Bioinformatics 2019-2020

Date: 15 January 2020

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2019-2020  Malignant Mesothelioma: a study on the interactome  Giulia Cassarà1, Ivan Colantoni1  1Group no. 13  Starting from existing knowledge about a pathological condition we explore the related information sources (DisGeNet datasets), collect the list of human genes of interest and get protein-protein interaction data. After getting the data of interest we will build: seed genes interactome - interactions that involve seed genes only; union interactome -all proteins interacting with at least one seed gene; intersection interactome - all proteins interacting with at least one seed gene confirmed by both Dbs used in the protein-protein interaction data acquisition phase. Using the service Enrichr, we find and report in tables charts of the overrepresented GO categories limiting to the first 10 for each main category, BP, MF, CL) and the the overrepresented pathways (KEGG 2019 Human) for the seed genes and the union interactome genes. |

The disease

**Malignant mesothelioma** is a cancer of the thin tissue (mesothelium) that lines the lung, chest wall, and abdomen. The major risk factor for mesothelioma is asbestos exposure.

In the section below we will see the genes involved in the disease (the so called seed genes) and the interactions between seed genes and non seed genes in the human organism. We collected the interaction data from two different PPI (Protein-Protein Interaction) sources and integrated together to build the seed genes interactome, the union interactome and the intersection interactome. Furthemore, we analyzed the network graph obtained by different interactomes using a Python library, NetworkX. We applied clustering methods for disease modules discovery. Finally, we found putative disease genes using DIAMOnD tool.

Seed genes

To get the list of the seed genes involved in the disease, we explored the **DisGeNet** website which has a search engine that helps users to find gene-disease associations

(GDAs). The gene-disease associations in **DisGeNET** is organized according to the types of source databases: for example, we get our GDAs from the CURATED dataset, which contains GDAs from UniProt, PsyGeNET, Orphanet, the CGI, CTD (human data), ClinGen, and the Genomics England PanelApp. From the browser we specified our disease of interest (Malignant Mesothelioma) and downloaded the dataset as tab separated text file.

Then, all of our data analysis is perfomed using a Python Library, Pandas, all configured under the Jupyter Lab framework. In Table 1 we show an example of the tab separated text file obtained from the procedure explained above. Of course, for the demonstration purpose we omitted some informations.

**Table 1. List of seed genes for the Malignant Mesothelioma disease obtained from the CURATED dataset of the DisGeNet database.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Gene\_id | UniProt | Gene\_Full\_Name | Protein\_Class |
| NAT2 | 10 | P11245 | N-acetyltransferase 2 | transferase |
| PARP1 | 142 | P09874 | poly(ADP-ribose) polymerase 1 | NaN |
| ANXA2 | 302 | P07355 | Annexin A2 | NaN |
| APOA1 | 335 | P02647 | Apolipoprotein A1 | NaN |

For all genes in the seed gene list, we have checked if the symbols were updated and approved on the HGNC website. All of the gene symbols were approved.

Then, we have collected the following information from Uniprot:

1. Official gene symbol.
2. Uniprot AC (a.k.a. ‘Uniprot entry’).
3. Protein name.
4. Entrez Gene ID (a.k.a ‘GeneID),
5. and a very brief description of its function

Every information was collected from Uniprot’s section Reviewed (Swiss-Prot). Anyway, for the last point, we had to clean the string of function’s description and truncate the string to keep it shorter and more readable.

In Table 2 we show an example of the tab separated csv file obtained from the procedure written above. Of course, for the demonstration purpose we omitted some informations.

We noticed that every gene has a function description in Uniprot Swiss Reviewed, except for the protein transducin beta like 1 X-linked receptor 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Symbol** | **Name** | **GeneID** | **UniprotAC** | **Function** |
| NAT2 | N-acetyltransferase 2 | 10 | P11245 | Participates in the detoxification of a plethora of hydrazine and arylamine drug |
| PARP1 | poly(ADP-ribose) polymerase 1 | 142 | Q6N069 | Auxillary subunit of the N-terminal acetyltransferase A (NatA) complex which displays alpha (N-terminal) acetyltransferase activit |
| ANXA2 | annexin A2 | 302 | Q9H2H9 | Functions as a sodium-dependent amino acid transporte |
| APOA1 | apolipoprotein A1 | 335 | P09874 | Poly-ADP-ribosyltransferase that mediates poly-ADP-ribosylation of proteins and plays a key role in DNA repair |

