Universiteit Leiden
The Netherlands

Opleiding Informatica

A Network Analysis Approach to studying PolyQ Protein Interactions

L.M.I. Janssens

Supervisors:
Katy Wolstencroft & Frank Takes

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

22/07/2020

Abstract

Polyglutamine (polyQ) diseases are a group of rare neurodegenerative diseases that are caused by a protein with an expanded glutamine (Q) repeat. Although the mutation is known for these diseases, the native function of the non-mutated proteins has not been fully elucidated for most. This is also the case for many other polyQ proteins, although they seem to be important in protein interactions. One of the methods to study protein functions is by analysing a protein-protein interaction network. As the disease proteins cause a neurodegenerative disease, the focus of this research is on brain tissue. Mouse (Mus musculus) brain cells are a common source for protein-protein interaction data because there is more in vivo data available than in human. However, some proteins that are polyQ in human are missing the polyQ domain in the mouse orthologue, with only two disease polyQ proteins in human also being polyQ in mouse. An interactionnetwork was constructed to find common topological and inter-actor functions among polyQ proteins-, disease proteins- and human orthologs that are not themselves a polyQ -protein. This research has shown that the disease proteins are more highly connected than other groups and have a different functional enrichment profile. polyQ proteins have a higher average degree than non-polyQ orthologues, but they in turn have higher centrality. Both groups have very similar enrichment profiles, showing that both polyQ and polyQ orthologues are important in the network and are involved in protein binding activities.

Contents

1 Introduction

This section provides an introduction that explains some of the background information necessary for this thesis. It shortly explains polyglutamine proteins, networks, the biological implications of some of the topological results and clustering of the network.

1.1 polyQ protein

Polyglutamine (polyQ) diseases are a family of neurodegenerative diseases that involve a protein with an expanded glutamine (CAG) repeat. To date, nine different polyQ diseases have been described. The most common of these diseases is Huntington disease (HD) caused by a CAG expansion in the *HTT* gene. The other eight diseases are: SCA1, SCA2, SCA6, SCA7, SCA17, Machado-Joseph disease (SCA3), DR-PLA and spinal and bulbar muscular atrophy. The exact mechanisms behind these diseases remains unknown. These diseases can have a severe effect on the health of a patient. For example in Huntington disease, the expanded CAG segment, at least more than 36 glutamines [Totzeck et al., 2017], leads to the production of extremely long Huntington proteins. These proteins accumulate in neurons, disrupting their function, see Figure 1. This process mainly affects the striatum and cerebral cortex. Another often occurring symptom is that with each new generation that carries the disease *HTT* gene, the polyQ tract grows in size [McColgan and Tabrizi, 2018]. This is associated with the sign of symptoms in an earlier stage of life. Although there have been many observations on the effects of the accumulated mutated Htt protein, relatively little is known about the natural Htt function. It has been found to interact with many other proteins. Studying these interactions, and those of related polyQ proteins may offer new insights into common functions.

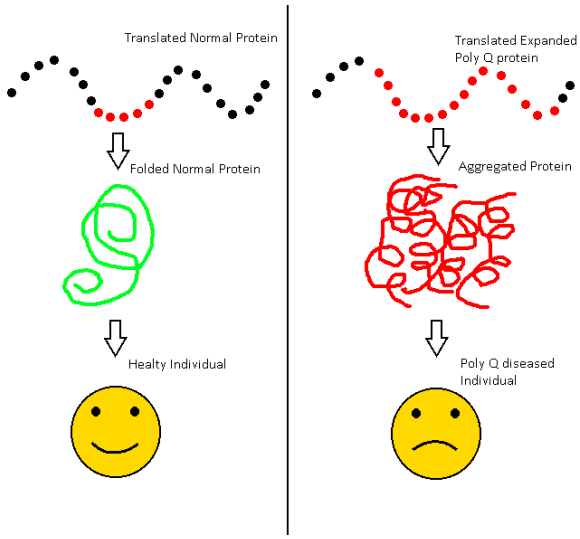


Figure 1: Cartoon of the possible working of a polyQ protein

Polyglutamine proteins are proteins that contain a polyglutamine tract. In healthy individuals, this polyQ tract can have a minimum of 8 out of 10 glutamines as described by [Totzeck et al., 2017]. This glutamine is encoded by CAA and CAG codons. This minimal length should be enough

for a polyQ tract to do its function but not be pathogenic. Still, according to the research of Totzeck et al, even a small change in the polyQ tract can have a pathogenic effect. With Huntingtin for example, a polyQ region of 34 glutamines is not pathogenic while a region of 36 glutamines is pathogenic. Figure 1 shows how a polyQ tract can cause an aggregate of proteins that then block the function of the protein as described by [Schaefer et al., 2012]. Cohen-Carmon and Eran Meshorer [Cohen-Carmon and Meshorer, 2012] describe that disrupted chromatin regulates may be directly involved with the pathophysiology of polyQ-related diseases. They describe that the mutant Htt has a stronger interaction with CBP, which leads to a CBP depletion and thus to hypoacetylation of histones.

1.2 Networks

To take a better look into the interactions between these disease proteins and the other proteins we create a protein-protein interaction network, because these interactions can potentially show cooperation between proteins that can drive a biological process [Robertson et al., 2011]. Proteins such as Htt interact with other proteins in a manner that can be described like the interactions in a social network. Just like the interactions between people can be visualized in an undirected Facebook network, protein-protein interaction data can be used to create an undirected protein-protein interaction (PPI) network.

1.2.1 Networks Topology

This undirected PPI network exists of nodes that represent protein and edges that represent interactions between two proteins. Complete protein-protein interaction networks have been shown to have scale-free like properties [Albert, 2005]. A network is scale free when the node-degree distribution follows a power law. If the network is scale-free, mathematical properties can be used to gain a new or better understanding of the protein's functions. One of the properties of a scale-free network is that they display the so-called small world property, which states that usually any two nodes are an average of six nodes away from each other in the network.
With this scale-free property there are some topological properties that are important for the description of PPI networks. For example the giant component in the network which is the largest connected group of nodes. Another method of grouping nodes is by finding their first neighbours (FN), which are al the nodes that are directly connected to the target node. The degree of a node is the number of directly connected nodes to the target node [Dong and Horvath, 2007]. The average shortest path, shortest path being the smallest distance between two nodes, indicates how closely connected the network is, or how connected a node is. The network centralization, also degree centralization, is an index of the connectivity distribution. The heterogeneity value, another network description value, indicates if a network contains hub nodes.
Another important method for network description or analysis is centrality analysis. This can be used to find out which protein in the network is the most important and why. There are different forms of centrality: degree centrality, betweenness centrality and closeness centrality. While the degree centrality gives a rough estimate of centrality because it is only a local measure, it still can be used to identify hub proteins in networks [Yu et al., 2007].
The betweenness centrality is based on the number of shortest paths in the graph that pass through the node [Yu et al., 2007]. The proteins in the network that have the highest betweenness are likely

to be involved in a signal pathway and can even have a crucial function [Yu et al., 2007]. Yu et al describe these proteins as bottleneck proteins in the network. According to Yu et al hubs are the top 20% with the highest degree distribution and the bottlenecks in the network are the top 20% of the nodes with the highest betweenness centrality value. Their findings also indicate that in PPI networks the degree measurement is a better indicator of essentiality than betweenness, because a PPI network is undirected. The closeness centrality measures the shortest paths from a particular node to all nodes it can give an indication of how relevant a protein can be for the other proteins in the network.

1.2.2 Clustering

Protein-protein interactions can be stable or transient [Nooren and Thornton, 2003]. Stable protein-protein interactions form permanent interactions that fulfil a certain biological role. These stable interactions are also called protein complexes. Transient protein-protein interactions are interactions that only occur for a short period to fulfil a biological role and then move on. To find these protein complexes community detection or clustering is used. There are specific protein-protein interaction clustering algorithms that are focussed on biological networks. There are several options for clustering biological networks like, MCODE, SR-MCL, DME, NeMo, etc.

The algorithm used in this research to find clusters (densely connected regions in the network that indicate protein complexes) is Molecular Complex Detection (MCODE) [Bader and Hogue, 2003]. The MCODE algorithm is a heuristic-based algorithm that operates in three phases [Bhowmick and Seah, 2016]: (1) vertex weighting, (2) complex prediction and (3) post-processing. In the first phase the nodes are weighted based on the local network density value using the highest k-core of the node neighbourhood. Besides the calculation of the k-core (minimum amount of degrees), the density of the neighbourhood is calculated and taken into the weighing. The second phase predicts molecular complexes using the node weights (calculated in the previous phase) in a greedy seed and extend manner. The seed gets expanded until a user specified threshold (vertex weight percentage) is reached. In the third and final stage the complexes are pruned and so called fluffed (extra nodes, not from the cluster are added) if necessary. Complexes can be filtered out if they do not meet a minimum node amount. After these steps the complexes are scored and ranked based on their density. In general, the clustering algorithm takes as an input the entire giant component and then finds densely-connected regions in this network. The algorithm outputs all the clusters that it could identify. There is no overlap between the clusters [Lin et al., 2007].

1.3 Related work

Over the last 10 years a lot of research about the function of polyQ protein has been done. In this section research that has a similar subject or method to this study will be discussed.
A study by [Tourette et al., 2014] looks into Huntington disease by creating a protein-protein interaction network. They study the network by analysing the network statistics, functional enrichment, and clustering. They start with a network based on the Htt-interacting proteins that are from an

earlier research that described a Y2H screen using Htt as bait. This resulted in 102 high confidence interactions with Htt. They then added the first inter-actors of these 102 inter-actors to form the HD-net network. After this they compared the network to a random network of proteins gathered from de Human Protein Resource Database (HPRD).
For the network topology analysis, they calculated the shortest path with Matlab package: MATLAB BGL. HD-net had a shortest path of 18.25. This network centred around Htt has 3235 interactions with 2141 proteins. All network metrics were based on the largest subgraph of HD-net. For the functional annotation of the network a total data set and subgroups of primary and secondary partners of Htt were analysed using Ingenuity Pathway Knowledge Base. For the functional analysis they used a p value of 0.01. The enrichment data indicated that Htt is involved in membrane dynamics, cell attachment and motility.

Another study by Schaefer et al analyses the function of CAG (Q) repeats in PPI networks in different species (total of 11) [Schaefer et al., 2012]. The interactions are not studied in a network but as a list of interactions and enrichment data. The human protein-protein interactions are from the HIPPIE database and the interactions from the other species are from the BioGRID database. They specified polyQ protein with a minimum length of 10 Q's with a maximum of one mismatch. The results relevant for this research is the finding that polyQ proteins are probably stabilizing protein-protein interactions through changes in their structural form. This alteration probably leads to the protein aggregation that is lethal in the disease varieties.

1.4 Problem statement and research question

Even though there has been a lot of research into the workings of the disease variants of the proteins, a lot still remains unknown about the workings of regular poly-Q protein. With this research we hope to find some new insights into the workings of poly-Q protein by studying the differences or commonalties between inter-actors in disease and normal variants. Because all the disease proteins have a neurological impact, only brain proteins will be studies to limit the amount of proteins. Since there is far more information about protein-protein interactions in the brain in mouse databases than there is in human databases, the choice was made to limit the research to mouse PPI data. This can be summarized using the following research statement (RS) and three research questions (RQ).

RS: Investigate the similarities and differences between polyQ protein-protein interactions in Mus *musculus* brain to find common inter-actor functions in normal and disease states.

To further address the research statement, it is best to divide it into individual research questions. The first method that is used to find commonalities is to look for common functional signatures in polyQ inter-actors. These polyQ inter-actors show can show the biological function of the included proteins. When looking for the polyQ inter-actors the following RQ1A can be asked.

