,its connected component is practically the whole network.

Uniprot Name	Node Degree	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Ratio betweenness/Node Degree
P05067	2210	0.0658	0.0576	0.4899	2.98E-05
Q14258	2184	0.0479	0.0889	0.4982	2.19E-05
Q15717	1824	0.0434	0.0647	0.4823	2.38E-05
Q92731	2437	0.0412	0.1009	0.4978	1.69E-05
P14866	1517	0.0381	0.0513	0.4797	2.51E-05
P01106	2119	0.0296	0.1180	0.4996	1.40E-05
P00533	1436	0.0214	0.0679	0.4806	1.49E-05
P04629	1976	0.0208	0.1064	0.4885	1.05E-05
P02545	826	0.0164	0.0477	0.4608	1.98E-05
P02545	826	0.0164	0.0477	0.4608	1.98E-05
Q5XPI4	716	0.0162	0.0258	0.4307	2.27E-05
P78317	1208	0.0157	0.0505	0.4581	1.30E-05
P62993	1092	0.0152	0.0574	0.4720	1.39E-05
P05412	1601	0.0140	0.1067	0.4871	8.74E-06
P0CG48	1116	0.0138	0.0641	0.4740	1.24E-05
O14980	1263	0.0137	0.0614	0.4665	1.08E-05
Q9UBU9	1144	0.0125	0.0516	0.4578	1.09E-05
P04637	1170	0.0124	0.0748	0.4753	1.06E-05
Q9H6Z9	1229	0.0103	0.0532	0.4590	8.34E-06
Q9HCE1	1030	0.0102	0.0463	0.4536	9.90E-06

Table 5: local measures of top 20 LCC U network nodes,sorted by betweeness centrality

Figure 1 shows the graph of the Seed Genes Interactome network.

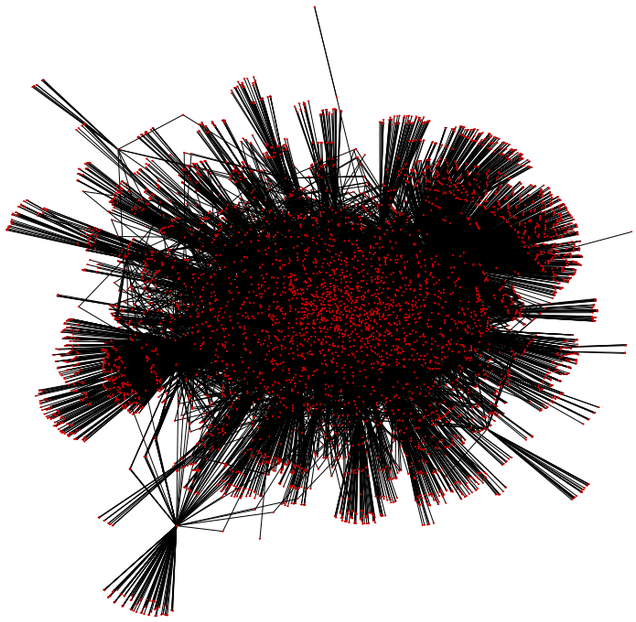

Figure 1: Seed Genes interactome network graph

Figure 2 shows the graph of the largest connected component, which is one and essentially is the whole Intersection Interactome network.

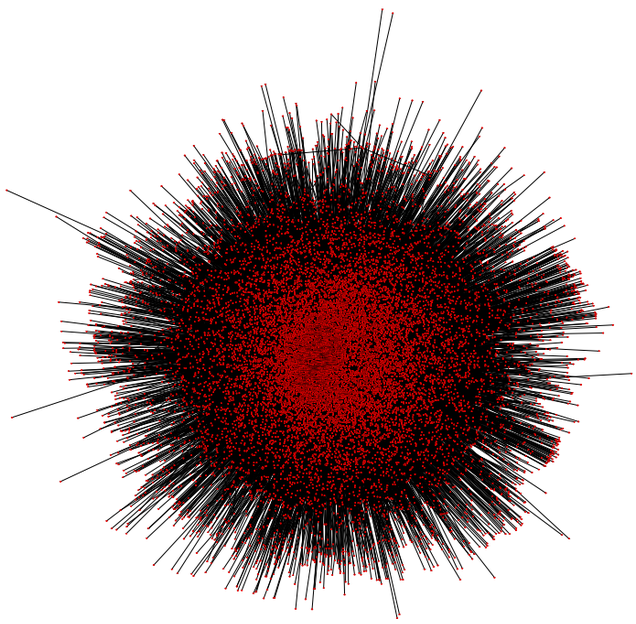

Figure 2: LCC I network graph

5 Enrichment Analysis

5.1 Enrichment analysis of seed, interactome genes and putative disease genes

.

We used Enrichr service to obtain overrepresented Gene Ontologies and pathways for several sets of genes: seed genes, union protein-protein interaction genes and putative disease genes discovered with Diamond tools from Biogrid interactions.