Summary on interaction data

For each seed gene, we collected all binary protein interactions from two different PPI sources:

* Biogrid Human, latest release available
* IID Integrated Interactions Database (experimental data only)

Table 3 Summarize the main results reporting:

no. of seed genes found in each different Dbs; total no. of interacting proteins, including seed genes, for each DB; total no. of interactions found in each DB.

|  |  |  |  |
| --- | --- | --- | --- |
| Source Database | No. of seed genes | No. of interacting proteins | No. of interactions |
| Biogrid | 104 of 109 | 17899 | 485380 |
| IID | 104 of 109 | 17278 | 270230 |

We have retrieved some informations from Uniprot about missing genes. These are the genes missing in Biogrid: **CCL27**, which has full name C-C motif chemokine 27 and is a Chemotactic factor that attracts skin-associated memory T-lymphocytes; **MIR125A, MIR126, MIR484, PWAR6** haven’t any Uniprot corrispondence.

These are the genes missing from IID: **GPR27** which is an Orphan receptor and possible candidate for amine-like G-protein coupled receptor. **MIR125A, MIR126, MIR484, PWAR6** haven’t any Uniprot corrispondence.

# Interactome data

We have Build and stored three tables from all Dbs:

1. seed genes interactome: interactions that involve seed genes only
2. union interactome: all proteins interacting with at least one seed gene.
3. intersection interactome: all proteins interacting with at least one seed gene confirmed by both DBs

In the format:

*interactor A gene symbol, interactor B gene symbol, interactor A Uniprot AC, interactor B*

*Uniprot AC, database source*

**Seed genes interactome**

We iteratively build a Pandas dataframe that contains the seed genes that interact in both Biogrid and IID, indicating the source db from which the interaction originates. We then saved the interactome as a tab-separated csv file. After eliminating the redundancies, we obtained a total number of 480 seed genes that interact on both databases.

**Union Interactome**

The construction of the union interactome was more cumbersome. We had to consider all the interactions in which one of the interactor is a seed gene. A nonseed gene interacts at least once with a seed gene; then, in the union interactome we also considered interactions between nonseeds.

In the Python implementation, we first built the dataframe containing seed-nonseed interactions for both databases.

Then for the nonseed-nonseed interactions we first built a list containing the gene symbols of the nonseeds.

Iterating on both dataframes we memorized the positions where both interactors are nonseeds, and at the end of the cycle we took the rows positioned in those indexes and we built a new dataframe, one for each database.

Finally, we merged all the tables obtained in the previous steps (the table with the seed-nonseed interactions and the two tables with the nonseed-nonseed interactions) and saved in the usual tabular csv format.

**Intersection Interactome**

The construction of the intersection interactome was pretty straightforward.

We splitted the table with all proteins interacting with at least one seed gene confirmed

by both DBs by the 'source\_db' column. We dropped the 'source\_db' column and merged again the two tables. In pandas, the default type of merge will use all columns and is inner, so it returns a new dataframe with values present in both dataframes.

# Enrichment analysis

To perform the analysis on Enrichr, one must pass a list of genes, one for each row, both for the union and for the seed interactome.

But first, we had to clean up the symbols: many symbols were separated by a ";". It was necessary to create an ad-hoc function that would separate those symbols, one for each line. Other symbols were the full name of the protein and therefore had to be deleted. After this preliminary cleaning phase we passed the list of symbols to HGNC which returned the list of approved symbols which were also good for Enrichr. Finally, we have collected the tables of interest from Enrichr -KEGG HUMAN 2019 and the Ontologies tables of the GO categories overrepresented-. The tables are shown below. For simplicity purpose, not all columns are listed in this report: the complete version of the tables can be consulted among the documents attached to the report.