RQ1 A: What are the common functional signatures in polyQ inter-actors in the network?

A part of the disease proteins in mice does not fit the polyQ protein criteria as mentioned in [Totzeck et al., 2017]. Only Tbp and Ar classify as a polyQ protein. The other disease proteins are only a polyQ protein in transgenic induced mice. This means that the possibility of a difference between these proteins should be explored.

RQ1 B: What are the differences between inter-actors of mouse polyQ proteins and mouse orthologues of human polyQ proteins that are not themselves polyQ proteins?

If any common inter-actors are found among the polyQ proteins, the next step is to find out if these common inter-actors or their mathematical properties differ from the normal polyQ proteins.

RQ 2: How do the polyQ disease protein inter-actors differ from ones not associated with disease?

To try and find answers to these questions, a protein-protein interaction network is created with mouse brain data from MGI's MouseMine database [Motenko et al., 2015]. These proteins are then imported into Cytoscape using the StringDB plugin. StringDB retrieves all the possible interactions above a certain threshold and turns them into a network in Cytoscape. This network was then analysed using functional enrichment (STRINGDB enrichment), statistical analysis, clustering, and network descriptions.

1.5 Thesis Overview

Section 1 gives a introduction to this research by explaining the basic principles needed. Section 2 provides an overview of the used materials and methods in this thesis. Section 3 is used to give an overview of the results. In section 4 the results of this thesis are discussed, and the conclusion will be given in section 5.

2 Materials and Methods

The aim of this section is to provide a background in order to give a full understanding of the used techniques and tools in this research. The subsections explain the concepts and implementation of all the different processes.

2.1 Data gathering and cleaning

To gather all mouse brain proteins, as described in RS 1.4 *MouseMine.org* was used. This is an online data warehouse created to provide MGI mouse data using the InterMine framework [Motenko et al., 2015]. This online application has an interface where you can filter all the mouse protein on different values, among which is expression tissue. When filtering on brain tissue, several types of tissue, like brain, cerebellum, cerebral cortex, midbrain, etc... are returned. This list of proteins is then downloaded as an .tsv file. This file contained a lot of duplicate data, because some of the proteins are expressed in multiple tissues, a small python script was used to remove duplicate proteins in the file. The script filtered on doubles based on the MGI:ID instance. The resulting file contained 9646 unique proteins.
There was then one issue of mapping. The STRINGDB tool that was used to import the PPI data into Cytoscape, see Figure 2, did not accept MGI symbols or MGI ID's as input but needs the input to be EnsemblProtein ID or Ensembl Gene ID. Some identifiers were not mapped to other identifiers. The optimal mapping strategy was between MgI-Symbol and Ensembl-protein ID. This resulted in the largest amount of nodes (9156 out of 9646) compared to the other possibilities.
A more compact and understandable view of this process is shown in Figure 2 where the entire process to eventually create the network in Cytoscape is visualized.
The mapping process was done using the Ensembl BIOMART tool [Kinsella et al., 2011]. This tool allows to map different ID's to other ID's or ortholog ID's. This list of Ensembl protein ID's was then used to create the network.

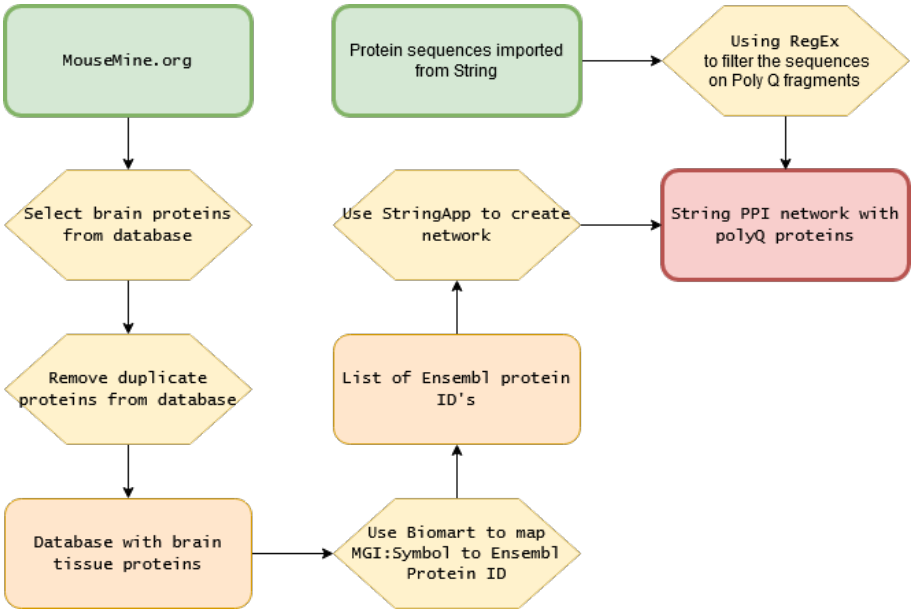


Figure 2: This workflow shows the data cleaning and gathering process and how the final network was created. The green coloured rectangles are the "source" resources for the network. The hexagons are actions performed on the data and the red coloured rectangle is the final network.

2.2 Network Creation

This subsection explains the process of how the network was created using Cytoscape and various other tools. First a brief concept of the process is given.

2.2.1 Concepts

The list of brain proteins from MGI is the source for the unique nodes in the network that is to be created. This list of brainproteins does not contain any information about the interactions. To build the network, all the different interactions, or edges, between the nodes should be identified. The most used source for this is StringDB [von Mering et al., 2005].
StringDB is a database that contains a collection of protein-protein interactions from different species that are quality controlled. The interactions are from high-throughput experimental data, databases and literature and from predictions. Databases that are used as source for the StringDB provide different types of evidence. PPI databases, such as BioGRID, MINT, DIP. Pathway databases, such as KEGG and reactome. Literature evidence from PubMed and genome resources from ENSEMBL/swissprot. Databases like COG provide phylogenetic evidence. The GeneOntology (GO) database is used by the StringApp as a source for enrichment data.

All these databases and their evidence channels are combined into a matter that integrates the probability from all these different channels into an evidence score. The list that follows contains all the proteins and the proteins they interact with, allong with other data gathered by string. The nodes now have besides their name data like, protein sequence, species, tissue expression score and more. The edges contain data on the different types of evidence and the score that they gained for

this evidence as well as the total evidence score. This list of protein interactions with all of the meta-data can then be visualized.

2.2.2 Implementation

The application that is used to visualize and analyse the network in this research is Cytoscape. Cytoscape is an open source software platform for visualizing protein interaction networks and biological pathways and combining these networks with annotations, gene expression profiles and other types of data [Shannon et al., 2003]. The version of Cytoscape used in this research is 3.7.2. The Cytoscape environment also allows for the installation of plugins using the build in Appstore. These plugins add new functionalities that range from algorithm layout to enrichment. The plugins that are used in this research are: StringApp (version: 1.5.5) [Doncheva et al., 2019], YfilesLayourAlgorithm (version: 1.1) [YWorks, 2020] and MCODE (version 1.6.1) [Bader and Hogue, 2003]. StringApp is the most critical plugin for this research.
StringApp is a plugin that acts as an API for the above mentioned StringDB. To import the list of brain proteins from MGI, StringApp was used. When importing proteins into Cytoscape using the stringApp, the user can specify a cut-off-score that specifies the minimum amount of evidence for an interaction between proteins. Alongside the proteins and their interactions stringApp imports the additional data from StringDB. The list of brain proteins described in section 2.1 passed through the StringApp plugin using the protein query option with a cut-off-score of 0.7 and specified on Mus *musculus*. The cut-off-score of 0.7 was chosen to guarantee interactions with a high confidence [von Mering et al., 2005]. The giant component of the resulting network was then exported to a separate network file.

2.3 Clustering

Another step in the nework creation was the clustering process. The networks giant component (GC), that that was identified while creating the network as described in the previous subsection 2.2 was clustered to find possible protein complexes using the MCODE algorithm. For the MCODE algorithm the parameters used were: k-score 3, node score cut-off of 0.3 and a degree cut-off of 3. These values are slightly increased compared to the standard values but are chosen after doing tests with different values and finding that these parameter values resulted in the clusters with the larger number of clusters with a minimum size of 3.

2.4 Network Enrichment

To answer the research-questions and research statement polyQ protein and disease proteins have to be identified in the network. This subsection explains the process of how disease proteins, polyQ proteins and the orthologs that are not a polyQ protein themselves (polyQO) were identified using Cytoscape and it's plugins. First a brief concept of the process is given.

2.4.1 Concepts

Identification of the disease protein follows out of identifying them by their protein name and labelling them. The polyQ protein were can be identified on the protein sequence that gets imported

from the StringDB. Because polyQ have a minimum of 8 out of 10 Q's on a domain, regular expression can be used on a basic filtering to identify all the different variants of that 10 amino acid long sequence.
Besides the disease proteins and the polyQ proteins the ortholog proteins still had to be identified (process is shown in Figure 3. To identify these proteins, different methods were used to ensure that all human polyQ orthologs were found. The polyQ2.0 database is a database that consists off all human polyQ proteins. The database contains the 9 disease proteins, 105 reviewed non-disease proteins and 146 unreviewed proteins. The unreviewed proteins are protein that have not been experimentally confirmed yet.

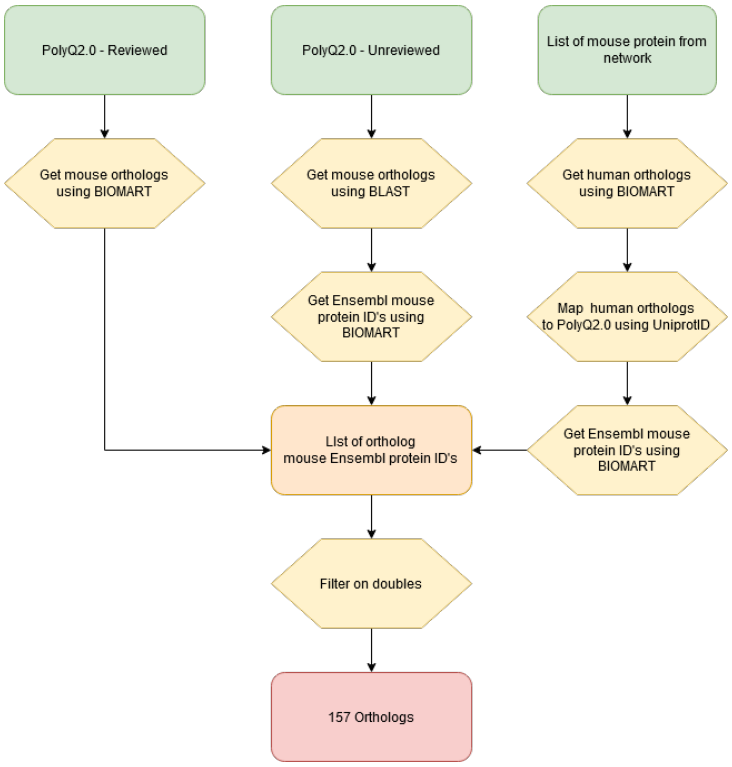


Figure 3: This Figure shows the control process for the ortholog proteins in the network. The green coloured rectangles are the input for the workflow. The hexagons are actions performed on the data and the red coloured rectangle is the end product.

