NTNU

Norwegian University of
Science and Technology

Applying Systems Biology to Give Lone Genes a Meaning

Building and analyzing a network around 25 genes

Marie Verdonck

Course BI3019

Group 7

In collaboration with

Emmy Wang

Kamil Szura

Thea Johanna Aandstad Hettasch

Content Page

List of abbreviations

ABCA1	ABC-binding cassette transporter 1
ABI1	Abi interactor 1
ACVR1B	Activin receptor type 1B
APAF1	Apoptotic Peptidase Activating Factor 1
APOE	Apolipoprotein E
BAD	BCL2 Associated Agonist Of Cell Death
BAK	BCL2 Antagonist/Killer 1
BAX	BCL2 Associated X
BBC3	BCL2 Binding Component 3
BCL2	BCL2 Apoptosis Regulator
BCL2L1	BCL2 Like 1
BCL2L11	BCL2 Like 11
BID	BH3 Interacting Domain Death Agonist
BIRC5	Baculoviral IAP repeat containing 5
CASP10	Caspase 10
CASP3	Caspase 3
CASP6	Caspase 6
CASP7	Caspase 7
CASP8	Caspase 8
CASP9	Caspase 9
CDKN1A	Cyclin Dependent Kinase Inhibitor 1A
CMS2	Consensus molecular subtype 2
CRC	Colorectal cancer
CSF1R	Colony stimulating factor 1 receptor
CYCS	Cytochrome C

DISC	Death Inducing Signaling Complex
DUSP1	Dual Specificity Phosphatase 1
DUSP3	Dual Specificity Phosphatase 3
DUSP6	Dual Specificity Phosphatase 6
DUSP9	Dual Specificity Phosphatase 9
ETS1	ETS Proto-Oncogene 1
FADD	Fas Associated Via Death Domain
FAS	Fas Cell Surface Death Receptor
FDR	False Discovery Rate
GO	Gene Ontology
GRB2	Growth Factor Receptor Bound Protein 2
ICAM1	Intercellular adhesion molecule 1
IKBKB	Inhibitor Of Nuclear Factor Kappa B Kinase Subunit Beta
IL10	Interleukin 10
IL10RA	Interleukin 10 Receptor Subunit Alpha
IL10RB	Interleukin 10 Receptor Subunit Beta
IL13	Interleukin 13
IL4	Interleukin 4
IL4RA	Interleukin 4 Receptor
IL6	Interleukin 6
IRAK1	Interleukin 1 Receptor Associated Kinase 1
IRAK4	Interleukin-1 Receptor-Associated Kinase 4
IRF1	Interferon regulatory factor 1
JAK1	Janus Kinase 1
JAK2	Janus Kinase 2
JUN	AP-1 Transcription Factor
MAP2K1	Mitogen-Activated Protein Kinase Kinase 1

MAP2K2	Mitogen-Activated Protein Kinase Kinase 2
MAP2K3	Mitogen-activated protein kinase kinase 3
MAP2K4	Mitogen-Activated Protein Kinase Kinase 4
MAP2K6	Mitogen-Activated Protein Kinase Kinase 6
MAP2K7	Mitogen-activated protein kinase kinase 7
MAP3K1	Mitogen-Activated Protein Kinase Kinase Kinase 1
MAP3K11	Mitogen-Activated Protein Kinase Kinase Kinase 11
MAP3K12	Mitogen-Activated Protein Kinase Kinase Kinase 12
MAP3K3	Mitogen-Activated Protein Kinase Kinase Kinase 3
MAP4K1	Mitogen-Activated Protein Kinase Kinase Kinase Kinase 1
MAPK14	Mitogen-Activated Protein Kinase 14
MAPK3	Mitogen-activated protein kinase 3
MAPK8	Mitogen-Activated Protein Kinase 8
MMP1	Matrix Metallopeptodase 1
MYC	MYC Proto-Oncogene
MYD88	MYD88 Innate Immune Signal Transduction Adaptor
NFKB1	Nuclear Factor Kappa B Subunit 1
P	Protein
PPI	Protein Protein Interactions
PRO	Protein Ontology
RB1	RB transcriptional corepressor 1
RELA	RELA proto-oncogene, NF-kB subunit
RIPK1	Receptor interacting serine/threonine kinase 1
SIF	Simple Interaction File
SMAC	Diablo IAP-Binding Mitochondrial Protein
SMAD2	SMAD Family Member 2
SMAD3	SMAD Family Member 3

SMAD4	SMAD Family Member 4
SOCS1	Suppressor Of Cytokine Signaling 1
SOCS3	Suppressor Of Cytokine Signaling 3
STAT3	Signal Transducer And Activator Of Transcription 3
STAT6	Signal Transducer And Activator Of Transcription 6
TAB1	TGF-Beta Activated Kinase 1 (MAP3K7) Binding Protein 1
TAB2	TGF-Beta Activated Kinase 1 (MAP3K7) Binding Protein 2
TAK1	Mitogen-Activated Protein Kinase Kinase Kinase 7
TF	Transcription Factor
TGF-β	Transforming Growth Factor Beta
TNFA	Tumor Necrosis Factor
TNFRSF1A	TNF Receptor Superfamily Member 1A
TP53	Tumor Protein P53
TRADD	TNFRSF1A Associated Via Death Domain
TRAF2	TNF Receptor Associated Factor 2
TRAF6	TNF Receptor associated factor 6
TRF4	Terminal Nucleotidyltransferase 4

Abstract

Systems Biology is revolutionizing our understanding and perception of (natural) systems. This type of research uses models and simulations, aiming to explain the complex behavior of a biological system. It integrates a wide range of data and requires an extensive range of research disciplines to collaborate. In this project, we examine how this applies to a set of 25 (not so?) arbitrary genes. Statistical tools were used for overrepresentation analysis on the genes (BiNGO, Reactome, Panther). Based on these results, a network was built in Cytoscape. Protein databases like UniProtKB provided extensive information about proteins while pathway databases like Reactome and WikiPathways provided specifics about pathways. Association databases like String were used to identify important interactors for the network. For completion of the network, we used the triple store BioGateway.

In this project, a network containing genes and proteins involved in the human immune system is built and analyzed. Important biological processes like interleukin signaling and cytokine production are collated with the apoptotic process of programmed cell death. With MCODE, clusters involved in these processes were identified and analyzed. The network is put into biological perspective by overlaying it with omics data comparing the expression levels of genes in the consensus molecular subtype 2 of colorectal cancer (CMS2 CRC) and normal tissue. Our findings indicate a lower expression of apoptotic genes while the overexpression of SOCS1 and SMAD3 promotes tumor formation in colorectal cancer cells by interacting with different genes and pathways like transforming growth factor-beta (TGF-β) signaling. A down- and up-regulated cluster were identified with ClusterMaker. The results manifest the dynamic and holistic nature of a biological system and show how this can be used in, for example, cancer treatment.

Introduction

Understanding a system/network of genes requires more than just taking cognizance of the individual constituents. The reductionistic attitude towards complex situations, pioneered by René Descartes, contemplates a system as the sum of its components (1). To unravel the complicated build-up of networks, scientists analyzed the isolated components. This reductionist approach and the holistic thinking of Systems Biology are significantly out of kilter. Smuts, Von Bertalanffy, and many other scientists realized an alternative approach was crucial for a correct/complete understanding and interpretation of biological systems. A system is more than just the sum of its components. Dynamical interactions between constituents lead to behavior that can't be elucidated by analysis of the separate isolated parts. These emergent properties are the preeminent reason for the holistic approach.

Von Bertalanffy's 'General Systems Theory' is based on the perception that all systems are made up of interlinked components whose behavior is different within the system compared to an isolated state. These conceptions are pillars of Systems Biology. On the other hand, Loeb's mechanistic approach in which everything, including living organisms, is cogitated about as complex machines, has no place in Systems Biology. Every organism is unique and shows idiosyncratic behavior.

Systems Biology is a model-driven field that operates on both static and dynamical models. The time dependence of network components and relations is analyzed by systematic perturbations in dynamical models (2). A static model focuses on the specific structure of a system and is more conceptual. Experimental data is analyzed and integrated into both models while new hypotheses are persistently formulated and put to the test.

To decipher complex systems, the collaboration of numerous scientific domains and integration of all knowledge from experimental data and prior research is required. This approach generates the potential for new and more pertinent experiments and practical applications of knowledge. Systems biology can be used to understand and manipulate the increasing complexity of nature. New models will be more accurate and relevant while different principles and relationships will be discovered. Essentially, Systems Biology is and will keep revolutionizing our understanding and perception of (natural) systems.

Gene ontology analysis aims to formalize the area of 'genetic' knowledge which is crucial when merging a lot of data (3). This covers biological processes, molecular functions, and

cellular components. A determined set of species-neutral terms with a clear-cut connotation is used to analyze/describe a system of genes and their products. The use of standardized terms grants scientists easy access to more detailed information about specific genes or proteins. Gene ontology is linked to wide-ranging databases and has become a fundamental part of bioinformatics. Other ontologies like protein ontology (PRO) can likewise be exploited to build a network of genes and proteins.

Driven by emerging new genome-scale technologies like genomics and proteomics, Systems Biology integrates multiple orthogonal datasets to get increased reliability of annotations. Single annotations are not sufficient for a complete view of gene or protein functions and pathways. Data integration reduces noise by lowering the number of false positives of negatives (4). Integrating omics data forms a bridge between genomics and Systems Biology and thus allows analysis of a network at the gene expression level. A Semantic Systems Biology approach merges model-based Systems Biology with data integration and analysis (5). This augments both the integration and sharing of data to develop new hypotheses.

Because networks lie at the core of biological systems, network theory is used to analyze these systems (6). Graph-theoretical concepts are very convenient for both the description and analysis of the biological system. It combines the biological and mathematical characteristics of a system and is performed to compare different complex networks. Based on characteristic relationships of network parameters, different types of networks can be distinguished from each other. In a scale-free network, the node-degree follows a power-law distribution and the clustering coefficient is independent of the node degree (7). In a random network, the node degree follows a Poisson distribution. A hierarchical network is a network composed of densely connected sub-networks. Different highly clustered neighborhoods are connected by only a few hubs. The clustering coefficient is proportional to the reciprocal of k and the node-degree follows a power-law distribution. Scale-free and hierarchical networks are robust. Loss of nodes does not shatter the network.

In this project, we tried to build a coherent network around 25 provided genes. This network was expanded as relevant as possible. The main biological processes and pathways are apoptosis, the I-kappaB kinase/NF-kappaB cascade, Interleukin-4, and interleukin-13 signaling, and the MYD88 cascade initiated on the plasma membrane. Once the network was complete, a thorough analysis was performed to interpret the biological content. Three important biological clusters were identified within the network, a SMAD2-SMAD3-SMAD4-JUN cluster, JAK-STAT cluster and a TRIKA2 cluster. Additionally,

our network-building approach was analyzed showing we successfully expanded the most overrepresented pathways and biological processes.

The network metrics obtained from network analysis, indicated scale-free behavior while, for a PPI, hierarchical behavior was expected.

Finally, the network was put in a biological perspective by analyzing the transcriptional activity of the genes in consensus molecular subtype 2 of colorectal cancer. A down- and up-regulated cluster were identified. As cancer cells can ignore self-destruction signals, we expect a lowered expression of apoptotic genes. In contrast, tumor-promoting genes were expected to have higher transcriptional activity. Higher expression of MYC and lower expression of APAF1 among others confirm this hypothesis.

Methods and approaches

In this project, Cytoscape is used for data visualization and graph-based analysis. This open-source software displays networks of molecular interactions and biological pathways. A network consists of nodes, entities like genes or proteins, and edges, also called links, which relationships between nodes. In this project, we focus on protein-protein interactions. The resulting network is a protein-protein-interaction network (PPI). Descriptors of nodes and edges are called attributes. These allow for data integration into the Cytoscape model. Edges and nodes are provided with annotations, expression data, and other types of information. Data representation is mostly static therefore the software doesn't support dynamical modeling. Cytoscape can be used for data analysis. Basic features are available but for additional tasks/inquiries, plugins can and should be exploited. These are computer programs interacting with the host application with specific features. One can download these apps to integrate and operate them in Cytoscape.

Relationships between genes or proteins can be non-directional or directional. A non-directional relationship indicates equivalence. This applies to protein binding and co-expression of genes. These are represented by lines in the Cytoscape network. Directional relationships indicate causality. In the network, a pointy arrow designates an activation, and a circle upregulation whereas a blunt arrow represents inactivation and a half-circle downregulation. Different colors were used to optimize the visual display of the network.

Data/Knowledge sources used

Different types of databases were exploited in this project. Deep databases deal with all data about a certain organism or topic while broad databases deal with one type of data for all of Life Sciences.

General information about the provided genes was retrieved in **GeneCards**. This includes function, aliases, and pathway involvement. Protein database **UniProtKB** was used to fetch extensive information about proteins covering functional descriptions, (official) identifiers, pathway involvement, molecular interactions, among others.

Overrepresentation analysis on different GO subtrees (biological process, molecular function, cellular component) was performed in the Cytoscape app **BiNGO**. To perform overrepresentation analysis on pathways both **Reactome** and **Panther** were used.

For network extension, different types of tools were operated. Pathway databases **Reactome**, **Pather**, and **WikiPathways** were used to get extensive information about pathways. This includes network members and dynamic relationships. Association databases **GeneMANIA** and **String** merge numerous information sources to create a network of the search terms and associated units. This was utilized to find the most important interactors for our network members, in addition to the tool **IntAct**. To identify and obtain more specific information about the interactions, **Signor** was used. To gather triples, data entities composed of subject-predicate-objects, triple store **BioGateway**, was employed. This tool was used specifically for the completion of the network.

When collecting non-curated data, literature and signal transduction pathway text mining tool **Litinspector** was used to verify the truthfulness of fetched relations between specific genes/proteins.

In network analysis **BiNGO**, **Reactome** and **ClueGO** were used for overrepresentation analysis. To calculate the most important network parameters, the Network Analyzer in **Cytoscape** was run. In the first analysis, the network was treated as undirected. This was compared to the second analysis in which the network was treated as directed. For a third analysis, PPI cancer data was imported and merged with our network. This merged network was analyzed as undirected and was opposed to our network to demonstrate the effects of further expansion. It was also used to show how selectively our network was built.

For the identification of highly interconnected regions within the network, Cytoscape plugin **MCODE** was used. It identified three biological clusters and important subpathways. When the network was overlaid with experimental data, **ClusterMaker** was operated to identify clusters based on their transcriptional activity in the colorectal cancer cells. Both tools were beneficial for analysis and visualization of specific biological data.

Please see Appendix I for a complete set of information on the data and knowledgde sources used in this project.

Procedures, selection, cleaning

The first step in this project was getting familiar with the software Cytoscape. Extensive tutorials are available online to explore and learn how to work with Cytoscape. These tutorials were performed on random gene lists. The purpose was solely to understand the program and to be able to operate on our gene list (once provided).

In pursuance of an easy, unambiguous, and effective group collaboration, we decided to use official identifiers for the members of our network. Ensembl IDs were adopted for genes and UniProt accession numbers for proteins. Other standards we agreed on as a group are related to the representation of our genes. Only proteins are shown in the Cytoscape network but we used the gene symbol of the genes encoding these proteins to mark them. We decided to make a distinction between transcription factors (TF) and all other types of proteins (P).

The first step to build the network was a thorough analysis of the 25 provided genes. Database GeneCards was used to gather function, aliases, and pathway involvement for each gene. The information about the pathways in which the genes are active and the interactions they undergo were relevant and crucial for network building. Parallel with this tool, UniProtKB was used to retrieve information about the encoded proteins. For this project function and interactions were the most relevant. UniProtKB was also used to retrieve the official identifiers, UniProt accession numbers, for the proteins in the network. Key information was collected in a shared group document. All official identifiers were stored. Furthermore, the focus was on the proteins encoded by the genes. Official names, functions, and pathways associated with the gene/protein were incorporated into the document.

An overrepresentation analysis was performed with BiNGO on the 25 original genes. In this analysis, we want to determine whether a certain GO term is significantly overrepresented in our list. This means that for each GO term, the tool statistically determines the probability of it being significantly more represented in our gene set than in a specific reference set. These analyzes were performed with a significance level (α) of 0.0001 to achieve a confidence of 99.9999%. The hypergeometric distribution was used for the statistical tests. To support a large number of highly correlated tests, the Benjamini & Hochberg False Discovery Rate (FDR) was applied. On September 22, the most recent annotation file for Homo Sapiens was downloaded from GeneOntology for these analyses. Additionally, an overrepresentation analysis was performed on the GO subtree molecular function to identify molecular-level activities performed by gene products and on cellular components to identify stable

macromolecular complexes. The results were used to support the selection of biological processes which were used as a foundation for the network. Lastly, the GO term cellular component was analyzed.

Afterward, an overrepresentation analysis was performed in Reactome and Panther to identify the most important pathways. This provided information about the pathways and the genes/proteins involved. The most relevant and opportune pathways were chosen to serve as the base for the network. Reactome also identifies genes/proteins related to specific pathways, so it was used to retrieve new members for the network.

To confirm the involvement of the genes in specific pathways and processes, text mining was performed. After all the misinformation was removed, the network was expanded. New, interacting partners were included and relationships between members were fetched using various sources and tools.

Once all biological data had been collected, it was processed into a network. This allows both visualization and analysis. First, our network-building approach was analyzed. For this purpose, the p-value, which represents the overrepresentation of certain biological processes and pathways, was examined. The p-values from the overrepresentation analysis on the original set of 25 genes were compared with the p-values of analyzes on the extended network. A lower value implies the process or pathway is relatively more overrepresented. For the final network, analysis was performed in BiNGO, Reactome, and ClueGO.

Afterward, several network metrics were examined. The **node degree** (*K*) indicates the connectivity of a node. This is the number of links to other nodes. In a directed network, there is a distinction between incoming and outgoing edges. This introduces the terms **in- and out-degree** (K_{in} and K_{out}). For the undirected network, we look at the **degree distribution** (*P(k)*). *P* is the probability that a node has exactly *k* links. Network characterization is done using a plot of *P(k)* versus *k.* In a random network, the node degree follows a Poisson distribution. The average node degree is, in that case, typical. In a scale-free and hierarchical network, on the other hand, the node degree follows a power-law distribution ($y = ax^b$, $P(k) \sim k^{-\gamma}$). The **degree exponent** (γ) is smaller when hubs have a more important role in the network. For $\gamma > 3$ networks behave like a random network, hubs are not relevant. Scale-free networks have $\gamma < 3$. γ equal to 2 indicates a hub and spoke network in which the most-connected hub links to a large fraction of all the nodes. If γ has a value

between 2 and 3, this indicates a hierarchy of hubs (5). The most-connected hub links only to a small fraction of all nodes.

The **clustering coefficient** (*C*) indicates the tendency of nodes to form a cluster. It is defined as $C_I = 2n_I / k(k–1)$. Where C_I is the clustering coefficient of node I and n_I is the number of links connecting k_I neighbors of node I. The average value $< C >$ gives the overall tendency of nodes to form clusters. In a scale-free network, this coefficient is independent of *k*. In a hierarchical network, the clustering coefficient is proportional to the reciprocal of *k*. The **shortest path** is the lowest number of links passed through to get from one node to another. Small shortest path lengths in biology can be an indication of shared function. The **mean path length** ($< l >$) is the average of all shortest paths. This parameter is used for the interpretation of network navigability. For the calculation, it is crucial to distinguish between undirected and directed networks. The **betweenness centrality** is the number of times a node acts as a connecting point in the shortest path length between two other nodes. It represents the involvement of a node in the network. The **closeness centrality** of a node is defined as the reciprocal of its farness to all other nodes. It is the inverse of the average shortest distance between the node and all other nodes in the network. A node with a high closeness centrality score has a short distance to other nodes. This metric is a measure of centrality in the network.

To put our network into perspective, we imported and merged a large batch of PPI cancer data with our network. Certain nodes are marginal or secondary in our network, while they are very important in the cancer process and might form hubs in the merged network. This merging analysis gives a good insight into how selectively our network was built. Some nodes might have greater importance in our network because we focussed on pathways involving this gene/protein. But the opposite also occurs. Another advantage of this analysis is that we can predict trends for larger networks. This allows us to determine whether we should expand our network to perform additional experiments.

To gain a better understanding of the biology that lies in our network, clusters were identified with MCODE, introducing modularity in the network. For every cluster, a GO Biological Process overrepresentation analysis was performed in ClueGO to get an understanding off the processes in which the cluster is involved. So within the network, several smaller systems were distinguished, with a concrete biological function or meaning.
To analyze how the network behaves at the gene expression level, the network was overlaid with gene expression data. The network was then organized in clusters, identified by

ClusterMaker, based on the expression levels of the genes in CMS2 CRC and normal tissue. Upregulated and downregulated gene clusters were identified and interpreted in the context of colorectal cancer. A ClueGO analysis was performed on these clusters to get an idea about the active and less active biological processes in colorectal cancer cells. Additionally, the expression level of all the genes from the network was examined. The ClusterMaker clusters were compared to the clusters identified using MCODE.

Results

Overrepresentation analysis BiNGO & Reactome

Overrepresentation analysis with BiNGO on the original gene, with the whole annotation as a reference set, showed diverse biological processes were overrepresented. A significance level of 0.0001, FDR correction, and hypergeometric test were used. The complete ontology file was obtained in GeneOntology.