Term	Overlap	P-value
negative regulation of programmed cell death (GO:0043069)	21/408	4.6106221729077685E-15
negative regulation of apoptotic process (GO:0043066)	21/485	1.3662665928150565E-13
regulation of apoptotic process (GO:0042981)	26/815	1.426139171652826E-13
regulation of myeloid cell differentiation (GO:0045637)	9/65	6.91119208017374E-11
hemopoiesis (GO:0030097)	9/76	2.9325312007058026E-10
cytokine-mediated signaling pathway (GO:0019221)	19/633	1.1699083417917615E-9
cellular response to cytokine stimulus (GO:0071345)	16/456	3.055917331746675E-9
positive regulation of peptidyl-tyrosine phosphorylation (GO:0050731)	9/116	1.3125961517755585E-8
regulation of cell proliferation (GO:0042127)	19/740	1.487303161729997E-8
regulation of transcription from RNA polymerase II promoter (GO:0006357)	26/1478	5.734414438765913E-8

Table 6: Enrichment analysis of seed genes with Gene Ontology: Biological Processes

Term	Overlap	P-value
focal adhesion (GO:0005925)	15/356	8.323076721180074E-10
chromatin (GO:0000785)	10/296	4.672334784275548E-6
nuclear periphery (GO:0034399)	5/78	6.455299915888626E-5
nuclear chromatin (GO:0000790)	8/253	6.951390384218093E-5
euchromatin (GO:0000791)	3/21	1.897307119999985E-4
nuclear matrix (GO:0016363)	4/59	2.912073151269398E-4
membrane raft (GO:0045121)	5/119	4.688340071652297E-4
secretory granule lumen (GO:0034774)	7/317	0.001673727130253139
tertiary granule (GO:0070820)	5/164	0.001971046178335233
tertiary granule lumen (GO:1904724)	3/55	0.0032762061040925196

Table 7: Enrichment analysis of seed genes with Gene Ontology: Cellular Components

Seed genes Result of Enrichment analysis for the seed genes is shown in the tables 6, 7, 8, 9 and bar-charts on the figures 3, 4, 5, 6.

Union interactome genes For the analysis of union interactome we extracted all unique genes involved in the interactions. We imported obtained gene set to Enrichr service and got similar results shown in the tables 10, 11, 12, 13, and figures 7, 8, 9, 10.

Putative disease genes For the task 2.4 we performed enrichment analysis of putative disease genes. In this project, except using MCL algorithm we also used DIAMOnD tool [4] to find a disease module. We analysed interactions from Biogrid dataset. First 30 genes (out of 200 obtained) are listed in the table 18. Then we uploaded it to the Enrichr service. The resulting overrepresented gene ontologies and pathways can be found in the tables 14, 15, 16, 17, and figures 11, 12, 13, 14.

6 Clustering

For the task 2.2 in order to obtain putative disease modules, I-LCC and U-LCC were clustered using the MCL algorithm with default inflation parameter = 2. Smaller inflation parameter leads to smaller number of clusters with bigger number nodes in each. The number of clusters, which contain at least 1 seed gene and number of nodes bigger than 10 are: 35 and 31 for I-LCC and U-LCC, respectively. Out of these clusters, we got 2 putative disease modules in each case, 4 in total. Their summaries can be seen in tables 19 and 20.

Term	Overlap	P-value
phospholipase inhibitor activity (GO:0004859)	3/8	8.407823424462166E-6
protein homodimerization activity (GO:0042803)	14/664	1.3534563134384025E-5
protein tyrosine kinase activity (GO:0004713)	7/147	1.5066032066694838E-5
C2H2 zinc finger domain binding (GO:0070742)	3/12	3.251469261094938E-5
calcium-dependent phospholipid binding (GO:0005544)	4/47	1.199394443336596E-4
RNA polymerase II regulatory region sequence-specific DNA binding (GO:0000977)	10/460	1.9486908340298398E-4
transcription regulatory region DNA binding (GO:0044212)	9/374	1.9501331832264844E-4
transmembrane receptor protein tyrosine kinase activity (GO:0004714)	4/61	3.3115174101903223E-4
transmembrane receptor protein kinase activity (GO:0019199)	4/62	3.5252863570958486E-4

Table 8: Enrichment analysis of seed genes with Gene Ontology: Molecular Functions

Term	Overlap	P-value
Acute myeloid leukemia	14/66	4.595197136052959E-19
Pathways in cancer	26/530	5.051750934617662E-18
Transcriptional misregulation in cancer	17/186	1.8294389292372543E-16
PI3K-Akt signaling pathway	17/354	7.153767146301172E-12
Hematopoietic cell lineage	8/97	5.3974150756531625E-8
Central carbon metabolism in cancer	7/65	5.918151369542834E-8
MAPK signaling pathway	12/295	6.703598169935355E-8
Apoptosis	9/143	8.134170802027952E-8
Cellular senescence	9/160	2.130487328483039E-7

