Bioinformatics 2018-2019
Date: 05 December, 2018
Project report

Bioinformatics@Data Science A.Y. 2018-2019

Network Biology project

Pochiraju Venkata Naga Sai Krishna Abhinay (1819771)

Rachuri Mani Niharika (1819748)

Group no. 7

Abstract

The seed genes list provided by the instructor are for Autoinflammatory diseases. This disease is mainly caused due to change in the genes that regulates the immune system. In this project we focused mainly on exploring the functionality of every gene in the list using official HGNC website, finding out all the interactions of the genes by rigorous searching in different databases (IID, BioGrid). We constructed network analysis graphs of all the genes, demonstrated how each gene interacts with one another and there by performed some set functions (union, intersection) on the network we built out how non-proteins are connected and built interactomes.

1 Basic introduction about the disease/process

The seed genes provided by the instructor belongs to Autoinflammatory disease. These diseases are a group of rare diseases characterized by seemingly unprovoked episodes of fever and inflammation and also known as 'periodic fever syndromes.' Autoinflammatory diseases involve abnormal activation of the innate immune system and are caused by changes in genes that regulate the innate immune system. These genetic changes can be passed from parents to their children, leading to multiple cases of disease in an extended family. The seed gene list for Autoinflammatory diseases, as given by the instructor are as follows:
ADCY3, CXCR2, DNMT3B, FAP, FOS, FUT2, GPR35, IFIH1, IL23R, IL27, IL2RA, KEAP1, LCE3B, NFKB1, NXPE1, OR5B21, OSMR, PTPN2, RAVER1, RNF186, RPS6KB1, SH2B3, TNFRSF6B, TYK2.

2 Seed genes

For the seed genes provided, we have used the official HGNC website (https://www.genenames.org/) and the official UniProt website (https://www.uniprot.org/) to collect some basic information about the Official Gene Symbol, Uniprot AC (Accession number), Protein Name, Entrez Gene ID and Functionality. The table below shows these information for all the seed genes.
The excel file with this data is also attached by name: Q2-SeedGenesInfo.xlsx.

Table 1. Seed Genes

Seed Gene	**Official Gene Symbol**	**Approved Name**	**Uniprot AC**	**Protien Name**	**Entrez Gene ID**	**Functionality**
ADCY3	ADCY3	adenylate cyclase 3	O60266 (ADCY3_HUMAN)	Adenylate cyclase type 3	109	It Catalyzes of G-protein signaling.
CXCR2	CXCR2	C-X-C motif chemokine receptor 2	P25025 (CXCR2_HUMAN)	C-X-C chemokine receptor type 2	3579	It is used as a Receptor for interleukin-8 and binding of IL-8.
DNMT3B	DNMT3B	DNA methyltransferase 3 beta	Q9UBC3 (DNM3B_HUMAN)	DNA (cytosine-5)-methyltransferase 3B	1789	It is Essential for the establishment of DNA methylation patterns during development.
FAP	FAP	fibroblast activation protein alpha	Q12884 (SEPR_HUMAN)	Prolyl endopeptidase FAP	2191	It Plays a role in tissue remodeling during development and wound healing.
FOS	FOS	Fos proto-oncogene, AP-1 transcription factor subunit	P01100 (FOS_HUMAN)	Proto-oncogene c-Fos	2353	It has an important role in signal transduction, cell proliferation and differentiation
FUT2	FUT2	fucosyltransferase 2	Q10981 (FUT2_HUMAN)	Galactoside 2-alpha-L-fucosyltransferase 2	2524	It Mediates the transfer of fucose to the terminal galactose .
GPR35	GPR35	G protein-coupled receptor 35	Q9HC97 (GPR35_HUMAN)	G-protein coupled receptor 35	2859	It acts as a receptor for kynurenic acid.
IFIH1	IFIH1	interferon induced with helicase C domain 1	Q9BYX4 (IFIH1_HUMAN)	Interferon-induced helicase C domain-containing protein 1	64135	It Innate immune receptor which acts as a cytoplasmic sensor of viral nucleic acid.
IL23R	IL23R	interleukin 23 receptor	Q5VWK5 (IL23R_HUMAN)	Interleukin-23 receptor	149233	Associates with IL12RB1 to form the interleukin-23 receptor.
IL27	IL27	interleukin 27	Q8NEV9 (IL27A_HUMAN)	Interleukin-27 subunit alpha	246778	It has an ability to functions in innate immunity.

