Gene-expression-profiling of Breast Cancer

Presented by : Rashmi Nagpal
Btech' 14085

Introduction:

- Gene expression profiling : Genetic microarray analysis of gene-expression dataset.
- It details the expression levels of thousands of genes in breast cancer & draws molecular portraits of breast cancer.

Problem Statement:

- Network analysis of molecular interactomes associated with Breast Cancer and further doing their biological interpretation.

Milestones:

Microarray dataset of Breast Cancer

Step 1: **Create a network associative with a gene-expression dataset.**

- Expression matrix - Correlation Matrix as it gives similarity between genes as dataset comprises of cell lines
- Used threshold of 0.75 as well as 0.60 in correlation matrix and drew graphical representation and performed functions on both graphs.

Step2:

Network Analysis.

- Degree distribution.
- Shortest path length.
- Clustering coefficient of each node.
- Centrality between each node.
- Closeness centrality
- Average clustering coefficient.
- Degree correlation coefficient.
- Closeness centrality
- Identified hubs.
- Clustering Modularity of graph.

- Betweenness centrality : Measure of centrality in a graph based on shortest paths.

```
>>> bet = nx.betweenness_centrality(G)
>>> bet
{'CAPZA1': 0.0, 'PPIL1': 0.0, 'BCAP31': 0.380938286221696, 'TTC9C': 0.14163090128755365, '
0, 'HSDL2': 0.0, 'GLOD4': 0.0, 'ACAT2': 0.0, 'MRE11A': 0.0, 'VMA21': 0.0, 'COPS2': 0.0, 'YKT6': 0
```

- Number of edges and nodes :

```
>>> nx.number_of_nodes(G)
1492
>>> nx.number_of_edges(G)
19772
```

- Clustering coefficient of each node : Gives the number of triangles.

```
>>> cc = nx.clustering(G)
>>> cc
{'BUB3': 0.525, 'BCAP31': 0.6421052631578947, 'TTC9C': 0.7636363636363637,
7435897435898, 'DAD1': 0.5470085470085471, 'GLOD4': 0.5151515151515151,
```

- Clustering Modularity : It represents the structure of networks or graphs the ones with higher modularity have dense connections between nodes within modules.

```
>>> p = community.best_partition(G)
>>> mod = community.modularity(p,G)
>>> mod
0.41845319696328553
```

- Identification of hubs :

```
>>> nd = sorted(G.degree_iter(),key = itemgetter(1),reverse = True)
>>> nd[0]
('OXSR1', 494)
>>> nd[0][0]
'OXSR1'
>>> for i in range(10):
        print("Top 10 hubs are",nd[i][0])


Top 10 hubs are OXSR1
Top 10 hubs are IQGAP1
Top 10 hubs are POLR2E
Top 10 hubs are RUVBL2
Top 10 hubs are SSR1
Top 10 hubs are KPNA3
Top 10 hubs are GLT25D1
Top 10 hubs are UBL4A
Top 10 hubs are TXLNA
Top 10 hubs are PGM1
```

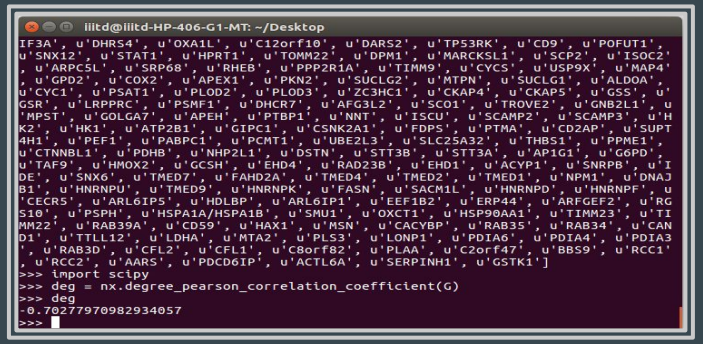
- Closeness centrality: Degree to which an individual is near to all other individuals in the network.

```
>>> close = nx.closeness_centrality(G)
>>> close
{'BUB3': 0.39738805970149255, 'BCAP31': 0.40626702997275205
ABCF1': 0.3924717030797578, 'DR1': 0.4010220548682087, 'TBC
1': 0.4082694414019715, 'PDAP1': 0.3556774809160305, 'PAFAI
```

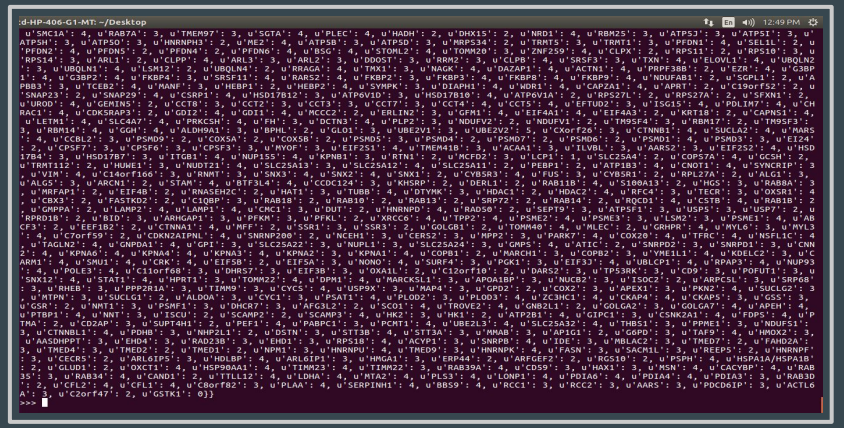
- Average clustering coefficient:

```
>>> avg = sum(cc.values())/len(cc)
>>> avg
0.5697701750637838
>>>
```

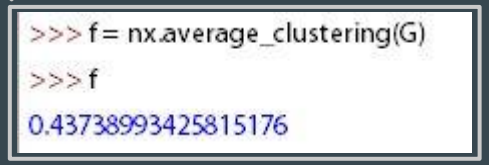
- Degree correlation coefficient :



- Shortest path length:



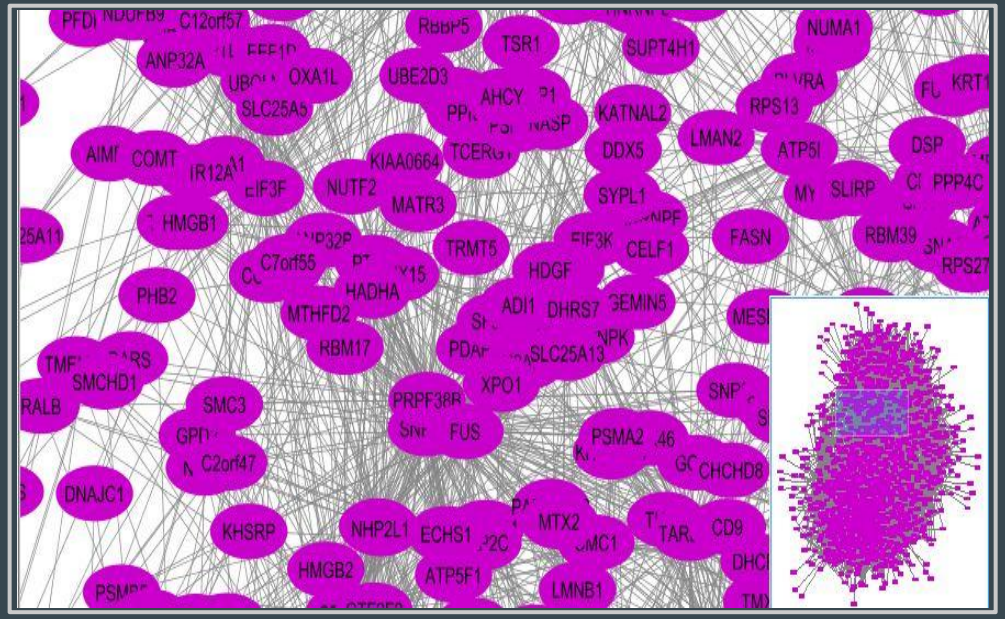
- Average clustering coefficient :



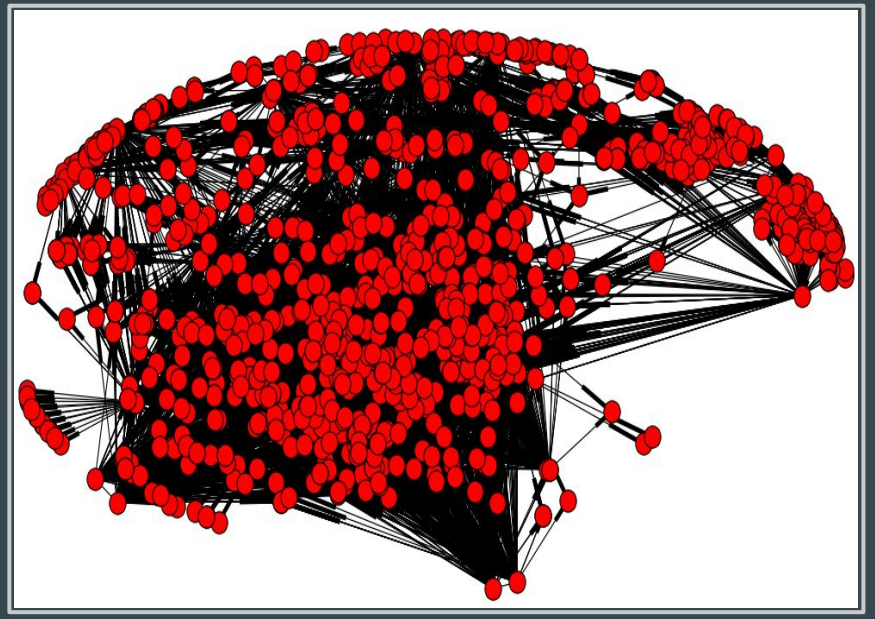
Step 3:

Network Visualization

- Used cytoscape and networkx tools to visualize network.



