Network Biology

Bioinformatics Project - Part 1.2

Silvia Basile, Tommaso Lanciano, group N. 11

Abstract

Networks are a powerful representation of interactomes, and its properties can be fully exploited to get significant results. Our starting point has been the interactomes found in our previous work:

- *Seed Genes Interactome* contains only interactions among seed genes;
- *Union Interactome* contains interactions that involve at least one seed gene, belonging to Biogrid and IID datasets;
- *Intersection Interactome* contains interactions that involve at least one seed gene, and that are present in both belonging to Biogrid and IID datasets.

We have conducted a explorative analysis of the networks, through the computation of many indices. Furthermore, we have looked up for structure in these networks that can have an important role in the context of the Chromosome Y. In this work we have also exploited DIAMonD, a tool that find putative disease proteins. All the files and codes involved in this work are available in the attached folder or at https://github.com/tlancian/BI_Homeworks.

Network measures

In this section we want to focus on graph indices, either global and local measures. In order to compute all the indices, we used the library `networkx` in Python, created to efficiently deal with graphs.

We developed those analysis only on graphs with more than 20 nodes, condition respected only by U and I interactomes. Even if the SGI graph will not be considered, we provided a graphical representation in **Figure 1**. It is clear that the network is very sparse, thus every indices would be really uninformative.

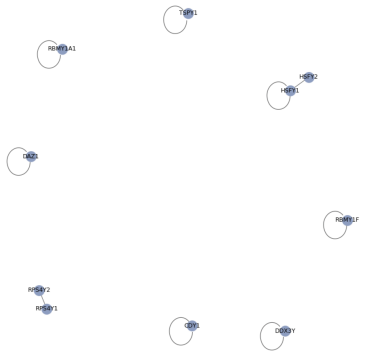

Figure 1: Seed genes interactome

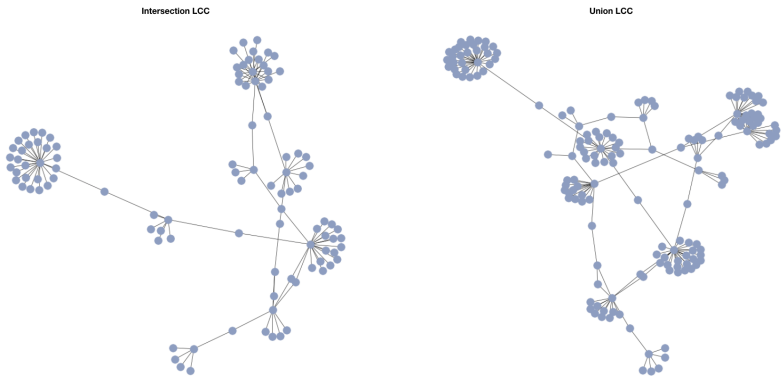

Figure 2: Intersection and Union LCC

Instead, in **Table 1** have been reported the main global indices for U and I. As a first remark, some of the measures does not appear in the table, since the network is not fully connected, thus we are dealing with more than one connected component per graph. Even though U has more nodes and edges, its indices resulted very similar to the one of I, meaning that the underlying structure of the network is quite common.

Table 1: Global measures for I and U

Measures	I	U
# of Nodes	199	288
# of Edges	204	302
# of Connected Components	12	11
# of Isolated Nodes	0	0
Average Path Length	-	-
Average Degree	2.05	2.1
Average Clustering Coefficient	0.0068	0.0046
Network Diameter	-	-
Network Radius	-	-
Centralization	0.1272	0.1014

In order to get more meaningful results, we focused our attention on the largest connected component (LCC) for both the interactomes. A graphical representation of the resulting networks is reported in **Figure 2**, and their relative global indiceshave been reported in **Table 2**.

In this scenario, is more evident how the 2 networks, shares a common structure. Except for the number of nodes, and the number of edges, all the indices are very close to each other. This can be related to some discrepancies between Biogrid and IID, in particular related to genes that in the network has a single connection to a central gene.The lack of this connection does not change substatially the structure of the network.

