Analysis of gene interactomes for Chron's disease

Group 3:

Pavan Kumar Alikana (1826777)
Dastagiri Dudekula (1826239)
Vamsi Gunturi (1794653)

Abstract:

The study of the biological network and its related complex dynamics is crucial to better understand human diseases. A disease represents not only a malfunction of a single gene but especially a consequence of inter-cellular network distortion. Network science simplifies the proteins interaction complexity to only two elements: the components/nodes and the interactions/edges. Using this model, network information can be gathered in order to examine the underlying disease mechanism. For this reason, the study of network properties and characteristics became interesting as the reflection of biological system activity comprehension. In this report we took the gene interactomes from the previous analysis we did on Chron's disease like Seed Gene Interactome (SGI), Interaction Interactome (I) and Union Interactome (U) and used these datasets to generate graphs using networkx python package and based on the observations we performed clustering of the genes using Louvain and MCL algorithms. Using the clustering information, we did Gene Ontology and Pathway analysis using Innate DB and put together putative disease proteins using DiAMOnD tool to conclude our analysis.

Introduction:

The first step in the network analysis is to gather technical information's about the network structure. Network properties provides insights about the organization of the biological system, the partition of the molecules, the functional structure. In the next sections we will report descriptions about a Chron's disease interactome related network.

Network measures (global and local) for SGI, I and U:

Initially from the previous analysis we took data of SGI (Seed Gene Interactome), II (Intersection Interactome), UI (Union Interactome) and created the graphs using networkx python package.

In Table I we put together all the global features of above graph. In particular we showed the global measures such as number of nodes, number of edges, number of connected components, number of isolated nodes, average path length (which represents the easiness of the proteins to communicate their reciprocal functions), Average degree (which reflects the network structure), Average clustering coefficient (which represents how much a graph tends to be divided into clusters and ranges from 0 to 1) , Network diameter, Network radius, Centralization (closer to 1 means more star-topology network so same connectivity in average)

Note: The 'largest connected components of the union (LCC-U) and intersection (LCC-I) presented only one component, so they more or less show the same features reported for the union and interaction interactome graphs respectively. So, we omitted these (LCC-U and LCC-I) graph properties from Table 1.

Global measures:

Graph Type	nodes	edges	CC	Isolated nodes	Avg path	Avg Degree	Avg cluster coefficient	Diameter	Radius	Centralization
SGI	16	24	8	0	1.33	3.01	0.583	2	1	0.00948
I	2038	3173	1	0	3.69	81.86	0.091	8	5	0.000125
U	2297	3715	1	0	3.65	89.07	0.096	8	4	9.9019

Table 1: Global properties

In Table 2, 3, 4 we put together all the network's local features. In particular, for the largest connected component of the intersection interactome (LCC-I) and the largest connected component of the union interactome (LCC-U) and largest connected component of seed genes interactome (LLC-SGI) we show the following local measures:

- Betweenness, shows how much nodes tend to be intermediaries between other neighbors, protein with high betweenness plays an essential role for the communication in the network
- Degree, represents the numbers of links connected to the nodes
- Closeness, this value shows how quickly or easily a particular node can communicate with other's nodes in the network so how much the information is spread along the system
- Eigenvalues, since not all connection are important in the same way, the eigenvalue rank highlights nodes that are connected to important neighbors, so proteins with high eigenvalues interact with several important proteins in the network
- Ratio, a high ratio means that nodes are connected with hubs instead of nodes with small connections

Interactome Graphs representation:

Figures 1,2 and 3 shows the graphical representation for SGI (Seed gene interactome), II (Interaction interactome), UI (Union interactome) respectively.

Note: Detailed figures can be found in plots folder, II.png, SGI.png, UI.png. Due to space constrains we included the graphs of LCC I and LCC U in Appendix A