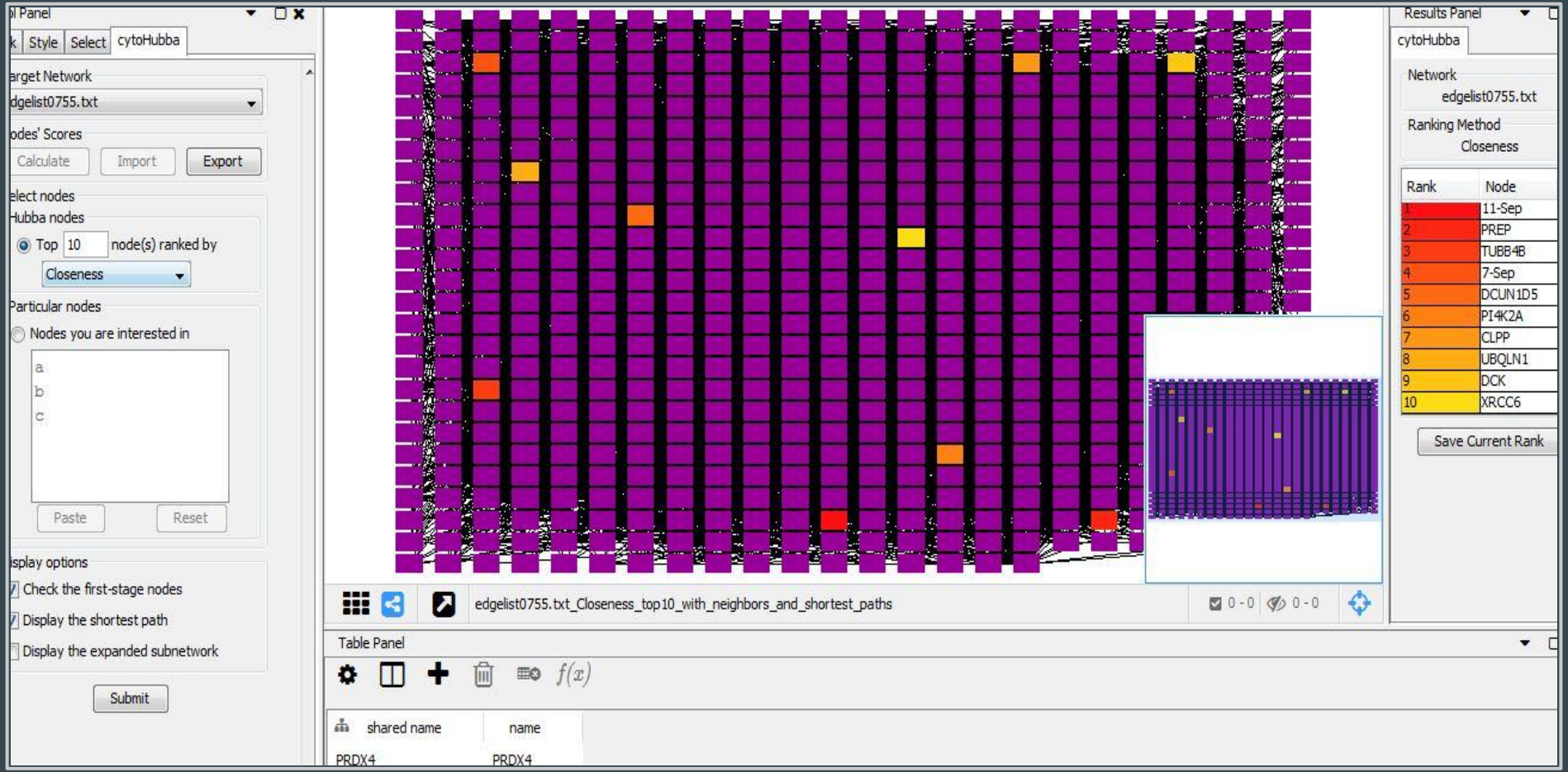
Using Cytoscape



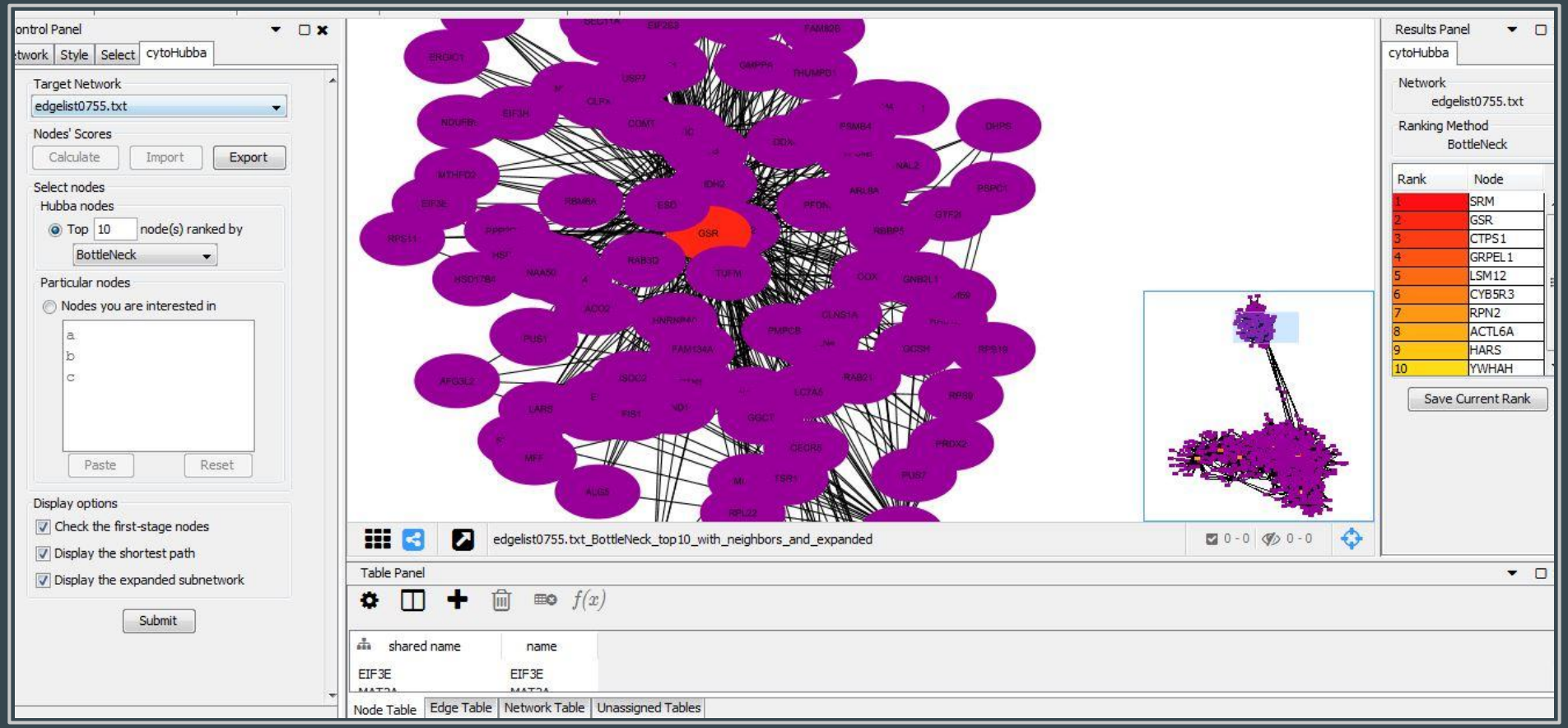
Using Networkx

Important Terms:

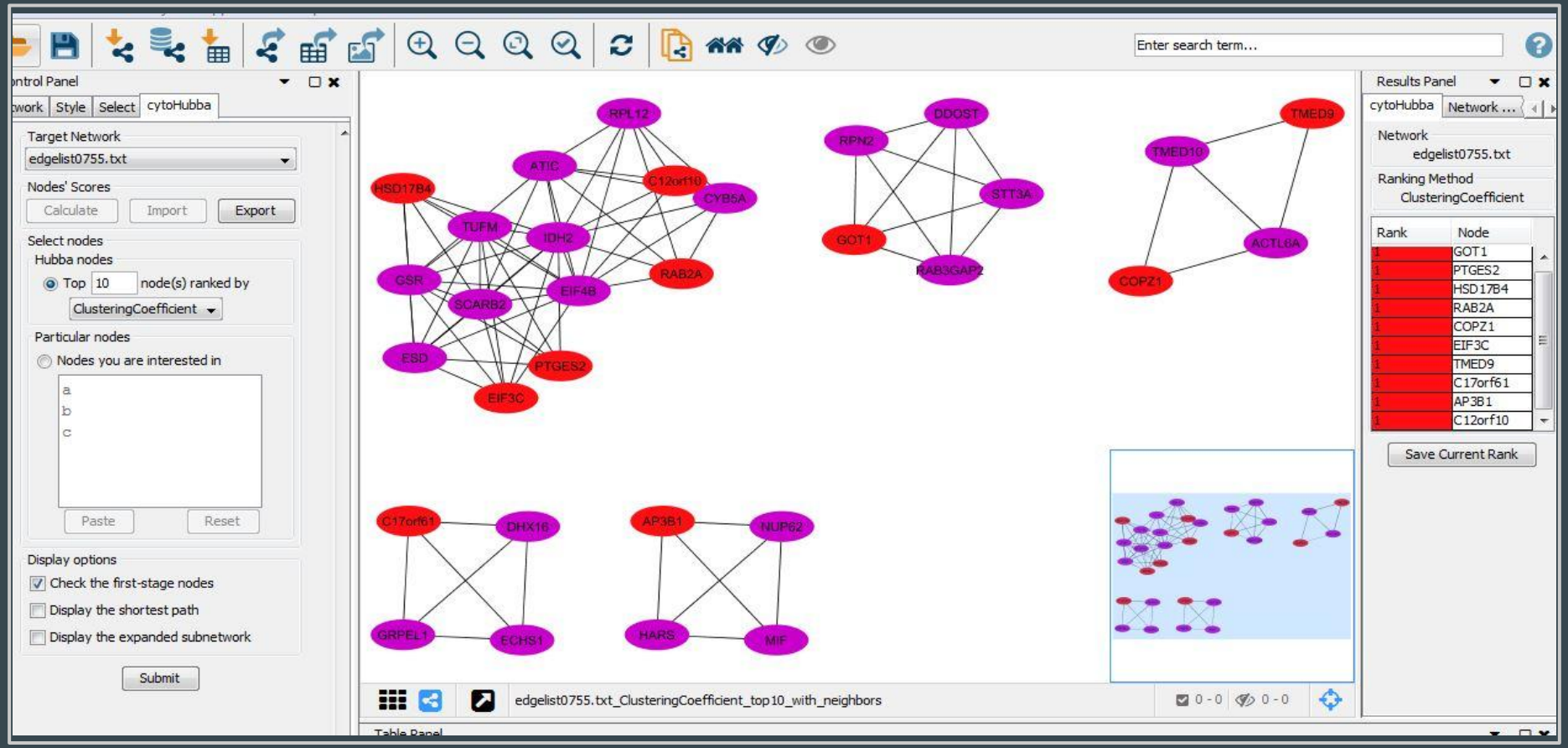
- Closeness Centrality: It is the reciprocal of sum of shortest path distances from u to all n-1 other nodes.Higher values of closeness indicates higher centrality.
- Bottleneck algorithm: In this method, for each node weight is defined, wherein weight is the number of nodes in the shortest path starting from node v where v is root node and passing through w.Hence, w is defined as bottleneck node in T_v
- DMNC algorithm: The score of node v is defined as the ratio of E/N by assuming that node belongs to maximum neighborhood component with a strong community structure.



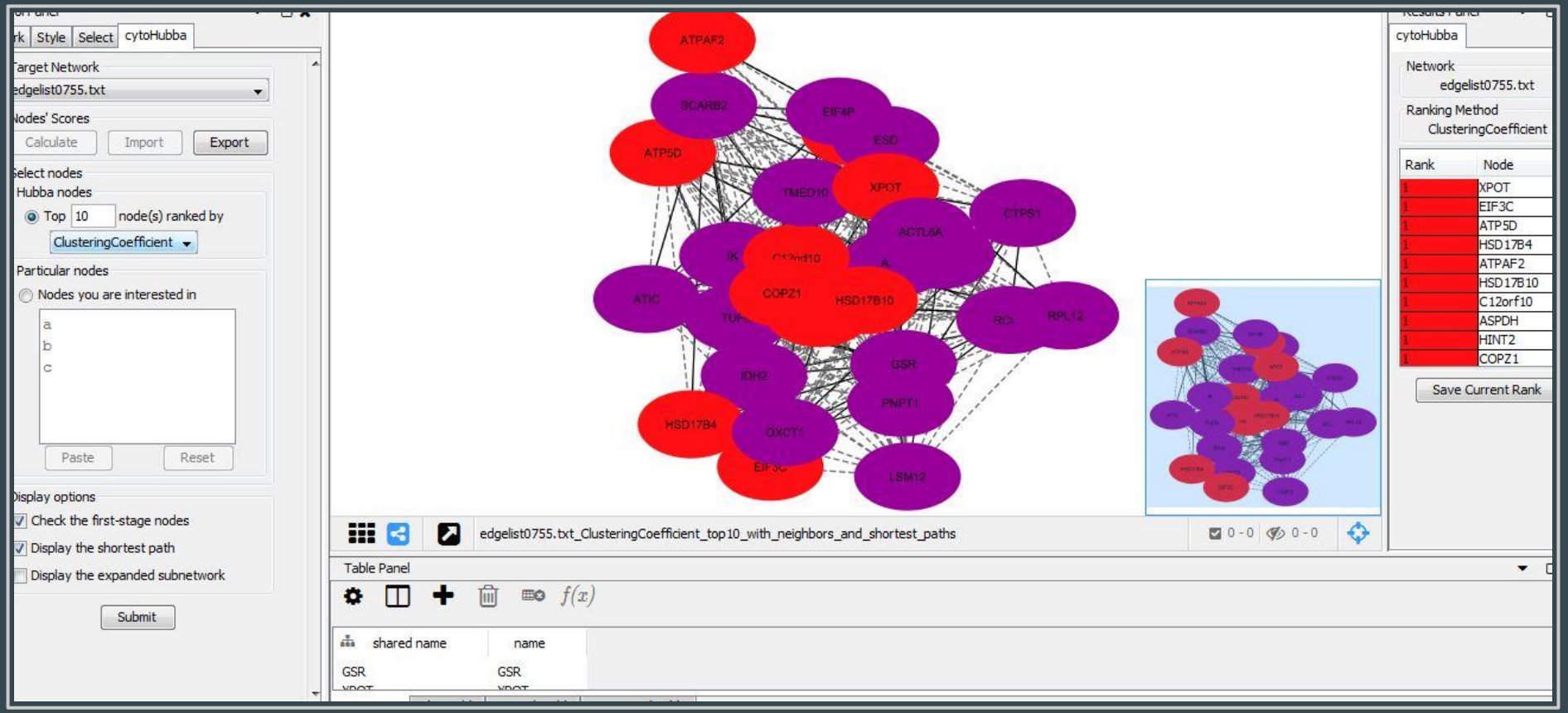
- Top 10 nodes on the basis of closeness and analysed it using cytohubba tool.



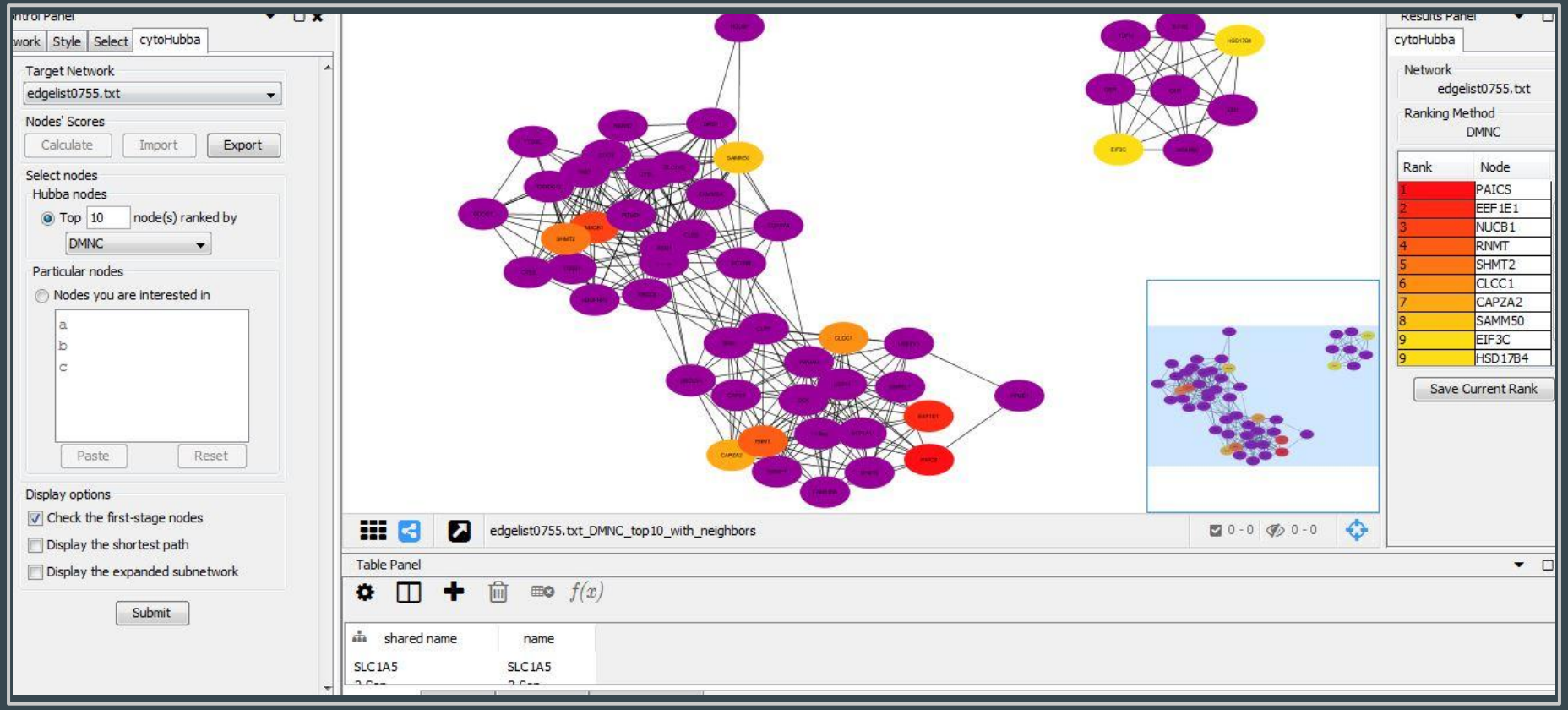
- Used Bottleneck topological feature.