Table 2: Global measures for I-LCC and U-LCC

Measures	I - LCC	U - LCC
# of Nodes	104	163
# of Edges	119	182
Average Path Length	5.98	5.93
Average Degree	2.288	2.23
Average Clustering Coefficient	0.013	0.008
Network Diameter	12	12
Network Radius	6	7
Centralization	0.2446	0.1797

Once obtained these networks, also local indices have been computed, in particular centrality measures. In **Table 3** (for I-LCC) and in **Table 4** (for U-LCC) have been listed the centrality measures for the 20 highest ranking genes for betweeness centrality. The betweeness is a centrality measure that underlines how many connections among nodes in the graph must pass through a single node.

The rankings are quite similar, but it is interesting to notice that many nodes that has an higher betweenness centraliity, has a small degree. In fact, they are exactly what betweenness look for, nodes that works like bridges among different sub-structures in the network. This information is well represented by the ratio Betweenness/Degree reported in the last column of the tables.

Table 3: Local indices for 20 highest ranking genes for betweenness, I-LCC

	degree	betweenness	eigenvector	closeness	ratio
DDX3Y	19	0,6473	0,0013	0,2512	0,0341
SRY	7	0,4683	0,0001	0,2146	0,0669
SMAD3	2	0,4466	0,0002	0,2320	0,2233
TBL1Y	27	0,4430	0,0002	0,1782	0,0164
HDAC3	2	0,3906	0,0000	0,1951	0,1953
RPS4Y1	9	0,3185	0,0005	0,2124	0,0354
RBMY1A1	18	0,2904	0,5221	0,1897	0,0161
TMSB4Y	5	0,2857	0,0181	0,2141	0,0571
POT1	2	0,2722	0,0033	0,2289	0,1361
TERF1	2	0,2399	0,0927	0,1988	0,1199
USP9Y	11	0,2274	0,0456	0,1758	0,0207
RPS4Y2	3	0,1866	0,0015	0,1977	0,0622
CSNK1E	2	0,1802	0,0081	0,1856	0,0901
CD81	2	0,1239	0,0003	0,2249	0,0619
IGSF8	2	0,1239	0,0003	0,2249	0,0619
CLK3	3	0,1107	0,1869	0,1761	0,0369
RBMY1F	18	0,1079	0,5213	0,1568	0,0060
RNF2	2	0,0933	0,0001	0,1785	0,0466
ZFY	5	0,0765	0,0000	0,1535	0,0153
CIRBP	2	0,0048	0,1791	0,1622	0,0024

Table 4: Local indices for 20 highest ranking genes for betweenness, U-LCC

	degree	betweenness	eigenvector	closeness	ratio
SRY	21	0,5359	0,0002	0,2324	0,0255
DDX3Y	28	0,5249	0,0029	0,2485	0,0187
SMAD3	2	0,3656	0,0005	0,2379	0,1828
TBL1Y	31	0,3370	0,0001	0,1806	0,0109
HDAC3	2	0,3114	0,0000	0,2035	0,1557
RBMY1A1	24	0,2659	0,5517	0,1858	0,0111
USP9Y	15	0,2467	0,0436	0,1901	0,0164
RPS4Y1	14	0,2381	0,0006	0,2088	0,0170
TMSB4Y	6	0,2351	0,0168	0,2077	0,0392
POT1	2	0,2234	0,0032	0,2231	0,1117
TERF1	2	0,1973	0,0911	0,1929	0,0986
AR	3	0,1740	0,0000	0,2069	0,0580
TSPY1	8	0,1515	0,0001	0,1940	0,0189
CDY1	6	0,1257	0,0016	0,1824	0,0209
CLK3	3	0,1243	0,1734	0,1767	0,0414
UBC	3	0,1202	0,0074	0,1841	0,0401
HIST2H2AC	2	0,1155	0,0003	0,1851	0,0577
RBMY1F	21	0,1073	0,4866	0,1555	0,0051
RPS4Y2	3	0,1067	0,0013	0,1952	0,0356
CSNK1E	2	0,1036	0,0072	0,1860	0,0518