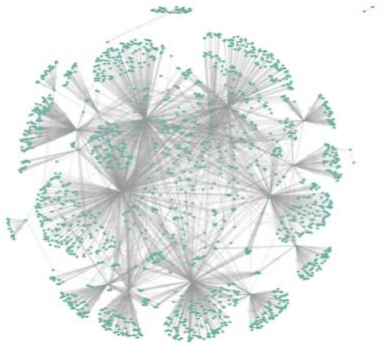

Figure 1: Intersection Interactome

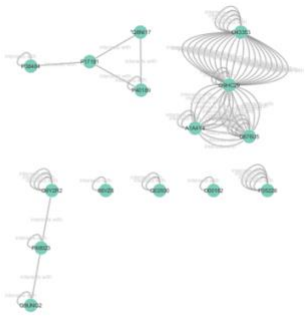

Figure 2: Seed-gene Interactome

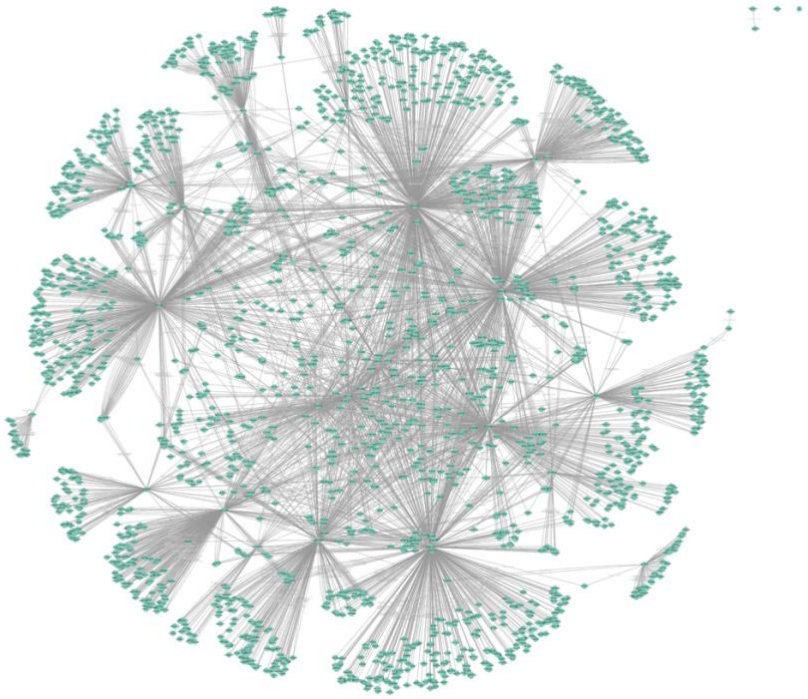

Figure 3: Union Interactome

LCC Local Measures:

LCC SGI:

	Betweenness	Degree	Closeness	Eigenvalues	Betweeness ratio
Q9HC29	1.0	1.0	1.0	1.0	0.4
Q676U5	0.0	0.8	0.75	0.8546	0.0
A1A4Y4	0.0	0.8	0.75	0.8546	0.0
O43353	0.0	0.6	0.6	0.4608	0.0

Table 2: Local measures of LCC (Largest connected component) for SGI (Seed gene interactome)

LCC I:

	Betweenness	Degree	Closeness	Eigenvalues	Betweenness ratio
O43353	1.0	1.0	0.9787	1.0	1.1924
Q9HC29	0.7807	0.8419	0.9815	0.8833	1.1058
P35228	0.6842	0.5809	0.8690	0.2311	1.4044
P40189	0.5768	0.64	0.8947	0.5298	1.0747
Q9Y2R2	0.4474	0.4476	0.8020	0.2144	1.1920
O00182	0.4207	0.2819	0.8039	0.1128	1.7797
Q676U5	0.2787	0.2495	0.8577	0.1114	1.3320
P17181	0.2765	0.3828	0.8778	0.3697	0.8613
P36969	0.2648	0.2152	0.7536	0.0349	1.4672

Table 3: Local measures of LCC (Largest connected component) for II (Interaction Interactome)

LCC U:

	Betweenness	Degree	Closeness	Eigenvalues	Betweenness Ratio
P48023	1.0	0.9295	0.8976	0.6708	1.4316
O43353	0.8540	1.0	0.9058	1.0	1.1365
Q9HC29	0.6564	0.8419	0.9080	0.8948	1.0376
P35228	0.5926	0.5809	0.8153	0.2375	1.3574
P40189	0.4594	0.6399	0.8381	0.5843	0.9553
Q9Y2R2	0.3339	0.4476	0.8322	0.3191	0.9926
O00182	0.2791	0.2819	0.7639	0.1297	1.3175
Q676U5	0.2466	0.2495	0.8086	0.1120	1.3152
P36969	0.2347	0.2152	0.7378	0.0365	1.4512

Table 4: Local measures of LCC (Largest connected component) for UI (Union Interactome)

Note: Above tables 3 and 4 only a fraction of local measures from LCC-I and LCC-U graphs are shown, for complete measures for all the nodes please refer to **data/lcc** folder and files LCC_I_table.csv, LCC_U_table.csv. Graphs for LCC-I and LCC-U can be found in Appendix A Figures 1 and 2. Due to space constrains we are including top 20 highest ranking genes for betweenness for LCC-I and LCC-U in Appendix A in tables 1 and 2.