The output of this analysis consisted of different biological processes (Table 1). GO ID, GO description, p-value, corrected-value, cluster frequency, total frequency, and genes connected to this GO term are available. This analysis immediately leads to the conclusion that the obtained genes are active in the immune system. It is also striking that cytokines are of importance in this gene set. The presence of TNF-α, IL10, and CSF1R, among others, is in agreement with this observation. The MAPK cascade and cell death are also strongly represented. This is supported by the presence of several MAP kinases in the gene set: MAP2K7, MAP2K3, and MAPK3. This indicates that the BiNGO software has performed a correct analysis.

Processes that were specific enough to establish an opportune starting point to build a network were selected. The biological process had to be restricted in extent. General/Basic processes, like signal transduction, are too extensive to serve as a foundation. The chosen processes allow for network extension, as an additional and crucial requirement. This is done by adding supplementary genes/proteins and edges. The selected biological processes from BiNGO are:

- Regulation of I-kappaB kinase/NF-kappaB cascade
- Positive and negative regulation of apoptosis
- Response to cytokine stimulus

Table 1: Overrepresentation Analysis GO Biological Process BiNGO (α = 0.0001)

notation: Curator = GO, Species or file = /Users/marieverdonck/Documents/NTNU/Courses/Systems Biology/Network/gene_association.goa_ref_human.gaf, Type = Consortium Ontology: Curator = unknown, Type = unkn

GO ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes
2682	regulation of immune system process	3.4661E-15	7.6705E-12	18/25 72.0%	1404/17824 7.8%	RB1 IL10 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA ICAM1 TNFA CASP8 E...
34097	response to cytokine	1.2066E-14	1.3351E-11	15/25 60.0%	803/17824 4.5%	ABCA1 CSF1R DUSP1 STAT3 RELA NFKB1 ICAM1 TNFA CASP8 ERK1 TRAF6 I...
10033	response to organic substance	2.9241E-14	2.1570E-11	21/25 84.0%	2662/17824 14.9%	MAP2K3 IL10 ABCA1 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA NFKB1 IC...
32268	regulation of cellular protein metabolic process	6.3090E-14	3.4905E-11	20/25 80.0%	2348/17824 13.1%	RB1 MAP2K3 IL10 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM...
9966	regulation of signal transduction	1.1020E-13	4.8775E-11	21/25 84.0%	2841/17824 15.9%	RB1 MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 ...
48522	positive regulation of cellular process	1.7828E-13	6.5756E-11	25/25 100.0%	5517/17824 30.9%	RB1 CSF1R CDKN1A ACVR1B RELA ICAM1 TNFA CASP8 ERK1 MYC RIPK1 MA...
51246	regulation of protein metabolic process	2.2743E-13	7.1327E-11	20/25 80.0%	2509/17824 14.0%	RB1 MAP2K3 IL10 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM...
44093	positive regulation of molecular function	2.5785E-13	7.1327E-11	17/25 68.0%	1494/17824 8.3%	RB1 MAP2K3 IL10 CSF1R CDKN1A STAT3 RELA ICAM1 TNFA CASP8 ERK1 M...
70887	cellular response to chemical stimulus	3.0860E-13	7.5880E-11	20/25 80.0%	2549/17824 14.3%	RB1 MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 ...
71310	cellular response to organic substance	1.2077E-12	2.1972E-10	18/25 72.0%	1965/17824 11.0%	MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 TNF...
10646	regulation of cell communication	1.3309E-12	2.1972E-10	21/25 84.0%	3212/17824 18.0%	RB1 MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 ...
23051	regulation of signaling	1.4717E-12	2.1972E-10	21/25 84.0%	3228/17824 18.1%	RB1 MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 ...
48518	positive regulation of biological process	1.5281E-12	2.1972E-10	25/25 100.0%	6011/17824 33.7%	RB1 CSF1R CDKN1A ACVR1B RELA ICAM1 TNFA CASP8 ERK1 MYC RIPK1 MA...
71345	cellular response to cytokine stimulus	1.5655E-12	2.1972E-10	13/25 52.0%	699/17824 3.9%	ABCA1 CSF1R DUSP1 STAT3 RELA NFKB1 ICAM1 TNFA ERK1 TRAF6 IRF1 RIP...
165	MAPK cascade	1.5886E-12	2.1972E-10	9/25 36.0%	180/17824 1.0%	MAP2K3 TNFA ERK1 MYC TRAF6 RIPK1 MAP2K7 NFKB1 MYD88
71222	cellular response to lipopolysaccharide	1.5886E-12	2.1972E-10	9/25 36.0%	180/17824 1.0%	IL10 ABCA1 TNFA ERK1 TRAF6 RELA NFKB1 MYD88 ICAM1
6915	apoptotic process	1.7851E-12	2.3238E-10	14/25 56.0%	904/17824 5.0%	RB1 IL10 CDKN1A ACVR1B NFKB1 TNFA CASP8 ERK1 MYC IRF1 BIRC5 RIPK1 ...
71219	cellular response to molecule of bacterial origin	2.4706E-12	2.9197E-10	9/25 36.0%	189/17824 1.0%	IL10 ABCA1 TNFA ERK1 TRAF6 RELA NFKB1 MYD88 ICAM1
51403	stress-activated MAPK cascade	2.5067E-12	2.9197E-10	7/25 28.0%	64/17824 0.3%	MAP2K3 TNFA ERK1 TRAF6 MAP2K7 NFKB1 MYD88
12501	programmed cell death	3.1628E-12	3.4996E-10	14/25 56.0%	943/17824 5.2%	RB1 IL10 CDKN1A ACVR1B NFKB1 TNFA CASP8 ERK1 MYC IRF1 BIRC5 RIPK1 ...
31098	stress-activated protein kinase signaling cascade	4.3343E-12	4.5676E-10	7/25 28.0%	69/17824 0.3%	MAP2K3 TNFA ERK1 TRAF6 MAP2K7 NFKB1 MYD88
8219	cell death	5.2476E-12	5.1727E-10	14/25 56.0%	979/17824 5.4%	RB1 IL10 CDKN1A ACVR1B NFKB1 TNFA CASP8 ERK1 MYC IRF1 BIRC5 RIPK1 ...
34612	response to tumor necrosis factor	5.3760E-12	5.1727E-10	9/25 36.0%	206/17824 1.1%	TNFA CASP8 ERK1 TRAF6 RIPK1 MAP2K7 RELA NFKB1 ICAM1
42221	response to chemical	6.3852E-12	5.8877E-10	22/25 88.0%	4023/17824 22.5%	RB1 MAP2K3 IL10 ABCA1 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA NFKB...
32496	response to lipopolysaccharide	6.6638E-12	5.8988E-10	10/25 40.0%	315/17824 1.7%	IL10 ABCA1 TNFA CASP8 ERK1 TRAF6 RELA NFKB1 MYD88 ICAM1
71216	cellular response to biotic stimulus	7.5764E-12	6.4487E-10	9/25 36.0%	214/17824 1.2%	IL10 ABCA1 TNFA ERK1 TRAF6 RELA NFKB1 MYD88 ICAM1
65009	regulation of molecular function	9.6625E-12	7.9197E-10	20/25 80.0%	3050/17824 17.1%	RB1 MAP2K3 IL10 CSF1R CDKN1A DUSP1 STAT3 RELA NFKB1 ICAM1 TNFA C...
33993	response to lipid	1.0367E-11	8.1933E-10	13/25 52.0%	812/17824 4.5%	IL10 ABCA1 CDKN1A DUSP1 STAT3 RELA NFKB1 ICAM1 TNFA CASP8 ERK1 T...
2237	response to molecule of bacterial origin	1.1207E-11	8.5524E-10	10/25 40.0%	332/17824 1.8%	IL10 ABCA1 TNFA CASP8 ERK1 TRAF6 RELA NFKB1 MYD88 ICAM1
9893	positive regulation of metabolic process	1.4928E-11	1.1012E-9	21/25 84.0%	3621/17824 20.3%	RB1 MAP2K3 IL10 CSF1R CDKN1A STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA ...
43408	regulation of MAPK cascade	1.6847E-11	1.1631E-9	12/25 48.0%	650/17824 3.6%	MAP2K3 CSF1R TNFA ERK1 DUSP1 MYC TRAF6 RIPK1 GRB2 MAP2K7 MYD88 ...
1903706	regulation of hemopoiesis	1.6849E-11	1.1631E-9	10/25 40.0%	346/17824 1.9%	RB1 IL10 TNFA CASP8 MYC TRAF6 IRF1 STAT3 RIPK1 ACVR1B
31325	positive regulation of cellular metabolic process	1.7344E-11	1.1631E-9	20/25 80.0%	3145/17824 17.6%	RB1 MAP2K3 IL10 CSF1R CDKN1A STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA ...
9628	response to abiotic stimulus	2.7749E-11	1.8061E-9	14/25 56.0%	1108/17824 6.2%	CDKN1A DUSP1 MMP1 RELA NFKB1 ICAM1 TNFA CASP8 ERK1 MYC IRF1 GRB...
51240	positive regulation of multicellular organismal process	3.3831E-11	2.1002E-9	15/25 60.0%	1388/17824 7.7%	RB1 IL10 ABCA1 CSF1R STAT3 ACVR1B RELA ICAM1 TNFA CASP8 MYC TRAF...
48583	regulation of response to stimulus	3.4806E-11	2.1002E-9	21/25 84.0%	3777/17824 21.1%	RB1 MAP2K3 IL10 ABCA1 CSF1R DUSP1 STAT3 ACVR1B RELA NFKB1 ICAM1 ...
32101	regulation of response to external stimulus	3.5114E-11	2.1002E-9	13/25 52.0%	895/17824 5.0%	RB1 IL10 CSF1R DUSP1 RELA NFKB1 TNFA CASP8 ERK1 TRAF6 IRF1 RIPK1 M...
1817	regulation of cytokine production	4.1915E-11	2.4410E-9	12/25 48.0%	703/17824 3.9%	MAP2K3 IL10 CSF1R TNFA CASP8 TRAF6 IRF1 STAT3 RIPK1 RELA NFKB1 MY...
6950	response to stress	4.7202E-11	2.6784E-9	20/25 80.0%	3315/17824 18.5%	MAP2K3 IL10 ABCA1 CSF1R CDKN1A DUSP1 STAT3 RELA NFKB1 ICAM1 TNF...
10604	positive regulation of macromolecule metabolic process	5.0266E-11	2.7810E-9	20/25 80.0%	3326/17824 18.6%	RB1 MAP2K3 IL10 CSF1R CDKN1A STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA ...
51247	positive regulation of protein metabolic process	5.8911E-11	3.1348E-9	15/25 60.0%	1443/17824 8.0%	MAP2K3 CSF1R CDKN1A STAT3 ACVR1B RELA ICAM1 TNFA CASP8 ERK1 MY...
9967	positive regulation of signal transduction	5.9495E-11	3.1348E-9	15/25 60.0%	1444/17824 8.1%	MAP2K3 IL10 CSF1R STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA CASP8 ERK1 ...
45639	positive regulation of myeloid cell differentiation	6.2564E-11	3.2190E-9	7/25 28.0%	100/17824 0.5%	RB1 TNFA CASP8 TRAF6 STAT3 RIPK1 ACVR1B
7154	cell communication	6.4003E-11	3.2190E-9	23/25 92.0%	5164/17824 28.9%	RB1 MAP2K3 IL10 ABCA1 CSF1R CDKN1A DUSP1 STAT3 ACVR1B RELA NFKB...
35556	intracellular signal transduction	6.7600E-11	3.3244E-9	15/25 60.0%	1457/17824 8.1%	RB1 MAP2K3 CSF1R CDKN1A DUSP1 RELA NFKB1 TNFA ERK1 MYC TRAF6 RI...
51173	positive regulation of nitrogen compound metabolic process	8.6978E-11	4.1844E-9	19/25 76.0%	2949/17824 16.5%	RB1 MAP2K3 IL10 CSF1R CDKN1A STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA ...
60548	negative regulation of cell death	9.2190E-11	4.3408E-9	13/25 52.0%	967/17824 5.4%	RB1 IL10 CDKN1A DUSP1 STAT3 RELA NFKB1 ICAM1 TNFA CASP8 MYC BIRC...
51239	regulation of multicellular organismal process	1.0259E-10	4.5942E-9	18/25 72.0%	2547/17824 14.2%	RB1 MAP2K3 IL10 ABCA1 CSF1R STAT3 ACVR1B RELA NFKB1 ICAM1 TNFA C...

Analysis of the GO subtrees molecular function showed the considerable presence of cytokine receptor binding, tumor necrosis factor receptor family binding, MAP kinase kinase activity, and death receptor binding (Table II.2). This is in accordance with the results obtained in the analysis of the GO biological process.

The analysis of GO cellular components manifested the extensive presence of our genes in the membrane microdomain, membrane raft, and cytosol (Table II.3). Furthermore, it emphasized the appearance of the ripoptosome and death-inducing signaling complex (DISC). This multi-protein complex, formed by clustered death receptors, has a crucial role in apoptotic events (8, 9). This conforms with previously obtained results.

An overrepresentation analysis of pathways was performed in Reactome (Table 2). The same criteria as for the biological processes from BiNGO were applied to the results from Reactome. The selected pathways had to be interconnected and effective for network extension. Signaling by interleukins and cytokine signaling in the immune system are very general pathways. Therefore they are not suited for this project. An overrepresentation analysis on Panther pathways was performed in Panther to help choose the processes (Table II.4).

The most defining overrepresented biological pathways we chose from Reactome are:

- Toll-Like Receptor 4 (TLR4) Cascade
- Interleukin-4 and interleukin-13 signaling
- MYD88 cascade initiated on the plasma membrane

The MYD88 cascade initiated on the plasma membrane and MAPK activation pathways are both subpathways of the Toll-Like Receptor 10 Cascade.

Table 2: Reactome Pathway Overrepresentation Analysis

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species name
Interleukin-4 and Interleukin-13 signaling	16	211	0.015	1.11E-16	2.44E-14	*21*	*47*	*0.003*	Homo sapiens
Signaling by Interleukins	26	643	0.045	1.11E-16	2.44E-14	*157*	*493*	*0.036*	Homo sapiens
Cytokine Signaling in Immune system	29	1,092	0.077	1.11E-16	2.44E-14	*201*	*708*	*0.052*	Homo sapiens
Immune System	34	2,684	0.188	6.11E-15	1.01E-12	*402*	*1,623*	*0.12*	Homo sapiens
Toll Like Receptor 4 (TLR4) Cascade	9	151	0.011	1.92E-9	2.19E-7	*58*	*95*	*0.007*	Homo sapiens
Toll Like Receptor 3 (TLR3) Cascade	8	102	0.007	1.99E-9	2.19E-7	*35*	*61*	*0.004*	Homo sapiens
TRIF(TICAM1)-mediated TLR4 signaling	8	107	0.007	2.89E-9	2.37E-7	*35*	*58*	*0.004*	Homo sapiens
MyD88-independent TLR4 cascade	8	107	0.007	2.89E-9	2.37E-7	*35*	*60*	*0.004*	Homo sapiens
Toll-like Receptor Cascades	9	188	0.013	1.26E-8	9.2E-7	*103*	*185*	*0.014*	Homo sapiens
Cellular Senescence	9	200	0.014	2.13E-8	1.41E-6	*26*	*90*	*0.007*	Homo sapiens
MyD88 cascade initiated on plasma membrane	7	94	0.007	3.13E-8	1.47E-6	*44*	*58*	*0.004*	Homo sapiens
Toll Like Receptor 10 (TLR10) Cascade	7	94	0.007	3.13E-8	1.47E-6	*44*	*59*	*0.004*	Homo sapiens
Toll Like Receptor 5 (TLR5) Cascade	7	94	0.007	3.13E-8	1.47E-6	*44*	*59*	*0.004*	Homo sapiens
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	7	101	0.007	5.08E-8	1.95E-6	*34*	*48*	*0.004*	Homo sapiens
MyD88 dependent cascade initiated on endosome	7	102	0.007	5.43E-8	1.95E-6	*47*	*63*	*0.005*	Homo sapiens
Toll Like Receptor 7/8 (TLR7/8) Cascade	7	103	0.007	5.8E-8	1.97E-6	*47*	*66*	*0.005*	Homo sapiens
Death Receptor Signalling	8	159	0.011	6.02E-8	1.99E-6	*49*	*93*	*0.007*	Homo sapiens
Toll Like Receptor 9 (TLR9) Cascade	7	106	0.007	7.03E-8	2.18E-6	*47*	*68*	*0.005*	Homo sapiens
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	7	118	0.008	1.45E-7	3.76E-6	*46*	*64*	*0.005*	Homo sapiens
Toll Like Receptor TLR6:TLR2 Cascade	7	118	0.008	1.45E-7	3.76E-6	*46*	*66*	*0.005*	Homo sapiens

Text-mining tool LitInspector was used to confirm the involvement of the genes in the pathways fetched through BiNGO and Reactome. To determine which pathways we wanted to use for the network, we created a group document in which we indicated which genes were active in each pathway (Figure 1). This is how we obtained an overview of which and how many genes would be included in the 'foundation' of the network. When creating this document, we integrated three additional pathways. These more general pathways are also results of the Reactome and BiNGO overrepresentation analyses: Generic Transcription Pathway[1], Positive regulation of transcription from RNA polymerase II promoter, and Fc epsilon receptor (FCERI) signaling. When we chose the biological processes we tried to ensure there was a certain connection with the other processes. Genes involved in several pathways or processes could serve as linking points in the network. Green cells in the spreadsheet indicate that the gene is only active in the corresponding pathway. After thorough analysis and consultation, we chose four pathways which are marked in bold.

[1] This includes, among others, subpathways transcriptional regulation by TP53 and transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer

Genes in blue are involved in at least one of the selected. Each member of the group was assigned a pathway or process to work on. These four pathways or processes served as the foundation for the network:

- Interleukin-4 and interleukin-13 signaling (Emmy Wang)
- MYD88 cascade initiated on the plasma membrane (Thea Johanna Aandstad Hettasch)
- Positive regulation of apoptosis (Marie Verdonck)
- Regulation of I-kappaB kinase/NF-kappaB cascade (Kamil Szura)

Gene/Process	Toll Like	**Interleukin-4 ;**	**Regulati**	Respons	Negative	**MyD88 ca**	**Positive**	Fc epsilc	Positive	Generic
ABCA1										
ABI1										
ACVR1B							x		x	
BIRC5		x			x					x
CASP8	x		x	x			x			x
CDKN1A		x			x		x			x
CSF1R										x
DUSP1							x			
ERK1/MAPK3	x					x		x		x
GRB2								x		
ICAM1		x								
IL10		x			x		x			
IRF1									x	
MAP2K3	x					x				
MAP2K7	x					x		x		
MMP1		x								
MYC		x								x
MYD88	x		x	x	x	x				
NFKB1	x				x	x		x		x
RB1									x	x
RELA	x		x	x	x	x		x	x	x
RIPK1	x		x	x			x			
STAT3		x		x					x	
TNFA		x	x		x		x		x	
TRAF6	x		x	x	x	x	x	x	x	

Figure 1: Process overview

Individual Research Process or Pathway

For each biological process or pathway, several databases and tools were used. The approach was similar, but every member of the group showed different preferences in tools and sources. In this report, the approach for the positive regulation of apoptosis is described.

As the biological process ‘Positive Regulation of Apoptosis’ is not very specific, further scrutiny of the given gene list was necessary. Reactome overrepresentation analysis showed that the pathway ‘Death Receptor Signalling’ was overrepresented by the original set of genes. Furthermore, the BiNGO analysis on the GO subtrees cellular component indicated the occurrence of ripoptosme and DISC. GO molecular function tumor necrosis factor receptor family binding and death receptor binding are overrepresented. This information was used to find a pathway that would support network extension. As pathway databases allow to find and download complete pathway modules, these knowledge base resources were utilized for this task. Thorough analysis in Reactome led to the pathway ‘TNFR1-Induced proapoptotic signaling’ that served as a foundation for part of the network (Figure II.1).

Visual pathway databases were used to add relevant genes or proteins to the network. Reactome, KEGG, and WikiPathways were used. From these tools, several genes were selected as new members of the network. Important genes or proteins include TRADD, TRAF2, RIPK1, TNFRSF1A (TNFR1), and FADD.

In total, 60 genes were added. Some of these genes are involved in more than one of the chosen processes or pathways. These genes function as connecting units in the network.

From pathway databases, regulatory connections were identified and stored in Google documents. At this point in the project, this happened in individual files. The group decided on a standard to save these interactions to facilitate the later merging of the four documents. The format adopted for the interaction is as follows:

> Gene1 (interaction type) Gene2
> Where gene 1 is the source of the interaction and gene 2 is the receiver.
> *Eg. BAD (inactivation) BCL2*

To find more interactions, interaction databases were used. IntAct, BioGRID, String, and GeneMANIA are the main tools. Searches were performed on individual genes or a list of the genes that were already included in the specific pathway/process. This led to a better understanding of the most important interactors. New members were identified and included in the network. Protein complexes were also detected. The interactions fetched through these tools are primarily of the type association. In the files, protein binding relationships were denoted as ‘physAssociation’.

Each member of the group tried to develop the assigned process or pathway as best as possible by adding many interactions. The four documents were then merged into one large document containing all the members of the network and the interconnections.