Disease Modules Detection

Another interesting task, is the one of community detection. For this, we have exploited two different algorithms:

- *Louvain*: it is a method to extract communities from large networks. Basically, first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated.
- *Markov Clustering (MCL)*: the main idea of this algorithm is that, starting from a node and then randomly travel to a connected node, it is more likely to stay within a cluster than travel between. The MCL algorithm is based on random walks evaluated through Markov Chains.

The results are shown in **Figure 3** and **Figure 4**, and they are nothing surprising, given that is was already visible in **Figure 2** that the networks had a well-defined community structure. To notice by the way, how the MCL clustering method produces a more refined partition of the network.

Applying an hypergeometric test, we can explore the modules in each configuration, looking for the one statistically enriched, i.e. the one that contains a significative number of seed genes.

Assuming the null hypothesis as "seed genes are not statistically overrepresented in a specific cluster", if the *p-value* will be ≤ 0.05 then we can reject the null hypothesis. We selected modules with more than 10 nodes and with a p-value ≤ 0.05 and we consider those modules as **putative**

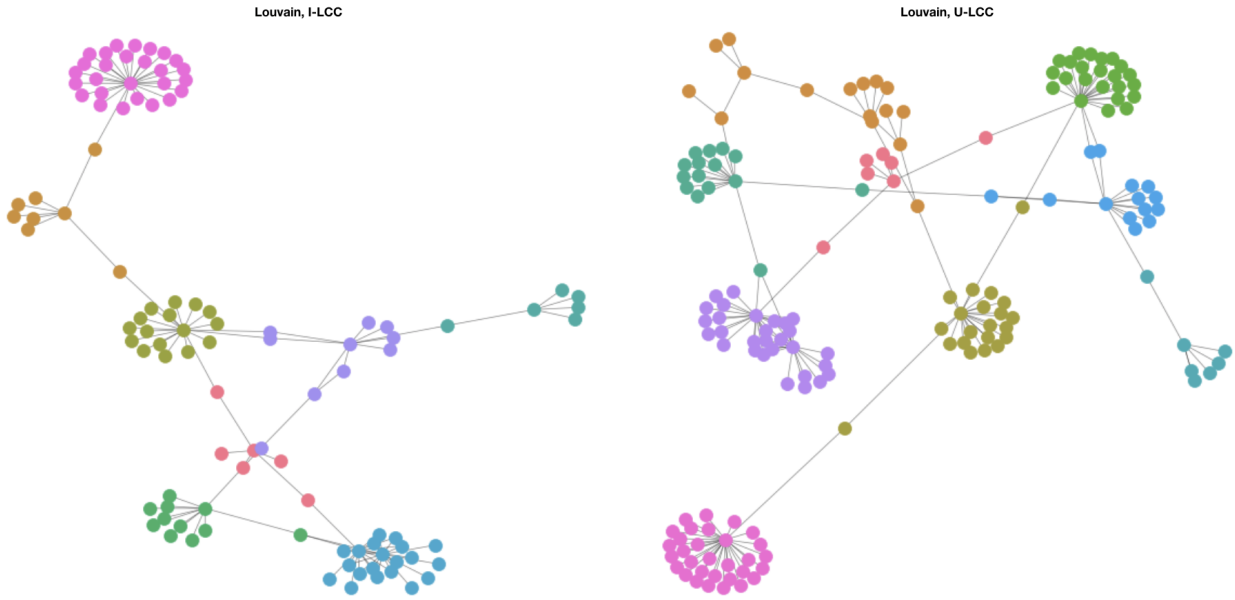

Figure 3: Louvain clustering