IL2RA	IL2RA	interleukin 2 receptor subunit alpha	P01589 (IL2RA_HUMAN)	Interleukin-2 receptor subunit alpha	3559	Receptor for interleukin-2.
KEAP1	KEAP1	kelch like ECH associated protein 1	Q14145 (KEAP1_HUMAN)	Kelch-like ECH-associated protein 1	9817	Acts as a substrate adapter protein for the E3 ubiquitin ligase complex.
LCE3B	LCE3B	late cornified envelope 3B	Q5TA77 (LCE3B_HUMAN)	Late cornified envelope protein 3B	353143	It is involved in innate cutaneous host defense (Probable).
NFKB1	NFKB1	nuclear factor kappa B subunit 1	P19838 (NFKB1_HUMAN)	Nuclear factor NF-kappa-B p105 subunit	4790	NF-kappa-B is a pleiotropic transcription factor present in almost all cell.
NXPE1	NXPE1	neurexophilin and PC-esterase domain family member 1	Q8N323 (NXPE1_HUMAN)	NXPE family member 1	120400	It Decreases viability and Increased caspase vaccinia virus (VACV).
OR5B21	OR5B21	olfactory receptor family 5 subfamily B member 21	A6NL26 (OR5BL_HUMAN)	Olfactory receptor 5B21	219968	Odorant receptor.
OSMR	OSMR	oncostatin M receptor	Q99650 (OSMR_HUMAN)	Oncostatin-M-specific receptor subunit beta	9180	Associates with IL31RA to form the IL31 receptor.
PTPN2	PTPN2	protein tyrosine phosphatase, non-receptor type 2	P17706 (PTN2_HUMAN)	Tyrosine-protein phosphatase non-receptor type 2	5771	It is Non-receptor that dephosphorylates receptor protein tyrosine.
RAVER1	RAVER1	ribonucleoprotein, PTB binding 1	Q8IY67 (RAVR1_HUMAN)	Ribonucleoprotein PTB-binding 1	125950	It Cooperates with PTBP1 to modulate regulated alternative splicing events.
RNF186	RNF186	ring finger protein 186	Q9NXI6 (RN186_HUMAN)	E3 ubiquitin-protein ligase RNF186	54546	It is a E3 ubiquitin protein.
RPS6KB1	RPS6KB1	ribosomal protein S6 kinase B1	P23443 (KS6B1_HUMAN)	Ribosomal protein S6 kinase beta-1	6198	Serine/threonine-protein kinase that acts downstream of mTOR signaling.
SH2B3	SH2B3	SH2B adaptor protein 3	Q9UQQ2 (SH2B3_HUMAN)	SH2B adapter protein 3	10019	It Links T-cell receptor activation signal

						to phospholipase C-gamma-1.
TNFRSF6B	TNFRSF6B	TNF receptor superfamily member 6b	O95407 (TNF6B_HUMAN)	Tumor necrosis factor receptor superfamily member 6B	8771	It is a decoy receptor that can neutralize the cytotoxic ligands.
TYK2	TYK2	tyrosine kinase 2	P29597 (TYK2_HUMAN)	Non-receptor tyrosine-protein kinase TYK2	7297	Probably involved in intracellular signal transduction.

3 Summary on interaction data

In this part we have retrieved all the binary protein interactions from two PPI sources (BioGRID Human, IID).

3.1 Data Sources

We collected the interactions from the first source (BioGRID) programmatically. The second source was accessed manually from the IID website (http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/).

List of interactions among non-seed genes: We have written a script that finds out the interactions among the non-seed genes. There are 250 interactions in this category. These interactions have been saved in non-SeedGeneInteractions.csv. Following are some of the first rows of these interactions.

Table 2. Non-Seed Gene Interactions

Gene A	Gene B
ADCY8	CALM2
ADCY8	KRAS
CREB1	RP5-1085F17.2
RP5-1085F17.2	RP4-811H24.2
RP5-1085F17.2	RP11-472F14.2
RP5-1085F17.2	RTF1
RP5-1085F17.2	CTR9
RP5-1085F17.2	POLR2A

3.1.1 Network Visualizations of Interactions:

We visualized these networks with python and Gephi. The following are some of the network visualizations of the interactions: (In the included files, there are 2 files named edges.xlsx and nodes.xlsx, which we used in Gephi to plot a Network Visualization). In all the network graphs, blue nodes are the seed genes, and the yellow nodes are the interacting genes (which are not seed genes). And there is an edge between nodes if there is an interaction between them.