Merged Document

To find additional network links, the tool BioGateway was used. All the members of the network were imported into Cytoscape. Firstly, gene regulation was examined. Relationships between transcription factors and genes encoding proteins in the network were sought. During these searches, a new and final member was added to the network, namely ETS1. This gene encodes Protein C-ets-1, a transcription factor, which directly controls the expression of cytokine and chemokine genes in a wide variety of different cellular contexts (10). Gene regulatory interactions were denoted as ‘upRegulation’ if the transcription factor stimulates transcription of the related gene or ‘downRegulation’ in case of repression. Next, regulatory interactions between all nodes in the network were studied. These can be both positive and negative. Finally, molecular interactions were retrieved. After consultation with the other group members, we decided to include only the molecular interactions of BioGateway into our network. The reason for this is that the network would otherwise be far too extensive, complicating analysis. A total amount of 279 edges was obtained.

At this point, the network was complete. All units were integrated into the network.

Network building

The SIF file containing all fetched interactions was imported into Cytoscape. To make the network interpretable and clear, the layout was adjusted. As mentioned before, in this network, we only consider proteins but they are indicated by the name of the gene encoding them.

We decided to clarify the distinction between transcription factors and other proteins by modifying the shape of the node. Transcription factors are represented as ovals while other types of proteins are shown as rectangles. Next, a distinction was made between genes from the original list and the units added during the project (Figure II.2). Original genes are indicated in purple while added genes are colored blue.

The different types of interactions can be distinguished from each other in two ways. Each interaction type has been given a distinctive color and type of line. Physical association is shown as a blue simple line. Activation is a light green arrow. Inactivation is a light red arrow ending in a vertical line. UpRegulation is a dark green line with a full circle at the end. DownRegulation, on the other hand, is a dark red line with a semicircle.

The applied layout makes a slight distinction between the chosen processes/pathways (Figure 2). At the bottom right you can find the process 'Positive Regulation of Apoptosis'. The pathway 'Interleukin-4 and interleukin-13 signaling' is shown at the top left. The distinction between the two other processes is less easy to make. Regulation of I-kappaB kinase/NF-kappaB cascade partly overlaps with the apoptosis part. NF-κB is a family of transcription factors regulating many important cellular behaviors, in particular inflammatory responses, cellular growth, and apoptosis, explaining the overlap (11). MYD88 cascade initiated on the plasma membrane is mainly found at the bottom left and top right. There was no focus on a strict distinction between the processes because this is not the case in a living cell, where there is always overlap. This is clearly visible in the network.

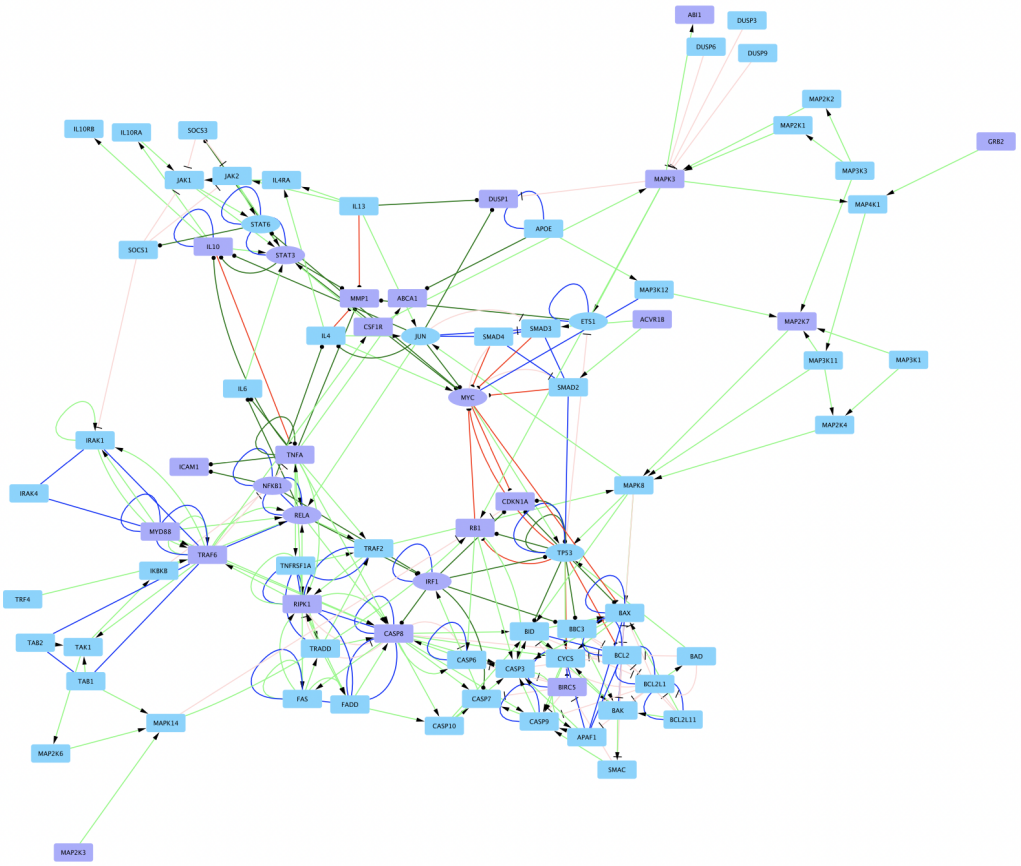


Figure 2: Extended Network in Cytoscape

In Cytoscape, a distinction is made between the node table and edge table. The node table contains the name of the node. This is the gene symbol of the gene encoding the protein. There is a column to indicate whether the protein is a transcription factor (TF) or other types of protein (P). A column with the gene ID from Ensembl, and a column with UniProt accession numbers are available. Other columns indicate whether the gene was part of the original list or was added later and contain information about molecular complexes of which the protein may be a part.

The edge table contains a column for ‘names’. These are given in the format of ‘Gene1 (interaction type) Gene 2’. Another column repeats the interaction type. The column ‘specific’ specifies how the interaction occurs. For example activation by phosphorylation. Two columns were inserted to indicate the source of the information. One column, ‘source’, either contains a tool like Signor or Uniprot or a name. This indicates manual curation by the mentioned group member. The column ‘specific source’ contains an URL, PubMed ID, or Signor ID. This allows external parties to be directly traced back to the source of information.

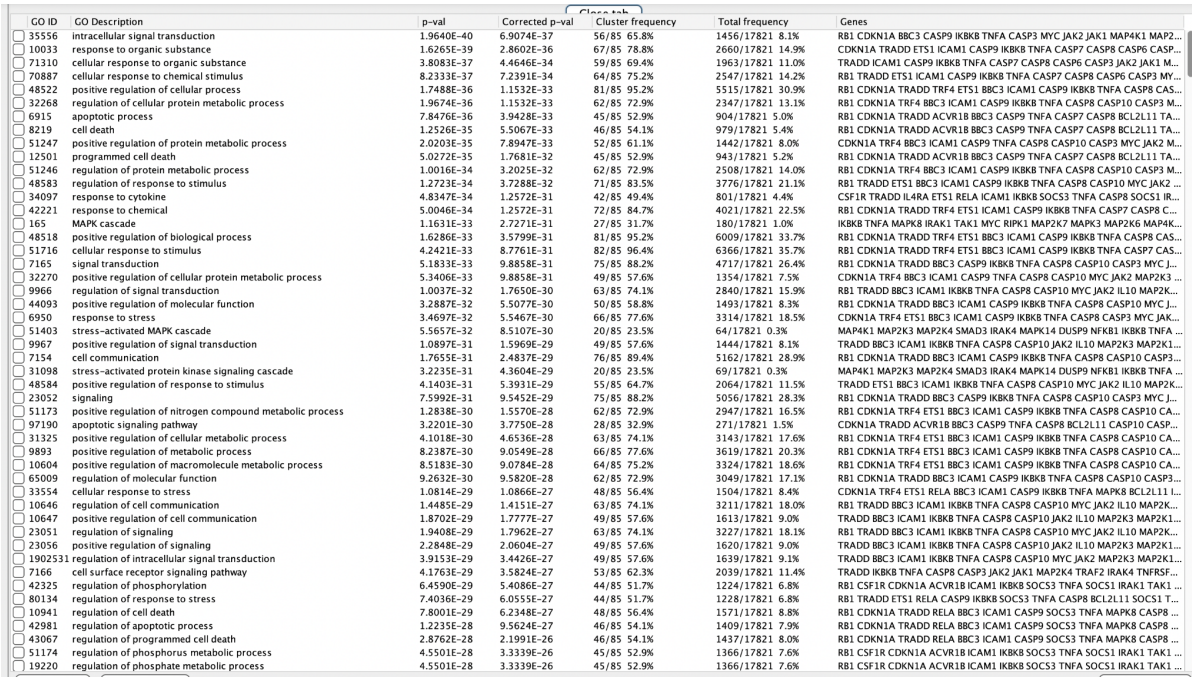
Data Analysis

BiNGO

An overrepresentation analysis was performed in BiNGO on the proteins from the extended network (Table 3). The same settings were used as for the analysis on the original gene set: significance level 0.0001, ontology and annotation file from GeneOntology, FDR correction. The reference was set to the full genome annotation. The process of apoptosis appears high in the list. This means that the associated p-value is very low. Cell death and programmed cell death also show a low p-value, as do response to cytokine and the MAPK cascade. In general, the terms cell death, programmed cell death, apoptosis, and other related terms appear more frequently than in the output of the analysis on the original gene list. This indicates that the network has adopted these terms or processes as a base and was built upon them. These outputs were expected. The MAPK cascade has also been included in the network and was expanded.

This output shows much lower p-values for the mentioned GO terms. The apoptotic process for example showed a corrected p-value of 2,328E-10 in Table 1 while Table 3 shows a corrected p-value of 3,9428E-33. This means that the chosen processes are more overrepresented in the entire network than in the original list. The goal was to use these processes as a foundation for the network and then extend them. Low p-values confirm that this was successful. The focus has shifted even more to these processes, which is in line with the aims of this project. This was confirmed by an overrepresentation analysis on the extended network in Reactome (Table II.5). It showed a reduced p-value for the MYD88 cascade initiated on the plasma membrane and Apoptosis. The FDR p-value for IL4 and IL13 signaling remained decreased while the non-corrected p-value was constant.

Table 3: Overrepresentation Analysis GO Biological Process BiNGO Extended Network (α = 0.0001)



GO ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes
35556	intracellular signal transduction	1.9640E-40	6.9074E-37	56/85 65.8%	1456/17821 8.1%	RB1 CDKN1A BBC3 CASP9 IKBKB TNFA CASP3 MYC JAK2 JAK1 MAP4K1 MAP2...
10033	response to organic substance	1.6265E-39	2.8602E-36	67/85 78.8%	2660/17821 14.9%	CDKN1A TRADD ETS1 ICAM1 CASP9 IKBKB TNFA CASP7 CASP8 CASP6 CASP...
71310	cellular response to organic substance	3.8083E-37	4.4646E-34	59/85 69.4%	1963/17821 11.0%	TRADD ICAM1 CASP9 IKBKB TNFA CASP7 CASP8 CASP6 CASP3 JAK2 JAK1 M...
70887	cellular response to chemical stimulus	8.2333E-37	7.2391E-34	64/85 75.2%	2547/17821 14.2%	RB1 TRADD ETS1 ICAM1 CASP9 IKBKB TNFA CASP7 CASP8 CASP6 CASP3 MY...
48522	positive regulation of cellular process	1.7488E-36	1.1532E-33	81/85 95.2%	5515/17821 30.9%	RB1 CDKN1A TRADD TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CAS...
32268	regulation of cellular protein metabolic process	1.9674E-36	1.1532E-33	62/85 72.9%	2347/17821 13.1%	RB1 CDKN1A TRF4 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CASP3 M...
6915	apoptotic process	7.8476E-36	3.9428E-33	45/85 52.9%	904/17821 5.0%	RB1 CDKN1A TRADD ACVR1B BBC3 CASP9 TNFA CASP7 CASP8 BCL2L11 TA...
8219	cell death	1.2526E-35	5.5067E-33	46/85 54.1%	979/17821 5.4%	RB1 CDKN1A TRADD ACVR1B BBC3 CASP9 TNFA CASP7 CASP8 BCL2L11 TA...
51247	positive regulation of protein metabolic process	2.0203E-35	7.8947E-33	52/85 61.1%	1442/17821 8.0%	CDKN1A TRF4 BBC3 ICAM1 CASP9 TNFA CASP8 CASP10 CASP3 MYC JAK2 M...
12501	programmed cell death	5.0272E-35	1.7681E-32	45/85 52.9%	943/17821 5.2%	RB1 CDKN1A TRADD ACVR1B BBC3 CASP9 TNFA CASP7 CASP8 BCL2L11 TA...
51246	regulation of protein metabolic process	1.0016E-34	3.2025E-32	62/85 72.9%	2508/17821 14.0%	RB1 CDKN1A TRF4 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CASP3 M...
48583	regulation of response to stimulus	1.2723E-34	3.7288E-32	71/85 83.5%	3776/17821 21.1%	RB1 TRADD ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 MYC JAK2 ...
34097	response to cytokine	4.8347E-34	1.2572E-31	42/85 49.4%	801/17821 4.4%	CSF1R TRADD IL4RA ETS1 RELA ICAM1 IKBKB SOCS3 TNFA CASP8 SOCS1 IR...
42221	response to chemical	5.0046E-34	1.2572E-31	72/85 84.7%	4021/17821 22.5%	RB1 CDKN1A TRADD TRF4 ETS1 ICAM1 CASP9 IKBKB TNFA CASP7 CASP8 C...
165	MAPK cascade	1.1631E-33	2.7271E-31	27/85 31.7%	180/17821 1.0%	IKBKB TNFA MAPK8 IRAK1 TAK1 MYC RIPK1 MAP2K7 MAPK3 MAP2K6 MAP4K...
48518	positive regulation of biological process	1.6286E-33	3.5799E-31	81/85 95.2%	6009/17821 33.7%	RB1 CDKN1A TRADD TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CAS...
51716	cellular response to stimulus	4.2421E-33	8.7761E-31	82/85 96.4%	6366/17821 35.7%	RB1 CDKN1A TRADD TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP7 CAS...
7165	signal transduction	5.1833E-33	9.8858E-31	75/85 88.2%	4717/17821 26.4%	RB1 CDKN1A TRADD BBC3 CASP9 IKBKB TNFA CASP8 CASP10 CASP3 MYC J...
32270	positive regulation of cellular protein metabolic process	5.3406E-33	9.8858E-31	49/85 57.6%	1354/17821 7.5%	CDKN1A TRF4 BBC3 ICAM1 CASP9 TNFA CASP8 CASP10 MYC JAK2 MAP2K3 ...
9966	regulation of signal transduction	1.0037E-32	1.7650E-30	63/85 74.1%	2840/17821 15.9%	RB1 TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 MYC JAK2 IL10 MAP2K...
44093	positive regulation of molecular function	3.2887E-32	5.5077E-30	50/85 58.8%	1493/17821 8.3%	RB1 CDKN1A TRADD BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 MYC J...
6950	response to stress	3.4697E-32	5.5467E-30	66/85 77.6%	3314/17821 18.5%	CDKN1A TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP3 MYC JAK...
51403	stress-activated MAPK cascade	5.5657E-32	8.5107E-30	20/85 23.5%	64/17821 0.3%	MAP4K1 MAP2K3 MAP2K4 SMAD3 IRAK4 MAPK14 DUSP9 NFKB1 IKBKB TNFA ...
9967	positive regulation of signal transduction	1.0897E-31	1.5969E-29	49/85 57.6%	1444/17821 8.1%	TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 JAK2 IL10 MAP2K3 MAP2K1...
7154	cell communication	1.7655E-31	2.4837E-29	76/85 89.4%	5162/17821 28.9%	RB1 CDKN1A TRADD BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CASP3...
31098	stress-activated protein kinase signaling cascade	3.2235E-31	4.3604E-29	20/85 23.5%	69/17821 0.3%	MAP4K1 MAP2K3 MAP2K4 SMAD3 IRAK4 MAPK14 DUSP9 NFKB1 IKBKB TNFA ...
48584	positive regulation of response to stimulus	4.1403E-31	5.3931E-29	55/85 64.7%	2064/17821 11.5%	TRADD ETS1 BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 MYC JAK2 IL10 MAP2K...
23052	signaling	7.5992E-31	9.5452E-29	75/85 88.2%	5056/17821 28.3%	RB1 CDKN1A TRADD BBC3 CASP9 IKBKB TNFA CASP8 CASP10 CASP3 MYC J...
51173	positive regulation of nitrogen compound metabolic process	1.2838E-30	1.5570E-28	62/85 72.9%	2947/17821 16.5%	RB1 CDKN1A TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CA...
97190	apoptotic signaling pathway	3.2201E-30	3.7750E-28	28/85 32.9%	271/17821 1.5%	CDKN1A TRADD ACVR1B BBC3 CASP9 TNFA CASP8 BCL2L11 CASP10 CASP...
31325	positive regulation of cellular metabolic process	4.1018E-30	4.6536E-28	63/85 74.1%	3143/17821 17.6%	RB1 CDKN1A TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CA...
9893	positive regulation of metabolic process	8.2387E-30	9.0549E-28	66/85 77.6%	3619/17821 20.3%	RB1 CDKN1A TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CA...
10604	positive regulation of macromolecule metabolic process	8.5183E-30	9.0784E-28	64/85 75.2%	3324/17821 18.6%	RB1 CDKN1A TRF4 ETS1 BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CA...
65009	regulation of molecular function	9.2632E-30	9.5820E-28	62/85 72.9%	3049/17821 17.1%	RB1 CDKN1A TRADD BBC3 ICAM1 CASP9 IKBKB TNFA CASP8 CASP10 CASP3...
33554	cellular response to stress	1.0814E-29	1.0866E-27	48/85 56.4%	1504/17821 8.4%	CDKN1A TRF4 ETS1 RELA BBC3 ICAM1 CASP9 IKBKB TNFA MAPK8 BCL2L11 I...
10646	regulation of cell communication	1.4485E-29	1.4151E-27	63/85 74.1%	3211/17821 18.0%	RB1 TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 MYC JAK2 IL10 MAP2K...
10647	positive regulation of cell communication	1.8702E-29	1.7777E-27	49/85 57.6%	1613/17821 9.0%	TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 JAK2 IL10 MAP2K3 MAP2K1...
23051	regulation of signaling	1.9408E-29	1.7962E-27	63/85 74.1%	3227/17821 18.1%	RB1 TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 MYC JAK2 IL10 MAP2K...
23056	positive regulation of signaling	2.2848E-29	2.0604E-27	49/85 57.6%	1620/17821 9.0%	TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 JAK2 IL10 MAP2K3 MAP2K1...
1902531	regulation of intracellular signal transduction	3.9153E-29	3.4426E-27	49/85 57.6%	1639/17821 9.1%	TRADD BBC3 ICAM1 IKBKB TNFA CASP8 CASP10 MYC JAK2 MAP2K3 MAP2K1...
7166	cell surface receptor signaling pathway	4.1763E-29	3.5824E-27	53/85 62.3%	2039/17821 11.4%	TRADD IKBKB TNFA CASP8 CASP3 JAK2 JAK1 MAP2K4 TRAF2 IRAK4 TNFRSF...
42325	regulation of phosphorylation	6.4590E-29	5.4086E-27	44/85 51.7%	1224/17821 6.8%	RB1 CSF1R CDKN1A ACVR1B ICAM1 IKBKB SOCS3 TNFA SOCS1 IRAK1 TAK1 ...
80134	regulation of response to stress	7.4036E-29	6.0555E-27	44/85 51.7%	1228/17821 6.8%	RB1 TRADD ETS1 RELA CASP9 IKBKB SOCS3 TNFA CASP8 BCL2L11 SOCS1 T...
10941	regulation of cell death	7.8001E-29	6.2348E-27	48/85 56.4%	1571/17821 8.8%	RB1 CDKN1A TRADD RELA BBC3 ICAM1 CASP9 SOCS3 TNFA MAPK8 CASP8 ...
42981	regulation of apoptotic process	1.2235E-28	9.5624E-27	46/85 54.1%	1409/17821 7.9%	RB1 CDKN1A TRADD RELA BBC3 ICAM1 CASP9 SOCS3 TNFA MAPK8 CASP8 ...
43067	regulation of programmed cell death	2.8762E-28	2.1991E-26	46/85 54.1%	1437/17821 8.0%	RB1 CDKN1A TRADD RELA BBC3 ICAM1 CASP9 SOCS3 TNFA MAPK8 CASP8 ...
51174	regulation of phosphorus metabolic process	4.5501E-28	3.3339E-26	45/85 52.9%	1366/17821 7.6%	RB1 CSF1R CDKN1A ACVR1B ICAM1 IKBKB SOCS3 TNFA SOCS1 IRAK1 TAK1 ...
19220	regulation of phosphate metabolic process	4.5501E-28	3.3339E-26	45/85 52.9%	1366/17821 7.6%	RB1 CSF1R CDKN1A ACVR1B ICAM1 IKBKB SOCS3 TNFA SOCS1 IRAK1 TAK1 ...