To identify the mouse orthologs for the reviewed proteins was a more straightforward task. This was completed using Ensembl's BIOMART tool. ENSEMBL maintains a list of accepted orthologues for model organism species that can be used to map the UniprotID's to EnsemblProteinIDs. To find the mouse orthologs of the unreviewed protein NCBI's protein BLAST [Madden et al., 1996] variant BLASTP was used. The Uniprot ID's from the polyQ2.0 database had to be converted to FASTA format using the uniprotmapping tool to prepare for the BLASTP process. The protein BLAST was performed using the default settings (Db: Non-redundant protein sequences, BLASTP, 100 Max target sequences, Exp. treshold: 10, Word size: 6, Matrix: BLOSUM62, Gap Costs: Existence: 11 Extension: 1), but limited the BLASTP to mouse proteins. The resulting output file from this BLASTP contained several types of identifiers that had to be mapped to EnsemblProteinID. As a

final check to minimize loss of information in the mapping process. All protein identified in the network were mapped to their human ortholog, and then compared to the PolyQ2.0 database by converting the human orthologs to a Uniprot ID. The proteins in the polyQ2.0 database were then added to the complete list of mouse ortholog EnsemblProteinID's (converted using BIOMART). The resulting list was then filtered for duplicate proteins and contained all unique polyQO proteins. After this first neighbour networks can be created of all the identified subgroups by adding all the direct inter actors of a target node into a subgroup.

2.4.2 Implementation

To identify the disease proteins, the Cytoscape select tool was used to filter on the DisplayName instance from the node's table. The disease proteins were then marked using the style options that Cytoscape offers. The same Cytoscape filtering was used for the polyQ protein. This filtering was done using a simplified form of Java regular expressions to mark all proteins in the network that have at least a polyQ trac of 10 with an allowed mismatch of 2 [Totzeck et al., 2017]. These proteins were again styled using the provided Cytoscape tool. The orthologs identified using the method described in Figure 3 were then identified in the network and again marked. Using the select tool the disease proteins were selected and with the built in First-Neighbours tool, the FN of these proteins were identified. Separate FN networks for each individual disease protein and one with all the disease proteins selected and their common FNs were created. Besides the disease protein subnetworks, there was a subnetwork created for all the polyQ protein in the GC and their common FN, and finally there was a subnetwork created for all the polyQO protein and their FN.

2.5 Network Topology

The network topology of all the subnetworks was calculated using the in Cytoscape build in NetworkAnalyzer Tool. The mean and median degrees of the polyQ, disease protein and ortholog groups in the GC were computed by exporting the network tables to Excel. The hub proteins and bottleneck proteins described in section 1 was calculated by using the select tool to find the marges that resulted in highest 20% of the total amount of proteins in the giant component.

2.6 Functional Enrichment

This subsection describes the concepts and implementation to functionally enrich the network. This is necessary to find out what possible function the proteins or subnetworks in the network have.

2.6.1 Concepts

For this step another layer of data is put over the network to be able to analyse the possible protein functions. In this case it is data from the GeneOntology (GO) database [Ashburner et al., 2000] [Consortium, 2019]. The GO database contains a unified "language" to describe the function, processes, and cellular processes in biology. This processes have different ID's or terms that are linked to certain protein. To enrich the network the proteinID's of the nodes in the network are linked with the proteinID's of the different GO-terms. The results is a network that can show

the possible function or process of every node in the network. This then has to be visualized in a manner that all the different GO-terms can be compared and analysed.

2.6.2 Implementation

Besides importing interactions and other protein data, the StringApp enables enrichment analysis over string networks that is calculated from GO annotation for networks with a maximum size around 1000 nodes. All the subnetworks were were functionally enriched with StringApp that uses a hypegeometric test with a significance value of 0.05. This tool enriches the network with different forms of functional enrichment terms. The enrichment process is visualized in Figure 4. All the enrichment data for the networks was exported to a csv file in order to get them to other tools for visualization. The GO terms for processes and functions were then with their respective p-values exported to the online REVIGO. REVIGO [Supek et al., 2011] is a tool that is able to visualise GO-terms in a lot of different manners (bubble plots, word clouds, tree-map, etc.) to make them more interpretable for the human eye then the StringApp allows.

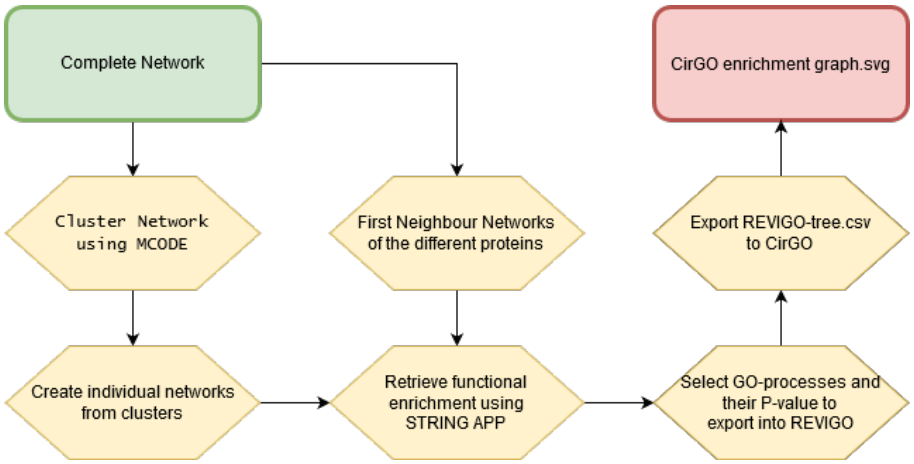


Figure 4: The enrichment process of the subclusters, in this example for GO_processes. The green coloured rectangles are the "start-files" for the workflow. The hexagons are actions performed on the data and the red coloured rectangle is the end product.

To make the enrichment data even more interpretable the REVIGO-treemap output (.csv) is put through another tool called CirGO to visualise the data in more interpretable pie charts with the most common processes or functions. These pie charts order the GO-terms based on their GO hierarchy. All the terms with the highest GO hierarchy are shown in the legend. The charts were then compared to check for commonalities between them.

3 Results

In this section the results of this research are shown. In the subsection, network, the general results of the network are discussed. The clustering subsection shows the clusters that were found. The network topology results off all the different subnetworks are shown in the next subsection. The last subsection shows all the results for the functional enrichment.

3.1 Network

The network totals 9168 nodes and 128904 interactions in this network. The GC that is 7582 nodes and 128809 interactions. Because all disease proteins are present in the GC, the rest of the results are based on the GC, see Figure 5. After filtering for polyQ proteins using the definition mentioned in the introduction 1 81 polyQ protein were identified in the network. Out of those 81 proteins, only 2 disease protein were identified as polyQ protein in Mus *musculus*, Ar and Tbp. The process of finding orthologs as described in Figure 3 started with a total of 260 human polyQ proteins. For the 114 reviewed and 146 unreviewed protein 157 mouse orthologs were found. Out of the 157 mouse orthologs 36 were identified in the network. Out of these 36, 12 turned out to be polyQO protein. The disease protein were checked using blast and seven were classified as a polyQO as wel.

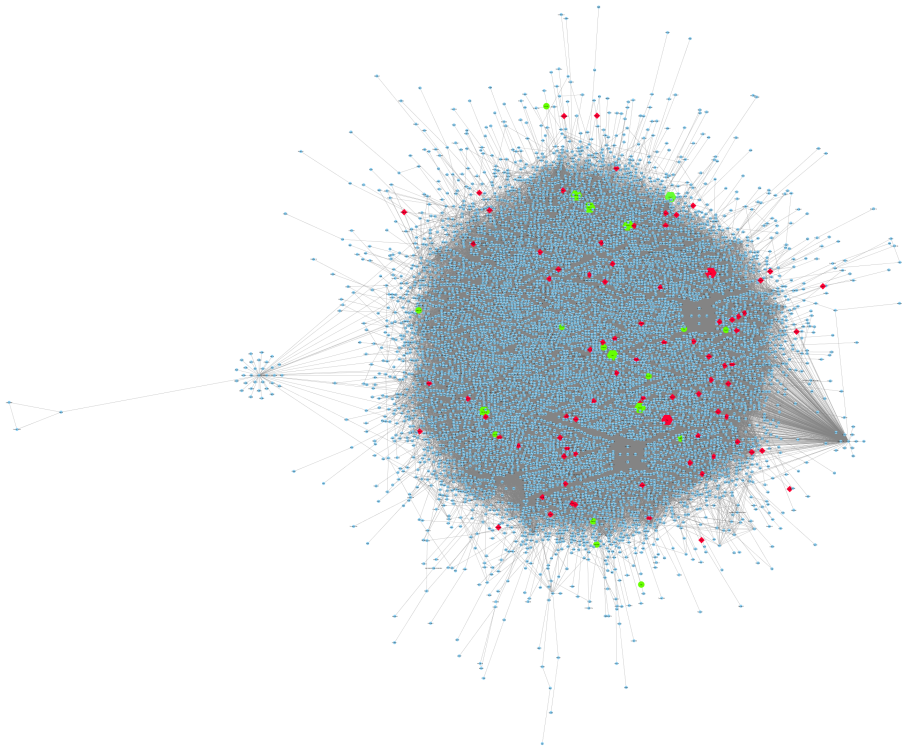


Figure 5: Giant component. The green nodes in the network are polyQO (19), the largest hexagons are disease proteins (9) and the red coloured nodes are polyQ protein (81).

3.2 Clustering

The MCODE clustering resulted in 85 clusters but not each of clusters had proteins that were of interest. The clusters that had either a disease protein, a polyQ protein or a Ortholog were marked as interesting. This resulted in the 15 clusters that are shown in the appendix Table 12. Cluster 4, 8 and 64 were chosen for their polyQ protein to continue with the enrichment process, besides these, clusters 7, 22 and 73 were chosen because they contain a polyQO. As can be seen in Table 12 there are not a lot of clusters that have polyQ protein and most of them only contain 1 polyQ. The three clusters that contain DP only have an average of 2 DP.

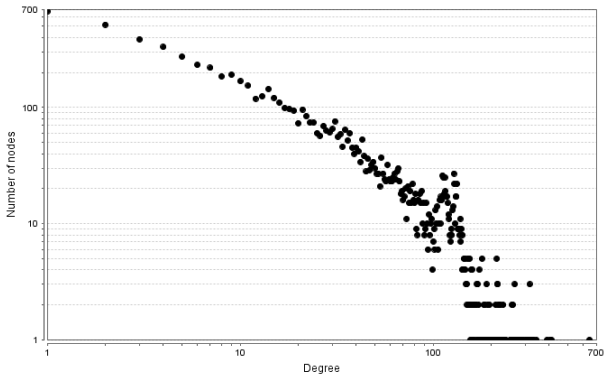
3.3 Network Topology

This section looks to analyse the various networks using their topological properties. The average number of neighbours in the GC is 33.69 neighbours and a network density of 0.004 as can be seen in Table 1. The GC of the network has a centralization of 0.083 and an average shortest path (path length) of 3.713) as shown in Table 1. This is a bigger pathlength than the one found in the FN networks but can be explained when looking at the shortest path distribution that is shown in Figure 6b.

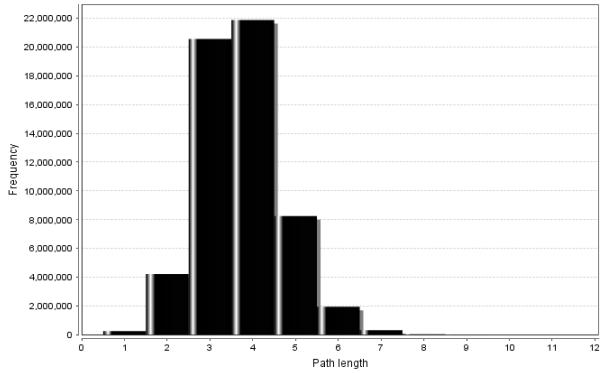
	Large Network	DP	Ortholog !Q	polyQ FN
#Nodes	7582	309	374	1529
clustering coefficient	0.45	0.58	0.51	0.62
Diameter	11	5	6	9
Radius	6	3	1	5
Centralization	0.08	0.32	0.20	0.23
Shortest paths	57479142	95172	137276	2336312
Path length	3.71	2.54	2.83	2.84
avg. #eighbours	33.98	25.07	15.94	58.15
Network Density	0.00	0.08	0.04	0.04
Heterogenity	1.35	0.84	0.85	0.90

Table 1: Table shows the different topological results for the FN networks of the disease protein (DP), polyQ and polyQO

The degree distribution of the GC is shown in Figure 6a, the mean degree or avg. number of neighbours of the GC is 33.978. The other values in Table 1 show that besides the GC, the polyQ FN network is much larger than the other FN networks and has a higher clustering coefficient and average number of neighbours (58.15).