Table 9: Enrichment analysis of seed genes with KEGG 2019

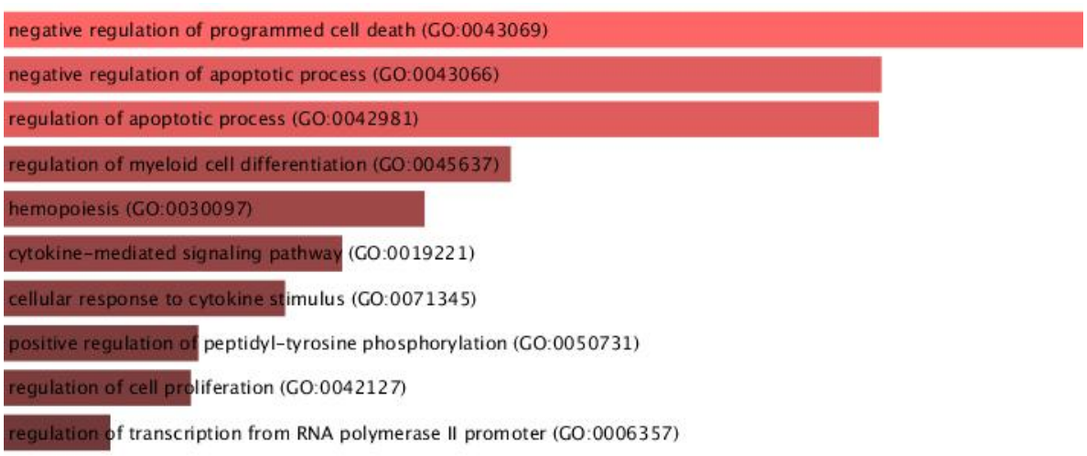

Figure 3: Seed genes: Overrepresented GO. Biological Processes.

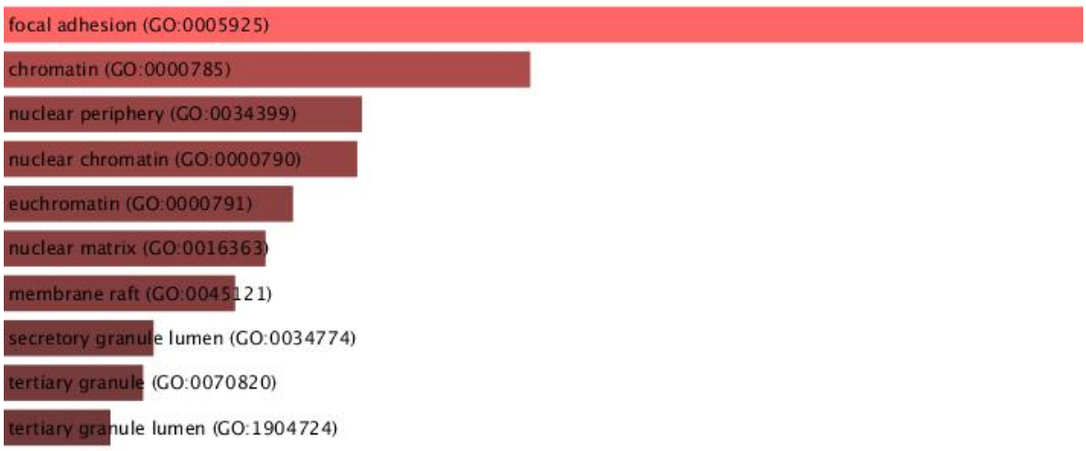

Figure 4: Seed genes: Overrepresented GO. Cellular Components.

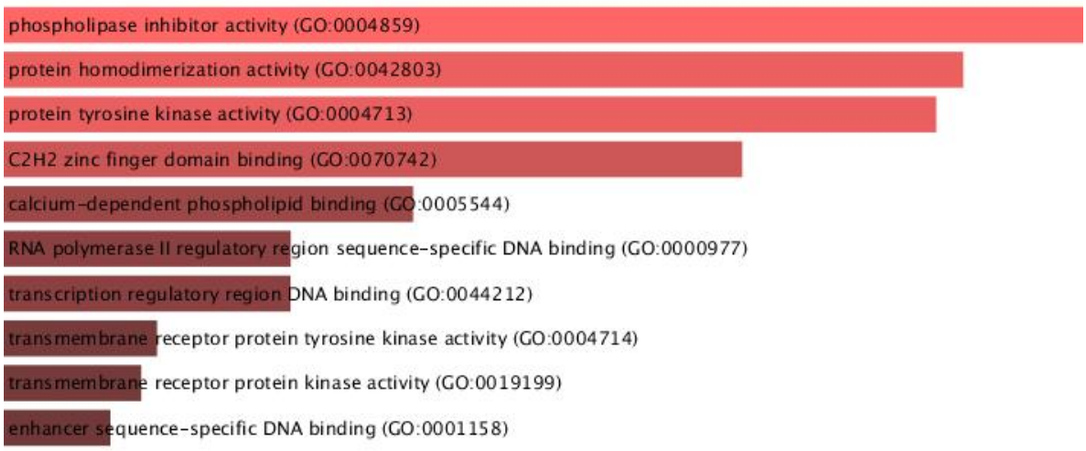

Figure 5: Seed genes: Overrepresented GO. Molecular Functions.