Another overrepresentation analysis was performed on the original set of genes. The same settings used as in the previous test were utilized, with one difference, the reference set in this case was the full network and not the genome. This test statistically determines whether a certain GO term occurs more often in the original gene set than would be expected based on the entire network. No results came back from this analysis, meaning that, compared to the entire network, no GO term is overrepresented in the original set. The genes that were added to the network are thus all involved in the same processes as the initial gene set. We, therefore, managed to extend the gene set in the biological processes that we had chosen in a previous stage of this project. This output confirms we adopted the correct approach.

Network Analyzer Cytoscape

Graph-based analysis with all edges considered to be undirected was performed in Cytoscape (Figure 3).

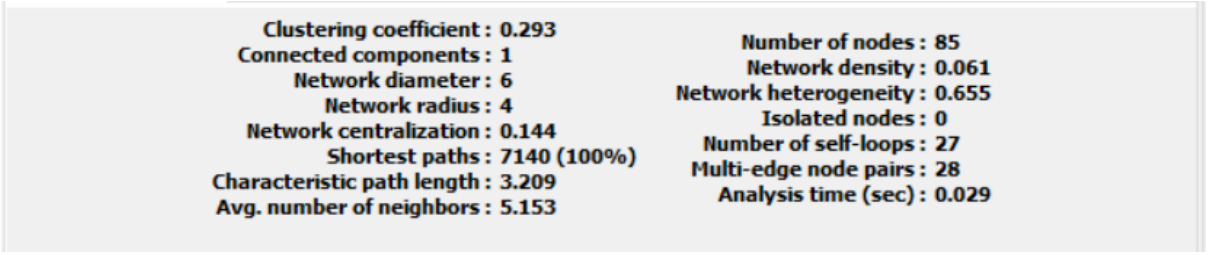


Figure 3: Network Analyzer Cytoscape (Undirected network)

Node Degree

For an undirected graph, the node degree indicates how many edges are incident to that node. For a directed graph, a distinction is made between the in-degree, which gives the number of edges to the node, and the out-degree, which gives the number of edges from the node. The average node-degree for the undirected network is 6.56.

The highest node degree is observed for CASP8, TP53, and RIPK1 (Table 4). Members with a high node degree are either very important proteins connecting several pathways or proteins that are central in a specific pathway.

For instance, TNFα is involved in Reactome pathways Death Receptor Signaling, IL10 signaling pathway, Interleukin-4, and 13 signaling, NF-kB (NFkB) Pathway, Cytokine signaling in immune system, TNFR1-induced NFkappaB signaling pathway, TNFR1-induced proapoptotic signaling, … (12).

BCL2 is a crucial and central element in the Bcl-2 Pathway which explains the high node degree of 15 (13).

Table 4: Nodes with Highest Node Degrees

Node	Degree
CASP8	26
TP53	21
RIPK1	20
TRAF6	18
TNFA	15
MYC	15
BCL2	15

Scale-Freeness

In a scale-free network, the node degree follows a power-law distribution ($P(k) \sim k^{-\gamma}$) where γ is the degree exponent (14).

The node degree distribution seems to fit the power-law fairly well (Figure 4). This power law is represented as $y = ax^b$. Applied to our network this gives:

a = 20.152 Correlation = 0.672

b = -0.836 R^2-value = 0.731

For this network, the degree exponent is equal to 0.836 which is smaller than 3. The network shows a typical scale-free behavior. Highly connected protein nodes, the hubs, are not hierarchical. For a PPI network, this means few proteins have many connections and many proteins have few connections. In Figure 4 the hubs can be seen at the bottom right (purple circle). These include CASP8, TP53, RIPK1, TRAF6, TNFA, MYC, and BCL2. They appear on the right because they have a high number of edges. Because there are only a few proteins with so many edges, they are very low on the vertical axis.

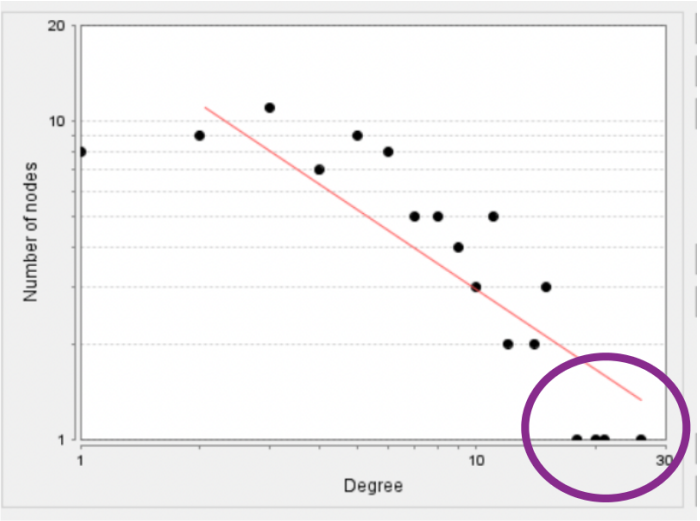


Figure 4: Node degree distribution for the own network

Hierarchical Behaviour

The hierarchical behavior of a network can be determined by looking at the node degree distribution and clustering coefficient. If the network shows hierarchical behavior, we expect the node degree distribution to follow the power-law. For the distribution of the clustering coefficient as a function of the node degree, the exponent of the power-law would be -1. Thus a higher clustering coefficient implies a lower node degree.

The average clustering coefficient plotted as a function of the number of neighbors clearly deviates from the power-law (Figure 5). Values for the trendline are:

a = 0.323	Correlation = 0.133
b = -0.062	R^2-value = 0.013

Parameter *b* has a negative value but as this absolute value is small, it seems like the node degree and average clustering coefficient are not strongly correlated. Hierarchical behavior can therefore be excluded. Analysis shows that the nodes with the highest clustering coefficient either interact with protein complexes or that they are part of a complex. TAB2 for instance is part of the TRIKA2 complex consisting of TAK1, TAB1, and TAB2. It also bridges TRAF6 to TAK1 to form the TRAF6-TAB2-TAK1 complex (15).

This clustering coefficient distribution could possibly change if protein complexes were reduced to one single node in the network. Further extension of the network also affects this distribution which does not follow the ideal trend (exponent equal to -1).

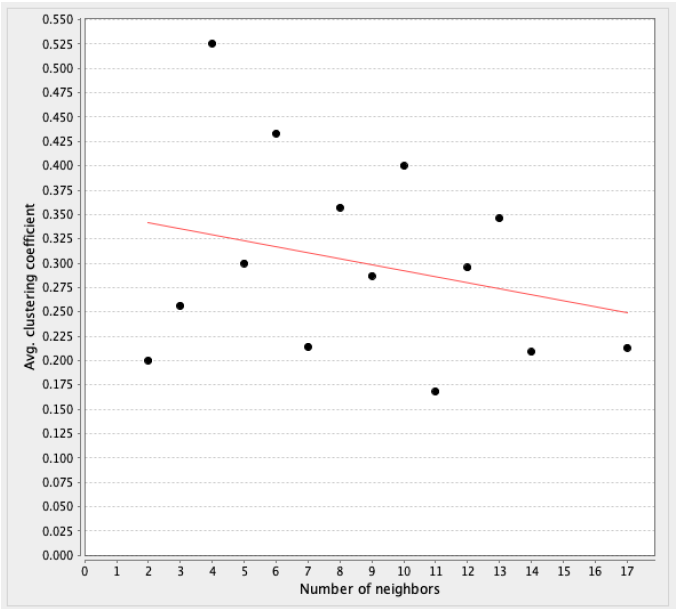


Figure 5: Average clustering coefficient distribution as a function of the number of neighbors

Betweenness Centrality

The betweenness centrality gives the fraction of all shortest paths that go through a node. Therefore a high betweenness centrality might indicate proteins that serve as important connecting points in the network. MYC for example is an important element in the transcriptional activity of the SMAD2/SMAD3:SMAD4 heterotrimer, but it is also involved in signaling by interleukins, MAPK family signaling cascade, and also plays a role in apoptosis (12). It connects the different subparts of the network, explaining the high betweenness centrality (Table 5).

Table 5: Nodes with Highest Betweenness Centralities

Node	**Betweenness centrality**
MAPK3	0.177
CASP8	0.16
MYC	0.15
TRAF6	0.144
RB1	0.120

Shortest Path

The characteristic path length is 3.209. Lower shortest path lengths indicate that the node is more central than average. The centrality of MYC and CASP8 was already indicated by their high betweenness centrality and high node degree and is now further rooted by their low shortest path length (Table 6).

Table 6: Nodes with Lowest Average Shortest Path Lengths

Node	Shortest Path
MYC	2.43
TP53	2.44
CASP8	2.46
JUN	2.46
RELA	2.48

Directed Network

Another analysis was performed in which the network was considered directed (Figure 6). Physical interactions are in theory bidirectional but we suspect Cytoscape cannot interpret these relationships correctly. Interactions in the form of (ProteinA physAssociation ProteinB) are treated as direct interactions. The analysis is therefore not entirely accurate.

In the directed network, the characteristic path length is longer. Even though all the nodes are connected in the network, the directed relations make it impossible to find a path between the nodes. A node without any outgoing or ingoing edges cannot serve as a connecting unit to find a path between nodes. The clustering coefficient is also lower for the directed network.

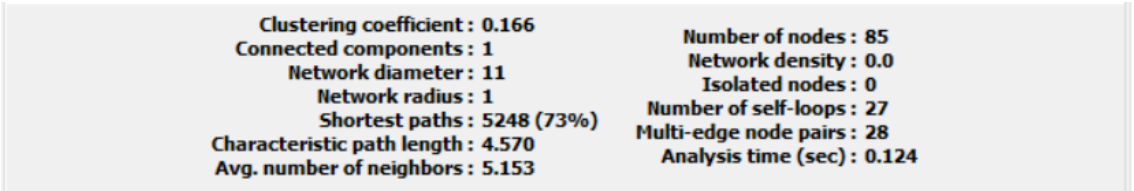


Figure 6: Network Analyzer Cytoscape (Directed network)

The in- and out-degree of nodes can give an indication of the activity of a node in the network (Table 7 and Table 8). ‘Passive’ nodes have a low out-degree while ‘active’ nodes have a high out-degree. TP53 for example has 21 edges of which 15 are outgoing. MYC has 15 edges of which 9 are incoming. This indicates MYC is a more passive node, dependent on many other nodes while TP53 is more active and is important for the activity of other nodes. This is explained by the fact that TP53 is a transcription factor (16).

CASP8 has an in-degree of 12 and an out-degree of 14. This means that CASP8 is an important link between different processes and pathways. It is central to the network. It depends on different nodes and at the same time, different nodes depend on CASP8.

Table 7: Nodes with Highest In-Degrees

Node	In-Degree
CASP8	12
BAX	11
BCL2	10
MYC	9

Table 8: Nodes with Highest Out-Degrees

Node	Out-Degree
TP53	15
CASP8	14
RIPK1	12
TNFA	11
TRAF6	10

Closeness Centrality

From the plot of closeness centrality against the number of neighbors, it can be seen that nodes with more neighbors tend to have a higher closeness centrality. These nodes serve as bridges between nodes in the network.

TNFα has a high closeness centrality which indicates it is a central node in the network (Table 9). It has a crucial role in the TNF signaling pathway which includes Death Receptor Signalling, Regulation of TNFR1 signaling, TNFR1-induced NFkappaB signaling pathway, TNFR1-induced proapoptotic signaling. It is also involved in the MAPK pathway and caspase cascade, explaining the high closeness centrality of TNFα (12).

Table 9: Nodes with Highest Closeness Centralities

Node	Closeness Centrality
TRAF6	0.37640449
RELA	0.36216216
RIPK1	0.34895833
TNFA	0.34358974
JUN	0.34010152

PPI from IntAct

A large batch of PPI cancer data was imported from IntAct. This network was merged with our network and network-based analysis was performed on this merged network which was treated as undirected. In the merged network, there are 5115 nodes, whereas our own network contains 85 nodes.

The observed clustering coefficient, shortest path lengths, and betweenness centrality are lower in the merged network. The node degree and closeness centrality are higher for the merged network (Table 10). The largest changes are observed for the average shortest path length and node degree. In the merged network, the node degree is on average 27 units bigger and the path length 0.5 units smaller.

Table 10: Comparison of Network Analysis on the Extended and the Merged Network. ASPL = Average Shortest Path Length, BC = Betweenness Centrality, ND = Node Degree, ClusCoef = Clustering Coefficient, CloseCent = Closeness Centrality

	Extended	**Merged**	**Extended**	**Merged**	**Extended**	**Merged**	**Extended**	**Merged**	**Extended**	**Merged**
Node	ASPL		BC		ClustCoef		CloseCent		ND	
MAP2K3	**601**	**310**	0	0.00747	0	0.00787	0.166	0.322	**1**	**104**
FAS	379	360	**0.00720**	**0.00115**	**0.750**	**0.145**	0.264	0.278	10	24
ABI1	**0**	**326**	0	0.00372	0	0.00672	0	0.307	1	42
CASP8	321	390	**0.211**	**0.00127**	0.129	0.152	0.312	0.256	26	57
TP53	437	303	0.0974	0.0411	0.110	0.0118	0.229	0.330	**21**	**323**
TRAF6	266	384	0.1853	5.79E-04	0.127	0.218	0.376	0.261	**18**	**18**

The node degree distribution for the merged network is characterized by a power-law with values:

a = 724.19 Correlation = 0.979

b = -1.293 R^2 = 0.838

The average clustering coefficient plotted as a function of the number of neighbors for the merged network follows the power-law (Figure 7). The clustering coefficient is proportional to the reciprocal of the number of neighbors. The merged network shows hierarchical behavior. Values for the trendline are:

a = 1.413 Correlation = 0.917

b = -1.066 R^2-value = 0.644

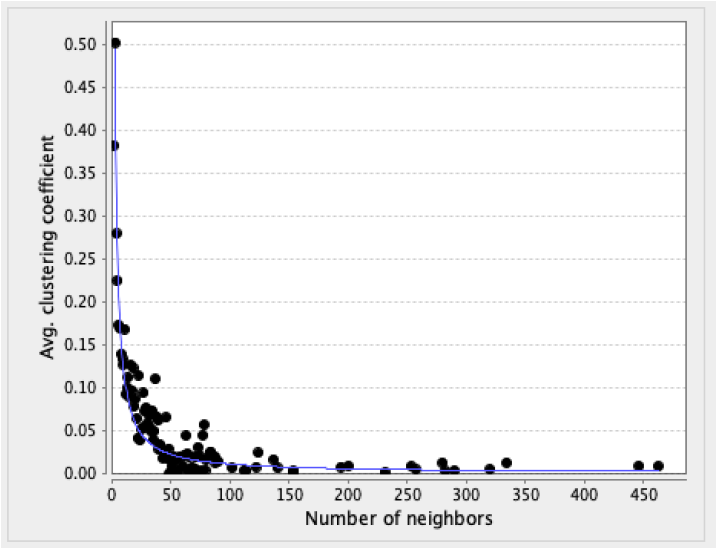


Figure 7: Average clustering coefficient distribution as a function of number of neighbors for the Merged Network

For some nodes like TRAF6, the node degree is the same as in our own network. This means that there were no additional interactors in the PPI data for these nodes. For other nodes like TP53 and MAP2K3, the node degree is significantly higher in the merged network.

The average shortest path length is longer in the merged network. This is due to the new interactions that have been imported into the network. For MAP2K3 the average shortest path length in the merged network is only half the value of our network. For other nodes, the difference is not as large. ABI1 has a higher value in the merged network. In our network, this node is marginal and has only one neighbor. All the metrics have a higher value in the merged network. This shows that when expanding our network, there was not much focus on this node.

MAP2K3 has a node degree of 1 in the extended network whereas it has a node degree of 104 in the merged network. This indicates that the node is a hub in the merged network. In our network, it only has one neighbor, indicating its low importance. Again, this is evidence of the selective development of our own network. For both ABI1 and MAP2K3 the closeness centrality is higher in the merged network.

The betweenness centrality is lower in the merged network. CASP8 and FAS have a higher betweenness centrality in the own network. Because this metric is an indication of the involvement of the nodes, it can be concluded that the process apoptosis, in which these two nodes play a crucial role, is more strongly emphasized in the own network. Many new members were added to the merged network and can act as connecting points. This creates a shortest path through other nodes. The clustering coefficient of FAS is higher in the own network, which indicates a higher tendency to form clusters. It tells how well connected the neighborhood of the FAS is. Again, this is an indication of the greater importance of the node in the own network.

ClueGO

From the GO biological process ClueGO analysis on the extended gene set, it is clear that as we expanded the network, the focus was shifted even more to apoptosis (Figure 8). The two most common GO terms in our network are both focused on apoptosis and make up at least 25% of the pie chart. The ClueGO indicates better that there are many processes related to growth. This is because related terms are grouped together and only the most significant term in each group is shown. MYD88 related processes are also present in this chart. Other overrepresented GO terms are MAP Kinase phosphatase activity and IL10 receptor activity. This is in line with our previous approach and the results from the BiNGO overrepresentation analysis on the extended gene set.

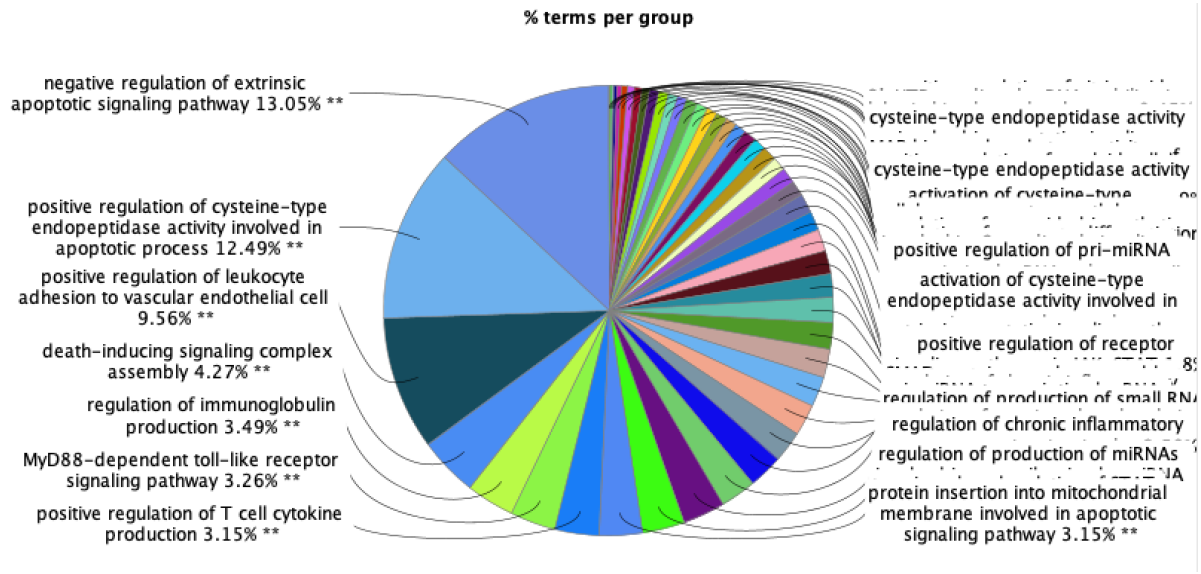


Figure 8: ClueGO analysis GO Biological Process on the Extended Network

MCODE

In MCODE several clusters were identified. The higher the complex score, the larger and denser the cluster is.

A first cluster (score 3.0) consists of nodes SMAD2, SMAD3, SMAD4, and JUN (Figure 9). SMAD2, SMAD3, and SMAD4 are colored blue because they form a complex named the SMAD2-SMAD3-SMAD4 complex. JUN directly binds to both SMAD3 and SMAD 4 which activates SMAD4 but inactivates SMAD3. This complex forms at the AP1 promoter site and regulates TGF-beta-mediated transcription (12). In this network, transforming growth factor β (TGF-β)-induced apoptosis is the main importance of this cluster. ClueGo analysis on the GO subtree biological process showed the overrepresentation of SMAD protein complex assembly (Figure 10). It also confirmed its role in TGF-β production.

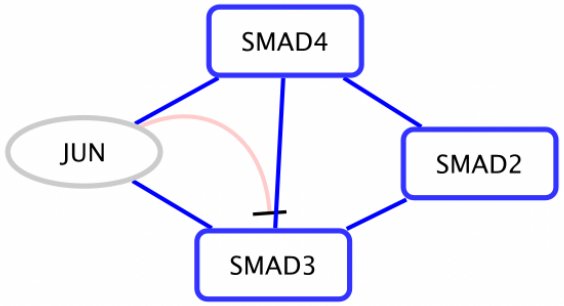


Figure 9: SMAD2-SMAD3-SMAD4-JUN Cluster

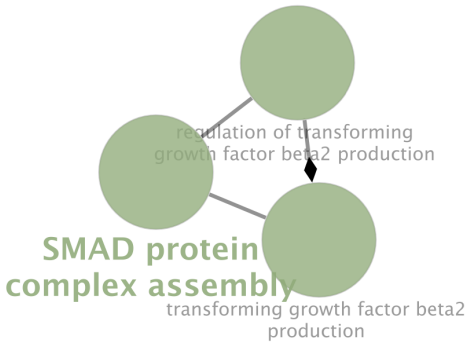


Figure 10: ClueGO Analysis cluster 1

The second cluster (score 3.5) contains seven proteins involved in the JAK/STAT pathway (Figure 11). This pathway transduces signals from interleukins, cytokines, and growth factors (17). These genes/proteins are also involved in the MAPK cascade. JAK1 and JAK2 activate STATS. STATs regulate the expression of proteins involved in the activation or inhibition of apoptosis (18). SOCS genes inhibit the JAK/STAT signaling pathway by inactivating JAKs (19). Proto-oncogene MYC bridges STAT3 and STAT6. ClueGO analysis on this cluster showed interleukin-6-mediated signaling pathway and regulation of response to interferon-gamma are overrepresented (Figure 12). This cluster is thus related to cell growth and proliferation.