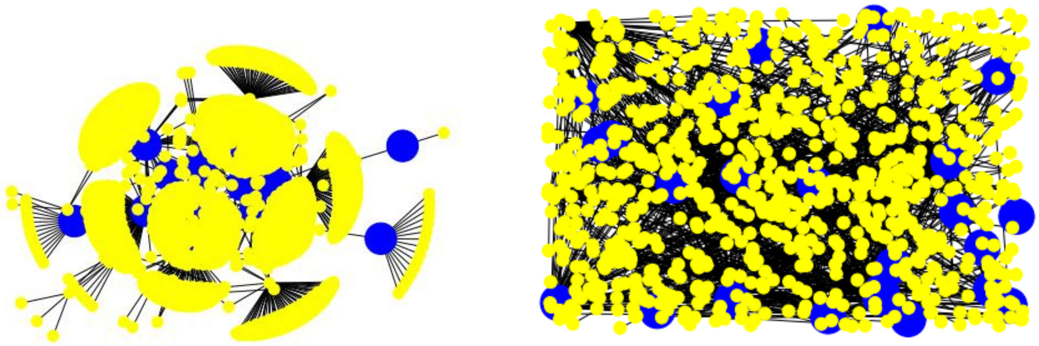

Figure 1: BioGRID interactions (2 different layouts)

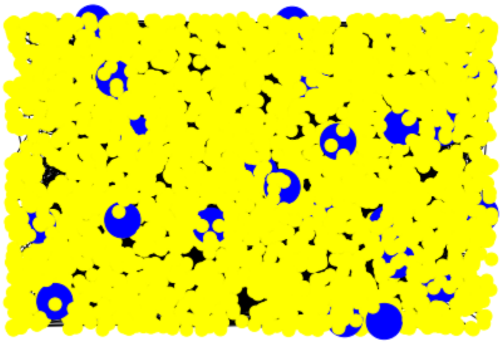

Figure 2: IID interactions

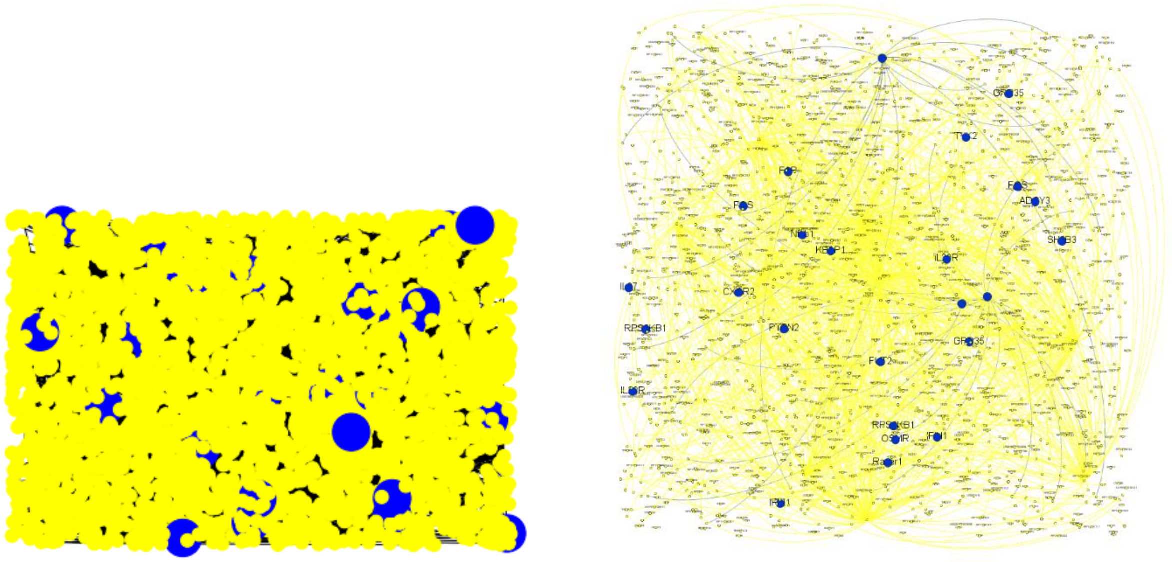

Figure 3: IID and BioGRID interactions combined (1st layout using python and 2nd layout using Gephi)