- On the basis of clustering coefficient, ranked top 10 nodes using built - in tool.



- Above graph depicts the nodes with similar clustering coefficient and their shortest path length.

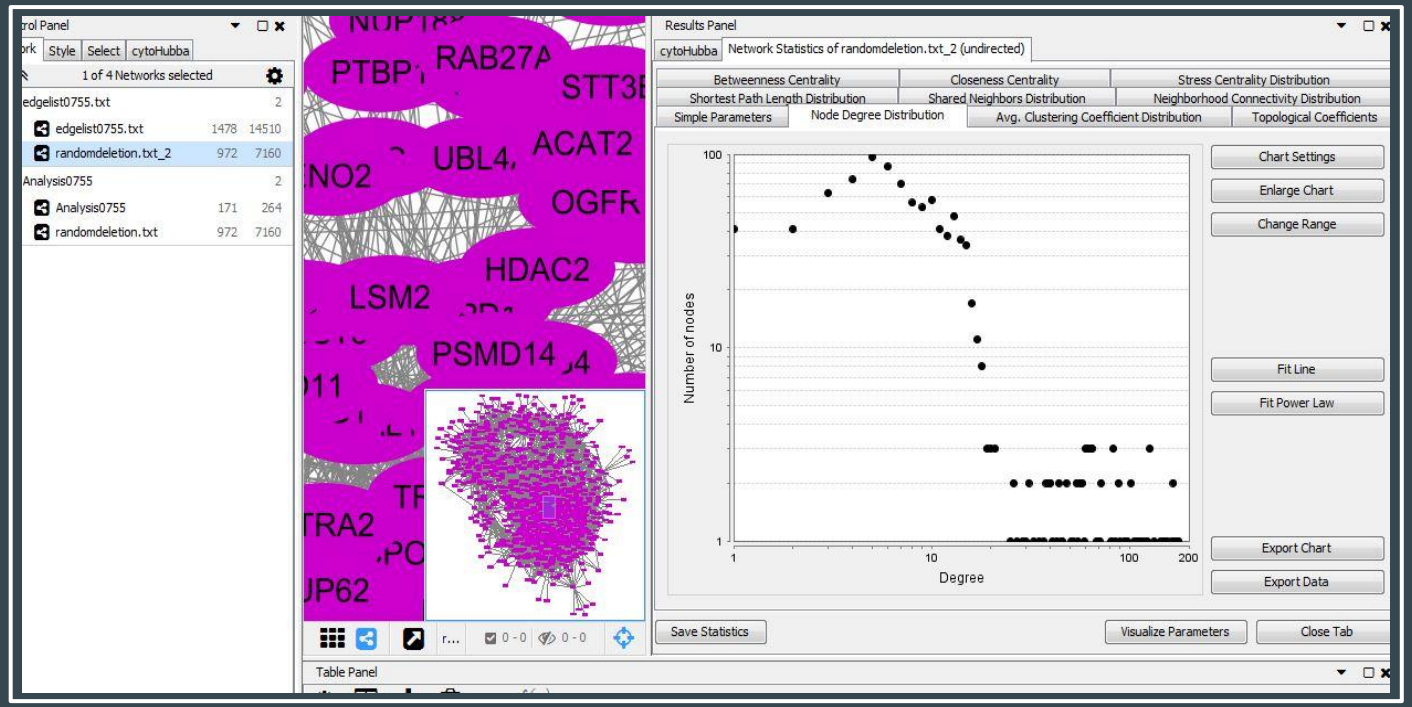


- Used DMNC algorithm to depict nodes of maximum nearest neighbor.

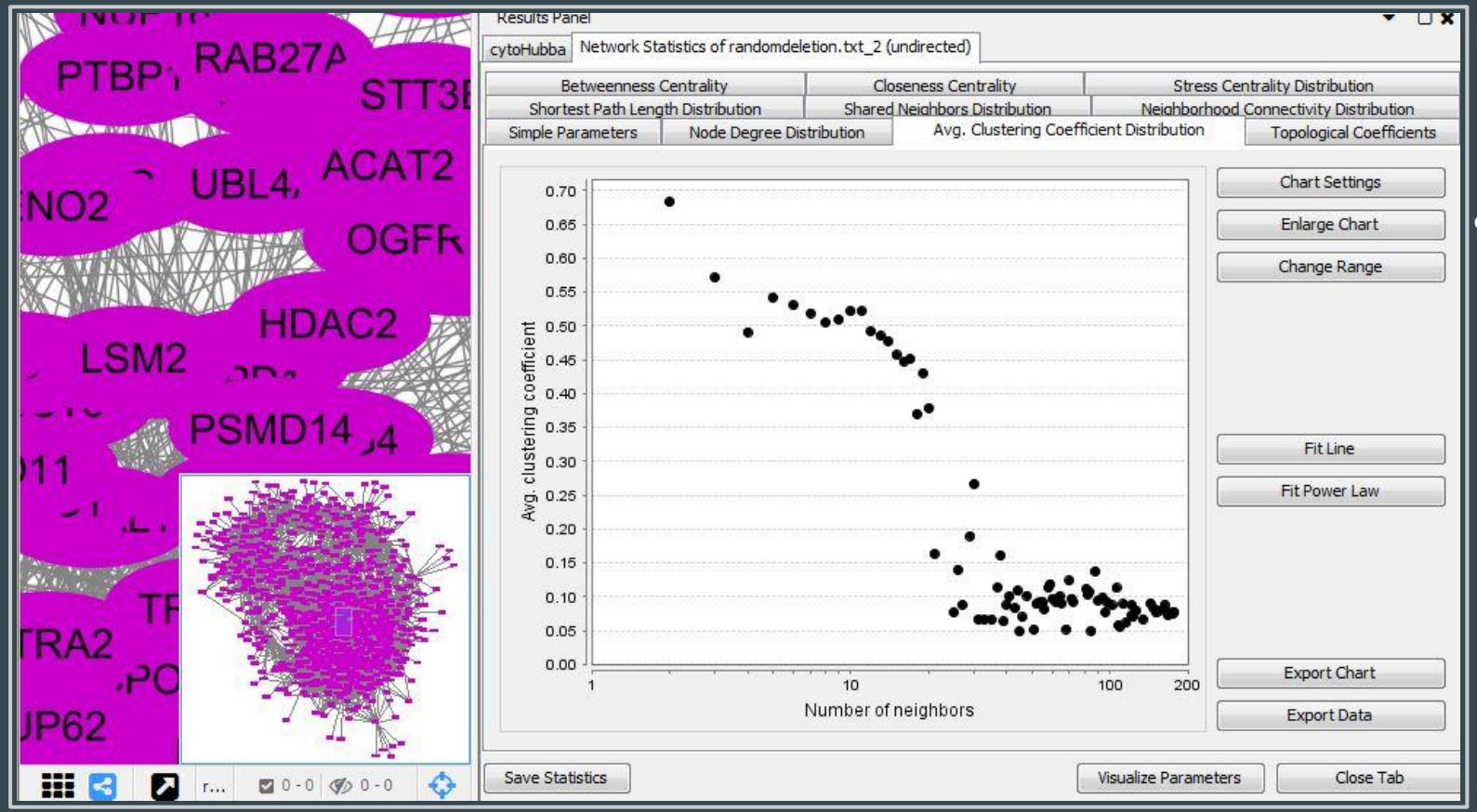
Step4:

Node Deletion

Random Deletion



- Follws power law distribution
- Some have extreme degrees and there is peak downfall for others.



- Most of the nodes have low clustering coefficient as compared to it's scale free counterpart.

Clustering coefficient by random deletion of nodes..