Clustering methods for disease modules discovery:

Making hypothesis in the context of the network allows us to discover relationship within it. Starting from the hypothesis that proteins involved in the same disease have a tendency to interact with each other, we can say that components associated with a specific disease have the tendency to cluster in the same network area. Consequentially a disease module represents an area in the network in which a group of nodes that suffered a perturbation can be related to a particular disease phenotype. In order to identify disease modules, it is useful to adopt some clustering methods. In this section we performed two different clustering methods: Markov clustering and Louvain methods.

MCL Markov Clustering simulates the flow diffusion in a graph. It is based on the idea of preserving flows where the current is strong and debasing flows where the current is weak. In this way if natural clusters are present in the network by penalizing current across different groups borders, the cluster structure of the graph will automatically appear. Thanks to the transfer matrix, we can obtain the probability description for a random walker to reach all elements in i'th rows and j'th columns in one step. This technique allows us to identify vertexes which are likely to be in the same community.

Louvain modularity idea is characterized by the comparison between the density of the connected the nodes within a community and the suitable density connection in a random graph. Louvain algorithm consists in two main steps: the first one is a local optimization modularity looking for small communities and the second one is made up of a nodes aggregation belonging to the same community.

In the Tables 5 and 6 we show the clustering details of these two algorithms on LCC-I and LCC-U. For each clustered partition we report information about modules which result having a p-value<0.5 in the hyper-geometric test (in other words the putative disease modules).

Algorithm	Index Module	Seed Genes found	Genes found	Seed To Gene Ratio	P Value
Louvain	6	4	163	0.024540	0.031746
Markov	2	1	29	0.034483	0.041126
Markov	6	1	19	0.052632	0.018545
Markov	9	1	26	0.038462	0.033601
Markov	10	1	31	0.032258	0.046467
Markov	17	1	22	0.045455	0.024543
Markov	20	1	3	0.333333	0.000363

Table 5: Clustering for LCC-I using both Louvain and Markov

Algorithm	Index Module	Seed Genes found	Genes found	Seed To Gene Ratio	P Value
Louvain	0	1	1	1.000000	0.000000
Markov	7	1	31	0.032258	0.043619
Markov	9	1	26	0.038462	0.031511
Markov	11	1	19	0.052632	0.017367
Markov	12	1	3	0.333333	0.000339
Markov	14	1	29	0.034483	0.038591
Markov	18	1	21	0.047619	0.021047
Markov	23	1	15	0.066667	0.010952
Markov	24	1	13	0.076923	0.008245

Table 6: Clustering for LCC-U using both Louvain and Markov

Note: Complete list of clustering on LCC-I and LCC-U graphs can be found in data/clustering/pdm.json file. Clustering modules for individual graphs using both Louvain and MCL can be found in folders Louvain Clusters and MCL clusters inside data/clustering folder.

Note: Results from clustering of LCC-I and LCC-U are included in data/clustering folder. This folder contains 3 .json files namely, louvain.json, markov.json and pdm.json which hold the details about clustering results.

Enrichment analysis on the disease modules:

Based on the putative disease modules that we obtained after the clustering of LCC-I and LCC-U (as shown in Tables 5 and 6) we performed overrepresented GO analysis (Gene Ontology) and overrepresented pathway analysis manually from the Innate DB website [6]. We did this analysis for all the putative modules (individual clusters of LLC's from either Louvain or MCL) with more than 10 nodes. We segregated this analysis in to 4 folders namely

- ⇨ Louvain_I and MCL_I for LCC-I graph clusters
- ⇨ Louvain_U and MCL_U for LCC-U graph clusters

Each of the above folders in turn contain 2 sub-folders namely Pathway (for overrepresented pathway analysis) and Gene ontology (overrepresented GO analysis) respectively.

Note: All these folders will be inside **data/ea** folder in our submission. We performed all this analysis manually by uploading the Uniprot Ids of particular genes found in individual clusters from Louvain or MCL clustering approach.

Putative disease proteins using the DIAMOnD:

For finding putative disease proteins for our seed gene list we used DIseAse MOdule Detection (DIAMOnD) algorithm which is derived from a systematic analysis of connectivity patterns of disease proteins in the Human Interactome.

For doing this analysis we provided the Diamond python script as input 2 txt files names seed gene list and PPI list from Bio Grid DB (which we obtained from previous analysis). The files are present in the folder data/diamond folder with names PPI.txt and genes.txt.

We restricted our putative protein list to 200. So, as a 3rd input to the Diamond script we gave 200 and we got the top 200 putative protein list which we saved in **data/diamond/first_200.txt** file. We applied following command for getting top 200 nodes,

python ./DIAMOnD.py BioGridPPI.txt SeedGene.txt 200

From the gene ids of top 200 nodes we performed overrepresented pathway analysis and overrepresented GO analysis manually from Innate DB site [6] and we included the results in data/diamond folder.