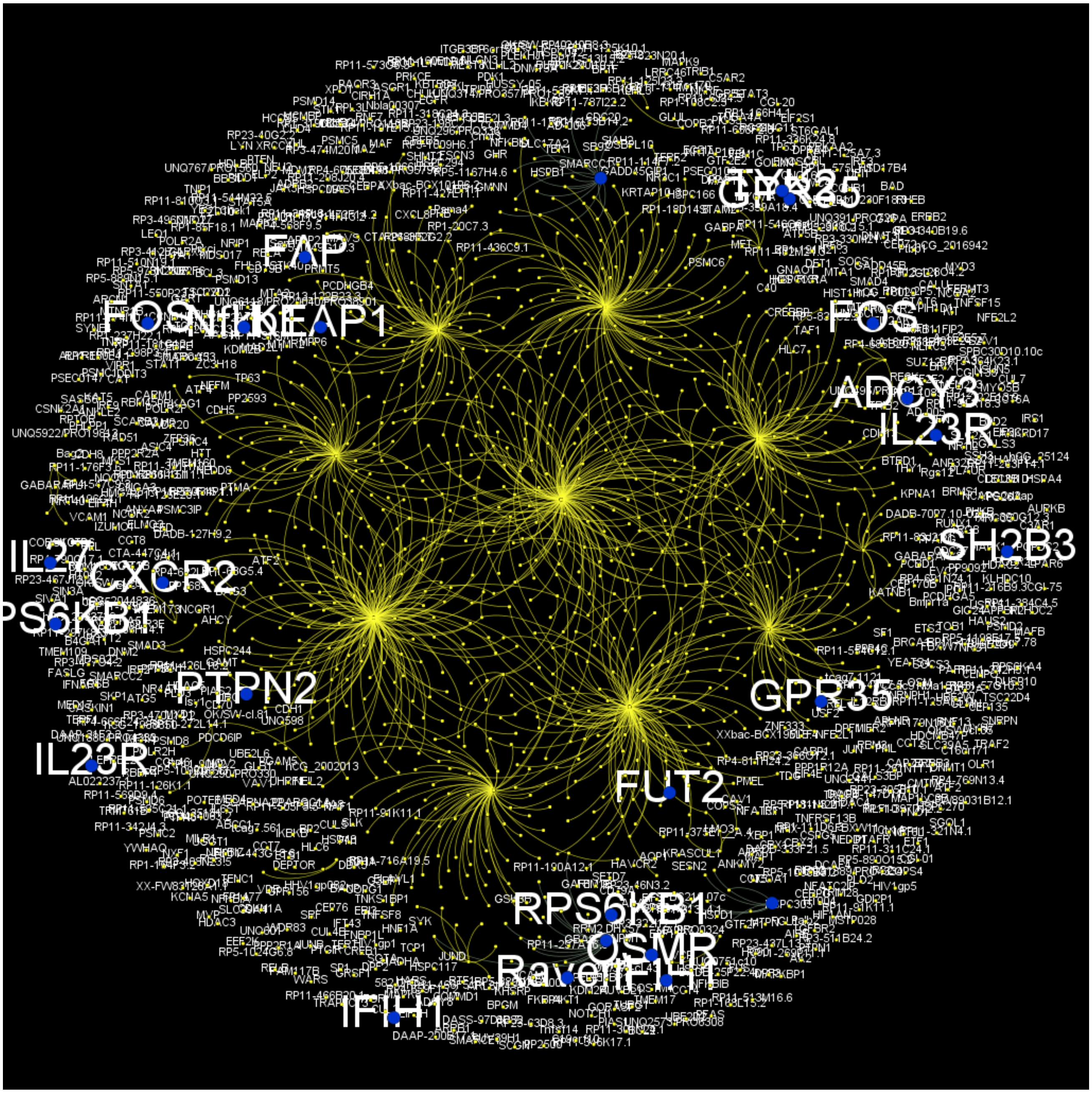

Figure 4: Different layout representing all the interactions (plotted using Gephi)

3.2 Interaction Data tables

The interaction data has been saved in 2 different files with names corresponding to the database, i.e., biogrid.xlsx, and IID.csv.