# Kegg Human

# Table 4 – Kegg Human for Union Interactome

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-Value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| Pathways in cancer | 332/530 | 2.3160184215191657e-68 | 7.133336738279031e-66 | 2.3304132973944296 | 362.9290980239513 |
| Proteoglycans in cancer | 148/201 | 1.1968845263714163e-43 | 1.8432021706119808e-41 | 2.7392797915185976 | 270.7269582669007 |
| MAPK signaling pathway | 190/295 | 6.589420977602359e-42 | 6.765138870338424e-40 | 2.396085552865214 | 227.20428017464832 |
| Neurotrophin signaling pathway | 100/119 | 6.681108708752214e-39 | 5.144453705739206e-37 | 3.12625050020008 | 274.8022167574134 |
| Human T-cell leukemia virus 1 infection | 151/219 | 7.713025893092007e-39 | 4.751223950144676e-37 | 2.5650956729723857 | 225.10743021521782 |
| Cellular senescence | 122/160 | 7.740662923139611e-39 | 3.973540300545002e-37 | 2.8366815476190474 | 248.9310923049112 |
| Chronic myeloid leukemia | 72/76 | 2.212717327939685e-36 | 9.735956242934615e-35 | 3.524436090225564 | 289.35212180706884 |
| MicroRNAs in cancer | 183/299 | 5.378868699539398e-36 | 2.0708644493226677e-34 | 2.2769350215002397 | 184.91122572831318 |
| Cell cycle | 100/124 | 7.722340351336125e-36 | 2.6427564757905847e-34 | 3.000192012288786 | 242.56236161726193 |
| FoxO signaling pathway | 104/132 | 1.5456864924884733e-35 | 4.7607143968644975e-34 | 2.9310966810966814 | 234.9420841108548 |

Table 5 – Kegg Human for Seed genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| Pathways in cancer | 28/530 | 3.897228444575711e-20 | 1.200346360929319e-17 | 9.693612601696383 | 433.22146927792573 |
| Cytokine-cytokine receptor interaction | 22/294 | 4.2162985923687685e-19 | 6.493099832247902e-17 | 13.730262747300756 | 580.9296017973247 |
| PI3K-Akt signaling pathway | 21/354 | 3.3004782409568536e-16 | 3.388490994049037e-14 | 10.884776862074327 | 388.01284210700084 |
| Gastric cancer | 14/149 | 7.624401769285048e-14 | 5.870789362349487e-12 | 17.240317714426453 | 520.7409939607603 |
| Melanoma | 10/72 | 6.135713154692977e-12 | 3.7795993032908735e-10 | 25.484199796126408 | 657.9229052213992 |
| Hepatocellular carcinoma | 13/168 | 7.547315642500119e-12 | 3.8742886964833937e-10 | 14.198339886413278 | 363.61705884563713 |
| MicroRNAs in cancer | 15/299 | 8.435036230126662e-11 | 3.711415941255731e-09 | 9.204995244085788 | 213.5194563996091 |
| Proteoglycans in cancer | 12/201 | 1.02050112873949e-09 | 3.928929345647037e-08 | 10.954402300424485 | 226.7886844049197 |
| JAK-STAT signaling pathway | 11/162 | 1.3552588265201067e-09 | 4.6379968729799205e-08 | 12.458942122550685 | 254.40254528793014 |
| Rheumatoid arthritis | 9/91 | 1.6385277794983662e-09 | 5.046665560854968e-08 | 18.146990624054844 | 367.10396056733885 |

Table 6 – GO Biological Process 2018 for Seed Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| cytokine-mediated signaling pathway | 33/633 | 1.3250962069344388e-23 | 6.761965943986441e-20 | 9.565633288403843 | 503.89816322760987 |
| regulation of cell proliferation | 33/740 | 1.728056361838507e-21 | 4.4091358072309494e-18 | 8.182494421026531 | 391.18288097824717 |
| negative regulation of cell proliferation | 22/363 | 3.806861592803352e-17 | 6.475471569358503e-14 | 11.120378092855155 | 420.42970764675295 |
| negative regulation of cellular process | 25/534 | 9.212689193067938e-17 | 1.1753088238056418e-13 | 8.590179706559463 | 317.17833889146505 |
| cellular response to cytokine stimulus | 23/456 | 3.725153514596662e-16 | 3.801891676997353e-13 | 9.254788347014324 | 328.78795621030633 |
| positive regulation of intracellular signal transduction | 21/479 | 1.3000676570450667e-13 | 1.1057075423168293e-10 | 8.044281856313804 | 238.6834145833249 |
| positive regulation of MAPK cascade | 17/289 | 3.2266561954769925e-13 | 2.352232366502728e-10 | 10.793308148947652 | 310.43885419871435 |
| positive regulation of protein phosphorylation | 18/412 | 9.256557917538748e-12 | 5.904526881650029e-09 | 8.01638906208248 | 203.66188622417422 |
| positive regulation of cell migration | 14/221 | 1.700397687971899e-11 | 9.641254890800668e-09 | 11.623562621943623 | 288.2361526936057 |
| positive regulation of leukocyte chemotaxis | 9/61 | 4.1505804953766564e-11 | 2.1180412267907072e-08 | 27.07174011129493 | 647.1550319766269 |