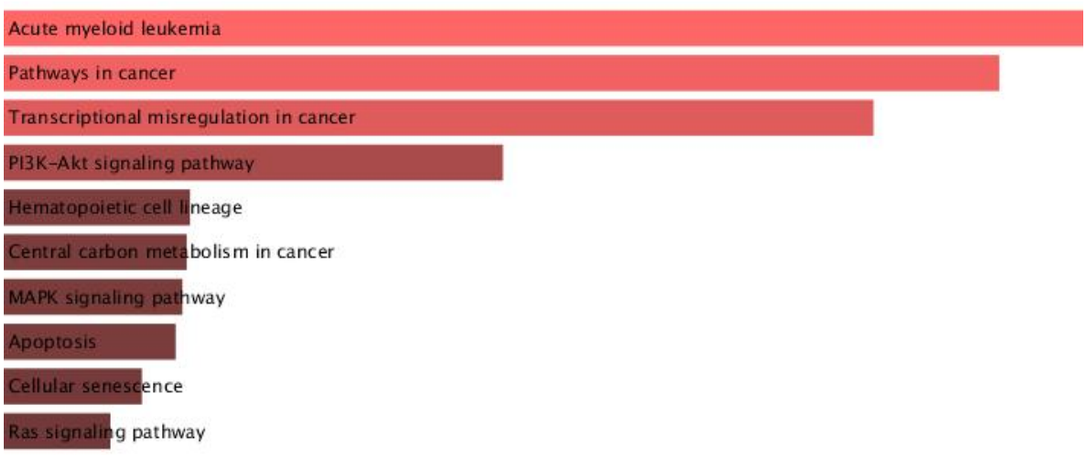

Figure 6: Seed genes: Overrepresented pathways. KEGG 2019.

Term	Overlap	P-value
positive regulation of transcription, DNA-templated (GO:0045893)	1093/1120	5.942802491953528E-54
regulation of transcription from RNA polymerase II promoter (GO:0006357)	1411/1478	1.720847470451196E-47
positive regulation of gene expression (GO:0010628)	760/771	8.114300985602406E-45
regulation of transcription, DNA-templated (GO:0006355)	1504/1598	5.6278185324593216E-39
cellular protein modification process (GO:0006464)	965/1001	1.7312599382041024E-38
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	824/848	9.625590044743214E-38
regulation of apoptotic process (GO:0042981)	789/815	6.404342024226732E-34
positive regulation of nucleic acid-templated transcription (GO:1903508)	498/502	2.88987969174245E-33
negative regulation of transcription, DNA-templated (GO:0045892)	786/813	5.142638497032214E-33
negative regulation of gene expression (GO:0010629)	605/618	1.0810611059126148E-31

Table 10: Enrichment analysis of union interactome genes with Gene Ontology: Biological Processes

Term	Overlap	P-value
nucleolus (GO:0005730)	665/676	7.354932396226679E-38
nuclear body (GO:0016604)	605/618	1.0810611059126148E-31
mitochondrion (GO:0005739)	974/1026	1.353399368590979E-29
microtubule organizing center (GO:0005815)	497/507	8.455850626868217E-27
centrosome (GO:0005813)	453/461	1.4967324506996085E-25
focal adhesion (GO:0005925)	354/356	5.462575506347883E-25
nucleoplasm part (GO:0044451)	400/407	9.727420825626622E-23
cytoskeleton (GO:0005856)	501/520	3.6310351544868434E-20
lysosome (GO:0005764)	409/422	2.182075785504756E-18
microtubule cytoskeleton (GO:0015630)	377/388	9.225117898592934E-18

Table 11: Enrichment analysis of union interactome genes with Gene Ontology: Cellular Components

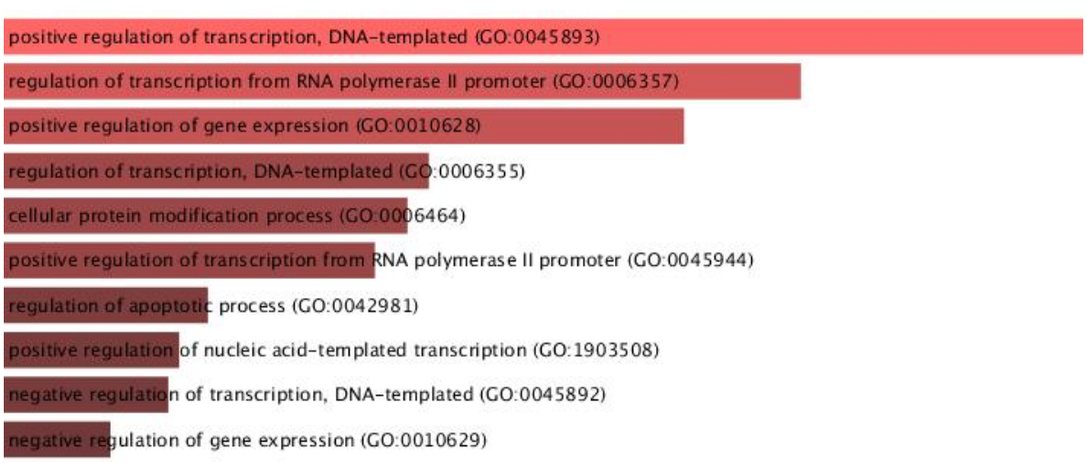

Figure 7: Union Interactome genes: Overrepresented GO. Biological Processes.