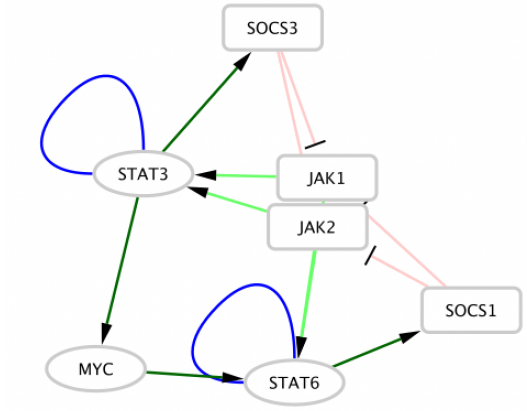


Figure 11: JAK-STAT Cluster

Figure 12: ClueGO Analysis cluster 2

The third cluster (score 2.0) was identified with fluff and contains the TRIKA2 complex (Figure 13). This complex is indicated in purple. TAB1 and TAB2 physically interact and both activate the third TAK1, the third member of the complex. TAK1 activates both IKBKB and MAP2K6. It is a crucial component in the MAPK cascade (20). IKBKB is involved in the negative regulation of the apoptotic process and positive regulation of I-kappaB kinase/NF-kappaB signaling (12). This cluster is therefore an important opposite of apoptotic promoting clusters or nodes. ClueGO analysis on this cluster showed I-kappaB phosphorylation and nucleotide-binding oligomerization domain containing signaling pathway are overrepresented (Figure 14). This indicates the cluster is important for the release and activation of NF-kappaB.

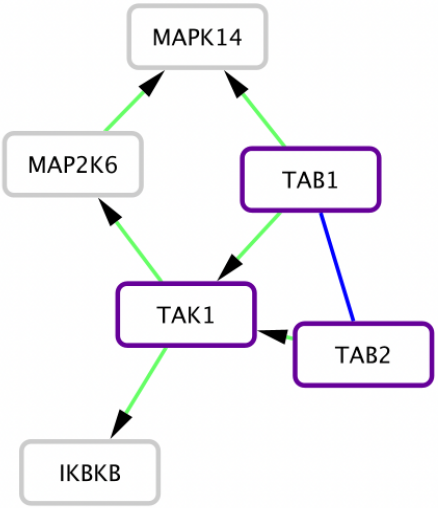


Figure 13: TRIKA2 Cluster

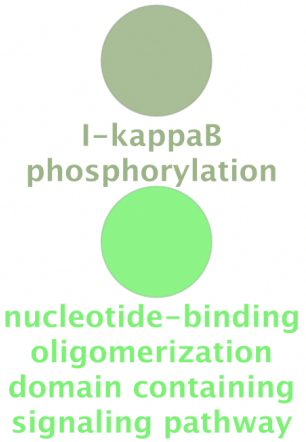


Figure 14: ClueGO Analysis cluster 3

ClusterMaker

When overlaying the network with gene expression data COAD_CMS2, some nodes were not covered. Differentially expressed genes when comparing the expression levels in the consensus molecular subtype 2 of colorectal cancer and normal tissue. Omics data results from experiments. If a node does not show differential expression in colorectal cancer cells and normal tissue, it might not be integrated into the data as this indicates the gene is irrelevant in this disease.

For a more visual display of this expression data, continuous mapping was applied. A red-colored node is more highly expressed in the case of colorectal cancer than in normal tissue. Blue indicates a lower expression. White means the expression is not affected or not measured.

Node clustering in ClusterMarker based on the fold changes identified a downregulated (Figure 15) and upregulated (Figure 16) cluster.

The downregulated cluster contains apoptosis-promoting genes. APAF1 for instance initiates apoptosis. Cytochrome c-mediated apoptosis is clearly less active in cancer cells. The lower transcriptional activity of JAK1 is consistent with the higher expression of SOCS1 and lower expression of STAT3. CASP9 is an apoptosis initiator whereas CASP3 is an executioner caspase (21). Both show lower expression in cancer cells. RB1 is a tumor suppressor gene which explains the low expression in cancer cells (12). SMAD3 and SOCS1 are both considered tumor suppressor genes (12, 22). However, these genes show a higher level of transcriptional activity in colorectal cancer cells.

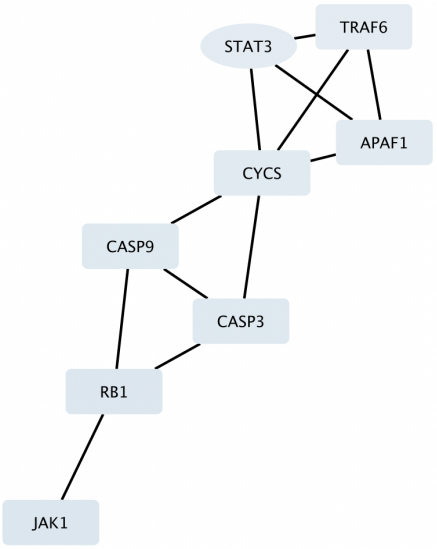


Figure 15: Downregulated Cluster

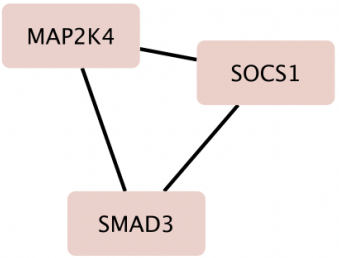


Figure 16: Upregulated Cluster

Other MAPKs like MAPK3, MAPK8, and MAP2K2 are more expressed in colorectal cancer cells. This indicates the MAPK cascade might play a crucial role in this disease. Proto-oncogene MYC and apoptosis inhibitor BIRC5 show a significant increase in transcriptional activity, while anti-apoptotic BCL2 and DUSP1, involved in the inactivation of MAPK activity, show a decrease. TP53 and SMAD4 have a higher expression. A complete overview of the transcriptional activity of the genes in the extended network can be found in Appendix II (Figure II.4).

ClueGO analysis on the downregulated cluster showed that interleukin-21-mediated signaling pathway (JAK1, STAT3), cysteine-type endopeptidase activity involved in the execution phase of apoptosis (CASP3, CASP9), and activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome C (CYCS, APAF1) are overrepresented (Figure 17). Thus confirming the lowered occurrence of apoptosis in colorectal cancer cells.

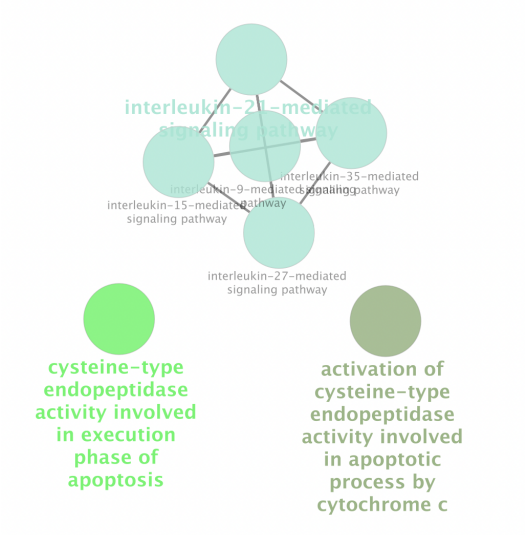


Figure 17: ClueGO Analysis Downregulated Cluster

ClueGO analysis was also performed on the upregulated cluster but this did not yield any result. The specificity of the test could have been lowered, but such analysis would not provide useful information. Therefore, no further test was performed.

Discussion

Network building

A first analysis of the 25 provided genes led to a very general idea about which biological processes and pathways were important for our list. The genes seemed to perform important functions of the human immune system. We expected to find processes related to growth and cell proliferation, as well as apoptosis. The three MAPK genes in the list were indicative of the MAPK cascade. These expectations were in accordance with reality. Grounded on overrepresentation analysis, the foundation of our network was interleukin-4 and interleukin-13 signaling, MYD88 cascade initiated on the plasma membrane, positive regulation of apoptosis, and regulation of I-kappaB kinase/NF-kappaB cascade.

The efficiency of network-building was confirmed by overrepresentation analysis in BiNGO, Reactome, and ClueGO. In each analysis, a lower corrected p-value is linked to the processes/pathways we selected to build the network. This means that in our extended network, these processes are more overrepresented than in the original gene list. We have consistently added genes/proteins involved in the selected processes/pathways. The analysis, additionally, shows that the focus has shifted towards apoptosis. This was expected as apoptosis is a broad biological process. We also concluded that, compared to the entire network, not a single GO term was overrepresented in the original set. All these analyses indicate we successfully built our network. The final extended network consists of 85 genes/proteins involved in the human innate immune system and 279 edges. These interactions are of the type activation, inactivation, physical association, upregulation, or downregulation.

Network analysis

Important properties of the network have been identified with the Network Analyzer in Cytoscape. CASP8 appeared to be the most central node. TP53, RIPK1, TRAF6, TNFA, MYC, and BCL2 are also very important. CASP8, RIPK1, TRAF6, and MYC were all included in the original gene list which is a further indication that the expansion of the network did not happen randomly but truly centered on the original gene list. The appearance of TP53 and BCL2 as central nodes is explained by their pathway involvement. TP53 plays a crucial role in, among others, apoptosis, MAPK pathway, and Interleukin-4 and 13 signaling (12). TP53 regulates the transcription of cell death genes (22). This includes death receptors and ligands, caspase activators, and caspases and genes involved in

cytochrome C release. BCL2 is a key regulator of caspase activating signals (23). It is involved in apoptosis, TGF-Beta pathway, and Interleukin-4 and 13 signaling (12).

Our network shows scale-free behavior. It is robust. Nodes with a low node-degree are not very important in the bigger network, removing them would not have great impact on the system. The most important property of scale-free networks is the appearance of hubs. Nodes with a degree that greatly exceeds the average node degree are crucial to connect and unify the network. CASP8, TP53, RIPK1, TRAF6, TNFA, MYC, and BCL2 have a high node degree. MYC, CASP8, and TRAF6 have a high betweenness centrality. CASP8 and MYC have a low shortest path length. The network metrics thus confirm these nodes appear as hubs in the network. The node degree follows a power-law distribution and the node exponent is smaller than 3. Hubs are not hierarchical.

The merged network, on the other hand, shows hierarchical behavior. We can conclude that adding more nodes and edges leads to the appearance of more sub-networks. There is more modularity in the larger network. We, therefore, expect that if we had expanded our own network more, it would have exhibited hierarchical behavior.

PPI networks are ‘small world’. Any two nodes in such a network have a shortest path of only a few links. Scale-free networks are ‘ultra small world’. Indicating the path length is much shorter than what is predicted by the small-world effect (24). The network is more cohesive. In our project, this is indicated by a clustering coefficient of 0.293 for the extended network while the merged network has a clustering coefficient of 0.175. The hubs in small world networks are not interlinked but connected to each other with a few connections. Our model confirms this; TP53 and CASP8 are connected through only one node, and so do MYC and BCL2.

Network clusters

SMAD3 and SMAD4 interact with JUN to activate transcription in response to TGF-β. These interactions are both functional and physical. Research shows SMAD and MAPK/JNK signaling converge at the AP1-binding sites (25). This means both processes are connected and explains why this cluster was identified in MCODE. AP1 regulates processes such as apoptosis and proliferation (26).

The JAK-STAT cluster is related to cell growth and proliferation. SOCS proteins have a critical role in the cytokine signaling pathway. SOCS1 and SOCS3 suppress the JAK-STAT pathway by binding to JAKs (27). The role of MYC in this cluster is to bridge STAT3 and STAT6. Dysregulation of the JAK-STAT pathway can lead to tumor development and therefore requires further research and analysis. The cluster identified with MCODE reflects the dynamic and interacting character of the genes and proteins. It should be considered and analyzed as a system. The interleukin-6-signaling pathway, identified with ClueGO, plays an important role in cancer. It is involved in growth regulation and occurs through STAT3. Deregulated overexpression of these genes can stimulate tumor development by inhibiting apoptosis (28). Interferon-gamma has antitumor and pro-apoptotic properties. It activates JAK1 and JAK2 and therefore positively regulates the JAK/STAT pathway (29).

TGF-β-activated kinase 1 (TAK1) has an essential role in the activation of both NF-κB and MAPKs. It activates IKBKB. TAB1 and TAB2 physically interact and both activate TAK1. They form the TRIKA2 complex. Overactivation of TAK1-TABs is related to cancer development (30).

Consensus molecular subtype 2 of colorectal cancer

In cancer cells, we expect to see high expression of growth-related genes. In contrast, apoptotic genes are expected to have a low expression. The clusters identified with ClusterMaker confirm the low activity of apoptotic genes. The downregulated cluster contains genes that are active in the JAK-STAT pathway. JAK1 and STAT3 have lower expressions. Thus, the JAK-STAT pathway is less active in colorectal cancer cells than in healthy tissue. This is consistent with the increased expression of SOCS1. Research has shown that CRC is related to SOCS1 overexpression which causes uncontrolled cell growth and resistance to death stimuli (31).

IL-21 has a tumor suppression and stimulation effect on CRC (32). Its role is thus far ambiguous and requires thorough analysis. A possible explanation for the confusing and dual action of IL-21 may be the emergence of certain properties through interaction with other genes or gene products. Within the framework of our network, all interactions can be analyzed and interpreted. Building a new network with different genes may be a good way to determine when IL-21 promotes or inhibits tumor growth. This is a perfect example of what Systems Biology is all about. A gene by itself is meaningless. It interacts with so many other entities and can only be interpreted in the context of a system. The genes in the downregulated cluster, JAK1, and STAT3, are related to the GO term IL-21-mediated

signaling pathway. This means that the interactions between these genes do not promote CRC. Instead, the expression of these genes is lower in colorectal cancer cells than in normal cells.

The TGF-β signaling pathway is commonly related to CRC. Mutations in SMAD4 promote tumor progression through disruption of TGF-β signaling (33). When mutations occur, the SMAD2-SMAD3-SMAD4 complex cannot form and therefore TGF-β-induced apoptosis doesn't happen. SMAD3 is involved in carcinogenesis but also functions as a tumor suppressor in certain cases. It inhibits cell proliferation and promotes apoptosis (34). SMAD3 has a regulating role in TGF-β-mediated immune suppression. The activated SMAD3-SMAD4 complex is essential for the repression of BCL2. This is in accordance with the findings. SMAD3 and SMAD4 both show higher expression in CRC while BCL2 has a much lower expression. The inactivation of the TGF-β signaling pathway activates MYC which has higher transcriptional activity and is a central player in CRC (35).

Higher expression of TP53 is common in colorectal cancers. In 50% of the cases, damage to the TP53 gene is observed (36). Mutations in this gene damage its tumor suppressor capacity and stimulate tumor growth.

Genetic alterations in the MAPK-signaling pathways are strongly associated with colorectal cancer (35). This confirms our hypothesis and is in accordance with our findings. MAP2K4, MAPK3, MAPK8, and MAP2K2 show a higher expression. MAPK3 and MAPK8 are associated with TP53 mutations (37).

APAF1 and CYCS are related to the activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome C. Both have lower transcriptional activity in CRC cells. Inhibits cytochrome-c dependent apoptosis in human CRC cells has a tumor-stimulating nature (38).

Additional experiments

Our model can be used to investigate pathways or interactions in other cancers or diseases. For each disease, key genes/proteins can be identified and interpreted within the framework of our network. The expression of the genes must be measured in patients and integrated into our model. This teaches us something about the interacting and dynamic nature of the relevant genes. Emergent properties responsible for the development of specific diseases can then be detected. If there is a suspicion that a gene may act as an oncogene, this hypothesis will be tested by overlaying the model with experimental data and running model simulations. The results can be used for the development of effective cancer therapies covering all (emergent) properties of the biological system.

STAT3 has shown both tumor-suppressing and promoting effects. It is therefore very interesting to use our model to investigate the conditions under which the gene suppresses tumor growth. For example, research may indicate what happens if we force higher expression on the STAT3 gene in CRC cells. Another interesting question is the effect of JAK1. This gene activates the JAK-STAT pathway and may affect the role of STAT3. The role of STAT3 and the JAK-STAT pathway can be analyzed in various cancers.

For concrete, more useful results, mathematical modeling should be used. This adds dynamics to the network. It can help to understand and predict the mechanisms underlying diseases, which is crucial for the development of new cures. Different computational models can be used. Logical models have a simplified dynamic. They are not able to represent the complex behavior of a biological system. Continuous models are theoretically more accurate than discrete systems and therefore preferred for modeling of complex systems (39). The accuracy of models is affected by the stochastic nature of gene expression which is not included in the model. Furthermore, the mean behavior of a population does not always apply to all single entities. Despite these limitations, mathematical modeling is very beneficial for the development of new therapies. It is, therefore, crucial to integrate this into a biological network when conducting experiments to find a cure.

Conclusion

Overrepresentation analysis on the GO Biological process in BiNGO and on pathways in Reactome and Panther showed that the 25 provided genes perform crucial functions in the human immune system. The foundation of our network was interleukin-4 and interleukin-13 signaling, MYD88 cascade initiated on the plasma membrane, positive regulation of apoptosis, and regulation of I-kappaB kinase/NF-kappaB cascade.

The network was successfully expanded on these processes/pathways. Overrepresentation analysis in BiNGO, Reactome, and ClueGO showed the selected processes and pathways were even more overrepresented in the extended network. During network building, the focus shifted towards apoptosis. This process, therefore, has an important role in the final network consisting of 85 nodes and 279 edges.

CASP8 is the most central node in the network. It is an apoptosis-initiator molecule. TP53, RIPK1, TRAF6, TNFA, MYC, and BCL2 were also found to be very important. The extended network is scale-free with non-hierarchical hubs. It is robust. Expanding the network could lead to hierarchical behavior.

In MCODE several clusters were identified. The SMAD2-SMAD3-SMAD4-JUN cluster regulates TGF-beta-mediated transcription. It is involved in growth and apoptosis. The JAK-STAT cluster is related to cell growth and proliferation. The TRIKA2 cluster is important for the release and activation of NF-kappaB. It has the potential to negatively regulate apoptosis.

The network was overlaid with gene expression data COAD_CMS2. In ClusterMarker a downregulated and an upregulated cluster was identified. Apoptosis-promoting genes have lowered expression in colorectal cancer cells while SMAD3, MAP2K4, and SOCS1 show a higher level of transcriptional activity.

Overexpression of SOSC1 is related to CRC. JAK1 and STAT3 do not promote colorectal cancer through the IL-21-mediated signaling pathway. Inactivation of the TGF-β signaling pathway activates MYC which has higher transcriptional activity and is a central player in CRC. SMAD3 and SMAD4 repress BCL2 and inactive the TGF-β signaling pathway. In 50% of the CRC cases, damage to the TP53 gene is observed. This gene shows higher expression in colorectal cancer cells. The MAPK-signaling pathways are strongly activated

and associated with CRC. The lower expression of APAF1 and CYCS inhibits cytochrome-c dependent apoptosis in human CRC cells.

The role of STAT3 in cancer development can be studied using our network and mathematical modeling. This could be very beneficial for the development of new therapies.

References

1. Trewavas A. A Brief History of Systems Biology. The Plant Cell. 2006 Oct;18(10):2420–30.

2. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science. 2001 May 4;292(5518):929–34.

3. Bard JBL, Rhee SY. Ontologies in biology: design, applications and future challenges. Nature Reviews Genetics. 2004 Mar;5(3):213–22.

4. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. Genome Biology. 2019 Jun 4;20(1).

5. Antezana E, Mironov V, Kuiper M. The emergence of Semantic Systems Biology. New Biotechnology. 2013 Mar;30(3):286–90.

6. Müller-Linow M, Hilgetag CC, Hütt M-T. Organization of Excitable Dynamics in Hierarchical Biological Networks. PLoS Computational Biology. 2008 Sep 26;4(9):e1000190.

7. Taylor R, Singhal M. Biological Network Inference and Analysis Using SEBINI and CABIN. In: Methods in Molecular Biology [Internet]. Totowa, NJ: Humana Press; 2009 [cited 2021 Nov 14]. p. 551–76. Available from: http://dx.doi.org/10.1007/978-1-59745-243-4_24

8. Arnhold J. Chapter 5 - Mechanisms of Cell Death. In: Cell and Tissue Destruction: Mechanisms, Protection, Disorders. Academic Press; 2020. p. 135–53.

9. Kim JW, Choi E-J, Joe CO. Activation of death-inducing signaling complex (DISC) by pro-apoptotic C-terminal fragment of RIP. Oncogene. 2000 Sep;19(39):4491–9.

10. The UniProt Consortium. ETS1 - Protein C-ets-1 - Homo sapiens (Human) [Internet]. ETS1 gene & protein. 2021 [cited 2021 Nov 14]. Available from: https://www.uniprot.org/uniprot/P14921
11. Barkett M, Gilmore TD. Control of apoptosis by Rel/NF-κB transcription factors. Oncogene. 1999 Nov;18(49):6910–24.