(a) Degree Distribution



(b) Shortest path distribution

Figure 6: (a)The degree distribution in the GC. (b)The shortest path distribution of the GC.

Using the values calculated by the NetworkAnalyzer tool, the average and mean values of degree, Betweenness centrality, Closeness centrality were calculated for the three most important subgroups in the GC. The degree values can be seen in Table 2 and the results for betweenness centrality and closeness centrality can be found in the same table.

Degree

	Mean	Median	Variance
DP	40	19	1468
PolyQ	32	13	2019
PolyQO	23	17	556
GC	34	15	2117

Betweeness

	Mean	Median	Variance
DP	1.41E-04	1.41E-04	1.35E-06
PolyQ	1.36E-04	1.41E-04	1.33E-06
PolyQO	1.00E-06	1.00E-06	1.65E-06
GC	1.44E-04	1.44E-04	1.85E-06

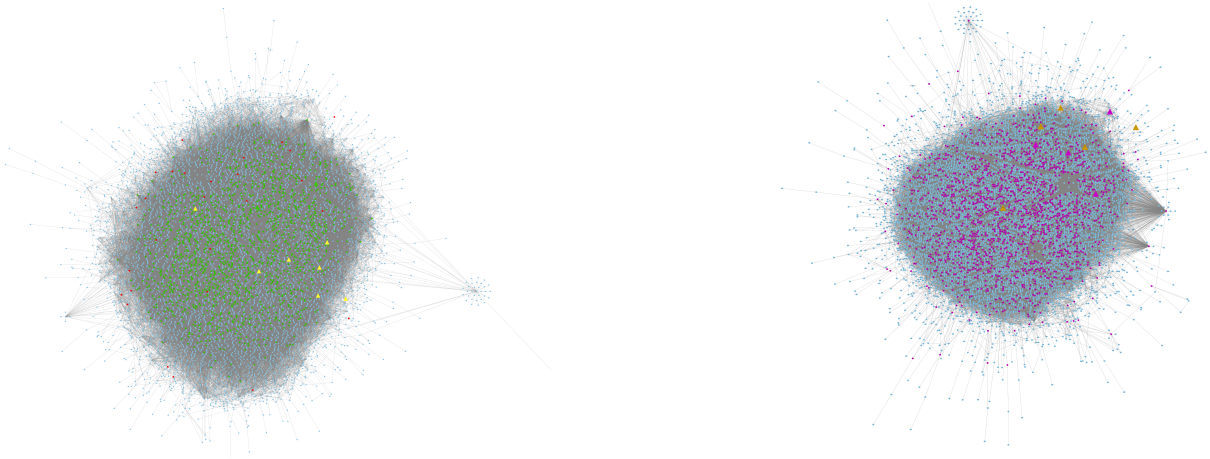
Closeness

	Mean	Median	Variance
DP	2.91E-01	2.91E-01	9.26E-04
PolyQ	2.79E-01	2.80E-01	1.43E-03
PolyQO	2.76E-01	2.76E-01	1.42E-03
GC	2.01E-01	2.01E-01	1.52E-03

Table 2: Median, mean and variance of the degree, betweenness- and closeness-centrality in three different subgroups of the GC.

To compare the different polyQ protein, disease protein and the polyQO, the mean and median of

the Degree, betweenness and closeness was calculated. The betweenness values can be found in Table 2. This table shows that the disease proteins have the highest mean and median betweenness centrality.
The Closeness centrality is shown in Table 2, this table shows again that the disease proteins are have the highest mean and median closeness centrality compared to the other functions. The order based on the mean stays the same between the two tables. The mean and median degree are shown in Table 2. In this graph the disease proteins have the highest mean and median. The polyQ protein have a lower median degree value then the mean value of the network.
Figure 7 shows the hub and bottleneck proteins in the network defined as the top 20% of the respective degree distribution and betweenness as described by [Yu et al., 2007]. The degree boundaries that approach 20% (1516 proteins) best are between 57 neighbours and 662 neighbours which leads to 1515 nodes. This definition of hub proteins makes it so that two of the disease proteins have a hub function, respectively Tbp and AR. For the betweenness centrality the boundary values were 4.045E-4 and 0.065, which resulted in 1516 nodes with 4 disease proteins. The disease proteins that can be marked as bottleneck proteins are: Ar, Tbp, Htt and Atxn7. Of the polyQ proteins 15 of them are hubproteins and 18 are bottleneck proteins. The polyQO have 2 hubproteins and 5 bottleneck proteins.



(a) Hubproteins (b) Bottleneck proteins

Figure 7: (a) Hub proteins in the GC. The green dots in the network represent hub proteins. The square formed nodes are disease proteins. (b) This graph shows the bottleneck proteins in the GC with betweenness centrality value between 4.1E-4 and 0.064.

To analyse the disease proteins, their different topological values were analysed in the GC and FN-networks. The network in the Figure shows how the disease proteins interact with each other in the GC shown in Figure 8. The Figure shows that most of the disease proteins are interconnected, only AR is not interacting with one of the disease proteins. The protein that is connected to the most disease proteins is Atxn7, it is connected to seven of the different disease proteins.

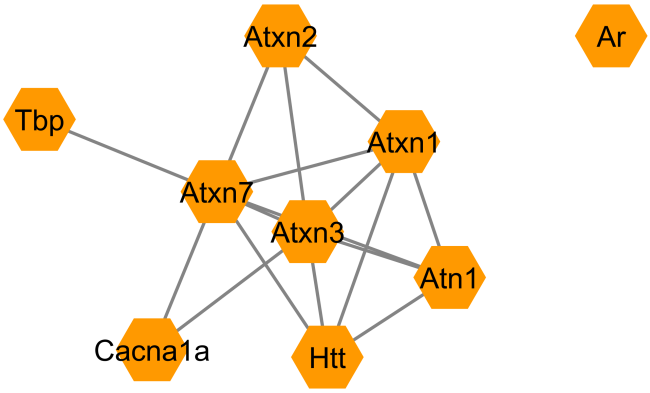


Figure 8: This network shows the interactions between the different disease proteins in the giant component.

In Table 3 the betweenness, closeness and degree of disease proteins in the GC are shown. The table shows that Tbp and Ar are the proteins with the highest degree among the disease proteins. Tbp has besides the high degree also the highest betweenness value. Ar, has the highest closeness value. Atn1, has the lowest betweenness, closeness and degree of all the disease proteins.

Ar	90	3.50E-03	3.57E-01
Atn1	5	5.58E-06	2.64E-01
Atxn1	10	2.97E-04	2.73E-01
Atxn2	19	2.05E-04	2.82E-01
Atxn3	14	1.54E-04	3.07E-01
Atxn7	17	6.80E-04	3.06E-01
Cacna1a	33	3.12E-04	2.95E-01
Htt	53	2.28E-03	3.45E-01
Tbp	122	1.93E-03	3.30E-01

Table 3: The 9 different disease proteins and their Betweenness centrality, closeness-centrality and degree value in the GC

The 9 different disease proteins were also analysed in their combined FN network. This network is shown in Figure 9. The FN networks of each of the different disease proteins are represented with their networks statistics in Table 10.

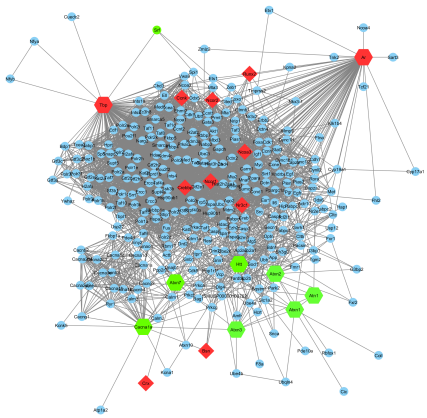


Figure 9: First neigbours network. The green nodes in the network are polyQO (8), The largest hexagons are disease proteins (9) and the red coloured nodes are polyQ protein (11).

The size of the FN-networks of course only differs by 1 from the degree values that the disease protein has in the GC. All the networks have a similar network diameter and network radius. The network with the highest mean degree (number of neighbours) is the Tbp network.
The chosen clusters: 4, 7, 8, 22, 64 and 73 were also analysed with NetworkAnalyzer, the results are in the appendix Table 12. Cluster 4 has 480 nodes, 7 has 273 nodes, 8 has 205 nodes, 22 has 47 nodes, 64 has 19 nodes and cluster 73 has 20 nodes.

3.4 Enrichment

The enrichment and analysis were done with two types of enrichment. Gene Ontology functions and Gene ontology process. The first enrichment was on the three different FN networks polyQ, disease protein and polyQO.
The GO_Process data is shown in the appendix Figure 10. The biggest term of the polyQFN network is the *positive regulation of cellular processes (43.3%)*. The biggest term of the Disease Protein FN is the *negative regulation of insulin-like growth factor receptor signalling pathway (44.3%)* and the biggest term of the polyQO network is *positive regulation of the cellular process, 39.2%*. The commonalities between these enriched networks are shown in appendix Table 8.

Go process term	**polyQ**	**PolyQO**	**DP**
positive regulation of the cellular process	X	X	
cellular process	X	X	
biosynthesis	X	X	
biological regulation	X	X	
multicellular organismal process	X	X	X
developmental process	X	X	

Table 4: GO process term enrichment for the FN networks of: polyQ, polyQO and DP. The X's show if a network is enriched with the according GO term. Only the terms that appeared in at least two networks are shown.

Between the polyQ proteins and the polyQO there are several overlaps. The overlap with the highest expression being: *positive regulation of the cellular process*. polyQ and polyQO have most of their processes in common. The disease protein FN only has *multicellular organismal process* in common with both of networks.

The GO function enrichment (shown in appendix, see Figure 14) resulted with *transcription regulatory region DNA binding (35.7%)* as the biggest term for the polyQO. *transcription factor binding, 34.5%* was the biggest for the polyQ protein network. *voltage-gated calcium channel activity (33.%)* was the biggest term for the disease protein. The commonalities between these enriched networks are shown in Table 9.

Go function term	**polyQ**	**PolyQO**	**DP**
transcription regulatory DNA binding	X	X	
transcription factor binding	X	X	
protein binding	X	X	X
transcription factor activity	X	X	
sequence specific DNA binding	X	X	
chromatin binding	X	X	
zinch ion binding	X	X	
transcription cofactor activity	X	X	
macromolecular complex binding	X	X	
binding	X	X	X
Volted-gated calcium channel activity		X	X

Table 5: GO function term enrichment of the FN networks for: polyQ, polyQO and DP. The X's in the table show which network is enriched with which term. Only the terms that appeared in at least two networks are shown.