Term	Overlap	P-value
RNA binding (GO:0003723)	1367/1387	7.541560873443412E-81
DNA binding (GO:0003677)	850/893	3.6398164502050835E-27
purine ribonucleoside triphosphate binding (GO:0035639)	393/396	1.2862364507338617E-26
ubiquitin-protein transferase activity (GO:0004842)	412/417	1.1575150521605017E-25
protein kinase activity (GO:0004672)	501/513	2.4832334956348853E-25
protein kinase binding (GO:0019901)	483/495	4.2452196182745625E-24
cadherin binding (GO:0045296)	312/313	3.1636561210520465E-23
kinase binding (GO:0019900)	408/418	1.2451512228478622E-20
ubiquitin-like protein ligase binding (GO:0044389)	295/297	1.594676851898165E-20
protein homodimerization activity (GO:0042803)	632/664	2.059304437250795E-20

Table 12: Enrichment analysis of union interactome genes with Gene Ontology: Molecular Functions

Term	Overlap	P-value
Pathways in cancer	518/530	1.6727120932753984E-26
Human T-cell leukemia virus 1 infection	219/219	1.1051025763360985E-17
PI3K-Akt signaling pathway	344/354	2.4616610848448734E-16
Human papillomavirus infection	319/330	5.375225737139749E-14
Endocytosis	239/244	2.8188455078351716E-13
MAPK signaling pathway	286/295	2.826653560786559E-13
Focal adhesion	196/199	4.209480949573358E-12
Apoptosis	143/143	8.935920280850417E-12
Cellular senescence	159/160	1.3846141931351268E-11
Epstein-Barr virus infection	197/201	3.0234379606732756E-11

Table 13: Enrichment analysis of union interactome genes with KEGG 2019

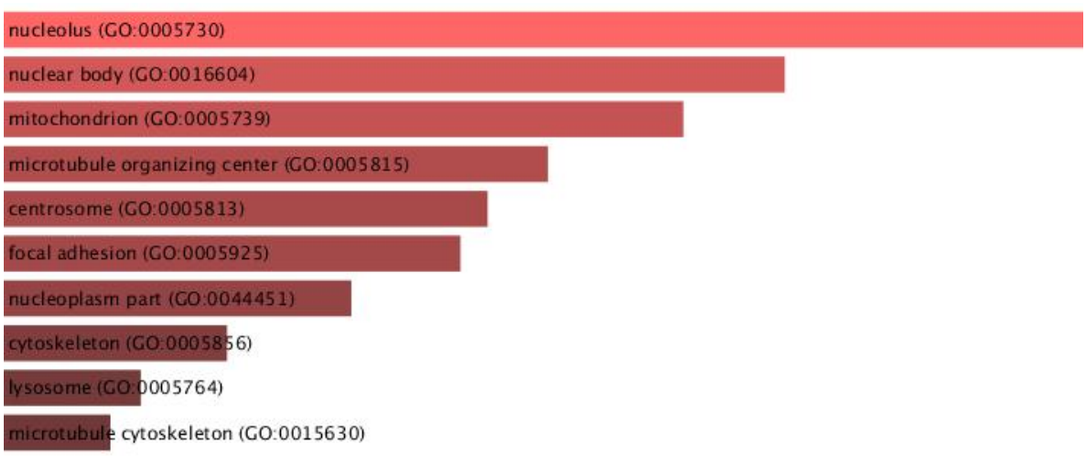

Figure 8: Union Interactome genes: Overrepresented GO. Cellular Components.

Figure 9: Union Interactome genes: Overrepresented GO. Molecular Functions.