12. Weizmann Institute of Science. GeneCards®: The Human Gene Database [Internet]. GeneCards. [cited 2021 Nov 14]. Available from: https://www.genecards.org

13. GeneTex. Bcl-2 Pathway [Internet]. GeneTex. [cited 2021 Nov 14]. Available from: https://www.genetex.com/MarketingMaterial/Index/Bcl-2_Pathway?utm_source=Genecards&utm_medium=referral&utm_campaign=Genecards_pathway

14. Broido AD, Clauset A. Scale-free networks are rare. Nature Communications. 2019 Mar 4;10(1).

15. Mizukami J, Takaesu G, Akatsuka H, Sakurai H, Ninomiya-Tsuji J, Matsumoto K, et al. Receptor Activator of NF-κB Ligand (RANKL) Activates TAK1 Mitogen-Activated Protein Kinase Kinase Kinase through a Signaling Complex Containing RANK, TAB2, and TRAF6. Molecular and Cellular Biology. 2002 Feb 15;22(4):992–1000.

16. National Institutes of Health (NIH). TP53 gene: MedlinePlus Genetics [Internet]. MedlinePlus. [cited 2021 Nov 14]. Available from: https://medlineplus.gov/genetics/gene/tp53/

17. Brooks AJ, Putoczki T. JAK-STAT Signalling Pathway in Cancer. Cancers. 2020 Jul 20;12(7):1971.

18. Mitchell TJ, John S. Signal transducer and activator of transcription (STAT) signalling and T-cell lymphomas. Immunology. 2005 Mar;114(3):301–12.

19. Croker BA, Kiu H, Nicholson SE. SOCS regulation of the JAK/STAT signalling pathway. Seminars in Cell & Developmental Biology. 2008 Aug;19(4):414–22.

20. Lee S-W. TAK1-dependent Activation of AP-1 and c-Jun N-terminal Kinase by Receptor Activator of NF-κB -BMB Reports. BMB Reports. 2002 Jan 1;35(4):371–6.

21. Li J, Yuan J. Caspases in apoptosis and beyond. Oncogene. 2008 Oct;27(48):6194–206.
22. Vazquez A, Bond EE, Levine AJ, Bond GL. The genetics of the p53 pathway, apoptosis and cancer therapy. Nature Reviews Drug Discovery. 2008 Dec;7(12):979–87.

23. Willis S, Day CL, Hinds MG, Huang DCS. The Bcl-2-regulated apoptotic pathway. Journal of Cell Science. 2003 Oct 15;116(20):4053–6.

24. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 2004 Feb;5(2):101–13.

25. Zhang Y, Feng X-H, Derynck R. Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-β-induced transcription. Nature. 1998 Aug 27;394(6696):909–13.

26. Karin M, Liu Z g, Zandi E. AP-1 function and regulation - PubMed. Current opinion in cell biology. 1997 Apr 1;9(2).

27. Croker BA, Kiu H, Nicholson SE. SOCS regulation of the JAK/STAT signalling pathway. Seminars in cell & developmental biology. 2008 Aug;19(4):414–22.

28. Guo Y, Xu F, Lu T, Duan Z, Zhang Z. Interleukin-6 signaling pathway in targeted therapy for cancer. Cancer Treatment Reviews. 2012 Nov;38(7):904–10.

29. Castro F, Cardoso AP, Gonçalves RM, Serre K, Oliveira MJ. Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion. Frontiers in Immunology. 2018 Jan 1;0.

30. Xu Y-R, Lei C-Q. TAK1-TABs Complex: A Central Signalosome in Inflammatory Responses. Frontiers in Immunology. 2021 Jan 1;0.

31. Tobelaim, Beaurivage, Champagne, Pomerleau, Simoneau, Chababi, et al. Tumour-promoting role of SOCS1 in colorectal cancer cells. Scientific Reports. 2015 Sep 22;5(1):1–13.

32. Li J, Huang L, Zhao H, Yan Y, Lu J. The Role of Interleukins in Colorectal Cancer. International Journal of Biological Sciences. 2020;16(13):2323–39.

33. Itatani Y, Kawada K, Sakai Y. Transforming Growth Factor-β Signaling Pathway in Colorectal Cancer and Its Tumor Microenvironment. International journal of molecular sciences. 2019 Nov 20;20(23):5822.

34. Millet C, Zhang YE. Roles of Smad3 in TGF-beta signaling during carcinogenesis. Critical reviews in eukaryotic gene expression. 2007;17(4):281–93.

35. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330–7.

36. Munteanu I, Mastalier B. Genetics of colorectal cancer. Journal of medicine and life. 2014;7(4):507–11.

37. Urosevic J, Nebreda AR, Gomis RR. MAPK signaling control of colon cancer metastasis. Cell Cycle. 2014 Sep 2;13(17):2641–2.

38. Sun Y, Tang XM, Half E, Kuo MT, Sinicrope FA. Cyclooxygenase-2 Overexpression Reduces Apoptotic Susceptibility by Inhibiting the Cytochrome c-dependent Apoptotic Pathway in Human Colon Cancer Cells. Cancer Research. 2002 Nov 1;62(21):6323–8.

39. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008 Sep 17;9(10):770–80.

Appendix

Appendix I

GeneCards

GeneCards is an integrated database of human genes. It allows for keyword searches of all known and predicted human genes. A simple one-gene-search in this tool issues a summary sheet on the inquired gene. The sheet contains various information. The function of the gene is summarized, but the structure, domains, and interactions are also provided. 32 interactions were fetched from GeneCards. Several official identifiers are likewise provided. Genomics, GeneHancer Regulatory Elements, and localization data are available but these do not contribute to this project. Relationships with drugs and chemical compounds, transcripts expression data, orthologs and paralogs, genomic variants, involvement in disorders, … All the information provided in these 'cards' is extracted from more than 150 other databases. It is a user-friendly tool that offers a clear view of all available data on a specific gene.

UniProtKB

UniProtKB stands for Universal Protein Resource Knowledgebase. It is a collaboration between three institutes: the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR). UniProtKB is an accessible resource that allows the retrieval of functional information about proteins. The protein knowledgebase allows you to look up proteins present in several organisms. Popular organisms are the human, rice, *A.thaliana*, the mouse, the zebrafish. In this project, we only look at human proteins. The result of a simple one-protein search is an informative file about the selected protein (and organism). This information can be either reviewed or unreviewed. The latter contains records that await manual annotation. Reviewed data, on the other hand, is extracted from literature and curator-evaluated computational analysis. The available information is diverse. Function, names and taxonomy, subcellular location, pathology and biotech, processing, expression, interactions, structure, domains, … 15 interactions were fetched from this tool.

BiNGO

The Biological Networks Gene Ontology tool (BiNGO) is an open-source tool that performs over-or underrepresentation analysis of GO terms. The Cytoscape app performs these analyses on simple gene lists or on networks. Results can be visualized both in a table and in a graph in Cytoscape. The main advantage of BiNGO is its ability to deal with and interact with molecular interaction graphs. In this project, BiNGO was used on the original gene set to get an idea about what biological processes are overrepresented in the gene list. It was also used to perform overrepresentation analysis on the extended network. The aim was to confirm that the network had effectively been expanded on the selected biological processes and pathways.

In BiNGO, the reference set can be the network or the whole annotation but there is also an option to customize this set. The organism or annotation being analyzed has to be selected in the settings. Several other parameters are available in the BiNGO panel to customize the analysis (Figure I.1). Overrepresentation, as well as underrepresentation analysis, can be executed. Both hypergeometric and binomial statistical tests are available. In this project, the hypergeometric distribution is used. A lot of highly correlated tests are executed, on the many GO terms. Most multiple testing methods don't support this and require a correction. P-values have to be adjusted. BiNGO disposes of Benjamini & Hochberg False Discovery Rate (FDR) and Bonferroni Family-Wise Error Rate (FWER). FDR is used in this project. Different ontology files are available for analysis. GO biological process, GO full, GO molecular function, GOSlim generic, … This can also be customized. Another important parameter is the significance level (α). This is the probability of rejecting the null hypothesis even though it is true. A significance level of 0.05 means that there is a 5% risk of wrongfully rejecting the null hypothesis. In BiNGO, the null hypothesis is that a specific GO term is not overrepresented. If the P-value in the BiNGO output is low, this indicates the null hypothesis is not correct. Therefore, a low P-value means the probability of a GO term being overrepresented in the gene set is high. Categories can be visualized in Cytoscape both after or before correction.

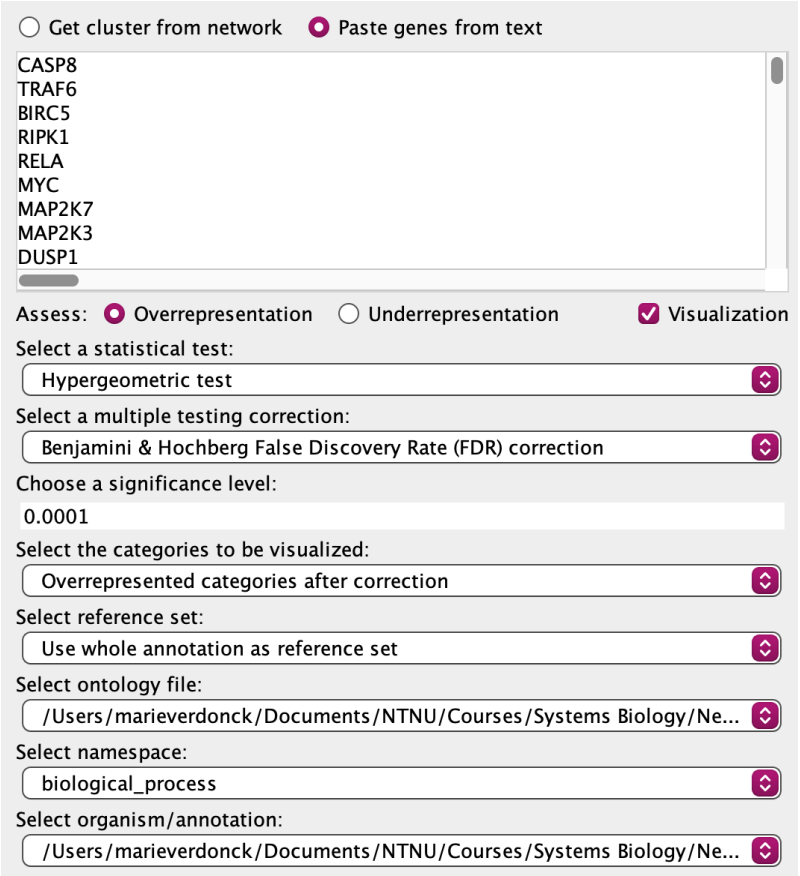


Figure I.1: Setting panel BiNGO

BiNGO analyses were performed with a significance level of 0.05, 0.001, and 0.0001. The amount of edges and nodes increases when the Stringency of the P-value cutoff is increased. As the number of nodes and edges was too high for 0.05 and 0.001, this project uses an increased Stringency of the P-value cutoff.

Signor

Signor is the Signaling Network Open Resource. This tool can be used for genes or proteins present in Homo sapiens, Mus musculus, or Rattus norvegicus. When searching for an entry, Signor provides an 'Entity Page' containing general information about the gene or protein. It also shows an interactive graphic visualization of the selected entity and its interactors (Figure I.2). Every edge and node is annotated. Signor also allows for multi-protein searches. These result in networks with interactions between the listed proteins. This feature is very useful for network building and was exploited in this project. Signor also has a pathway page that can be used to look up manually curated pathways. This functionality proved to be very effective in the extension of our network. Crucial genes of specific pathways were easily obtained. Furthermore, searches by organisms and PMIDs

are possible. We used this tool to perform both single gene/protein searches and multiple entities searches. The filter ‘type’ in Figure X allows fetching different types of interactions (transcription, direct and indirect relationships, binding, downregulation, upregulation). The SIGNOR score reflects the functional relevance of the relationships. The higher the score, the stronger the evidence. We focused on direct interactions with a threshold score of 0.5. A total of 132 interactions were retrieved from this tool.

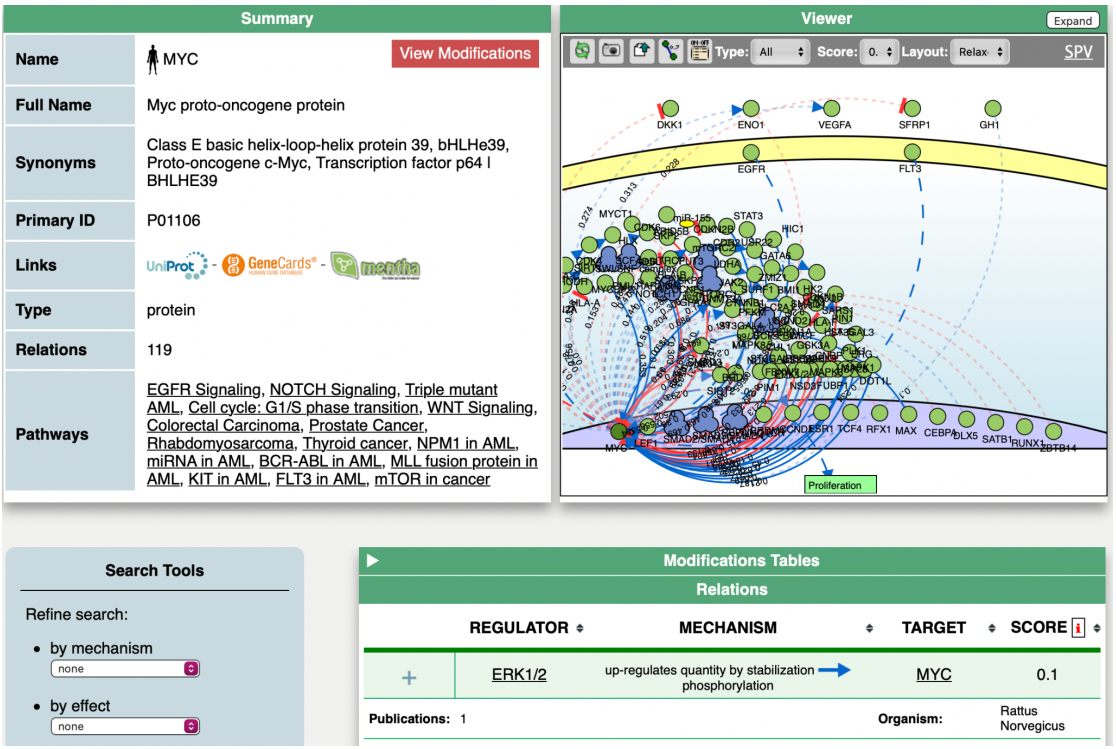


Figure I.2: Entity Page of MYC in Signor

IntAct

IntAct can be used online or as an app in Cytoscape. It provides data about molecular interactions between protein and genes. Analyses can be performed on single entities or lists of search terms. IntAct retrieves interactions linked to the terms of interest. IntAct distinguishes five major types of interaction. These are direct interaction, physical association, and association, proximity, and colocalization. IntAct is an important tool for network extension that allows for the detection of the most important interactors for the gene list. These can then be included in the network. IntAct provides interaction scores that represent the confidence degree for the correctness of an interaction. This value depends on the amount of experimental evidence supporting the interaction. Batch searches were performed to include the most interactors to the individual pathways/processes. In Figure I.3 a search was performed on the genes (from the original gene list) involved in positive

regulation of apoptosis. The filter on the interactor species was set to Homo Sapiens, interaction type to direct interaction, physical association, and association, interactor type protein, and gene were considered, neglecting dna, mrna and lncrna (Figure I.4). The score was set to 0.5. 41 interactions were fetched from IntAct.

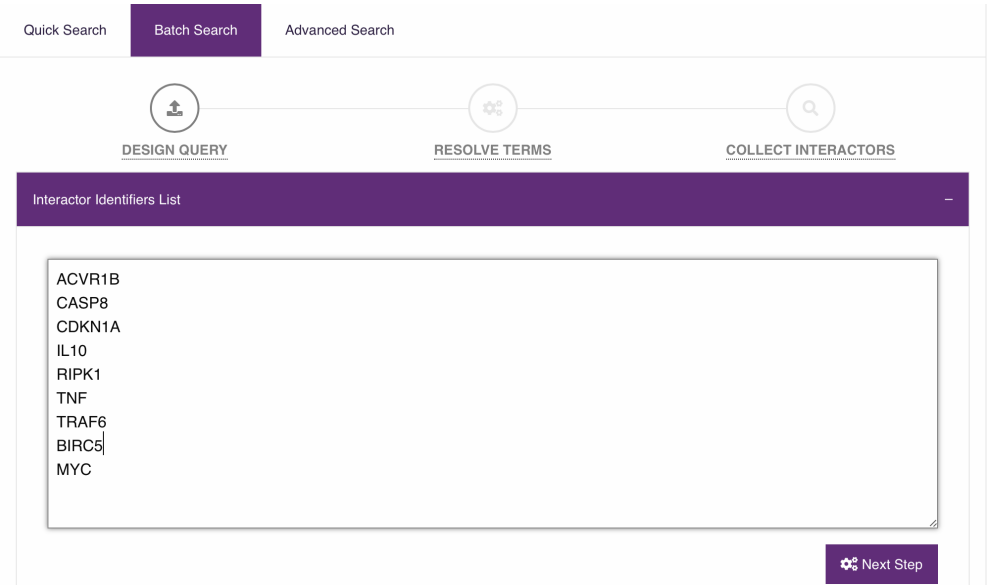


Figure I.3: Batch Search for genes/proteins involved in positive regulation of Apoptosis

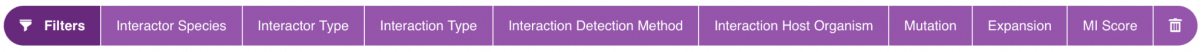


Figure I.4: Filters for IntAct Results

BioGrid

BioGRID, the Biological General Repository for Interaction Datasets, is a database containing genetic and protein interactions. Searches can be performed by gene/protein, publication, or chemical. We have performed searches on gene/protein (Figure I.5). The results include curated interactions in different organisms. A single-gene search results in a summary list containing different interactors with evidence supporting the interaction. 8 interactions were retrieved for the network. Another educational functionality is the list with gene ontology terms related to the gene. The list for GO Biological Process shows the involvement of a gene in a specific process and makes it easier to find rightful interactors which are also part of that biological process.

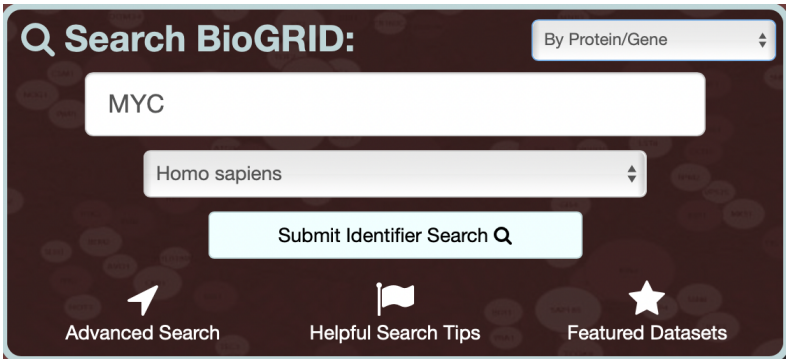


Figure I.5: Search in BioGRID

BioGateway

BioGateway is a Semantic Systems Biology database. It aggregates bio-ontologies and other sources of biological information from different sources such as Signor and IntAct. When new hypotheses concerning certain biological systems are generated through queries and interpretation of the results, we refer to the practice as Semantic Systems Biology. BioGateway aims to make this approach more accessible and common. A query can be built consisting of a set of questions. By adding additional lines, the search is further specified. BioGateway filters results based on GO terms and many different relation types. Fetched relations can be from or to selected nodes. The emerging genes, proteins, or relations can be imported into Cytoscape as a network. One can perform a query on a single gene or protein but BioGateway also allows for bulk searches. The data provided in BioGateway is curated and the original source of the data can easily be accessed.

We decided to add only interactions 'Molecularly interacts with'. We ran the query on the entire network and included self-loops (Figure I.6). This resulted in 184 interactions. These were curated through text mining. A total of 38 interactions were conserved and integrated into the network. An additional node, ETS1, was fetched through BioGateway. A bulk search was performed on the initial gene set to find all the proteins they encode. Then common relations to all these proteins were fetched, selecting ‘Regulators: involved in regulation of’. Selecting the ‘most in common’ button resulted in the gene ETS1 (Figure I.7).