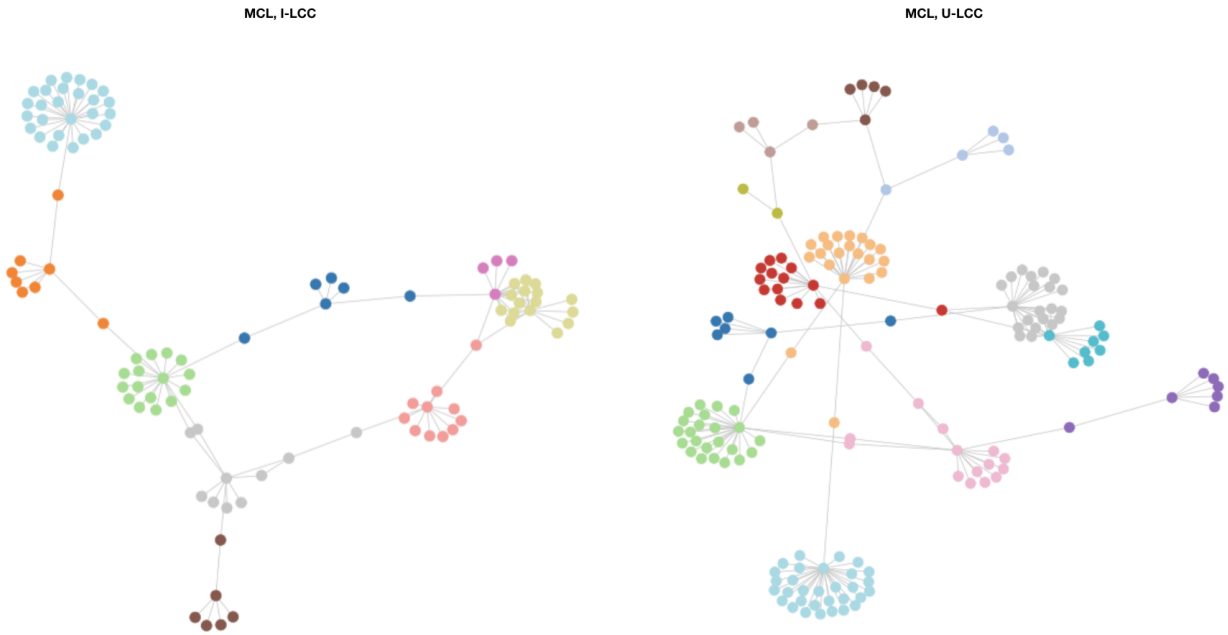

Figure 4: MCL clustering

disease modules. The results of the hypergeometric test are shown in **Table 5**. We found only a module that respected those properties, that is the N. 1 of the U-LCC.

Table 5: Candidates Putative disease module - I-LCC

Id	Algorithm	Seed Genes	Total Genes	SG Ratio	P-Value
1	Louvain	1	16	0,0625	0,827069
2	Louvain	1	11	0,09	0,690455
3	Louvain	2	20	0,1	0,612131
4	Louvain	2	10	0,2	0,246081
5	Louvain	1	27	0,037	0,957971
6	MCL	1	16	0,0625	0,827069
7	MCL	1	11	0,09	0,690455
8	MCL	2	10	0,2	0,246081
9	MCL	1	16	0,0625	0,827069
10	MCL	1	27	0,037	0,957971

Table 6: Candidates Putative disease module - U-LCC

Id	Algorithm	Seed Genes	Total Genes	SG Ratio	P-Value
1	Louvain	4	16	0,25	0,03425
2	Louvain	1	21	0,047	0,867177
3	Louvain	1	23	0,043	0,89216
4	Louvain	1	14	0,071	0,730976
5	Louvain	2	15	0,13	0,377092
6	Louvain	2	29	0,068	0,753428
7	Louvain	1	31	0,032	0,954638
8	MCL	1	21	0,047	0,867177
9	MCL	1	23	0,043	0,89216
10	MCL	1	14	0,071	0,730976
11	MCL	2	15	0,13	0,377092
12	MCL	1	21	0,047	0,867177
13	MCL	1	31	0,032	0,954638

Enrichment Analysis

In this section, an enrichment analysis over the only putative disease module, has been conducted. For this task, we have again exploited the tool released by InnateDB. In **Table 6** and **Table 7** we report the results obtained either for the overrepresented GO categories and overrepresented pathways.