Parameters :

- Size of giant cluster:

```
>>> gcc = sorted(nx.connected_component_subgraphs(G),key = len, reverse = True)
>>> gcc[0]
<networkx.classes.graph.Graph object at 0x0000000007426DD8>
>>> len(gcc[0])
1191
```

- List of isolated vertex:

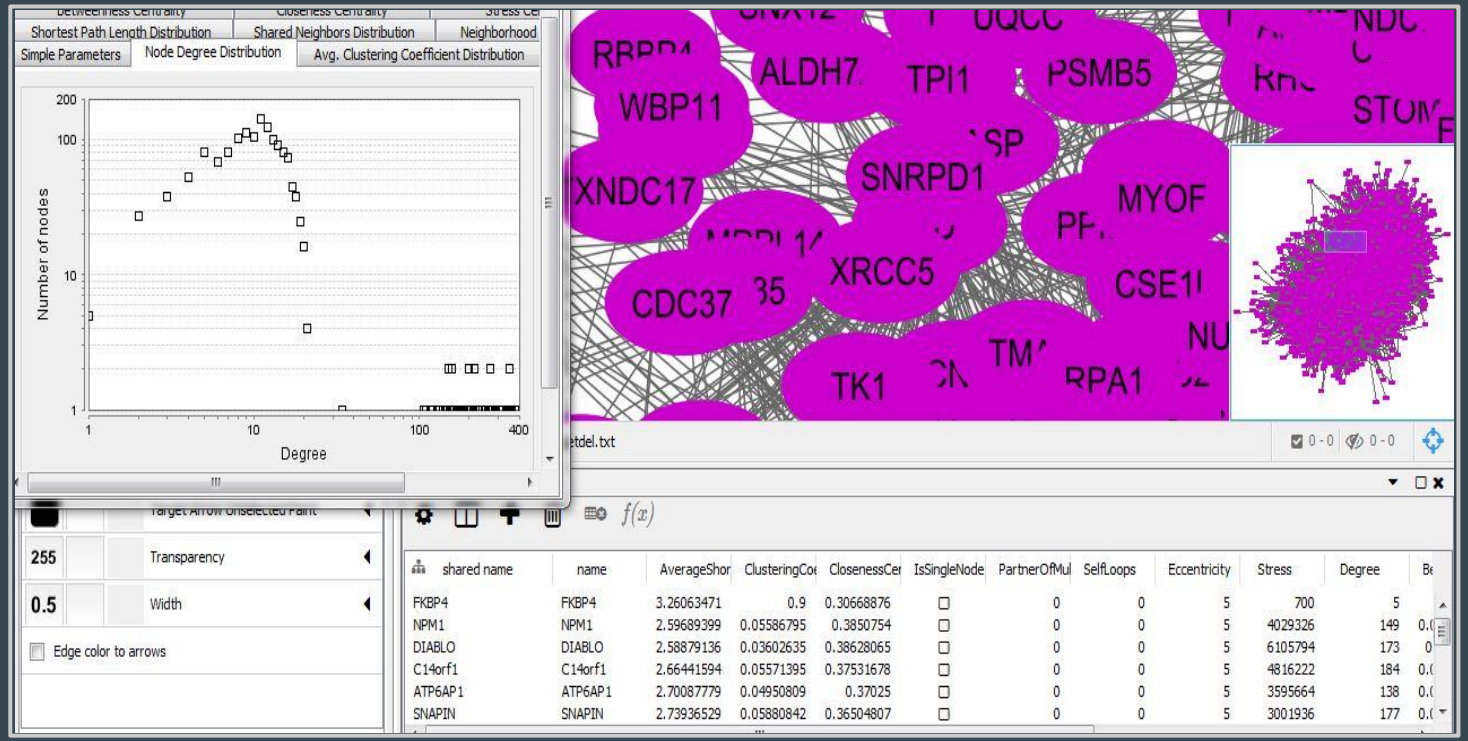
```
>>> ls = nx.isolates(G)
>>> ls
['STUB1']
>>> type(ls)
<class 'list'>
```

- Shortest Path Length:

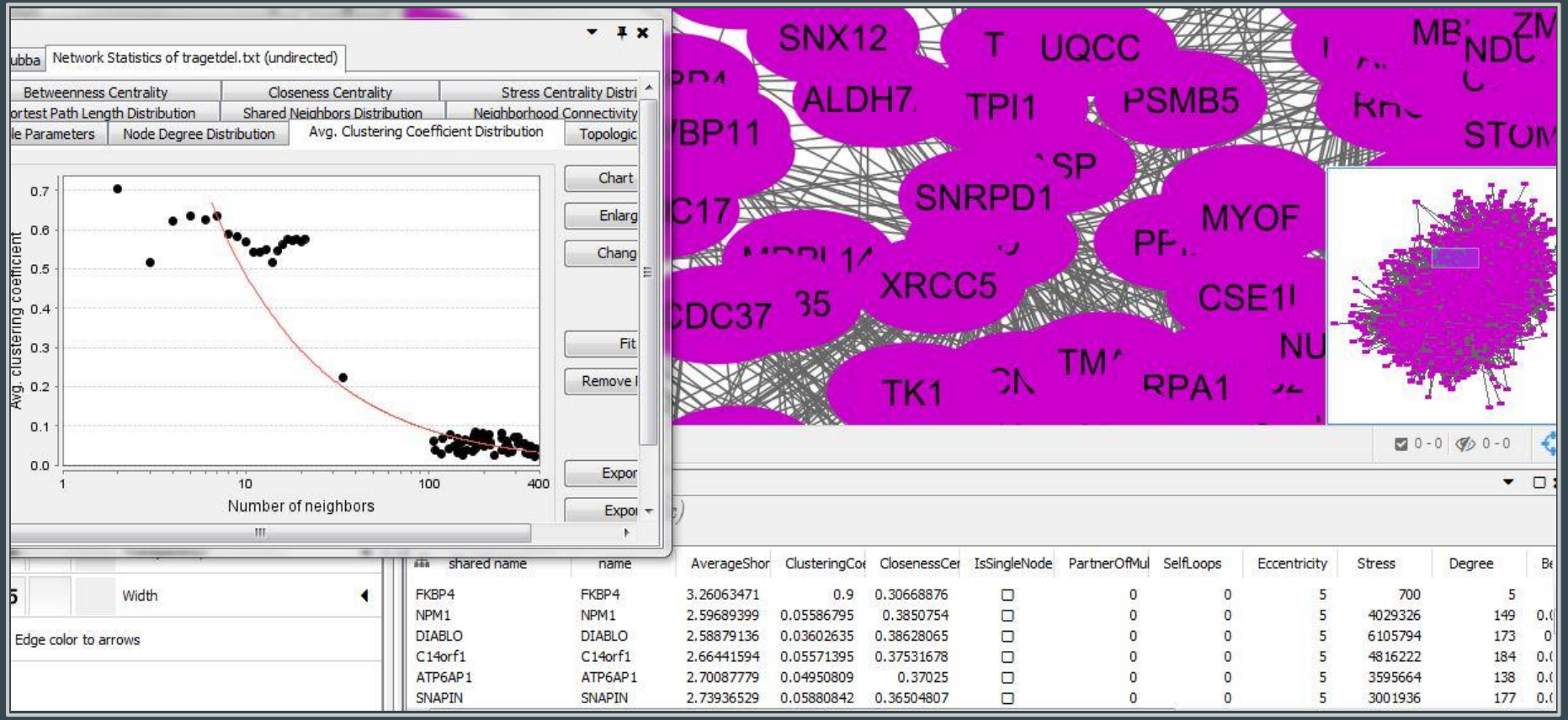
```
>>> sh = nx.average_shortest_path_length(G)

>>> sh
2.7483731624482513
```

Step4: Node Deletion

Targeted Deletion



- Some have extreme degrees and there is peak downfall of degrees when hubs are deleted.
- There is a point of fc where no more hubs are present.



Clustering Coefficient distribution of targeted attacks on gene dataset.

Parameters :

- Size of giant cluster:
- List of isolated vertex:
- Shortest Path Length:

```
>>> len(G.nodes())
1482
>>> sh = nx.average_shortest_path_length(G)
>>> sh
2.76992147954158
>>> gcc = sorted(nx.connected_component_subgraphs(G),key = len, reverse = True)
>>> gcc[0]
<networkx.classes.graph.Graph object at 0x0000000004DAF080>
>>> len(gcc[0])
1482
>>> ls = nx.isolates(G)
>>> ls
[]
```

NB : Hence, in case of targeted attacks, degree distribution falls at steeper rate in comparison to random deletion counterpart.Hence, this biological network too is an example of scale-free graph.

Step5:

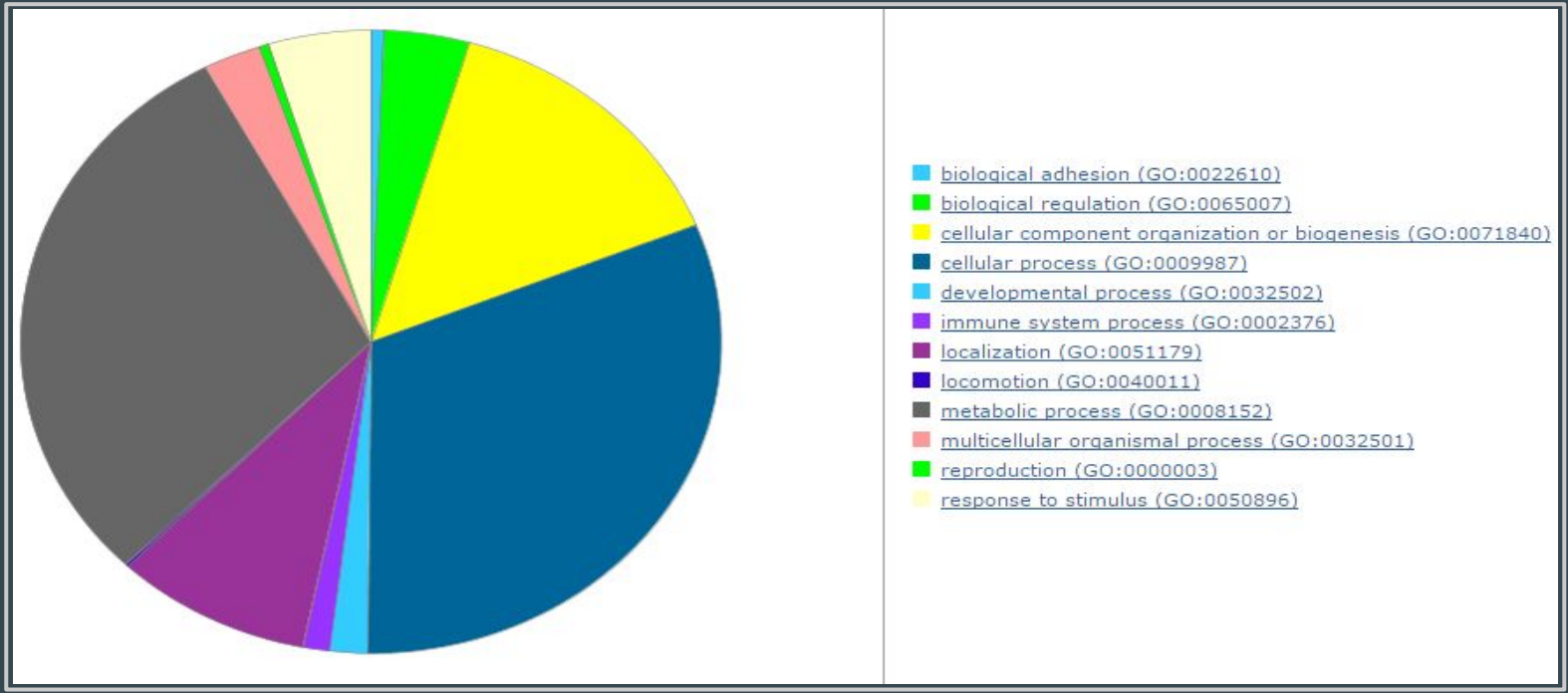
Gene Ontological Enrichment Analysis

GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
9987	cellular process	4.9078E-10	5.0158E-7	143/147 97.2%	4978/6204 80.2%	VPS29 RPL5 RPL30 RNH1 NUP188 ECI1 SRP14 ENO2 SMC3 SMC4 PHB2 SMC2 NFU1 LSM12 NRD1..
6457	protein folding	3.4583E-7	1.7672E-4	14/147 9.5%	113/6204 1.8%	CCT3 CCT2 PHB2 EMC1 TCP1 EMC2 EMC3 PIN1 EMC4 CCT8 PLP2 CCT7 CCT5 CCT4
44085	cellular component biogenesis	1.0386E-6	3.5381E-4	44/147 29.9%	903/6204 14.5%	RPL5 RPL30 NUP188 RPN2 MCM7 TAF9 RPN1 USO1 SRP54 ARL2 SRP14 SMC3 XPO1 SRP72 NUP8..
70972	protein localization in endoplasmic reticulum	2.1995E-6	5.6197E-4	8/147 5.4%	38/6204 0.6%	RER1 SRP72 SRP54 SRP68 SEC62 SRP14 SIL1 SEC63
33365	protein localization in organelle	3.7776E-6	7.0573E-4	15/147 10.2%	157/6204 2.5%	RTN1 NUP188 PAM16 SRP54 SRP68 SRP14 XPO1 RER1 SRP72 NUP85 FAR1 VPS35 SEC62 SIL1 S..
31365	N-terminal protein amino acid modification	4.8338E-6	7.0573E-4	5/147 3.4%	12/6204 0.1%	NAA50 NAA10 NMT1 NAA25 NAA38
18409	peptide or protein amino-terminal blocking	4.8338E-6	7.0573E-4	5/147 3.4%	12/6204 0.1%	NAA50 NAA10 NMT1 NAA25 NAA38
45047	protein targeting to ER	7.4443E-6	8.5354E-4	7/147 4.7%	32/6204 0.5%	SRP72 SRP54 SRP68 SEC62 SRP14 SIL1 SEC63
34621	cellular macromolecular complex subunit organization	7.5854E-6	8.5354E-4	27/147 18.3%	459/6204 7.3%	RPL5 NUP188 RPN2 MCM7 TAF9 RPN1 USO1 SRP54 SRP14 SRP72 NRD1 NUP85 RPL38 RPS27A U..
6614	SRP-dependent cotranslational protein targeting to membrane	8.7160E-6	8.5354E-4	6/147 4.0%	22/6204 0.3%	SRP72 SRP54 SRP68 SRP14 SIL1 SEC63
51641	cellular localization	9.4529E-6	8.5354E-4	37/147 25.1%	757/6204 12.2%	VPS29 ARF1 RTN1 NUP188 ARL3 USO1 SRP54 ARL2 SRP14 SMC4 PHB2 SMC2 SNX3 NFU1 XPO1 ...
6617	SRP-dependent cotranslational protein targeting to membrane, signal seque...	1.0022E-5	8.5354E-4	4/147 2.7%	7/6204 0.1%	SRP72 SRP54 SRP68 SRP14
6613	cotranslational protein targeting to membrane	1.1563E-5	8.5834E-4	6/147 4.0%	23/6204 0.3%	SRP72 SRP54 SRP68 SRP14 SIL1 SEC63
84	S phase of mitotic cell cycle	1.1758E-5	8.5834E-4	5/147 3.4%	14/6204 0.2%	MCM7 MCM3 MCM5 MCM6 MCM2
6461	protein complex assembly	1.3926E-5	8.8953E-4	18/147 12.2%	242/6204 3.9%	RPN2 MCM7 TAF9 RPN1 USO1 CMC1 SRP54 SRP68 SRP14 SCO1 SRP72 SUB1 VMA21 MCM3 ACP..
70271	protein complex biogenesis	1.3926E-5	8.8953E-4	18/147 12.2%	242/6204 3.9%	RPN2 MCM7 TAF9 RPN1 USO1 CMC1 SRP54 SRP68 SRP14 SCO1 SRP72 SUB1 VMA21 MCM3 ACP..
51320	S phase	1.7302E-5	9.9247E-4	5/147 3.4%	15/6204 0.2%	MCM7 MCM3 MCM5 MCM6 MCM2
43933	macromolecular complex subunit organization	1.8049E-5	9.9247E-4	29/147 19.7%	538/6204 8.6%	RPL5 NUP188 RPN2 MCM7 TAF9 RPN1 USO1 SRP54 SRP14 SRP72 NRD1 NUP85 ACP2 RPL38 BUB..
33036	macromolecule localization	1.9969E-5	9.9247E-4	35/147 23.8%	720/6204 11.6%	VPS29 ARF1 RTN1 NUP188 GDI1 ARL3 USO1 SRP54 ARL2 SRP14 SNX3 XPO1 SRP72 NUP85 RPS3..
22616	DNA strand elongation	2.0618E-5	9.9247E-4	7/147 4.7%	37/6204 0.5%	FEN1 RFC4 MCM7 MCM3 MCM5 MCM6 MCM2
6271	DNA strand elongation involved in DNA replication	2.0618E-5	9.9247E-4	7/147 4.7%	37/6204 0.5%	FEN1 RFC4 MCM7 MCM3 MCM5 MCM6 MCM2
34622	cellular macromolecular complex assembly	2.1364E-5	9.9247E-4	21/147 14.2%	324/6204 5.2%	RPL5 RPN2 MCM7 TAF9 RPN1 USO1 CMC1 SRP54 SRP68 SRP14 SCO1 SRP72 SUB1 VMA21 MCM...
45184	establishment of protein localization	2.3885E-5	1.0513E-3	29/147 19.7%	546/6204 8.8%	VPS29 ARF1 RTN1 NUP188 GDI1 ARL3 USO1 SRP54 ARL2 SRP14 SNX3 XPO1 SRP72 NUP85 VPS3..
6267	pre-replicative complex assembly	2.4688E-5	1.0513E-3	5/147 3.4%	16/6204 0.2%	MCM7 MCM3 MCM5 MCM6 MCM2
46907	intracellular transport	2.8361E-5	1.1594E-3	31/147 21.0%	610/6204 9.8%	VPS29 ARF1 RTN1 NUP188 ARL3 USO1 SRP54 ARL2 SRP14 SNX3 XPO1 SRP72 NUP85 RPS3 VPS3..
8104	protein localization	3.2245E-5	1.2675E-3	31/147 21.0%	614/6204 9.8%	VPS29 ARF1 RTN1 NUP188 GDI1 ARL3 USO1 SRP54 ARL2 SRP14 SNX3 XPO1 SRP72 NUP85 VPS3..
34975	protein folding in endoplasmic reticulum	3.4763E-5	1.2688E-3	4/147 2.7%	9/6204 0.1%	EMC1 EMC2 EMC3 EMC4
6474	N-terminal protein amino acid acetylation	3.4763E-5	1.2688E-3	4/147 2.7%	9/6204 0.1%	NAA50 NAA10 NAA25 NAA38
15031	protein transport	4.2827E-5	1.4872E-3	27/147 18.3%	505/6204 8.1%	VPS29 ARF1 RTN1 NUP188 GDI1 ARL3 USO1 SRP54 ARL2 SRP14 SNX3 XPO1 SRP72 NUP85 VPS3..
18193	peptidyl-amino acid modification	4.3655E-5	1.4872E-3	11/147 7.4%	108/6204 1.7%	NAA50 FEN1 NAA10 PPA1 ALG5 NMT1 NAA25 NAA38 DTD1 SEC63 LSM2
724	double-strand break repair via homologous recombination	4.8926E-5	1.6075E-3	7/147 4.7%	42/6204 0.6%	MCM7 RPA1 MCM3 RPA2 MCM5 MCM6 MCM2
6886	intracellular protein transport	5.0332E-5	1.6075E-3	20/147 13.6%	317/6204 5.1%	ATG3 UPF1 RTN1 NUP188 MDH2 ARL3 USO1 PAM16 SRP54 ARL2 SRP68 SRP14 XPO1 SRP72 NU...
34613	cellular protein localization	5.2763E-5	1.6341E-3	22/147 14.9%	371/6204 5.9%	ATG3 UPF1 RTN1 NUP188 MDH2 ARL3 USO1 PAM16 SRP54 ARL2 SRP68 SRP14 XPO1 RER1 SRP...
65003	macromolecular complex assembly	7.6988E-5	2.3142E-3	23/147 15.6%	408/6204 6.5%	RPL5 RPN2 MCM7 TAF9 RPN1 USO1 CMC1 SRP54 SRP68 SRP14 SCO1 SRP72 SUB1 VMA21 MCM...
6268	DNA unwinding involved in replication	8.7729E-5	2.5101E-3	4/147 2.7%	11/6204 0.1%	MCM7 RPA1 RPA2 MCM6
70727	cellular macromolecule localization	8.8419E-5	2.5101E-3	22/147 14.9%	384/6204 6.1%	ATG3 UPF1 RTN1 NUP188 MDH2 ARL3 USO1 PAM16 SRP54 ARL2 SRP68 SRP14 XPO1 RER1 SRP...
6091	generation of precursor metabolites and energy	1.1923E-4	3.2934E-3	17/147 11.5%	259/6204 4.1%	TPI1 MDH1 MDH2 IDH1 IDH2 COX5B ENO2 COX5A ADH5 PPA1 PGK1 COX2 ACO2 CYC1 ACP1 PG..