3.2.1 BioGRID

While getting the data from the BioGRID database, we found out that there are two functions available in python to get the interactions with the given protein/gene. One function (query function) provides Entrez Gene/Locuslink to provide the interactions. And the second function (BioGRID class) provides the interac-tions directly. These two kinds of interactions are in different sheets of the same notebook biogrid.xlsx

Following are some of the first rows from both the files:

Table 3. BioGRID 1

SeedGene	GeneA	GeneB	Interactor Desc 1	Interactor Desc 2
ADCY3	entrez gene/locuslink:805	entrez gene/locuslink:114	biogrid:107256|entrez gene/locuslink:CALM2	biogrid:106627|entrez gene/locuslink:ADCY8
ADCY3	entrez gene/locuslink:805	entrez gene/locuslink:114	biogrid:107256|entrez gene/locuslink:CALM2	biogrid:106627|entrez gene/locuslink:ADCY8
ADCY3	entrez gene/locuslink:154	entrez gene/locuslink:109	biogrid:106663|entrez gene/locuslink:ADRB2	biogrid:106623|entrez gene/locuslink:ADCY3
ADCY3	entrez gene/locuslink:154	entrez gene/locuslink:109	biogrid:106663|entrez gene/locuslink:ADRB2	biogrid:106623|entrez gene/locuslink:ADCY3
ADCY3	entrez gene/locuslink:1080	entrez gene/locuslink:114	biogrid:107506|entrez gene/locuslink:CFTR|entrez gene/locuslink:tcag7.78	biogrid:106627|entrez gene/locuslink:ADCY8

Table 4. BioGRID 2

SeedGene	GeneA	GeneB
ADCY3	ADCY3	CD79B
ADCY3	ADCY8	tcag7.78
ADCY3	ADCY3	SLC17A2
ADCY3	ADCY3	LGALS3
ADCY3	ADCY3	ANKMY2
ADCY3	ADCY3	CNGA3

3.2.2 IID (Integrated Interaction Database)

We downloaded the interactions for IID from the IID website (http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/), but later we found out that the data from the website that we downloaded includes experimentally detected PPIs from 9 curated databases (BioGRID, IntAct, I2D, MINT, InnateDB, DIP, HPRD, BIND, BCI). So, we did all the analysis in the further steps using two different data. One, all the interactions that we downloaded from the IID website, and the other considering only the interactions which have “iid” in the Sources column. Below are some of the first few lines from both the data. The file IID.xlsx indicates all the interactions that we retrieved for the IID website, and the file iidOnly.csv has all the interactions which include the source database as iid.

Following are some of the first rows from both the files:

Table 5. IID 1

Query ID	Query UniProt	Partner UniProt	Query Symbol	Partner Symbol	Species	Evidence Type	Detection Methods	PubMed IDs	Sources
ADCY3	O60266	Q2Y0W8	ADCY3	SLC4A8	human	exp	affinity chromatography	28514442	biogrid
ADCY3	O60266	P07101	ADCY3	TH	human	pred	-	25402006	iid
ADCY3	O60266	P17931	ADCY3	LGALS3	human	exp	affinity chromatography	########	biogrid
ADCY3	O60266	Q8NI99	ADCY3	ANGPTL6	human	pred	-	21836163	iid
ADCY3	O60266	O00624	ADCY3	SLC17A2	human	exp	affinity chromatography	28514442	biogrid

Table 6. IID 2 (IID only)

Gene A	Gene B
ADCY3	TH
ADCY3	ANGPTL6
ADCY3	GUCY2C
ADCY3	PRKAR2B
ADCY3	PDE4C

3.3 Summary of Main results

As we did all the analysis using two sets of data, we added a column for IID, which shows the summary and statistics in both cases (Full data from the IID website, and subset of this full data with IID as source).

Table 7. Summary

Database	Num of Seed Genes in DB	Num of interacting proteins in DB	Total num of interactions found
BioGRID	21	897	990
IID – Full Data	23	3211	4906
IID Only	22	2548	3925

The python script for the 3rd part is labeled: Q3.py

4 Interactomes data

This part of the question has been implemented programmatically, code and files included (The files have been included which show all the interactions from BioGrid+IID+8DBs as BioGRIDandIIDcombined.xlsx and interactions from BioGRID+IID as BioGRIDandIIDcombined1.xlsx)

4.1 Seed Genes interactome

In this part, we found the interactions which include seed genes only. As mentioned above, we did the analy-sis for two types of IID data. In the full data (BioGRID+IID+8DBs from IID website), there is a total of 25 in-teractions which include only the seed genes and have been saved in the file, 4-1-SeedGeneInteractome.csv. And, in the subset of full data BioGRID+IID, there are only 19 interactions which include only the seed genes, which has been saved in the file, 4-1-SeedGeneInteractome1.csv. The interesting thing that we observed here in both the files is that all these interactions are only from the IID database even though we considered the interactions from both the databases.