Table 7 – GO Biological Process 2018 for union interactome

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| focal adhesion | 246/356 | 2.0307521217463597E-63 | 9.057154462988762e-61 | 2.582736819118132 | 372.82956497885874 |
| nuclear body | 312/618 | 2.8626122970109088E-37 | 6.383625422334327e-35 | 1.8869533505215432 | 158.7756401506712 |
| nuclear chromosome part | 223/392 | 6.3003903982783755E-37 | 9.366580392107184e-35 | 2.1262476210816974 | 177.23344878903697 |
| chromatin | 182/296 | 1.7710370351351232E-36 | 1.9747062941756622e-34 | 2.298130685348028 | 189.18556079506217 |
| nucleolus | 329/676 | 3.2856822475828046E-35 | 2.930828564843861e-33 | 1.819048366782076 | 144.43408463071447 |
| cytosolic part | 114/159 | 2.5530499276119735E-32 | 1.8977671128582333e-30 | 2.67980239983357 | 194.9433893639449 |
| cytosolic ribosome | 89/124 | 1.3972500602057802E-25 | 8.902478955025399e-24 | 2.682645993212001 | 153.52815549613666 |
| cytoskeleton | 244/520 | 1.6517829255469964E-23 | 9.208689809924503e-22 | 1.753805902563144 | 92.00045179987413 |
| chromosome | telomeric region (GO:0000781) | 86/124 | 4.106995466029021e-23 | 0.0 | 2.592219723777889 |
| nuclear speck | 159/296 | 4.484949237649398E-23 | 2.000287359991632e-21 | 2.0077075767600903 | 103.31408203448869 |

Table 8 – GO Cellular Component 2018 for Seed Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| endoplasmic reticulum lumen | 10/270 | 2.243366742502301E-6 | 0.0010005415671560263 | 6.795786612300374 | 88.39641733225105 |
| secretory granule lumen | 10/317 | 9.260444955473252E-6 | 0.002065079225070535 | 5.788209417416722 | 67.08394905862652 |
| Golgi lumen | 5/98 | 1.9905230370730948E-4 | 0.029592442484486673 | 9.361542782250515 | 79.77853337459462 |
| membrane raft | 5/119 | 4.890641338098951E-4 | 0.0545306509198033 | 7.709505820676895 | 58.76969334749621 |
| micro-ribonucleoprotein complex | 2/7 | 6.07108833817942E-4 | 0.05415410797656043 | 52.42463958060288 | 388.2989507254963 |
| cytoplasmic vesicle lumen | 5/129 | 7.060183179056919E-4 | 0.052480694964323095 | 7.111869710546903 | 51.60279763008014 |
| RISC-loading complex | 2/9 | 0.001033366784759019 | 0.06584022657178892 | 40.77471967380224 | 280.32346928993604 |
| RISC complex | 2/9 | 0.001033366784759019 | 0.05761019825031531 | 40.77471967380224 | 280.32346928993604 |
| RNAi effector complex | 2/9 | 0.001033366784759019 | 0.05120906511139139 | 40.77471967380224 | 280.32346928993604 |
| endoribonuclease complex | 2/13 | 0.0022073067613513895 | 0.09844588155627197 | 28.228652081863093 | 172.64593272108507 |
| integral component of plasma membrane | 17/1463 | 0.002352100148907491 | 0.09536696967388555 | 2.1321025666752367 | 12.904437078779413 |

Table 9 – GO Cellular Component 2018 for union interactome

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| focal adhesion | 251/356 | 4.5346082209039613E-67 | 2.022435266523167E-64 | 2.6229768592830394 | 400.68978123491956 |
| nuclear body | 318/618 | 1.1458242188789406E-39 | 2.5551880081000377E-37 | 1.914297272306981 | 171.64487992538255 |
| nuclear chromosome part | 225/392 | 9.696848035164027E-38 | 1.441598074561052E-35 | 2.1353407434402336 | 181.98747404685267 |
| chromatin | 184/296 | 1.6808370213764038E-37 | 1.8741332788346904E-35 | 2.3125804375804377 | 195.82088567023192 |
| nucleolus | 335/676 | 2.0728760143590152E-37 | 1.8490054048082416E-35 | 1.8436091152437306 | 155.72359910650673 |
| cytosolic part | 118/159 | 1.3052605504227113E-35 | 9.702436758142153E-34 | 2.7609314165917938 | 221.7692637732338 |
| cytosolic ribosome | 92/124 | 4.867179530715983E-28 | 3.101088672427612E-26 | 2.7601766513056836 | 173.5871452195253 |
| cytoskeleton | 247/520 | 2.07532927998706E-24 | 1.156996073592786E-22 | 1.7671130952380953 | 96.36407418439946 |
| nuclear speck | 161/296 | 6.989001469895008E-24 | 3.4634385061924156E-22 | 2.023507882882883 | 107.88879542793433 |