Table 7: GO Over-Represented Analysis for Putative Disease Module

Pathway Name	Pathway p-value (corrected)
eukaryotic translation elongation factor 1 complex	0,000346
nucleus	0,000477
nucleosome assembly	0,000559
DNA binding	0,001175
seminiferous tubule development	0,001265
nucleosome	0,001684
sex differentiation	0,003306
translation elongation factor activity	0,003436
PcG protein complex	0,004128
POU domain binding	0,012156

Table 8: Pathway Over-Represented Analysis for Putative Disease Module

Pathway Name	Pathway p-value (corrected)
RNF mutants show enhanced WNT signaling and proliferation	5,42E-06
TCF dependent signaling in response to WNT	5,42E-06
XAV939 inhibits tankyrase, stabilizing AXIN	5,42E-06
misspliced LRP5 mutants have enhanced beta-catenin-dependent signaling	5,42E-06
Signaling by WNT in cancer	7,34E-06
Cellular responses to stress	9,64E-06
M Phase	1,8E-05
Signaling by Wnt	1,97E-05
Packaging Of Telomere Ends	2,18E-05
Condensation of Prophase Chromosomes	5E-05

Putative Disease Proteins detection [DIAMOnD]

DIAMOnD is a tool released by Susan Dina Ghiassian, Joerg Menche & Albert-Laszlo Barabasi, and it is more properly the implementation of the algorithm they present in their paper "A DIseAse MOdule Detection (DIAMOnD) Algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the Human Interactome".

It allows to detect the full disease module around a set of known disease proteins. In this work we have exploited it, using as network the latest release of the Biogrid interactome, and as seed file, the seed genes.

In **Table 8** we report the first 40 proteins, according to the rank of the DIAMOnD algorithm. In the attached folder, there you can find also the over-represented analysis for the proteins returned by DIAMOnD, joined with the seed genes. In **Table 9** and **Table 10** the Top-10 results for both analysis.

Table 9: Top-40 Genes according to DIAMOnD Algorithm

Rank	Entrez Gene ID
1	56122
2	1198
3	23500
4	27350
5	27316
6	80742
7	3190
8	1153
9	91746
10	202559
11	5935
12	10656
13	26121
14	10657
15	3178
16	51067
17	3276
18	70439
19	10236
20	3183
21	8175
22	8452
23	3609
24	11338
25	1660
26	9782
27	56257
28	3608
29	55832
30	6625
31	22913
32	10492
33	10949
34	3192
35	1655
36	3575
37	6428
38	9775
39	23363
40	3187

Table 10: Table: GO Over-Represented Analysis for Putative Disease Proteins

Pathway Name	Pathway p-value (corrected)
poly(A) RNA binding	1,4E-142
gene expression	1,5E-115
viral transcription	4,7E-111
translational termination	2,5E-107
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1,7E-105
translational elongation	1,8E-102
SRP-dependent cotranslational protein targeting to membrane	1,3E-100
viral life cycle	2,36E-96
translational initiation	3,43E-93
RNA binding	4,46E-91

Table 11: Pathway Over-Represented Analysis for Putative Disease Proteins

Pathway Name	Pathway p-value (corrected)
Eukaryotic Translation Termination	1,29E-98
Peptide chain elongation	1,29E-98
Viral mRNA Translation	1,29E-98
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	3E-98
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	2,04E-97
Nonsense-Mediated Decay (NMD)	2,04E-97
Eukaryotic Translation Elongation	1,16E-96
Formation of a pool of free 40S subunits	1,31E-93
L13a-mediated translational silencing of Ceruloplasmin expression	2,16E-91
GTP hydrolysis and joining of the 60S ribosomal subunit	7,32E-89