The polyQ proteins and the polyQO have both in a large amount *transcription regulatory DNA binding* in common. The disease protein network does not have a lot in common with polyQ or polyQO.
After this enrichment, the subnetworks of the disease proteins were enriched to see if there are commonalities between them. The CirGO diagrams of this enrichment are shown in the appendix, see Figure 11 and Figure 12. The most represented GO terms in these networks:

- Atxn1: Negative regulation of insulin-like growth factor receptor signalling pathway (43.4%)
- Atxn2: Regulation of mRNA metabolism (68%)
- Atxn3: ERAD pathway (69.7%)
- Atxn7: Positive regulation of nucleic acid-templated transcription (32.8%)
- Atn1: Negative regulation of insulin-like growth factor receptor signaling pathway (36.4%)
- Htt: Regulation of cell death (49%)

- Cacna1a: Regulation of ion transmembrane transport (44%)
- Ar: Negative regulation of cellular process (54.2%)
- Tbp: Transcription from RNA polymerase II promotor (65.6%)

Some of the disease proteins (Atxn2, Atxn3, Ar and Tbp) have one GO-term that takes up more then 50% of their processes. The commonalities between these enriched networks are shown in Table 6.

GO process term	**Atxn1**	**Atxn2**	**Atxn3**	**Atxn7**	**Atn1**	**Htt**	**Cacna1a**	**Ar**	**Tbp**
oganic cyclic compound metabolism	X			X					
protein deubiquitination	X				X				
social behavour	X				X				
behavior	X		X	X	X	X	X		
nervous system development		X				X			
intracellular transport			X	X					
locomotory behaviour			X	X			X		
response to stimulus			X				X		
multi-organism process			X		X	X			
primary metabolism				X	X				
cellular process				X					X
localization					X	X	X		
multicellular organismal process							X	X	
developmental process						X		X	
negative regulation of insulin like growth factor signalling pathway	X				X				
neuron apoptotic process	X				X				

Table 6: This table shows the GO process terms from the disease protein FN networks. The network is enriched with the term when there is an X present. Only the terms that appeared in at least two networks are shown.

Besides the direct commonalities there are some terms that have to do with *nervous tissue development*, *synapse organization* or *brain tissue development* that all have involved with the brain as an organ. The proteins that have this as a term are: Htt, Cacna1a Atxn1 and Atxn2.

When enriching the networks with GO_function instead of process all networks were enriched except for the Atn1-FN network because the network is very small and does not contain enough information for enrichment analysis. The CirGO diagrams of this enrichment are shown in appendix Figure 15 and Figure 16. The most represented GO terms in these networks:

- Atxn1: protein binding (45.9%)
- Atxn2: mRNA binding (72.9%)
- Atxn3: BAT3 complex binding (64.9%)
- Atxn7: enzyme binding (29.6%)
- Htt: enzyme binding (46.7%)
- Cacna1a: voltage-gated calcium channel activity (33%)
- Ar: protein domain specific binding (38.7%)
- Tbp: transcription regulatory region binding (39.3%)

The commonalities between these enriched networks are shown in Table 7.

GO function term	**Atxn1**	**Atxn2**	**Atxn3**	**Atxn7**	**Htt**	**Cacna1a**	**Ar**	**Tbp**
protein binding	X	X	X	X	X	X		X
protein C-terminus binding	X	X						
binding	X	X	X	X	X	X	X	X
chromatin binding	X			X	X			X
enzyme binding				X		X		
transcription coactivator activity				X				X
drug binding						X	X	
macromolecular complex binding							X	X

Table 7: GO function terms for the disease protein FN networks. The X's in the table show if the GO term is present in the network. Only the terms that appeared in at least two networks are shown.

Table 7 shows that all of the disease proteins have *binding* as a common term. And that all except for Ar have *protein binding* as a common term.
The clusters 4, 7, 8, 22, 64 and 73 were also enriched with GO_process and GO_function. The CirGO diagrams of this enrichment are shown in the appendix Figure 13. The most represented GO process terms in these networks:

- Cluster 4: Negative regulation of macromolecule metabolism (56.8%)
- Cluster 7: Regulation of localization (44.6%)
- Cluster 8: Regulation of TOR signalling (63.1%)
- Cluster 22: Regulation of signal transduction (61.4%)

- Cluster 64: Negative regulation of cellular metabolism (53.1%)
- Cluster 73: Spliceosomal snRNP assembly (50.1%)

The commonalities between these enriched networks are shown in Table 8.

GO process term	**Cl. 4**	**Cl. 7**	**Cl. 8**	**Cl. 22**	**Cl. 64**	**Cl. 73**
Response to stimulus		X		X	X	
Cellular process	X	X	X	X	X	
primary metabolism	X			X		
nitrogen compound metabolism	X			X		
metabolism	X			X		
signaling		X		X		
cell communication		X		X		
biological regulation	X	X		X		
cellular pcomponent organization or biogenesis	X	X	X			

Table 8: GO process terms for the selected clusters. The network is enriched with the term when there is an X present. Only the terms that appeared in at least two networks are shown.

Table 8 shows that besides *cellular process* the clusters do not have any enrichment that is shared among more then half of them.
The CirGO diagrams of the GO function terms are shown in appendix Figure 17. The most represented GO function terms in these networks:

- Cluster 4: Transcription factor binding (31.0%)
- Cluster 7: Ephrin receptor binding (25.6%)
- Cluster 8: Microtubule binding (27.5%)
- Cluster 22: Enzyme binding (29.3%)
- Cluster 64: NADP-retinol dehydrogenase activity (26.9%)
- Cluster 73: Chaperone binding (44.5%)

The commonalities between these enriched networks are shown in Table 9.

GO function term	**Cl. 4**	**Cl. 7**	**Cl. 8**	**Cl. 22**	**Cl. 64**	**Cl. 73**
oxidoreductase activity			X		X	
catalytic activity		X			X	
protein binding	X	X	X	X		X
binding	X	X	X	X		X
transcription factor activity, sequence-specific DNA binding	X					X
macromolecular complex binding	X	X	X			X
chromatin binding		X	X	X		
transferase activity	X	X		X		
molecular transducer activity	X			X		
hydrolase activity		X	X			
zinc ion binding	X	X				
transcrip cofactor activity	X	X				

Table 9: GO function terms for the selected clusters. The network is enriched with the term when there is an X present. Only the terms that appeared in at least two networks are shown.

The clusters shown in Table 9 show that they just like, the disease protein networks, have *protein binding and binding* as a common term. The percentages from appendix Figure 17 show that only Cluster 8, 22 and 73 have *protein binding* as a highly ranked common term.

4 Discussion

The results of the topological analysis indicate that brain proteins are quite highly connected. In the entire network there is one giant component that contains almost 80% of all nodes. Besides the high amount of interactions, the component has a large amount of edges as well. The GC's degree distribution indicates a power-law just like the shortest path distribution shows an almost "small-world" effect which may indicate a "scale-free ness" of the GC. This strengthens the other topological results.

There are 81 polyQ proteins in the GC which is only a small amount of the proteins present in the GC. But beside this they do have a betweenness- and closeness-centrality that is on average higher than that of the GC. Which may indicate that they have a slightly more important role in the network. This indication is contradicted by the degree, where the polyQ have a lower median and average degree than that of the GC. The hubprotein and bottleneck protein results show that out of the 81 polyQ protein 15 (19%) are hubprotein and 18 (22%) are bottleneck proteins. These results strengthen the implication that the polyQ protein have an important role in the network.

The enrichment analysis of the polyQ FN network shows that the polyQ protein and their neighbours are active in a lot of different processes and functions, with the most common GO-process being the positive regulation of the cellular process and the most common GO-function being the transcription factor binding. A lot of the polyQ GO-functions show some form of binding function.

The polyQO are represented by 12 regular proteins and 7 DP in the GC. The topological analysis shows that their betweenness- and closeness-centrality are higher than the average of the network and the polyQ protein. Which may indicate that they have a slightly more important role in the network than the polyQ. Although the large proportion of disease proteins in this group may skew these findings (as disease proteins polyQ and PolyQO appear to be a distinct group). The mean degree shows the polyQO with a lower degree than the GC and the polyQ whereas the median shows the polyQO above the polyQ and the GC. Two (11%) of the polyQO are classified as a hubprotein and 5 (26%) as a bottleneckprotein. Which implicates the important role as bottleneck protein of the polyQO.

Enrichment of the polyQO indicates that most of the polyQO protein have process of positive regulation of cellular processes as the most common GO-process. The most common GO-function being transcription regulatory region DNA binding. A lot of the polyQO GO-functions show some form of binding function

Comparing the topological results of the polyQ and the polyQO shows that the polyQO have a higher betweenness and closeness than the polyQ. When considering the degree values, the difference should be based on the median values. The median values of the degree are too far off the mean values which indicates outliers in the data. The closeness, betweenness and degree (median) considered together implies that the polyQO have a more important function in the network than the polyQ. The large proportion of disease proteins in this group may affect these results. But a higher percentage of the polyQ protein can be labeled as hubprotein compared to polyQO. The polyQO have more bottleneckproteins
The enrichment results show that the polyQ and the polyQO have a lot of common terms. This over-

lap is over most of the enriched terms. The polyQ protein have 12 enriched main GO-functions, 10 of these are in common with the GO-functions of the polyQO. The GO-process terms have a smaller overlap with about 6 out of the 12 for polyQ and 6 out of the 10 for polyQO. These results suggest that the polyQO protein in the network have a similar function in brain tissue as polyQ protein.

The results of the disease proteins show that two (Ar and Tbp) out of the nine are classified as a polyQ protein, the other seven disease protein are polyQO. Figure 8 shows that all except Ar have some sort of interaction among each other, this close interaction is not seen more generally in polyQ proteins. The topological analysis of the disease proteins shows that their betweenness- and closeness-centrality measured by mean and median is much higher compared to the other subgroups. The DP also have the highest mean and median degree. This demonstrates the important role of the DP in the network.
According to the hub- and bottleneck protein analysis the DP have 2 hubproteins and 4 bottleneck proteins. The 2 hubproteins being the only polyQ proteins (Ar and Tbp). Which may suggest that the polyQ domain has a part in this.

The enrichment analysis shows that Htt is involved in the GO-process regulation of cell death and the GO-function enzyme binding. The indication that Htt is involved in membrane dynamics, cell attachment and motility as shown by [Schaefer et al., 2012] is not found in the enrichment of Htt. This may be because protein annotation has changed since 2012, there is also more knowledge abhout Htt PPI.
The FN network of all the DP together do not show an overlap with the polyQ or polyQO networks other than the GO-functions binding and protein binding. The GO-function term that was the most expressed term, voltage-gated calcium channel activity and the GO-process term is the negative regulation of insulin-like growth factor receptor signalling pathway. These enrichment results, taken together with the fact that the disease proteins are highly interconnected, indicate a more specific set of functions that are disrupted in polyQ diseases.
The enrichment results of the FN networks of the individual DP shows that there is only one GO-process shared by most of them and that is behaviour. Atxn1 and Atn1 also have a lot of GO-processes in common, which may indicate that they are closely related. As expected of brain proteins, the results show a lot of terms that are involved with the brain as an organ.
The GO-function results show that all of the DP have a binding role and all even have a protein-binding function. AR is not labeled as protein-binding, this is probably due to an error in the REVIGO or CirGO, protein domain specic binding, the function of Ar is a subterm of protein binding, which is a subterm of binding. This indicates the protein-binding function, that is described [Totzeck et al., 2017] to be a possible function of polyQ regions.
Another commonality between Atxn1, Atxn7, Htt and Tbp is chromatin binding which is a proposed connection between polyQ proteins as described by [Cohen-Carmon and Meshorer, 2012]. This conflicts with the claim of the paper because only Tbp is a polyQ protein out of these 4 proteins.