Note: The first 30 genes coming from the DIAMOnD tool can be found in Appendix B, Table 3

Foot Notes:

- ⇨ We observed a high degree of inter connectivity for both Union and Intersection interactome, the presence of only one Largely connected component (LCC) for both union and intersection interactome datasets reflects that.

- ⇨ We also observed that the mapping between gene id, gene name and uniport id doesn't always result in the same output always. One common data set coupled with some standardization would be an ideal solution for these kinds of discrepancies.

- ⇨ The clustering results using Louvain and MCL are not the same for each run. So we observed some randomness in the implementation of these algorithms. Even the DiAMOnD algorithm yielded different gene ranks after each run.

- ⇨ We faced difficulties in correlating the results of Gene Ontology and Pathway analysis from Innate DB website, so we observed that the results include other parameters as well which makes them mostly not relevant for our analysis and correlation.

- ⇨ We observed that the clustering coefficient for different graphs is inversely proportional to number of nodes, so we can infer that it is difficult for a disease protein to propagate in a large network of other non-disease genes.

Bibliography:

[1] MCL - a cluster algorithm for graphs.
https://micans.org/mcl/

[2] The Louvain method for community detection in large networks
https://perso.uclouvain.be/vincent.blondel/research/louvain.html

[3] DIAMOnD: DIseAse MODule identification algorithim
Github repository: https://github.com/barabasilab/DIAMOnD

[4] Photocollage
https://www.photocollage.com/

[5] ConnectiviPy
https://connectivipy.readthedocs.io/en/latest/tutorial.html

[6] Innate DB
https://www.innatedb.com/

[7] Uniprot
https://www.uniprot.org/uploadlists/
https://www.uniprot.org/uniprot/

[8] NetworkX
https://networkx.github.io/documentation/networkx-1.10/overview.html

[9] Cytoscape
https://cytoscape.org/

[10] Hypergeometric test
http://systems.crump.ucla.edu/hypergeometric/index.php

[11] Biostars
https://www.biostars.org/p/22/

Appendix A: Network measures for SGI, I and U

Graphs for Largest connected components (LCC) of SGI, I and U:

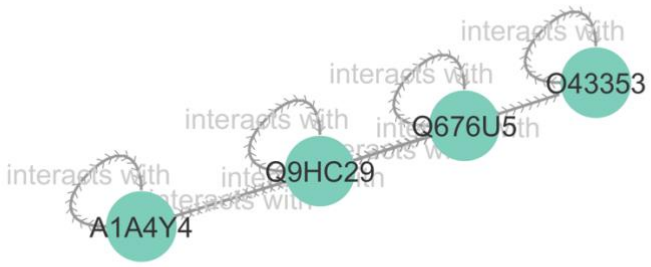

Figure 1: LCC for Seed Gene Interactome(SGI)

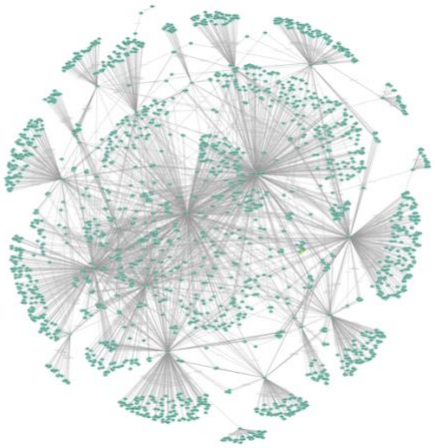

Figure 2: LCC for Intersection Interactome (I)

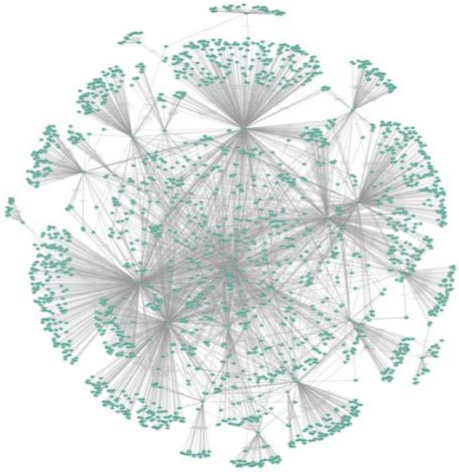

Figure 3: LCC for Union Interactome (U)

20 highest ranking genes (for betweenness):