Table 10 – GO Molecular Function 2018 for Seed Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| cytokine activity | 16/155 | 2.4729931254453526E-16 | 2.846415087387601E-13 | 18.940514945250072 | 680.6450624059376 |
| chemokine activity | 8/46 | 1.3174298501626127E-10 | 7.581808787685836E-8 | 31.910650179497406 | 725.9726581438525 |
| chemokine receptor binding | 8/49 | 2.2464562489679202E-10 | 8.618903808540254E-8 | 29.956936903201644 | 665.5381975150697 |
| cytokine receptor binding | 11/137 | 2.2491254011031932E-10 | 6.471858341674439E-8 | 14.732471706957744 | 327.2864186687703 |
| growth factor activity | 8/69 | 3.806975911266136E-9 | 8.763658547734645E-7 | 21.273766786331606 | 412.42240525569827 |
| CCR chemokine receptor binding | 6/38 | 5.46275429125203E-8 | 1.0479383648718477E-5 | 28.971511347175277 | 484.48269335895384 |
| growth factor receptor binding | 7/92 | 7.036776737660981E-7 | 1.1570471464353985E-4 | 13.960909453530117 | 197.78344245668825 |
| CXCR chemokine receptor binding | 4/17 | 1.8808008431372562E-6 | 2.706002213063727E-4 | 43.17323259579061 | 569.1878204612692 |
| phosphatidylinositol-4,5-bisphosphate 3-kinase activity | 6/68 | 1.8972531588132644E-6 | 2.4263759842156304E-4 | 16.189962223421478 | 213.304426719709 |
| phosphatidylinositol bisphosphate kinase activity | 6/71 | 2.449916455653453E-6 | 2.8198538404571245E-4 | 15.505879312572684 | 200.32753534651505 |
| phosphatidylinositol 3-kinase activity | 6/76 | 3.6581738178341934E-6 | 3.8277800584792334E-4 | 14.485755673587638 | 181.3406058707963 |

Table 11 – GO Molecular Function 2018 for Union interactome

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Overlap** | **P-value** | **Adjusted P-value** | **Odds Ratio** | **Combined Score** |
| RNA binding | 784/1387 | 2.1033101487881945e-129 | 2.420909981255212e-126 | 2.102859889449652 | 623.056282269345 |
| ubiquitin-like protein ligase binding | 218/297 | 1.9553527383479043e-63 | 1.125305500919219e-60 | 2.7306798140131474 | 394.2891122920125 |
| cadherin binding | 223/313 | 5.7003348126335006e-61 | 2.1870284564470527e-58 | 2.650521071048228 | 367.67277076419657 |
| ubiquitin protein ligase binding | 208/284 | 2.7866643968572264e-60 | 8.01862680195667e-58 | 2.724681421864521 | 373.63627249426 |
| protein kinase binding | 289/495 | 2.377291220000058e-50 | 5.4725243884401337e-48 | 2.172017797017797 | 248.18190583988815 |
| kinase binding | 247/418 | 2.137297724191196e-44 | 4.1000494675734443e-42 | 2.198322510822511 | 221.0505653886766 |
| protein kinase activity | 286/513 | 2.2513515012897213e-44 | 3.701865111406383e-42 | 2.074050867910517 | 208.44670290557488 |
| transcription regulatory region DNA binding | 223/374 | 4.872974981030829e-41 | 7.010992753958106e-39 | 2.2182168321874207 | 205.89995317705942 |
| transcription coactivator activity | 183/291 | 2.683954168284042e-38 | 3.43247916410548e-36 | 2.33953117329406 | 202.39504655411312 |
| DNA binding | 408/893 | 4.7041898858607056e-35 | 5.4145225586256724e-33 | 1.6997280435130382 | 134.3499459614465 |