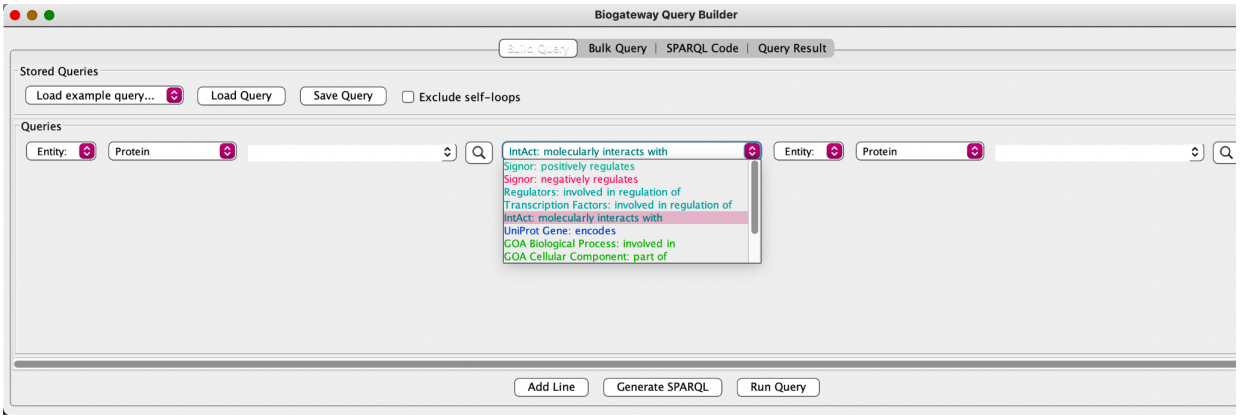


Figure I.6: Query Builder BioGateway

From Node	Relation Type	To Node
ETS1_HUMAN	involved in regulation of	MMP1_HUMAN
ETS1_HUMAN	involved in regulation of	MYC_HUMAN
ETS1_HUMAN	involved in regulation of	ABI1_HUMAN
ETS1_HUMAN	involved in regulation of	TRAF6_HUMAN
ETS1_HUMAN	involved in regulation of	IL10_HUMAN
ETS1_HUMAN	involved in regulation of	MP2K7_HUMAN
ETS1_HUMAN	involved in regulation of	MP2K3_HUMAN
ETS1_HUMAN	involved in regulation of	RIPK1_HUMAN
ETS1_HUMAN	involved in regulation of	CASP8_HUMAN
ETS1_HUMAN	involved in regulation of	NFKB1_HUMAN
ETS1_HUMAN	involved in regulation of	ACV1B_HUMAN
ETS1_HUMAN	involved in regulation of	CDN1A_HUMAN
ETS1_HUMAN	involved in regulation of	BIRC5_HUMAN
ETS1_HUMAN	involved in regulation of	TF65_HUMAN
ETS1_HUMAN	involved in regulation of	ABCA1_HUMAN
ETS1_HUMAN	involved in regulation of	STAT3_HUMAN
ETS1_HUMAN	involved in regulation of	RB_HUMAN
ETS1_HUMAN	involved in regulation of	GRB2_HUMAN
ETS1_HUMAN	involved in regulation of	ICAM1_HUMAN
ETS1_HUMAN	involved in regulation of	DUS1_HUMAN
ETS1_HUMAN	involved in regulation of	IRF1_HUMAN

Figure I.7: Query result 'Regulators: involved in regulation of'

GeneMANIA

GeneMANIA is a tool that analyses gene lists and creates an associative network containing the interactions between specific genes and their most important interactors. Using a large set of association data, GeneMANIA finds genes that are functionally similar to the queried genes. Networks can be created based on several types of data. Co-expression, co-localization, pathway involvement, genetic interactions, physical interactions, … Customizing the query results in more specific and useful networks. In this project, the focus is set on the genetic and physical interactions as well as on pathway involvement (Figure I.8). Queries can be performed on single genes or gene lists. In order to fetch the relations, GeneMANIA relies on data sources such as GEO, BioGRID, Ensembl, and NCBI. It is also available as a Cytoscape app that supports analysis of larger gene lists.

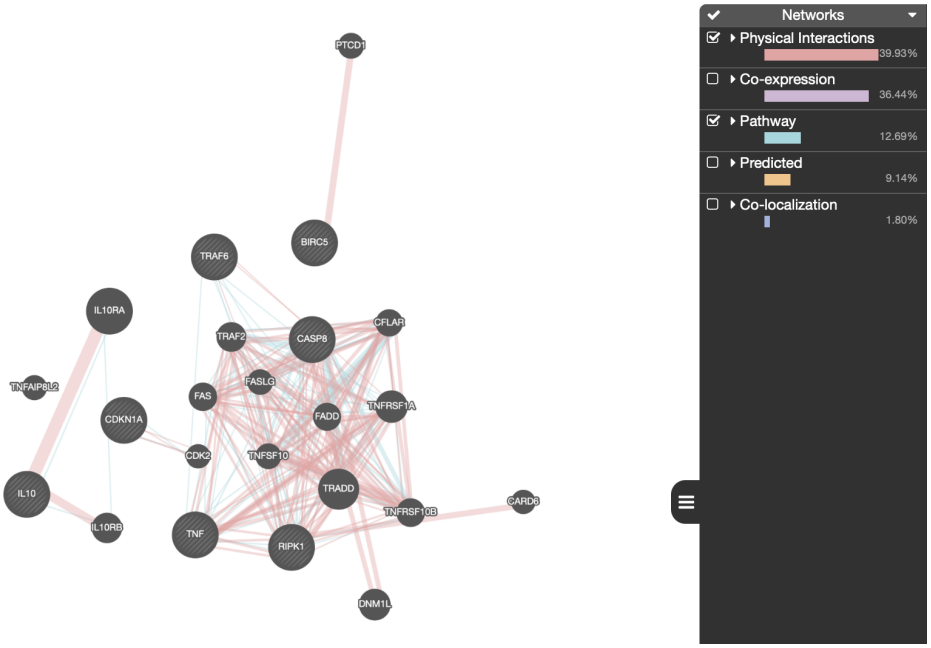


Figure I.8: GeneMANIA analysis of genes/proteins involved in positive regulation of Apoptosis

String

String is an associative database of known and predicted protein-protein interactions. The interactions are either physical or functional. Searches can be performed on single proteins or lists of several proteins. String returns a network of associated proteins. The edges can be visualized in evidence mode or confidence mode. In evidence mode, the lines indicate the type of evidence for the association. In confidence mode, the degree of confidence for the prediction of the interaction is given by the thickness of the line. Therefore String can be used to find new, reliable members for the network. For analysis, the species filter was set to Homo Sapiens. String can also be used for enrichment analysis of GO terms and pathways (Figure I.9).

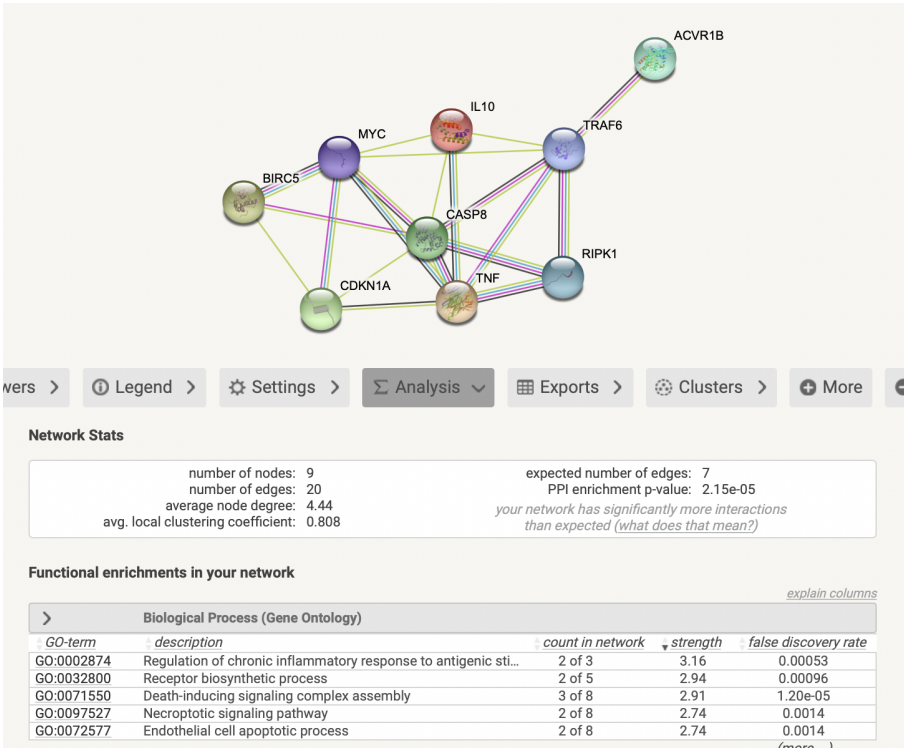


Figure I.9: String analysis on genes/proteins involved in positive regulation of Apoptosis

Reactome

Reactome is an open-source pathway database. It is manually curated. Reactome aims to provide visualization and analysis of biological pathways. Its analysis tool merges pathway identifier mapping, overrepresentation, and expression analysis. In this project, Reactome is used for overrepresentation analysis on the initial gene set. Reactome also helps in network extension as it identifies genes related to specific pathways.

For the analysis, the Project to human option was selected (Figure I.10). All non-human identifiers were thus converted to their human equivalents. The interpretation and selection of pathways were centered on the p-values. Pathways with the lowest p-value are the most overrepresented. Reactome uses a Binomial Test.

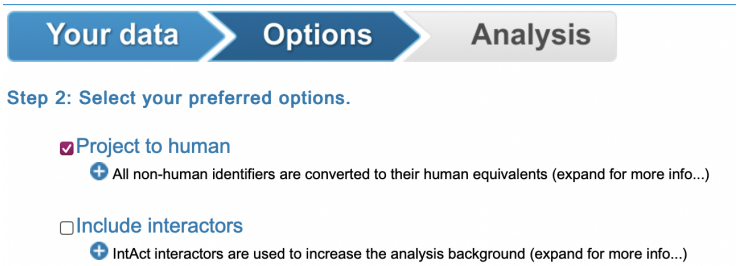


Figure I.10: Analysis options Reactome

Panther

The Protein Analysis Through Evolutionary Relationships Classification System classifies genes and proteins according to families and subfamilies, molecular functions, biological processes, and pathways. This facilitates analysis which, in Panther, can be performed on a single gene or protein, or an entire list. Several types of analysis are available. Functional classification can be viewed in gene lists or in graphic charts. Statistical overrepresentation tests can be executed on diverse annotation sets such as Reactome or Panther pathways, GO biological process or GO molecular function. Furthermore, a statistical enrichment test can be performed. Panther was primarily used to perform overrepresentation analysis on our gene list. Additionally, functional classification based on biological processes and pathways was utilized in this project.

For overrepresentation analysis, the default settings were used (Figure I.11). The whole-genome list for Homo Sapiens genes was used as a reference. A Fisher's Exact test was performed and several annotation data sets were used (Panther pathways, GO molecular function, GO Biological process, …). The False Discovery Rate was used as a correction.

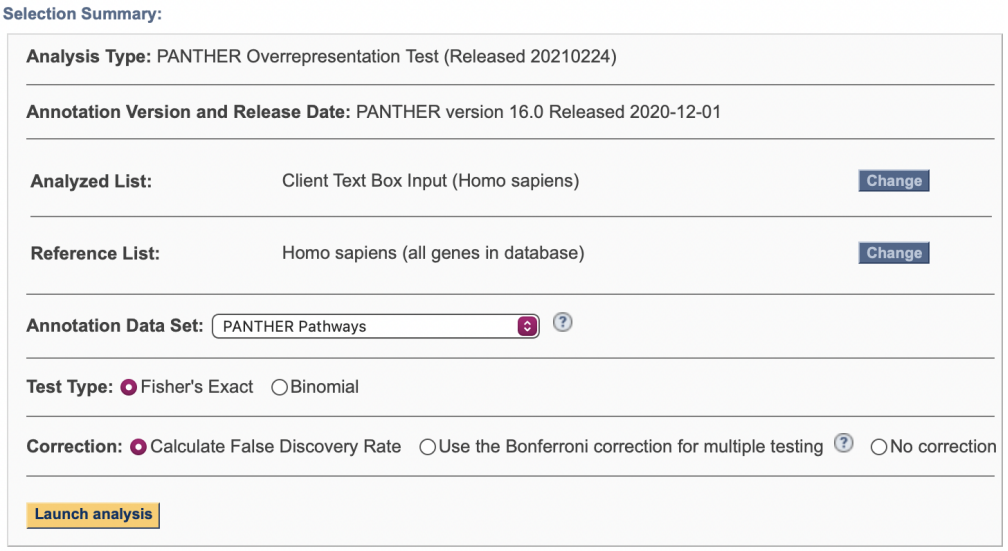


Figure I.11: Selection Summary Panther Overrepresentation Analysis

WikiPathways

Wikipathways is a pathway database. It is primarily used for the visualization of a specific pathway. It offers an intuitive visual display and description of the reactions and members of the pathway. 5 interactions were imported into the network.

LitInspector

LitInspector is used for literature and signal transduction pathway mining. It is based on manually curated databases of pathway names and components, but also general pathway keywords. LitInspector is used for data curation. The literature search tool allows for text mining within the NCBI's PubMed database. Searches can be performed on two or more genes to look up articles mentioning both. These articles are used to verify the truthfulness of fetched relations between specific genes. Adding more genes in the query will yield more specified results.

ClueGo

ClueGo creates a functionally organized GO term network for a gene set. It integrates GO terms in addition to KEGG/BioCarta pathways. The output of the analysis can be visualized in networks and charts. In networks, the GO terms are represented as nodes. The size of these nodes is an indication of how significantly overrepresented the terms are. To facilitate the interpretation of the results, functionally related groups of GO terms are indicated in the same color. The results are also represented as a table in which the most significant term in each group is selected. ClueGO is thus a tool that allows overrepresentation analysis with an easily interpretable visual output.

ClueGO was used for functional analysis (Figure I.12). GO Biological Process (all types of evidence) was analyzed. Network specificity was set close to detailed (Figure I.13).

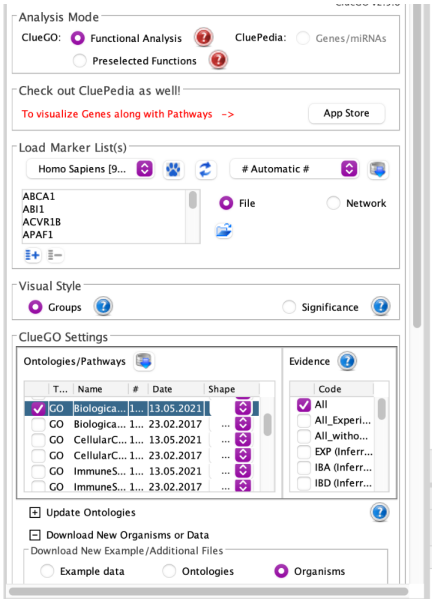


Figure I.12: Parameters ClueGO

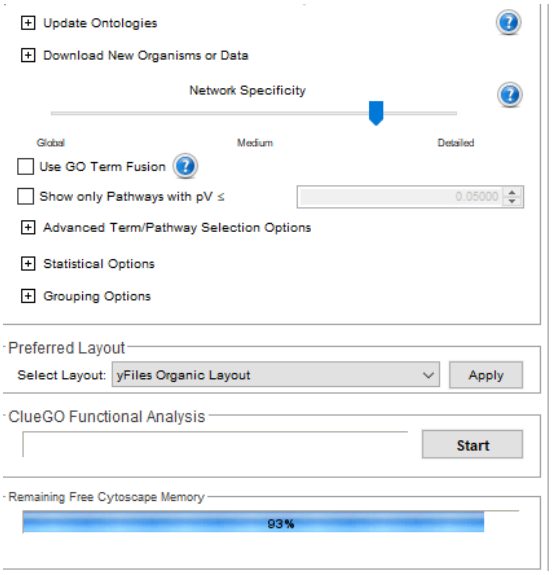


Figure I.13: Parameters ClueGO (Continued)

ClusterMaker

ClusterMaker is a multi-algorithm Cytoscape plugin that clusters nodes based on the correlation of attributes. These clusters can be used for the analysis and visualization of specific biological data. In this project, clusterMaker was used to cluster nodes based on the log fold change. Identification of up-and downregulated clusters facilitated the analysis and interpretation of the transcriptional activity of the genes in specific situations. This was applied to colorectal cancer cells.

A correlation network was created based on the node attribute logFC (Figure I.14). The distance metric was set to Euclidean distance. Nodes and edges without data were ignored. Edges were only created between nodes if the distance between them was lower than 0.01.

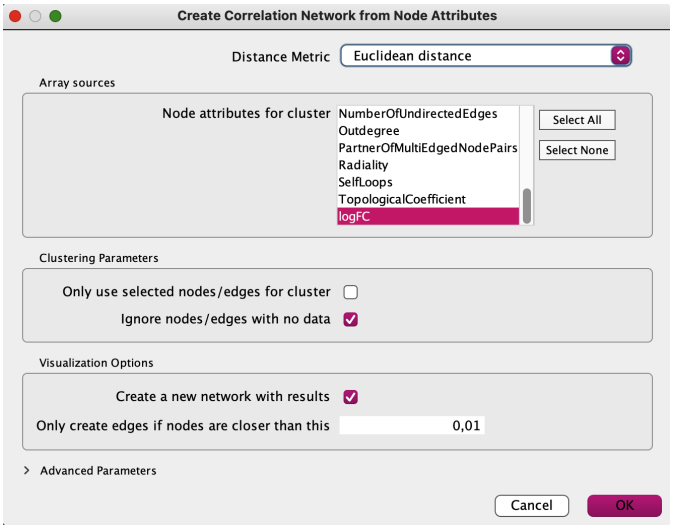


Figure I.14: Parameters ClusterMaker

MCODE

MCODE is a graph-based algorithm that finds highly interconnected regions within a network. These clusters can have different meanings and can facilitate the analysis of the network. In our PPI network, MCODE has been used for the identification of protein clusters. These were then analyzed and placed in a biological perspective. Various protein complexes have been detected with this tool as well as important subpathways. MCODE is available as a plugin in Cytoscape.

We ran two analyses. The loops were included both times. One analysis was run without fluff and with a degree cutoff of 2 (Figure I.15). This means that, for a node to be scored, it needs a minimum of two connections. The haircut was not activated. We wanted to see if a cluster was regulated by a transcription factor that could be only connected to the cluster by one

edge. The Node Score Cutoff was set to 0.2. This parameter influences the cluster size. Only nodes with a node score that deviates by less than 0.2 from the cluster's sed node's score, will be added to the cluster. Thus this small value of 0.2 creates rather small clusters. Increasing this value would increase the size of the clusters. The K-Core was set to 2. This is the default value. A higher value will reject smaller clusters and is thus not desired. The Maximum Depth is set to 100, a high value so that our clusters are not limited by distance. This led to the SMAD2-SMAD3-SMAD4-JUN cluster and the JAK-STAT cluster.

The second analysis was run with the fluff and a degree cutoff of 3 (Figure I.16). In order to not have too expanded clusters, the Node Density Cutoff was 0.1. Other parameters were the same. This gave the TRIKA2 cluster.

MCODE returns clusters with a score, node status, and cluster name. A higher complex score indicates a larger, more dense complex.

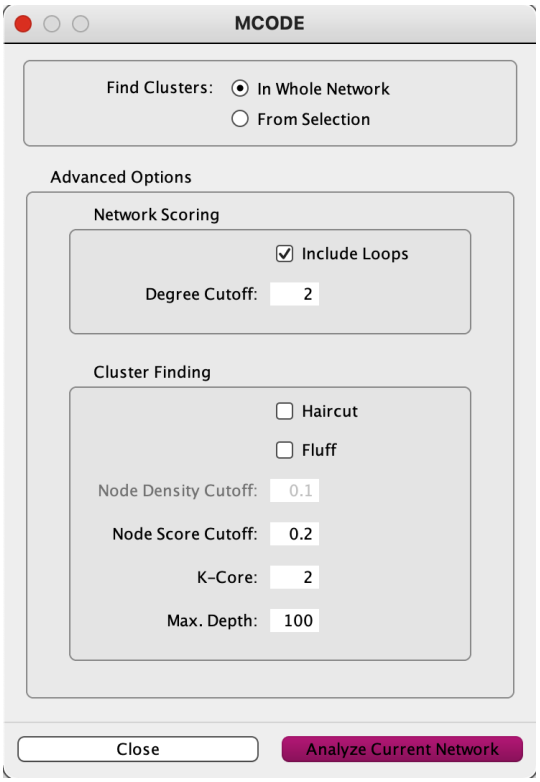


Figure I.15: MCODE Analysis 1

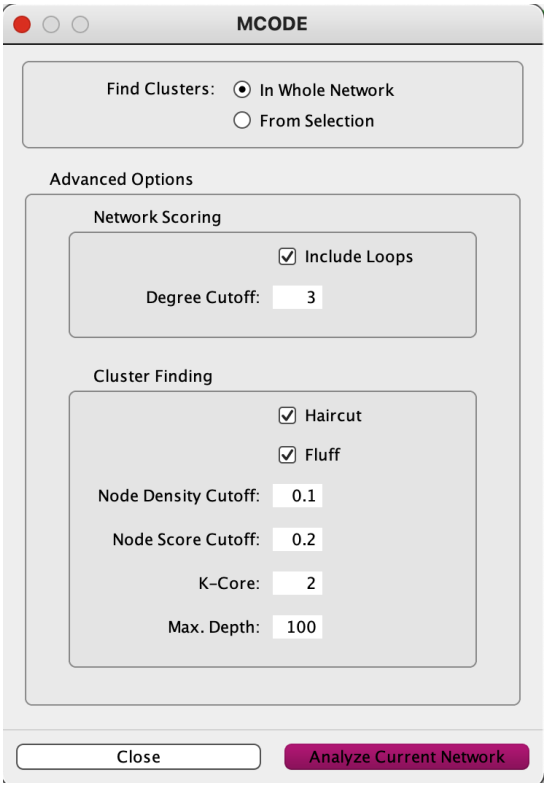


Figure I.16: MCODE Analysis 2

Appendix II

Group collaboration

Due to busy schedules, it was challenging to come together to work on our project. As a group, we agreed on communication through shared spreadsheets and docs. We made a drive folder so that everyone could easily access the most up-to-date information. Once completed, these files were imported to Cytoscape. This online collaboration was a good solution for us. As we decided to use Cytoscape individually, we all got familiarized with the software and were all able to operate it. We shared our .cys files with each other when deemed necessary.