Largest Connected Component of Interaction Interactome (LCC I) –

	Betweenness	Degree	Closeness	Eigenvalues	Betweenness Ratio
Q15465	7.5180	0.0057	0.6503	0.0254	0.01568
P01579	7.5180	0.0038	0.6367	0.0129	0.0235
Q9UNN8	7.5180	0.0038	0.6367	0.0129	0.0235
O43353	1.0	1.0	0.9787	1.0	1.1924
Q9HC29	0.7807	0.8419	0.9815	0.8833	1.1058
P35228	0.6842	0.5809	0.8690	0.2311	1.4044
P40189	0.5768	0.64	0.8947	0.5298	1.0747
Q9Y2R2	0.4474	0.4476	0.8020	0.2144	1.1920
O00182	0.4207	0.2819	0.8039	0.1128	1.7797
Q676U5	0.2787	0.2495	0.8577	0.1114	1.3320
P17181	0.2765	0.3828	0.8778	0.3697	0.8613
P36969	0.2648	0.2152	0.7536	0.0349	1.4672
Q02930	0.1997	0.1714	0.7874	0.0327	1.3894
7Z699	0.1986	0.1847	0.7833	0.05	1.2820
O95267	0.1957	0.1676	0.7004	0.0267	1.3928
Q8IV20	0.1434	0.1161	0.7642	0.0161	1.4725
Q8NI17	0.1242	0.2285	0.8218	0.212	0.6480
P27824	0.1130	0.0038	0.6059	0.0038	35.3811
Q9UKF2	0.1108	0.0685	0.4860	0.0001	1.9282
P38484	0.09217	0.1866	0.8161	0.1713	0.5888

Table 1: 20 highest ranking genes (for betweenness) in LCC I

Largest Connected Component of Union Interactome (LCC U) –

	Betweenness	Degree	Closeness	Eigenvalues	Betweenness Ratio
P24821	9.7678	0.0057	0.6239	0.0379	0.0227
Q13651	9.7678	0.0057	0.6239	0.0379	0.0227
Q03405	9.7678	0.0057	0.6239	0.0379	0.0227
P48551	9.7678	0.0057	0.6239	0.0379	0.0227
Q15465	5.534	0.0057	0.6121	0.0262	0.0128
P01579	5.534	0.0038	0.601	0.0132	0.0193
Q9UNN8	5.534	0.0038	0.601	0.0132	0.0193
P48023	1.0	0.9295	0.8976	0.6708	1.4316
O43353	0.854	1.0	0.9058	1.0	1.1365
Q9HC29	0.6564	0.8419	0.908	0.8948	1.0376
P35228	0.5926	0.5809	0.8153	0.2375	1.3574
P40189	0.4594	0.6399	0.8381	0.5843	0.9553
Q9Y2R2	0.3339	0.4476	0.8322	0.3191	0.9926
O00182	0.2791	0.2819	0.7639	0.1297	1.3175
Q676U5	0.2466	0.2495	0.8086	0.112	1.3152
P36969	0.2347	0.2152	0.7378	0.0365	1.4512
P17181	0.22	0.3828	0.8243	0.4014	0.7648
Q02930	0.1735	0.1714	0.7504	0.0393	1.3469
7Z699	0.1624	0.1847	0.7445	0.0633	1.1702
O95267	0.1623	0.1676	0.683	0.0374	1.2888

Table 2: 20 highest ranking genes (for betweenness) in LCC U

Appendix B: Find putative disease proteins using the DIAMOnD tool

Top 30 nodes after applying DIAMOnD tool -

Rank	Gene ID	Gene Name
1	**55**	TTTY8B
2	**331**	ENOSF1
3	**1386**	CLDN5
4	**2026**	TMEM191C
5	**2065**	DSG3
6	**2885**	SETD4
7	**4869**	CBLN2
8	**5580**	RBFOX2
9	**5682**	LRR1
10	**5719**	ZNF770
11	**5777**	MYH9
12	**5879**	C14orf183
13	**6774**	FAM53A
14	**7097**	ITPKA
15	**7127**	HBA1
16	**7203**	POLR2F
17	**7416**	SPTBN5
18	**8061**	CDAN1
19	**8737**	TPSD1
20	**9019**	SNX13
21	**9690**	CYP2D6
22	**9902**	B2M
23	**10392**	SPATA5L1
24	**10395**	BIK
25	**10971**	PLEKHA8
26	**22823**	ZNF358
27	**23131**	GOLGA2P1
28	**51150**	GNPDA1
29	**51526**	NR2C1
30	**51702**	GPR156

Table 3: Top 30 nodes from Bio Grid PPI and Seed genes list using DiAMOnD tool