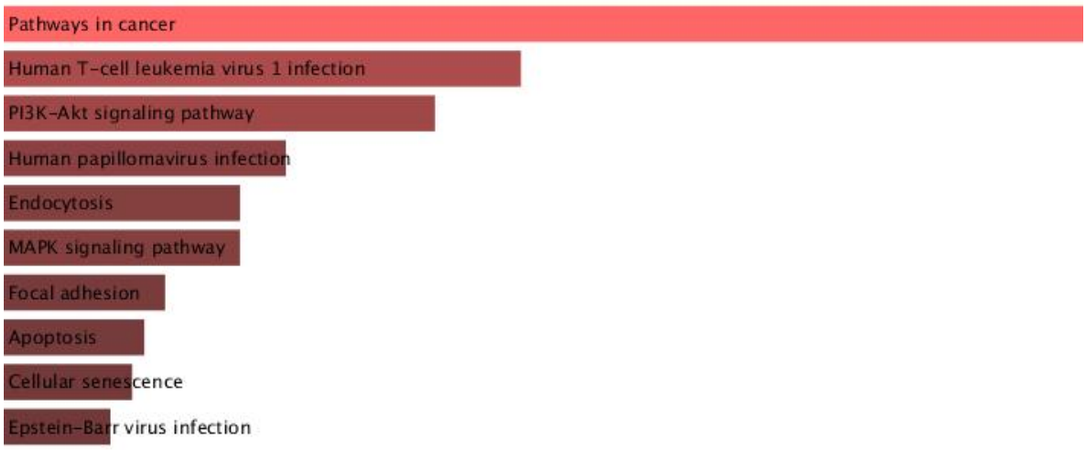

Figure 10: Union Interactome genes: Overrepresented pathways. KEGG 2019

Term	Overlap	P-value
regulation of transcription from RNA polymerase II promoter (GO:0006357)	129/1478	6.947014887630558E-96
positive regulation of transcription, DNA-templated (GO:0045893)	116/1120	1.6373997604043286E-92
negative regulation of transcription, DNA-templated (GO:0045892)	97/813	7.472548594684964E-81
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	98/848	1.979285299111918E-80
positive regulation of gene expression (GO:0010628)	93/771	1.6698736438605161E-77
regulation of transcription, DNA-templated (GO:0006355)	118/1598	3.4187227790021366E-77
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	77/565	5.8671031973547306E-67
positive regulation of nucleic acid-templated transcription (GO:1903508)	74/502	9.919742718668944E-67
negative regulation of gene expression (GO:0010629)	78/618	3.0198709269781617E-65
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	62/512	7.685693338450172E-50

Table 14: Enrichment analysis of DIAMOnD putative disease genes with Gene Ontology: Biological Processes

Term	Overlap	P-value
focal adhesion (GO:0005925)	15/356	8.323076721180074E-10
chromatin (GO:0000785)	10/296	4.672334784275548E-6
nuclear periphery (GO:0034399)	5/78	6.455299915888626E-5
nuclear chromatin (GO:0000790)	8/253	6.951390384218093E-5
euchromatin (GO:0000791)	3/21	1.897307119999985E-4
nuclear matrix (GO:0016363)	4/59	2.912073151269398E-4
membrane raft (GO:0045121)	5/119	4.688340071652297E-4
secretory granule lumen (GO:0034774)	7/317	0.001673727130253139
tertiary granule (GO:0070820)	5/164	0.001971046178335233
tertiary granule lumen (GO:1904724)	3/55	0.0032762061040925196

Table 15: Enrichment analysis of DIAMOnD putative disease genes with Gene Ontology: Cellular Components

Term	Overlap	P-value
transcription regulatory region DNA binding (GO:0044212)	72/374	1.344985905364838E-73
RNA polymerase II regulatory region sequence-specific DNA binding (GO:0000977)	61/460	1.9851139352342884E-51
transcription coactivator activity (GO:0003713)	50/291	1.2700990864818726E-47
DNA binding (GO:0003677)	71/893	7.681703110801254E-45
transcription regulatory region sequence-specific DNA binding (GO:0000976)	45/292	1.5070104542394387E-40
regulatory region DNA binding (GO:0000975)	41/224	4.1571680285018175E-40
core promoter proximal region sequence-specific DNA binding (GO:0000987)	42/278	1.940004141164843E-37
RNA polymerase II core promoter proximal region sequence-specific DNA binding (GO:0000978)	39/262	1.7130023162793916E-34
RNA polymerase II transcription factor binding (GO:0001085)	28/121	1.4754402502974654E-30
transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding (GO:0000982)	36/280	1.6098308559649883E-29

Table 16: Enrichment analysis of DIAMOnD putative disease genes with Gene Ontology: Molecular Functions

Term	Overlap	P-value
Pathways in cancer	58/530	6.526141013464866E-44
Cellular senescence	34/160	1.2633266108985507E-35
Human T-cell leukemia virus 1 infection	37/219	7.823079776653987E-35
Thyroid hormone signaling pathway	30/116	2.6703208372167324E-34
Viral carcinogenesis	35/201	1.9322345203567846E-33
Hepatitis B	31/163	6.440999538739641E-31
Prostate cancer	26/97	2.858095864423978E-30
Cell cycle	28/124	3.1231192416818707E-30
FoxO signaling pathway	27/132	6.351614816656627E-28
Pancreatic cancer	21/75	4.496050418385471E-25

Table 17: Enrichment analysis of DIAMOnD putative disease genes with KEGG 2019

Rank	Gene Symbol	Rank	Gene Symbol
1	RB1	16	CDKN2C
2	RAF1	17	KDELR2
3	PML	18	NF2
4	GLIS2	19	BECN1
5	EPHA2	20	LATS2
6	ARNT	21	AKT1
7	MAP2K5	22	TERT
8	CDKN2B	23	RASSF1
9	FGFR4	24	FZR1
10	PDGFRA	25	CCNE1
11	ERBB2	26	TEAD2
12	MAP2K3	27	CDKN2A
13	CDK4	28	MAP3K5
14	STK11	29	ARAF
15	GRM1	30	TSC1

Table 18: Top-30 genes obtained by DIAMOnD tool from Biogrid PPI.