The clustering results show that there are not a lot of polyQ or disease proteins packed together in clusters or even present in clusters. There are only 3 clusters with DP and 12 clusters with polyQ with an average of 2 polyQ per clusters. Because there is no cluster with a lot of polyQ or polyQO proteins, the conclusion can be made that the polyQ and polyQO are dispersed in the network.
These reasons may indicate a less important role in protein-protein interactions for polyQ regions

than assumed so far. If the polyQ region would have a big influence on protein-protein interactions, a lot of clusters with polyQ protein would be expected.
The functional enrichment of clusters with polyQO or DP show that these clusters do have protein-binding and binding as a common GO-function term and Cellular process as a GO-process. Except for this the enrichment does not show big common terms.

The research interactions from StringDB are from different evidence sources and not only from experimentally sourced interactions. The evidence for the interactions can also be from ortholog proteins. This means that not every interaction shown in the network is sure to be in mouse brain proteins. It could even mean that annotation is compared about polyQs (in mouse) with other polyQs (in human). But due to the different evidence scores that are combined and the high cut-off score, the interactions can be used in for the predictions/indications made in this study. This is also true for the GO annotations. A further study where only mouse-derived annotation and no sequence similarity results may show some differences.
It is beyond the scope of this study to make a conclusion about the polyQ protein functions. The PPI network cannot provide direct evidence for individual polyQ protein functions, but patterns across the network indicate binding is important. The fact that the polyQOs also show binding functions indicate that it is not necessarily the polyQ domain that confers the binding function.

5 Conclusion

In order to investigate the research-statement several research-questions were proposed. By analysing common functional signatures in polyQ proteins, this thesis has found that polyQ share a lot of different GO-processes and GO-functions. The biggest GO-process being: positive regulation of cellular processes and GO-function: transcription factor binding.
The differences between inter-actors of mouse polyQ proteins and mouse orthologs of human polyQ proteins that are not polyQ proteins (polyQO) turned out to be small. The main difference being that polyQO have a more important role in the network based on their network topology. The enrichment results were almost identical for GO-function and very similar on GO-process.
To answer the third question polyQ disease protein enrichment were compared to ones not associated with disease. The enrichment results show that the DP really stand out from the rest of the protein. They do not have a lot of GO-functions or GO-processes in common. The GO-functions that they have in common have protein binding and binding as a common GO-function and multicellular organismal process as GO-process. These are to be expected from protein that have a possible protein binding function. This assumption that they have this role is strengthened by their topological analysis. The DP have a very high degree, closeness and betweeness value compared to the polyQ protein and the regular network protein.

This research aimed to find similarities and differences between polyQ protein-protein interactions in Mus *musculus* brain to find common inter-actor functions in normal and disease states using network analysis and functional enrichment of the network. This research has shown that the disease protein and regular polyQ protein are more connected than normal polyQ protein and have a very different enrichment profile.
The current state of polyQ research leaves a lot of possibilities to be researched. For example, this research could be redone by comparing the results to a network containing only experimentally confirmed mouse interactions accompanied by an experimental study that compares the interactions in the brain between the polyQO and brain proteins and polyQ proteins and brain proteins. This would make sure that the interactions found are all real mouse interactions and not from orthologs.

Another good addition to the study of polyQ proteins could be a comparative study that compares polyQ proteins in different PPI networks across different species. This could lead to new insights in the functions and processes that polyQ proteins are involved in. To add to this information it could be useful to compare human polyQO and polyQ protein functions and processes in the other species. This researches could provide extra data for the results found in this paper.

6 Acknowledgements

Many thanks go to Katy Wolstencroft for the helpful advice and expertise she provided for this project as well as to Frank Takes for his advice on how to work with the topological side of networks.

References

Réka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, 2005. ISSN 00219533. doi: 10.1242/jcs.02714.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, and Gavin Rubin, Gerald M. Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.Gene.

Gary D. Bader and Christopher W.V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:1–27, 2003. ISSN 14712105. doi: 10.1186/1471-2105-4-2.

Sourav S. Bhowmick and Boon Siew Seah. Clustering and Summarizing Protein-Protein Interaction Networks: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):638–658, 2016. ISSN 10414347. doi: 10.1109/TKDE.2015.2492559.

Dorit Cohen-Carmon and Eran Meshorer. Polyglutamine (polyQ) disorders. *Nucleus*, 3(5):433–441, 2012. ISSN 1949-1034. doi: 10.4161/nucl.21481.

The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019. ISSN 13624962. doi: 10.1093/nar/gky1055.

Nadezhda T. Doncheva, John H. Morris, Jan Gorodkin, and Lars J. Jensen. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *Journal of Proteome Research*, 18(2): 623–632, 2019. ISSN 15353907. doi: 10.1021/acs.jproteome.8b00702.

Jun Dong and Steve Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1:1–20, 2007. ISSN 17520509. doi: 10.1186/1752-0509-1-24.

Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, 2011:1–9, 2011. ISSN 17580463. doi: 10.1093/database/bar030.

Chuan Lin, Young Rae Cho, Woo Chang Hwang, Pengjun Pei, and Aidong Zhang. Clustering Methods in a Protein-Protein Interaction Network. *Knowledge Discovery in Bioinformatics: Techniques, Methods, and Applications*, pages 319–355, 2007. doi: 10.1002/9780470124642.ch16.

T.L. Madden, R.L. Tatusov, and J. Zhang. Applications of network BLAST server. *Computer methods for macromolecular sequence analysis*, 266:131–141, 1996.

P. McColgan and S. J. Tabrizi. Huntington's disease: a clinical review. *European Journal of Neurology*, 25(1):24–34, 2018. ISSN 14681331. doi: 10.1111/ene.13413.

H. Motenko, S. B. Neuhauser, M. O'Keefe, and J. E. Richardson. MouseMine: a new data warehouse for MGI. *Mammalian Genome*, 26(7-8):325–330, 2015. ISSN 14321777. doi: 10.1007/s00335-015-9573-z.

Irene M.A. Nooren and Janet M. Thornton. Diversity of protein-protein interactions. *EMBO Journal*, 22(14):3486–3492, 2003. ISSN 02614189. doi: 10.1093/emboj/cdg359.

Amy L. Robertson, Mark A. Bate, Steve G. Androulakis, Stephen P. Bottomley, and Ashley M. Buckle. PolyQ: A database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Research*, 39(SUPPL. 1), 2011. ISSN 03051048. doi: 10.1093/nar/gkq1100.

Martin H. Schaefer, Erich E. Wanker, and Miguel A. Andrade-Navarro. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Research*, 40 (10):4273–4287, 2012. ISSN 03051048. doi: 10.1093/nar/gks011.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models. *Genome Research*, 13(22):426, 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.metabolite.

Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0021800.

Franziska Totzeck, Miguel A. Andrade-Navarro, and Pablo Mier. The protein structure context of polyQ regions. *PLoS ONE*, 12(1):2–11, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0170801.

Cendrine Tourette, Biao Li, Russell Bell, Shannon O'Hare, Linda S. Kaltenbach, Sean D. Mooney, and Robert E. Hughes. A large scale huntingtin protein interaction network implicates RHO GTPase signaling pathways in huntington disease. *Journal of Biological Chemistry*, 289(10): 6709–6726, 2014. ISSN 1083351X. doi: 10.1074/jbc.M113.523696.

Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(DATABASE ISS.):433–437, 2005. ISSN 03051048. doi: 10.1093/nar/gki005.

Haiyuan Yu, Philip M. Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):713–720, 2007. ISSN 1553734X. doi: 10.1371/journal.pcbi.0030059.

YWorks. yFiles Layout Algorithms for Cytoscape, 2020. URL https://www.yworks.com/products/yfiles-layout-algorithms-for-cytoscape.

7 Appendix

	Tbp	Ar	Cacna1a	Htt	Atn1	Atxn1	Atxn2	Atxn3	Atxn7
#Nodes	123	91	34	54	6	11	20	15	18
clustering coefficient	0.75	0.64	0.81	0.70	0.77	0.43	0.62	0.72	0.81
Diameter	2	2	2	2	2	2	2	2	2
Radius	1	1	1	1	1	1	1	1	1
Centralization	0.73	0.85	0.64	0.89	0.40	0.82	0.90	0.78	0.77
Shortest paths	15006	8190	1122	2862	30	110	380	210	306
Path length	1.72	1.83	1.60	1.68	1.27	1.67	1.81	1.68	1.68
avg. #neighbours	34.20	15.32	13.06	7.44	3.67	3.27	3.70	4.53	5.44
Network Density	0.28	0.17	0.40	0.14	0.73	0.33	0.20	0.32	0.32
Heterogenity	0.63	0.78	0.65	0.96	0.34	0.82	1.03	0.68	0.65

Table 10: Results of the topological analysis from the first-neighbour networks of the individual disease proteins

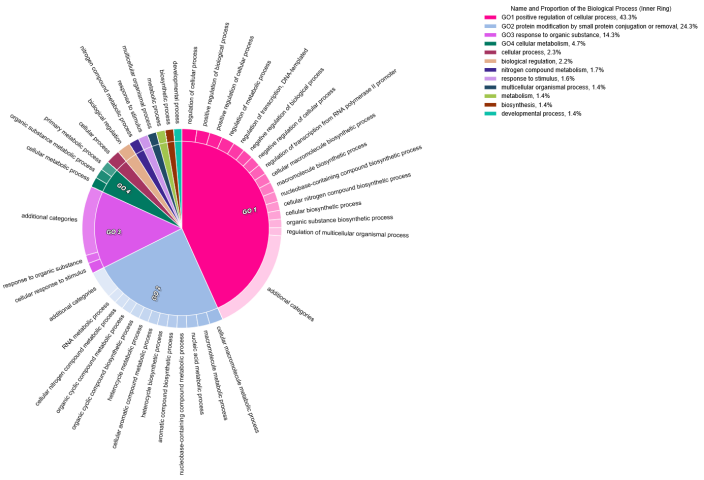
	Network	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 7	Cl. 8	Cl. 22
Number of Nodes	7582	272	231	480	291	273	205	47
Disease Protein	9	0	0	1	0	0	2	0
polyQ	81	1	1	9	4	3	4	1
Ortholog	53	0	0	3	1	1	4	1
Cl. ing coefficient	0.45	0.97	0.93	0.85	0.86	0.85	0.87	0.82
Diameter	11	4	4	6	8	12	12	11
Radius	6	2	2	3	5	6	6	6
Centralization	0.08	0.26	0.22	0.11	0.09	0.08	0.10	0.19
Shortest paths	57479142	73712	53130	229920	84390	74256	41820	2153
Path length	3.71	2.27	2.34	3.26	3.80	5.26	4.89	4.88
Avg. # neighbours	33.98	70.40	42.04	36.83	20.82	13.12	11.80	6.72
Network Density	0.00	0.26	0.18	0.08	0.07	0.05	0.06	0.15
Heterogenity	1.35	0.17	0.26	0.31	0.57	0.49	0.43	0.65

	Cl. 24	Cl. 25	Cl. 41	Cl. 43	Cl. 47	Cl. 64	Cl. 73	Cl. 77
Number of Nodes	118	33	7	103	70	19	20	16
Disease Protein	0	0	0	0	0	3	0	0
polyQ	1	1	1	1	1	0	1	1
Ortholog	0	0	0	0	1	0	1	1
Cl. ing coefficient	0.67	0.76	0.83	0.71	0.71	0.69	0.70	0.50
Diameter	15	7	3	19	15	8	7	8
Radius	8	4	2	10	8	4	4	4
Centralization	0.07	0.12	0.23	0.07	0.07	0.20	0.20	0.20
Shortest paths	13806	1056	42	10506	4830	342	380	240
Path length	6.76	3.46	1.38	7.97	6.82	3.49	3.24	3.70
Avg. # neighbours	6.02	5.46	4.00	4.51	4.06	3.79	3.60	3.38
Network Density	0.05	0.17	0.67	0.04	0.06	0.21	0.19	0.23
Heterogenity	0.52	0.39	0.33	0.38	0.43	0.44	0.44	0.47