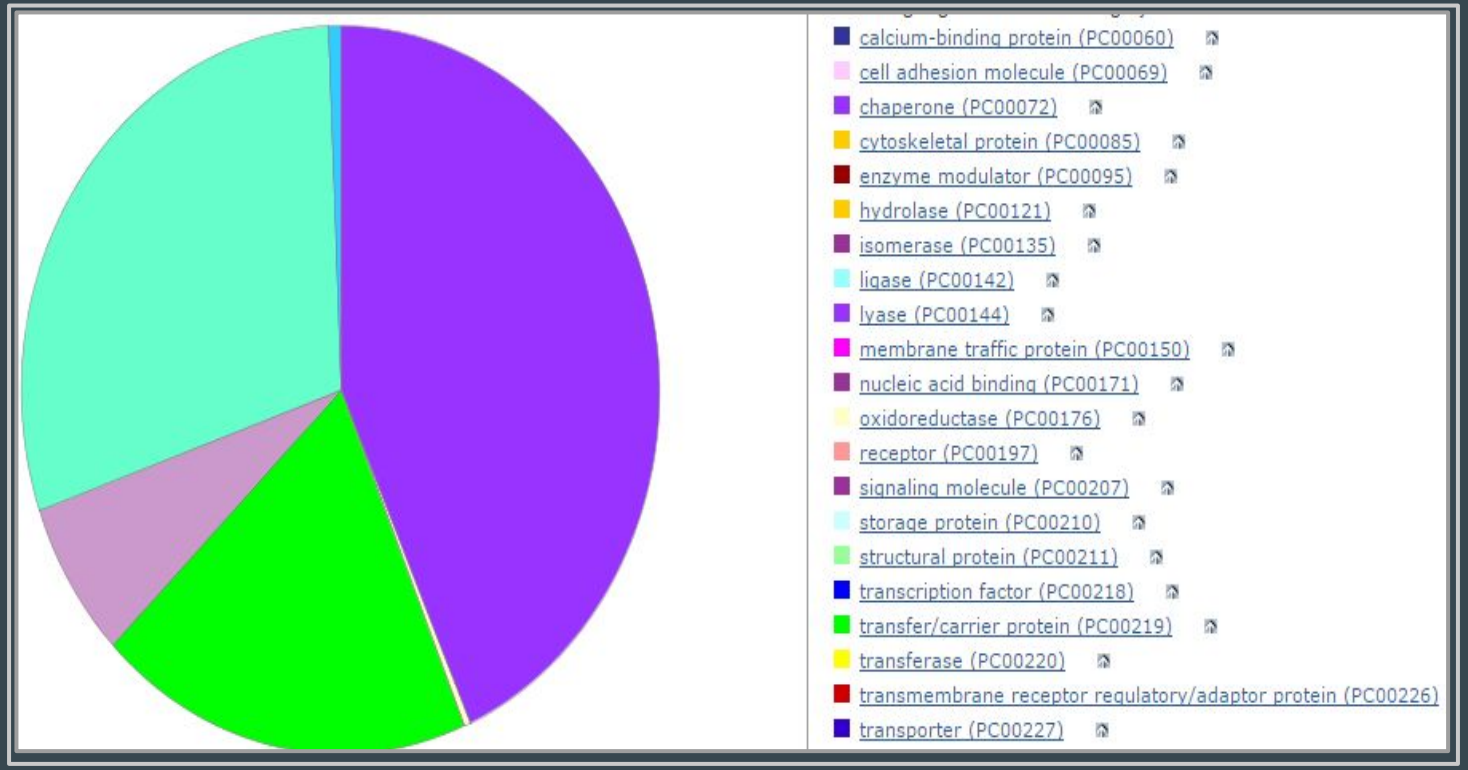
Step6:

Biological Interpretation of nodes

- Using panther

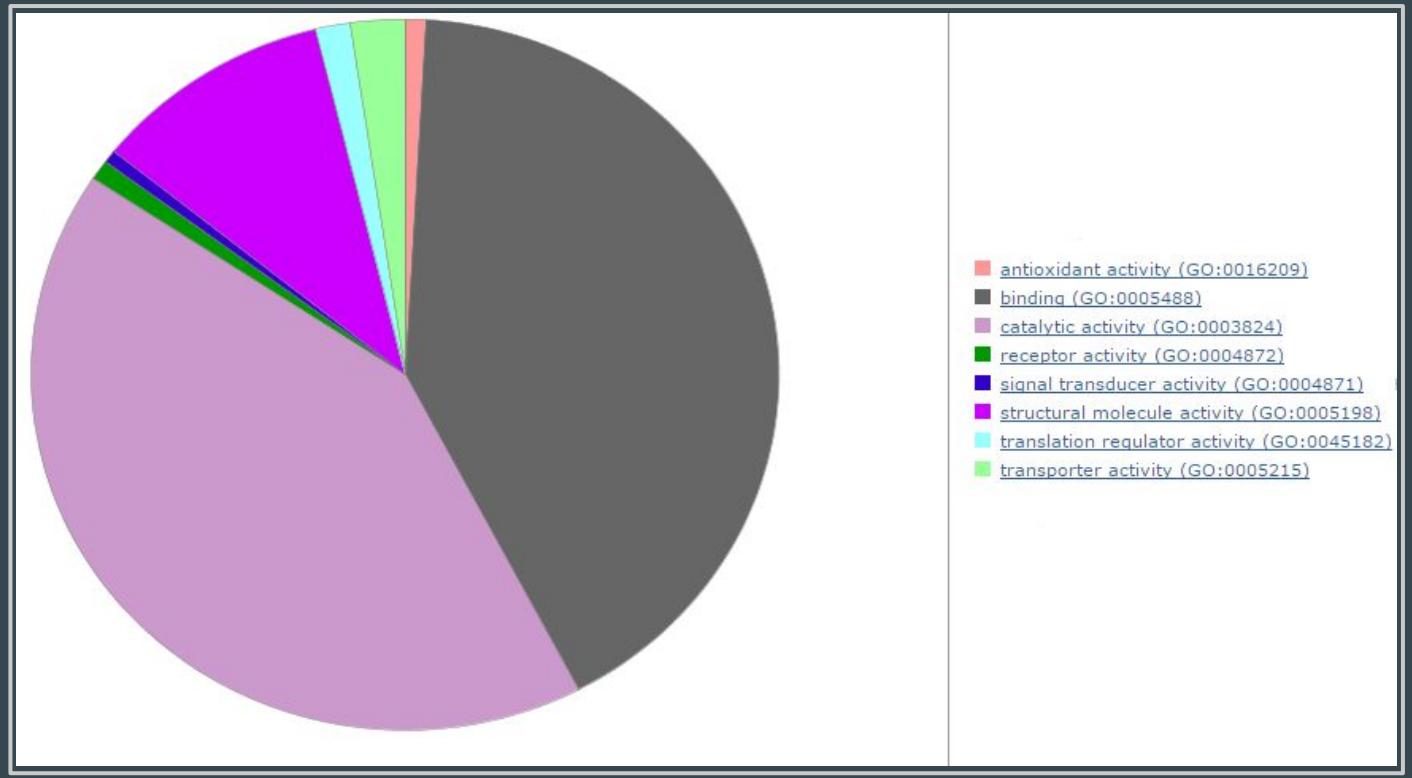


Biological analysis.