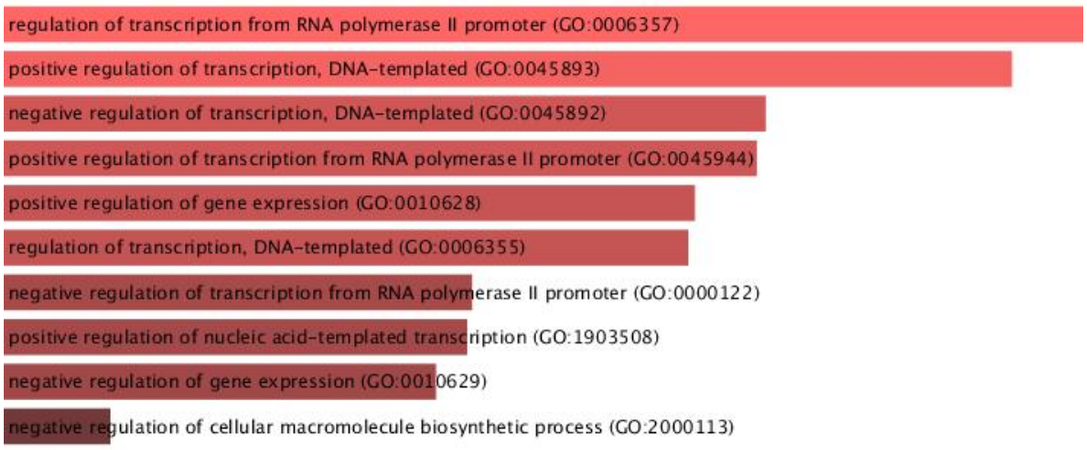

Figure 11: Putative disease genes(DIAMOnD): Overrepresented GO. Biological Processes.

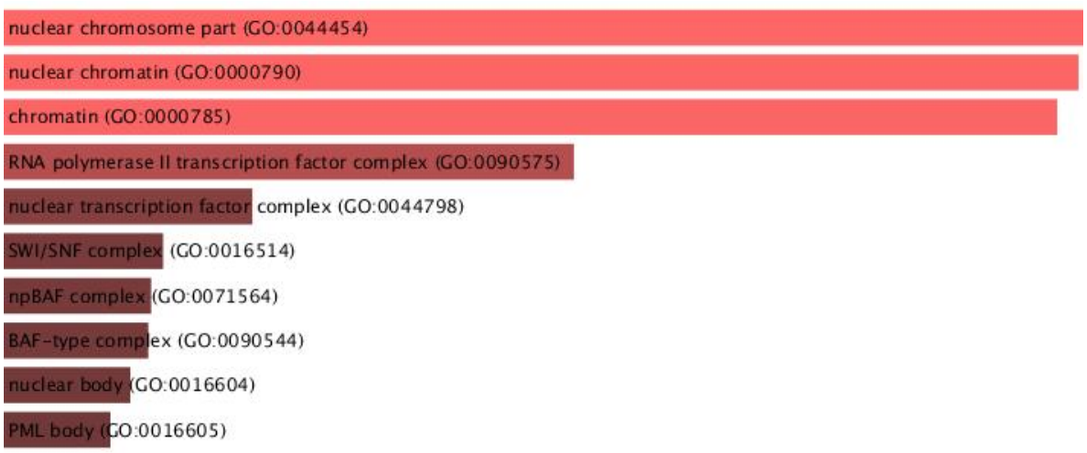

Figure 12: Putative disease genes(DIAMOnD): Overrepresented GO. Cellular Components.

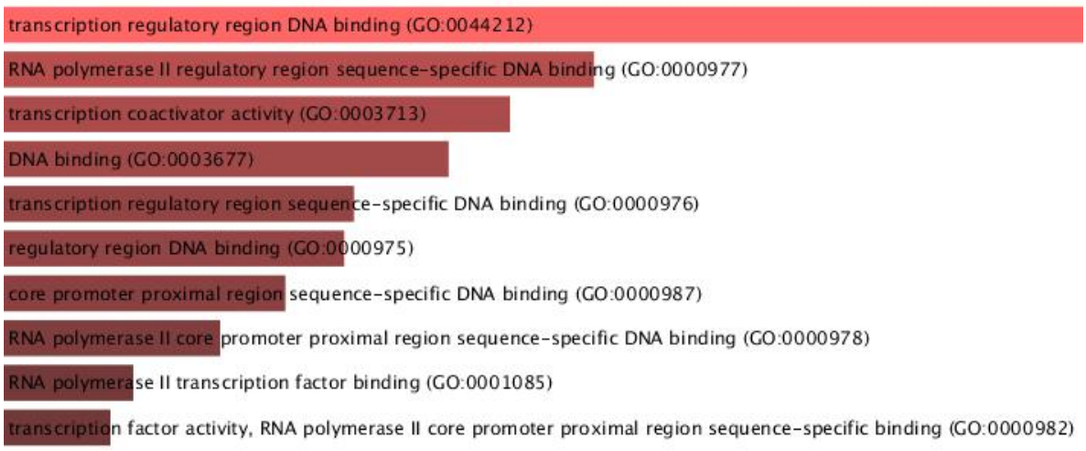

Figure 13: Putative disease genes(DIAMOnD): Overrepresented GO. Molecular Funcitons.