After the final presentation, I did not attend the lectures anymore. The other group members decided to change plans and use a different cluster than mentioned and analyzed in the final presentation without notifying me. I used the TRIKA2 cluster from the final presentation in this report. They used a smaller cluster containing only three nodes. I was not aware until two days before the deadline and therefore did not adapt my report after this inconvenience.

Additional tables and figures

Table II.1 contains the 25 genes provided to us to start this project. ERK1 is referred to as MAPK3 in this report.

Table II.1: Original Gene List

ABCA1	STAT3
ABI1	RB1
ACVR1B	MYD88
CSF1R	CASP8
ICAM1	TRAF6
IL10	BIRC5
IRF1	RIPK1
GRB2	RELA
NFKB1	MYC
ERK1 (MAPK3)	MAP2K7
TNFA	MAP2K3
MMP1	DUSP1
CDKN1A	

Table II.2 shows the result of an overrepresentation analysis on the GO term molecular function for the original gene list. The same reference, annotation set, correction, and significance level have been used as other analyses in BiNGO. The most relevant results are cytokine receptor binding, tumor necrosis factor receptor family binding, MAP kinase kinase activity, and death receptor binding. These results are in accordance with the results obtained for overrepresentation analysis on the GO biological process.

Table II.2: Overrepresentation Analysis GO Molecular Function BiNGO (α = 0.0001)

GO ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes
19899	enzyme binding	4.4011E-10	1.1619E-7	16/25 64.0%	2042/18232 11.2%	RB1 ABCA1 MAP2K3 CDKN1A DUSP1 STAT3 ACVR1B RELA TNFA CASP8 ...
5126	cytokine receptor binding	1.7107E-9	2.2582E-7	8/25 32.0%	268/18232 1.4%	IL10 TNFA CASP8 TRAF6 STAT3 RIPK1 GRB2 MYD88
32813	tumor necrosis factor receptor superfamily binding	5.7999E-9	5.1039E-7	5/25 20.0%	49/18232 0.2%	TNFA CASP8 TRAF6 RIPK1 MYD88
31625	ubiquitin protein ligase binding	1.0363E-7	6.8394E-6	7/25 28.0%	296/18232 1.6%	RB1 CDKN1A CASP8 TRAF6 RIPK1 ACVR1B RELA
44389	ubiquitin-like protein ligase binding	1.5823E-7	8.3545E-6	7/25 28.0%	315/18232 1.7%	RB1 CDKN1A CASP8 TRAF6 RIPK1 ACVR1B RELA
42802	identical protein binding	3.5522E-7	1.5306E-5	13/25 52.0%	1954/18232 10.7%	RB1 CSF1R STAT3 RELA NFKB1 TNFA CASP8 TRAF6 BIRC5 RIPK1 GRB2 M...
19900	kinase binding	4.0584E-7	1.5306E-5	9/25 36.0%	760/18232 4.1%	RB1 MAP2K3 CDKN1A DUSP1 TRAF6 STAT3 GRB2 MAP2K7 RELA
4708	MAP kinase kinase activity	1.8333E-6	6.0500E-5	3/25 12.0%	18/18232 0.0%	MAP2K3 MAP2K7 MAPK3
19901	protein kinase binding	2.1407E-6	6.2795E-5	8/25 32.0%	677/18232 3.7%	MAP2K3 CDKN1A DUSP1 TRAF6 STAT3 GRB2 MAP2K7 RELA
5123	death receptor binding	2.9800E-6	7.8673E-5	3/25 12.0%	21/18232 0.1%	CASP8 RIPK1 MYD88

Table II.3 shows the result of an overrepresentation analysis on the GO term cellular component for the original gene list. For this analysis, a significance level of 0.05 was used because a significance level of 0.0001 did not yield any useful results. The confidence of the results is thus lower than for other BiNGO analyses. The other parameters were the same as in previous analyses. It shows membrane microdomain, membrane raft, and cytosol are overrepresented. Ripopsotome and death including signaling complex are also overrepresented, indicating cell death is important for our gene list.

Table II.3: Overrepresentation Analysis GO Cellular Component BiNGO (α = 0.05)

GO ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes
98857	membrane microdomain	3.8581E-6	3.3373E-4	6/25 24.0%	335/18979 1.7%	ABCA1 TNFA CASP8 RIPK1 ICAM1 MAPK3
45121	membrane raft	3.8581E-6	3.3373E-4	6/25 24.0%	335/18979 1.7%	ABCA1 TNFA CASP8 RIPK1 ICAM1 MAPK3
97342	ripoptosome	2.4907E-5	1.4363E-3	2/25 8.0%	6/18979 0.0%	CASP8 RIPK1
5829	cytosol	4.3050E-5	1.8619E-3	17/25 68.0%	5371/18979 28.2%	RB1 MAP2K3 CDKN1A STAT3 ACVR1B RELA NFKB1 CASP8 IRF1 TRAF6 AB...
31264	death-inducing signaling complex	5.9631E-5	2.0632E-3	2/25 8.0%	9/18979 0.0%	CASP8 RIPK1
5768	endosome	2.2413E-4	5.9997E-3	7/25 28.0%	996/18979 5.2%	ABCA1 TNFA TRAF6 RIPK1 GRB2 MYD88 MAPK3
32991	protein-containing complex	2.4276E-4	5.9997E-3	16/25 64.0%	5423/18979 28.5%	RB1 CSF1R CDKN1A STAT3 ACVR1B RELA NFKB1 CASP8 MYC TRAF6 ABI...
5654	nucleoplasm	6.9147E-4	1.4953E-2	13/25 52.0%	4040/18979 21.2%	RB1 MAP2K3 CSF1R CDKN1A STAT3 RELA NFKB1 CASP8 MYC IRF1 BIRC5 ...
5634	nucleus	1.1809E-3	2.2699E-2	18/25 72.0%	7584/18979 39.9%	RB1 MAP2K3 CSF1R CDKN1A DUSP1 STAT3 RELA NFKB1 CASP8 MYC IRF...
35525	NF-kappaB p50/p65 complex	1.3172E-3	2.2788E-2	1/25 4.0%	1/18979 0.0%	RELA
31981	nuclear lumen	1.5688E-3	2.4673E-2	13/25 52.0%	4386/18979 23.1%	RB1 MAP2K3 CSF1R CDKN1A STAT3 RELA NFKB1 CASP8 MYC IRF1 BIRC5 ...
98797	plasma membrane protein complex	1.9587E-3	2.6793E-2	5/25 20.0%	702/18979 3.6%	CASP8 TRAF6 RIPK1 GRB2 ACVR1B
1990682	CSF1-CSF1R complex	2.6328E-3	2.6793E-2	1/25 4.0%	2/18979 0.0%	CSF1R
70436	Grb2-EGFR complex	2.6328E-3	2.6793E-2	1/25 4.0%	2/18979 0.0%	GRB2
35189	Rb-E2F complex	2.6328E-3	2.6793E-2	1/25 4.0%	2/18979 0.0%	RB1
70557	PCNA-p21 complex	2.6328E-3	2.6793E-2	1/25 4.0%	2/18979 0.0%	CDKN1A
71159	NF-kappaB complex	2.6328E-3	2.6793E-2	1/25 4.0%	2/18979 0.0%	RELA
785	chromatin	4.0281E-3	3.8715E-2	6/25 24.0%	1205/18979 6.3%	RB1 MYC IRF1 STAT3 RELA NFKB1
1902554	serine/threonine protein kinase complex	4.9431E-3	4.2956E-2	2/25 8.0%	80/18979 0.4%	CDKN1A ACVR1B
43235	receptor complex	4.9660E-3	4.2956E-2	4/25 16.0%	536/18979 2.8%	CSF1R TRAF6 RIPK1 ACVR1B
33256	I-kappaB/NF-kappaB complex	5.2590E-3	4.3324E-2	1/25 4.0%	4/18979 0.0%	NFKB1
43228	non-membrane-bounded organelle	6.1744E-3	4.4902E-2	13/25 52.0%	5066/18979 26.6%	RB1 CDKN1A STAT3 RELA NFKB1 CASP8 MYC IRF1 TRAF6 ABI1 BIRC5 GR...
43232	intracellular non-membrane-bounded organelle	6.1744E-3	4.4902E-2	13/25 52.0%	5066/18979 26.6%	RB1 CDKN1A STAT3 RELA NFKB1 CASP8 MYC IRF1 TRAF6 ABI1 BIRC5 GR...
9986	cell surface	6.2567E-3	4.4902E-2	5/25 20.0%	920/18979 4.8%	ABCA1 CSF1R TNFA ACVR1B ICAM1
5694	chromosome	6.4888E-3	4.4902E-2	7/25 28.0%	1771/18979 9.3%	RB1 MYC IRF1 STAT3 BIRC5 RELA NFKB1
1902911	protein kinase complex	6.9007E-3	4.5916E-2	2/25 8.0%	95/18979 0.5%	CDKN1A ACVR1B
31265	CD95 death-inducing signaling complex	7.8785E-3	4.8678E-2	1/25 4.0%	6/18979 0.0%	CASP8
32133	chromosome passenger complex	7.8785E-3	4.8678E-2	1/25 4.0%	6/18979 0.0%	BIRC5

Table II.4 is a table with the results of the overrepresentation analysis in Panther. This analysis was performed in addition to Reactome and BiNGO. It helped for pathway/process selection.

Table II.4: Overrepresentation Analysis Panther Pathways Original Gene set

	Homo sapiens (REF)	Client Text Box Input					
PANTHER Pathways	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
Toll pathway-drosophila	2	1	.00	> 100	+	3.63E-03	3.03E-02
Toll receptor signaling pathway	58	6	.07	85.22	+	1.19E-10	6.62E-09
Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade	33	3	.04	74.89	+	1.09E-05	2.03E-04
p38 MAPK pathway	41	3	.05	60.28	+	2.02E-05	2.81E-04
Ras Pathway	73	5	.09	56.42	+	3.41E-08	1.14E-06
Apoptosis signaling pathway	118	8	.14	55.85	+	1.54E-12	1.28E-10
p53 pathway feedback loops 2	51	3	.06	48.46	+	3.75E-05	4.47E-04
Interleukin signaling pathway	86	5	.10	47.90	+	7.42E-08	2.07E-06
Oxidative stress response	56	3	.07	44.13	+	4.89E-05	5.45E-04
B cell activation	70	3	.08	35.31	+	9.26E-05	9.09E-04
EGF receptor signaling pathway	141	5	.17	29.21	+	7.87E-07	1.88E-05
T cell activation	86	3	.10	28.74	+	1.67E-04	1.55E-03
Gonadotropin-releasing hormone receptor pathway	231	8	.28	28.53	+	2.64E-10	1.10E-08
FGF signaling pathway	123	4	.15	26.79	+	1.57E-05	2.38E-04
CCKR signaling map	172	5	.21	23.95	+	2.04E-06	4.25E-05
PDGF signaling pathway	147	4	.18	22.42	+	3.10E-05	3.98E-04
Angiogenesis	175	4	.21	18.83	+	6.02E-05	6.29E-04
Inflammation mediated by chemokine and cytokine signaling pathway	255	5	.31	16.15	+	1.33E-05	2.21E-04
Integrin signalling pathway	193	3	.23	12.81	+	1.67E-03	1.47E-02
Unclassified	17977	6	21.82	.27	-	8.28E-13	1.38E-10

In order to expand the network, a pathway has been chosen to work out the 'Positive Regulation of Apoptosis' process. In Reactome, we searched for an overrepresented pathway related to the positive regulation of apoptosis. TNFR1-Induced proapoptotic signaling, a sub-pathway of Death Receptor Signaling, was selected (Figure II.1).

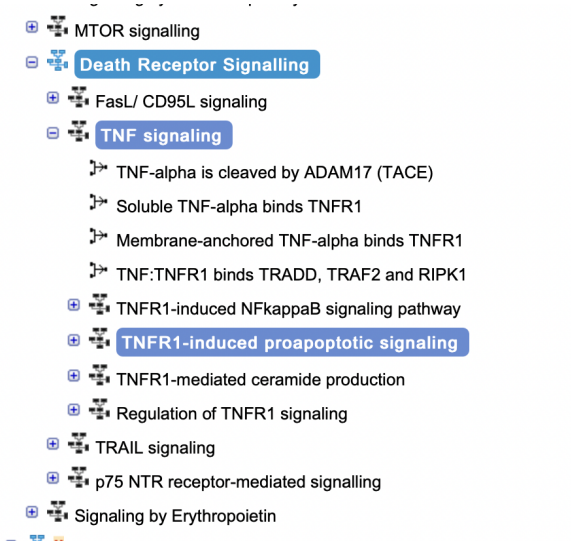


Figure II.1: Pathway Selection in Reactome

In Figure II.2 the 85 genes of the network are shown. The original nodes are purple, the added nodes blue. Transcription factors are oval, while other types of proteins are rectangular.

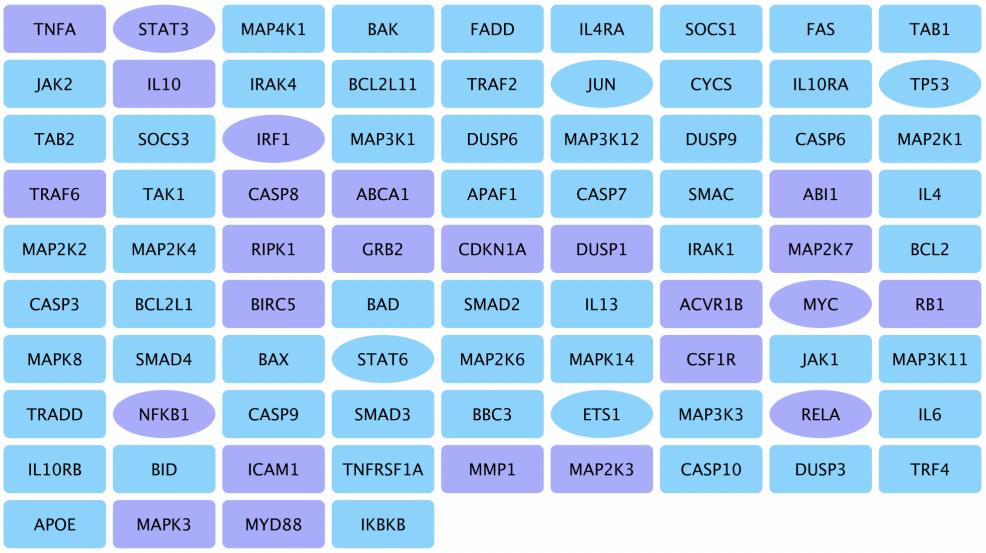


Figure II.2: Original (Purple) and Added (Blue) Genes in Cytoscape

In addition to BiNGO and ClueGO, an overrepresentation Analysis was performed in Reactome. The results of this analysis are given in Table II.5. They confirm we successfully expanded the network on the selected pathways and biological processes.

Table II.5: Overrepresentation Analysis Reactome Extended Network

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species name
Interleukin-4 and Interleukin-13 signaling	35	211	0.015	1.11E-16	2.22E-15	44	47	0.003	Homo sapiens
MyD88 cascade initiated on plasma membrane	23	94	0.007	1.11E-16	2.22E-15	53	58	0.004	Homo sapiens
Toll Like Receptor 10 (TLR10) Cascade	23	94	0.007	1.11E-16	2.22E-15	53	59	0.004	Homo sapiens
Toll Like Receptor 5 (TLR5) Cascade	23	94	0.007	1.11E-16	2.22E-15	53	59	0.004	Homo sapiens
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	23	101	0.007	1.11E-16	2.22E-15	43	48	0.004	Homo sapiens
MyD88 dependent cascade initiated on endosome	23	102	0.007	1.11E-16	2.22E-15	56	63	0.005	Homo sapiens
Intrinsic Pathway for Apoptosis	23	64	0.004	1.11E-16	2.22E-15	54	62	0.005	Homo sapiens
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	24	118	0.008	1.11E-16	2.22E-15	55	64	0.005	Homo sapiens
Toll Like Receptor 7/8 (TLR7/8) Cascade	23	103	0.007	1.11E-16	2.22E-15	56	66	0.005	Homo sapiens
TRIF(TICAM1)-mediated TLR4 signaling	24	107	0.007	1.11E-16	2.22E-15	49	58	0.004	Homo sapiens
MAP kinase activation	19	69	0.005	1.11E-16	2.22E-15	27	32	0.002	Homo sapiens
Toll Like Receptor TLR6:TLR2 Cascade	24	118	0.008	1.11E-16	2.22E-15	55	66	0.005	Homo sapiens
Toll Like Receptor TLR1:TLR2 Cascade	24	121	0.008	1.11E-16	2.22E-15	55	66	0.005	Homo sapiens
Toll Like Receptor 9 (TLR9) Cascade	23	106	0.007	1.11E-16	2.22E-15	56	68	0.005	Homo sapiens
MyD88-independent TLR4 cascade	24	107	0.007	1.11E-16	2.22E-15	49	60	0.004	Homo sapiens
Toll Like Receptor 2 (TLR2) Cascade	24	121	0.008	1.11E-16	2.22E-15	55	68	0.005	Homo sapiens
Toll Like Receptor 3 (TLR3) Cascade	24	102	0.007	1.11E-16	2.22E-15	49	61	0.004	Homo sapiens
Apoptosis	29	192	0.013	1.11E-16	2.22E-15	111	141	0.01	Homo sapiens
Interleukin-17 signaling	19	77	0.005	1.11E-16	2.22E-15	27	35	0.003	Homo sapiens
TP53 Regulates Transcription of Cell Death Genes	17	83	0.006	1.11E-16	2.22E-15	52	68	0.005	Homo sapiens

Figure II.3 shows the result of the Network Analyzer on the Merged network when it is considered to be undirected. This is used for comparison with our own extended network. The trends for the parameters are different for the merged network because there are many more nodes. This provides good insight into a real biological system.

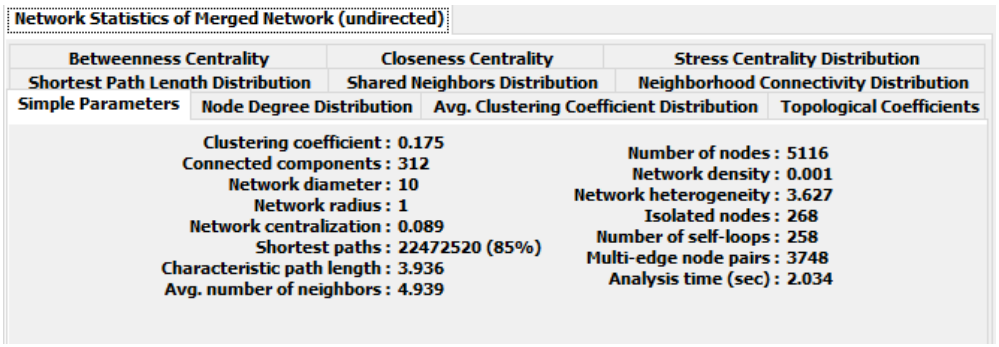


Figure II.3: Network Analyzer Cytoscape (Merged Undirected network)

In Figure II.4 the 85 nodes are shown in red, blue, or white. The genes are overlaid with omics data of CMS2 CRC. They are colored according to the scale given in figure II.5 which gives a red color for genes that are more expressed in colorectal cancer, and blue for genes with a lower expression. The expression of genes in white is not affected in colorectal cancer cells.

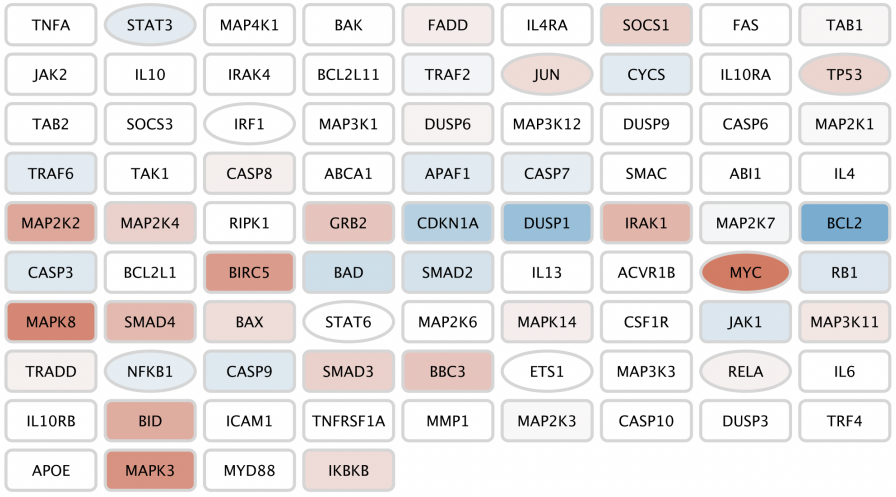


Figure II.4: Genes overlaid with Omics data CMS2 CRC

Figure II.5: Color Scale Omics Data (based on logFC)