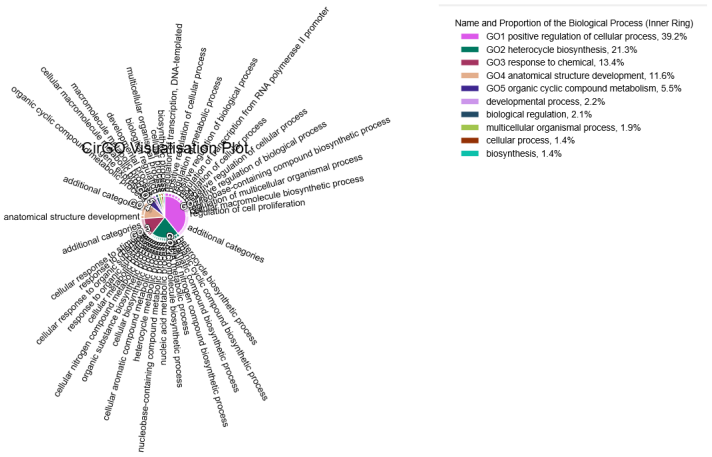
Table 11: The different generated clusters and their topological values. The values of the Network (giant component) are included for comparison

Cl.	Dis. Prot.	polyQ prot.	Orthologs
Cl. 2	-	Aak1	-
Cl. 3	-	Numa1	-
Cl. 4	Tbp	Tbp, Ncoa6, Kmt2c, Ncor1, Rpa1, Crebbp, Ccnk, Ncoa3, Kmt2d	Kmt2d, Crebbp, Ncoa
Cl. 5	-	Med15, Ncor2, Phf21a, Med12	med15
Cl. 7	-	Arid1a, Phc1, Smarca2	arid1b
Cl. 8	Cacna1a, Ar	Lin7a, Nr3c1, Mef2a, Ar	Ar, Mef2a, Cacna1a, Lin7a
Cl. 22	-	Kat6a	Mga
Cl. 24	-	Clock	-
Cl. 25	-	Sry	-
Cl. 41	-	Mkl1	-
Cl. 43	-	Gls	-
Cl. 47	-	Amotl1	Amotl1
Cl. 64	Atxn3, Atxn1, Atn1		-
Cl. 73	-	Abcf1	Pou5f1
Cl. 77	-	Cdx2	Cdx2

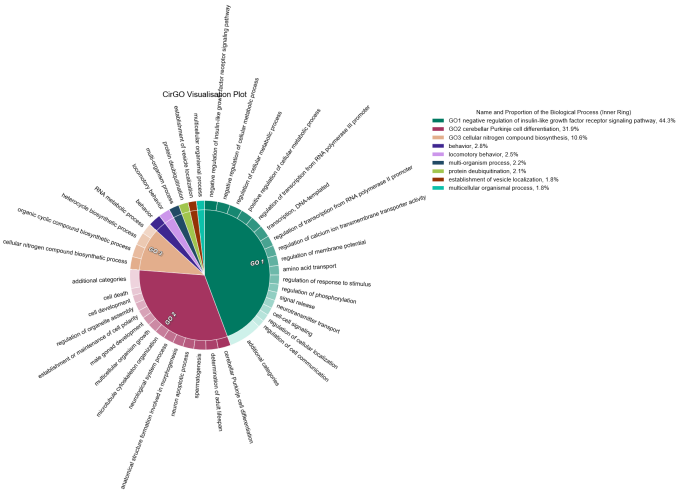
Table 12: The different disease proteins/orthologs that are present in the clusters.



(a) polyQ

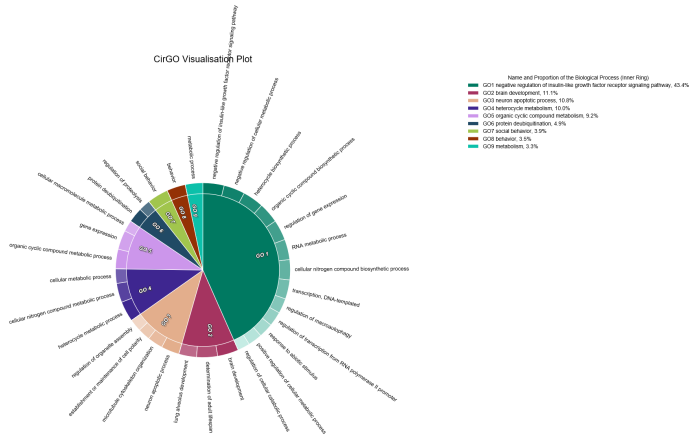


(b) PolyQO

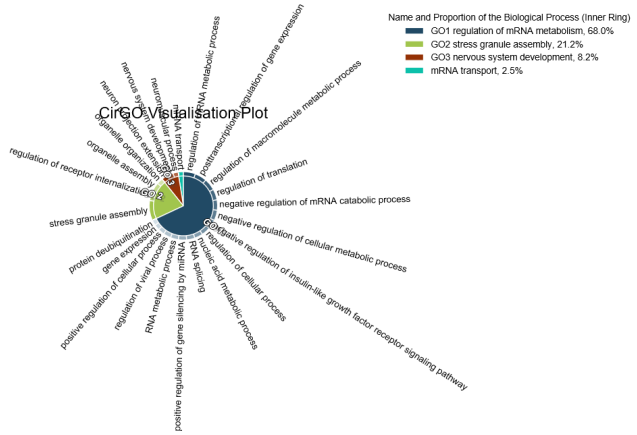


(c) Disease Proteins

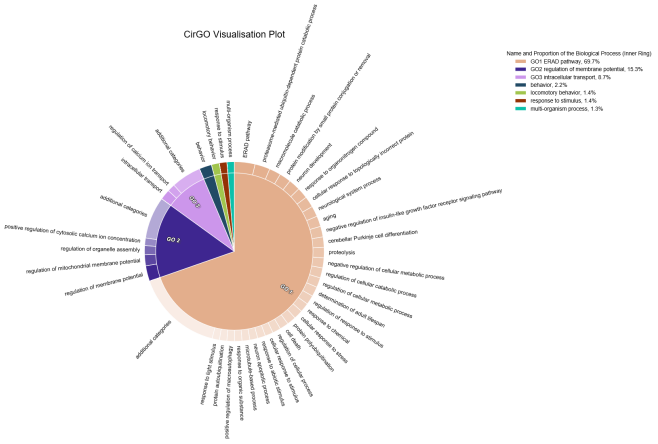
Figure 10: The CirGO GO_process enrichment of the FN networks for the DP, polyQ and polyQO



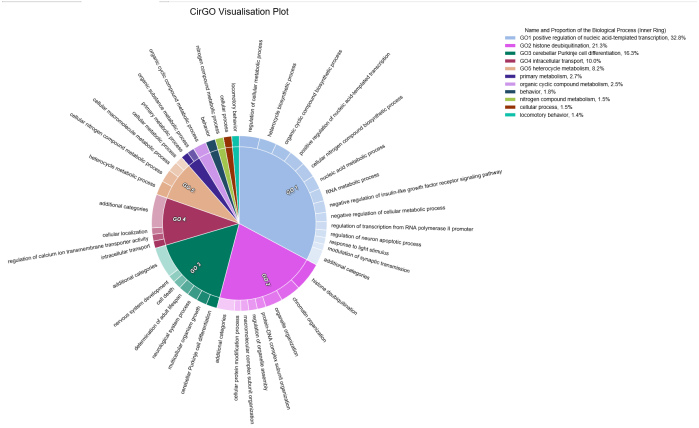
(a) Atxn1



(b) Atxn2

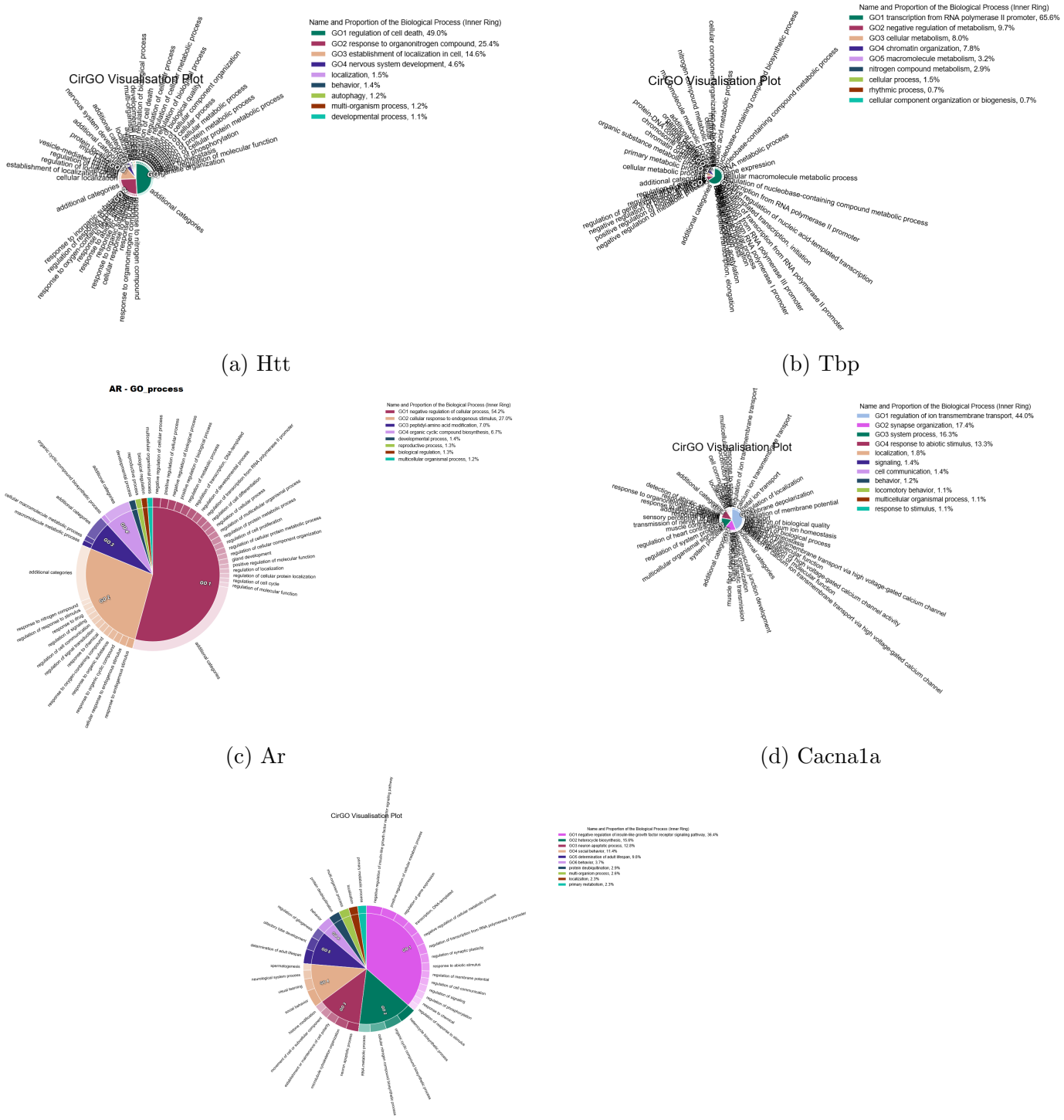


(c) Atxn3



(d) Atxn7

Figure 11: The CirGO GO_process enrichment of 4 of the different disease protein's first neighbour networks.