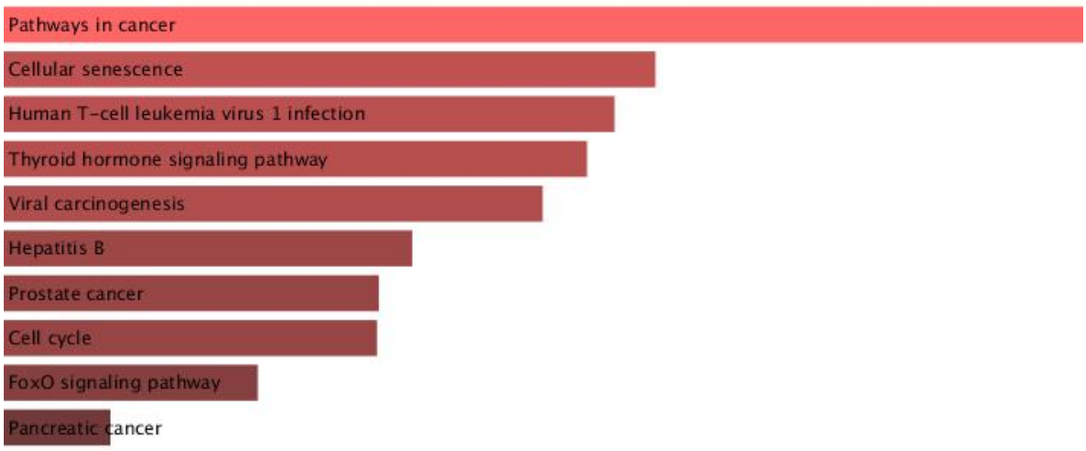

Figure 14: Putative disease genes(DIAMOnD): Overrepresented Pathways. KEGG 2019

Module ID	No. seed genes	Total no. genes	No. seed genes / Total genes	p-value
9	2	31	0.0645	0.0167
13	2	23	0.0869	0.0094

Table 19: U-LCC putative disease modules clustered using MCL algorithm

Module ID	No. seed genes	Total no. genes	No.seed genes / Total genes	p-value
9	2	31	0.0645	0.0173
15	2	25	0.08	0.0115

Table 20: I-LCC putative disease modules clustered using MCL algorithm

7 Notes and comments

7.1 Notes

For part 1.1.b): In order to get a short description of the function of each gene, we parsed the respective description from the Uniprot website, and kept only the first sentence of each description. In that way, we keep a both short and meaningful description of each function, although some of them consist of more than 20 words.

For part 1.2: After downloading the Biogrid and IID protein interaction databases, we noticed that both of them some missing values for either UniprotAC , Official Gene Symbol, Entrez Id, Uniprot Id and also we observed that some columns have multiple values for UniprotAC and Official Gene Symbol, or even text instead of symbols. We have chosen to take the first element in the case of valid multiple symbols and remove the rows for the case of missing either UniprotAC or Official symbol. We didnt take the missing id values into account. In that way, we performed the analysis only for the proteins that have valid symbol for both Uniprot and official one, so that our results have better explanatory power.Furthermore, because of the fact that we have many raw data,from both Biogrid and IID datasets, we can perform this data processing and cleaning process and still maintain a large amount of data for performing the rest of the analysis.(See table 2)

For part 2.1.b): We have noticed that all of the SGI,U,I interactomes are connected graphs,with obviously 1 connected component,which is the whole network. That is why in this paper when we mention LCC-U and LCC-I, we also mean U and I respectively.

For part 2.1.b.ii): As mentioned before, we have only one connected component to all of our interactomes, which means that we must calculate the centrality measures, using a big amount of nodes and edges. We used the networkx library for Python 3.0 for our project and we found out that the running time for closeness centrality, using the networkkx function of closeness centrality, is really huge and could not be run in our systems. That is why we used a created function of calculating closeness centrality of a graph, which was found online, in the following link. click here for the code

7.2 Comments

For part 2.1.b.i): After observing the global measures and seeing the graphical visualization, it becomes clear that the seed genes interactome network has quite a larger value of centralization, which means it is tightly organized around its center. The union and intersection interactome have far lower centralization value and it has some nodes that are far from its center. It can be seen in the figures 1 and 2.

For part 2.1.b.ii): We can observe that the higher ranked nodes of the Union and Intersection interactome, in terms of betweeness centrality, are the same for both of the networks.See tables 5 and 4.

References

[1] Wikipedia,Acute myeloid leukemia
https://en.wikipedia.org/wiki/Acute$_m yeloid_l eukemia$

[2] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Börner, 2016. Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4938516/

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon,2002, Network Motifs: Simple Building Blocks of Complex Networks. https://www.ncbi.nlm.nih.gov/pubmed/12399590

[4] A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. https://doi.org/10.1371/journal.pcbi.1004120