4.2 Union interactome

In this part, we found all the protein interactions with at least one seed gene. There is a total of 4913 interac-tions involving at least one seed gene in the data with all the interactions (BioGRID+IID+8DBs from IID web-site) and have been saved in the file, 4-2-UnionInteractome.csv. And, there is a total of 3932 interactions for BioGRID+IID only and has been saved in the file, 4-2-UnionInteractome1.csv.

4.3 Intersection interactome

This part of the question highlights the interactions that involve at least one seed gene confirmed by both da-tabases (BioGRID and IID). This also has been implemented programmatically and the results have been saved in the file, 4-3-intersectionInteractome.csv. Note: we used only the interactions from BioGRID and IID (only) for this as we were supposed to find the intersection interactome confirmed by BioGRID and IID. There are a total of 129 interactions in this interactome.

Note: The format for the files for these 3 interactomes are: interactor A Gene Symbol, interactor B Gene Symbol, interactor A Uniprot AC, interactor B Uniprot AC, Database source. As the intersection interactome covers only the interactions confirmed by both DBs, the column Database source will be excluded. To get the Uniprot ACs for all these Gene symbols, we tried uniprot api_idmapping (https://www.uniprot.org/help/api_idmapping), but there were some problems getting them. So, we download-ed the mappings from (https://www.uniprot.org/uploadlists/). These results have been saved in the file uniprot-list.xlsx.

The python script for the 4th part is labeled: Q4.py.

5 Enrichment analysis

In this part of the project we were supposed to find Gene Ontology categories and Pathways for the 3 interactomes that we developed above in the 4th part. So, using excel, we get all the unique set of all the Genes that are involved in the interactions for the data from the tables of 4th part, and use the online GUI of InnateDB to perform these tasks.

Note: We tried to perform these tasks programmatically using the KEGG module from bioservices. Specifically using the function get_pathway_by_gene(), but we could get only id and the pathway name. But using the GUI of InnateDB (http://www.innatedb.com/redirect.do?go=batchPw and http://www.innatedb.com/redirect.do?go=batchGo) we get more detailed results. So, we proceeded to use the online portal.

First, we take all the unique proteins/genes from the interactions in 4.1, 4.2 and 4.3 and save them in files: 4-1PA.xls, 4-2PA.xls and 4-3PA.xls respectively. Then using the above links and uploading these files we get pathways and GO categories for all the three interactomes respectively.
The results were downloaded and named: 5-1-GO-SeedGeneInteractome.xls, 5-1-GO-UnionInteractome.xls, 5-1-GO-IntersectionInteractome.xls, 5-2-Pathway-SeedGeneInteractome.xls, 5-2-Pathway-UnionInteractome.xls and 5-2-Pathway-IntersectionInteractome.xls respectively.

6 Notes and comments

- We were provided with 24 seed genes. Out of which we couldn't find the interactions for: 3 seed genes in BioGRID database (LCE3B, NXPE1 and OR5B21), and 2 seed genes in IID database (NXPE1 and RNF186).
- There is a huge difference in the number of interactions for our seed genes in BioGRID (990) and IID (3925) databases.
- The code snippets and all the supporting tables are in the attached folder.
- Some of the interactions were between the same seed genes, although we couldn't find why that happens.
- The data returned from the IID website includes data about interactions from not only IID but 8 other databases (BioGRID, IntAct, I2D, MINT, InnateDB, DIP, HPRD, BIND, BCI).
- We downloaded the interactions for IID from http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/, but the site was down later. But the site to download now has been changed to http://178.128.224.72/search_by_proteins/
- We came across an interesting python-based library called GOATOOLS. The documentation can be found here (https://github.com/tanghaibao/goatools) which would be very useful for Gene Ontology analysis.

References

- Walter, & Eliza Hall Institute of Medical Research. (2018, August 07). Autoinflammatory diseases | The Walter and Eliza Hall Institute of Medical Research. Retrieved from https://wehi.edu.au/research-diseases/immune-disorders/autoinflammatory-diseases
- Rivas, J. D., & Fontanillo, C. (n.d.). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. Retrieved from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000807