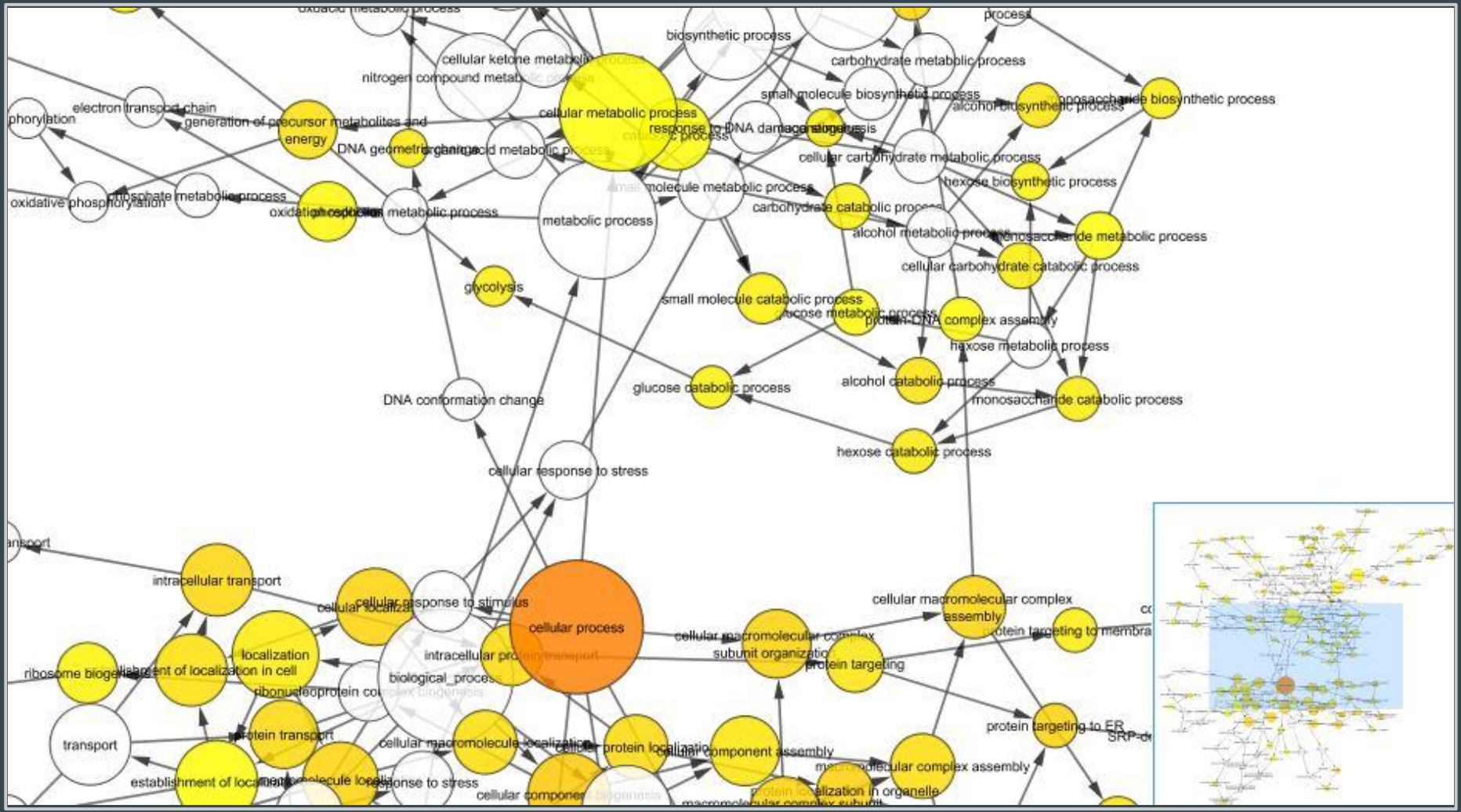
Cellular analysis.

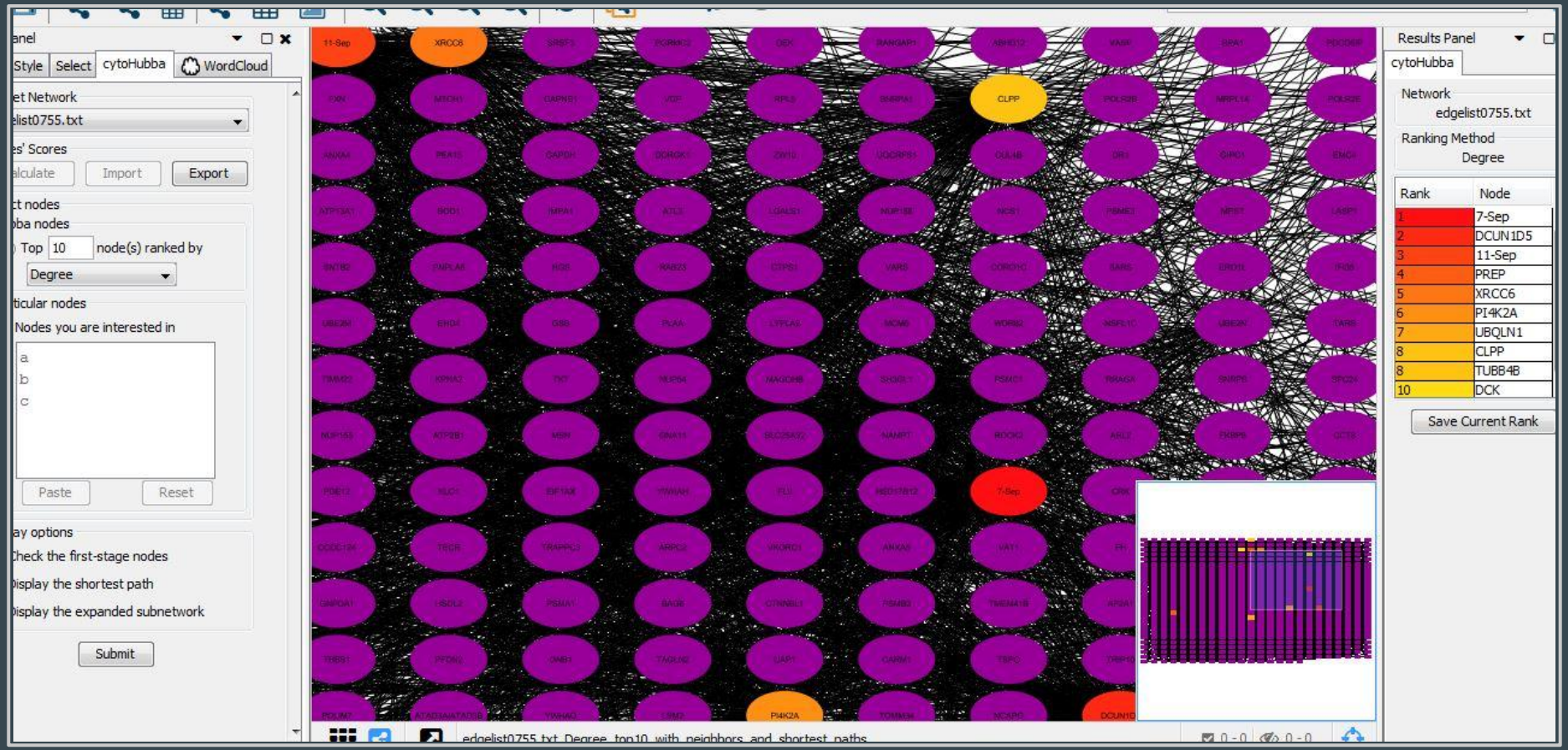
- Majority of breast cancer genes possess “lyase” which is an enzyme which catalyzes the joining of specific molecule.
- Also, there exists transfer/carrier protein cellular function in most of genes.



Catalytic activity

- Most of the genes possess similar catalytic activity like PPIB, CTSB,GSPT1, TXNDC17 to name a few.


biosynthetic process
carbohydrate metabolic process
cellular metabolic process
small molecule biosynthetic process
monosaccharide biosynthetic process
alcohol biosynthetic process
electron transport chain
generation of precursor metabolites and energy
response to DNA damage stimulus
cellular carbohydrate metabolic process
hexose biosynthetic process
oxidative phosphorylation
phosphate metabolic process
oxidation reduction
metabolic process
small molecule metabolic process
carbohydrate catabolic process
alcohol metabolic process
monosaccharide metabolic process
cellular carbohydrate catabolic process
glycolysis
small molecule catabolic process
protein-DNA complex assembly
hexose metabolic process
DNA conformation change
glucose catabolic process
alcohol catabolic process
monosaccharide catabolic process
hexose catabolic process
cellular response to stress
intracellular transport
cellular response to stimulus
cellular process
cellular macromolecular complex assembly
localization
intracellular protein transport
cellular macromolecular complex subunit organization
protein targeting
ribosome biogenesis
establishment of localization in cell
biological_process
protein targeting to ER
transport
protein transport
cellular macromolecule localization
cellular component assembly
establishment of localization
response to stress
protein localization in organelle



- Top 10 nodes which are central to network

Biological Significance of central nodes:

- **PREP :** Also known as prolyl endopeptidase which is involved in maturation and degradation of peptide hormones.
- **XRCC6 :** X-Ray Repair Cross Complementing 6 is a protein coding gene which is single stranded DNA dependent ATP-dependent helicase.
- **TUBB4B:** Its a protein coding gene which is a major constituent of microtubules.
- **dCK:** deoxycytidine is an enzyme which is encoded by DCK gene in humans.It plays vital role in drug resistance and sensitivity especially in cancer.
- **UBQLN1:** Ubiquiline1 plays a vital role in regulation of protein degradation mechanisms and pathways.GO annotations included with this gene is identical protein binding and protein domain specific binding.