(a) Htt

(b) Tbp

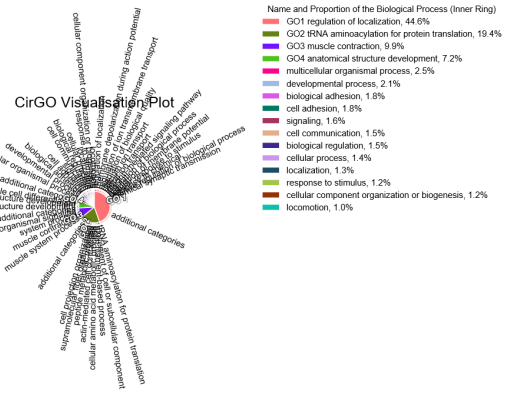
(c) Ar

(d) Cacna1a

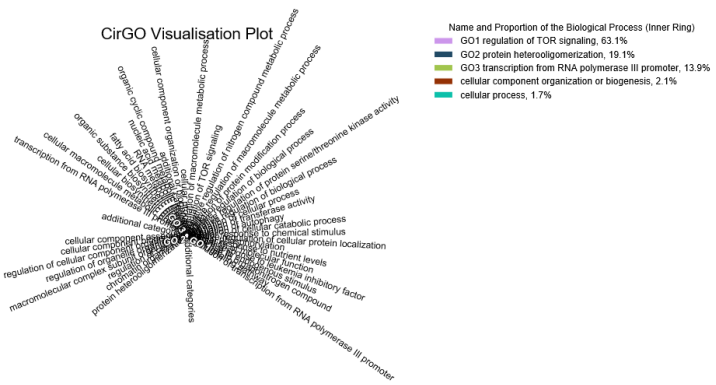
(e) Atn1

Figure 12: The CirGO GO_process enrichment of 5 of the different disease protein's first neighbour networks.

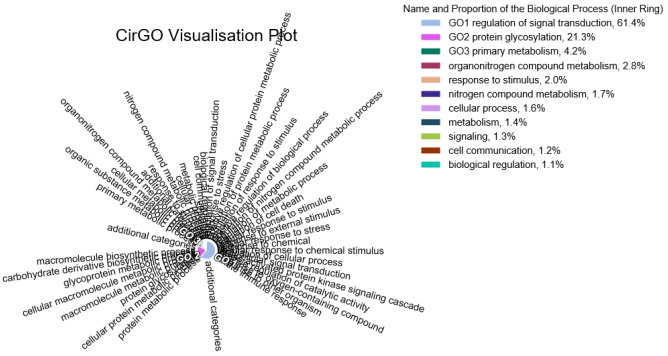
(a) Cluster 4



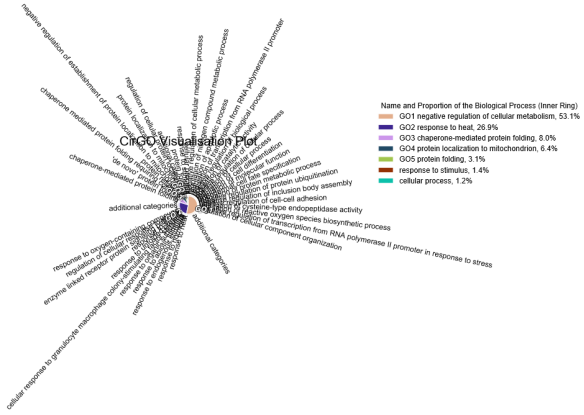
(b) Cluster 7



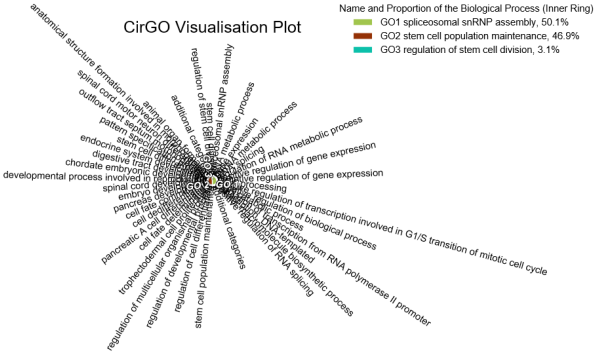
(c) Cluster 8



(d) Cluster 22

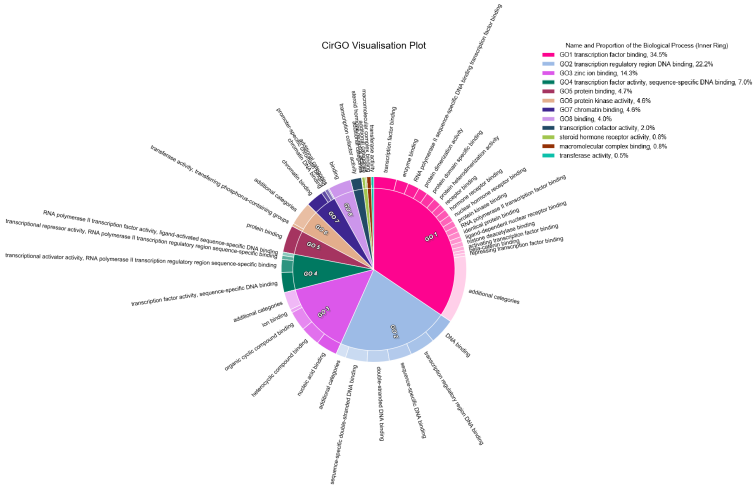


(e) Cluster 64

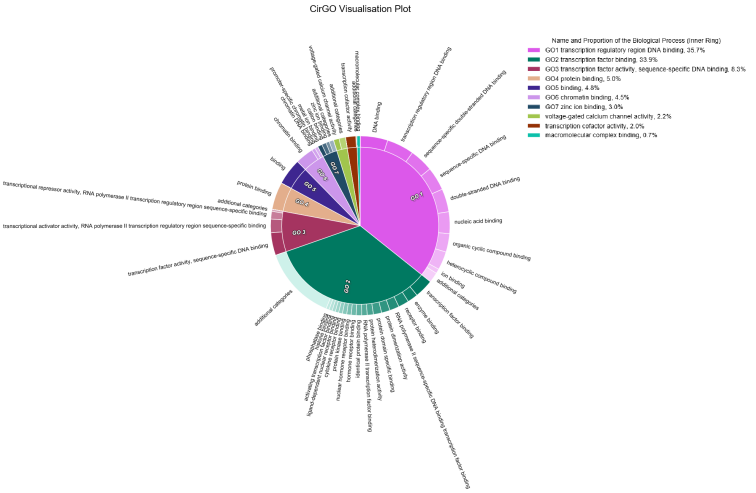


(f) Cluster 73

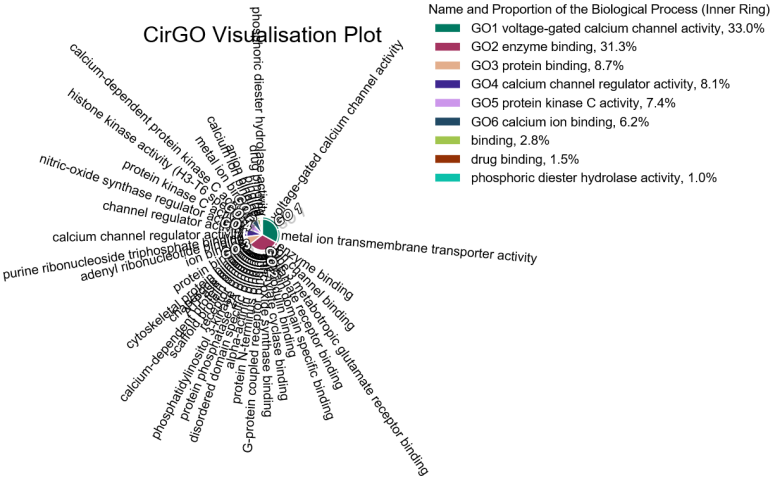
Figure 13: The CirGO GO_process enrichment of the clusters



(a) polyQ

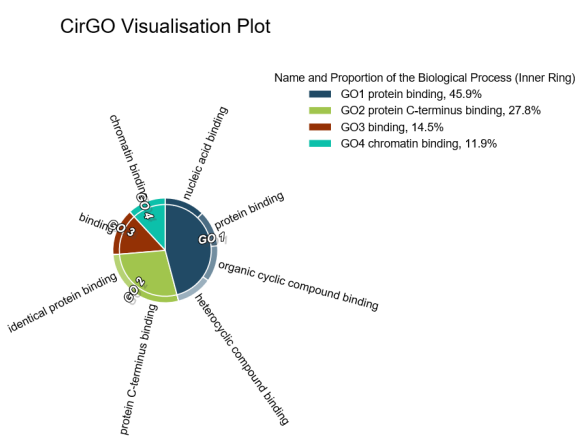


(b) PolyQO

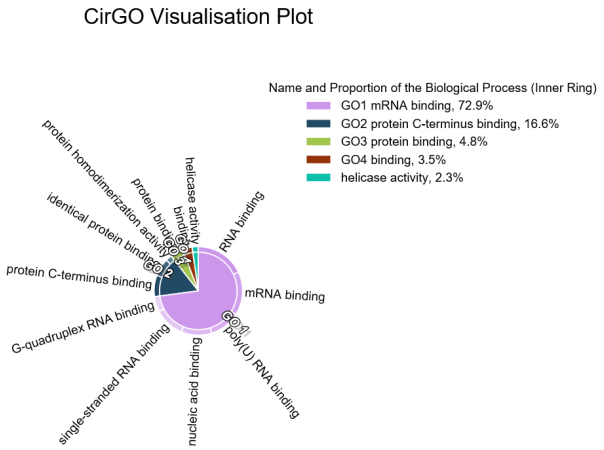


(c) Disease Proteins

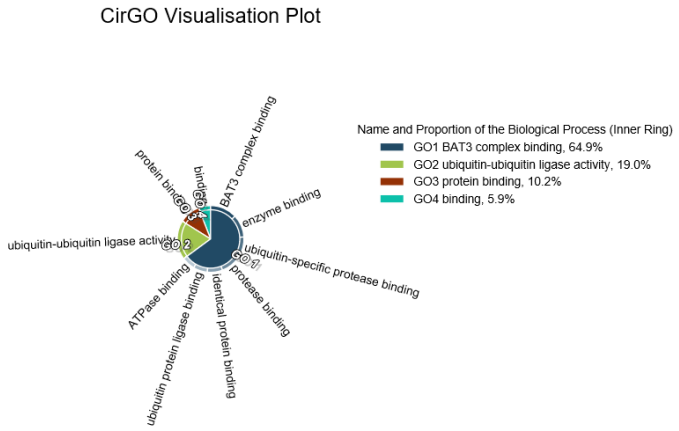
Figure 14: The CirGO GO_function enrichment of the FN networks for the DP, polyQ and polyQO



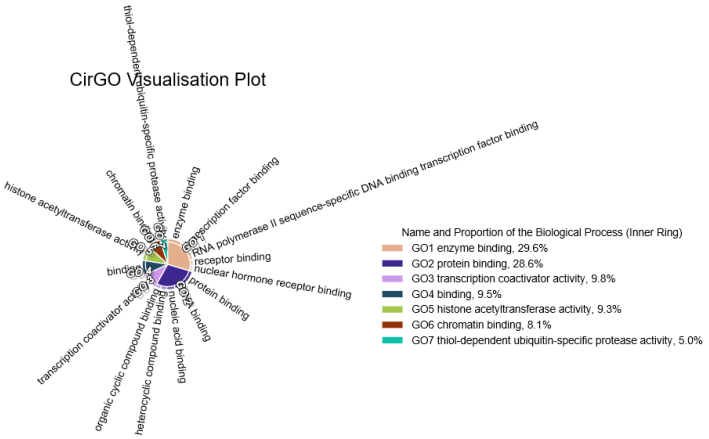
(a) Atxn1



(b) Atxn2

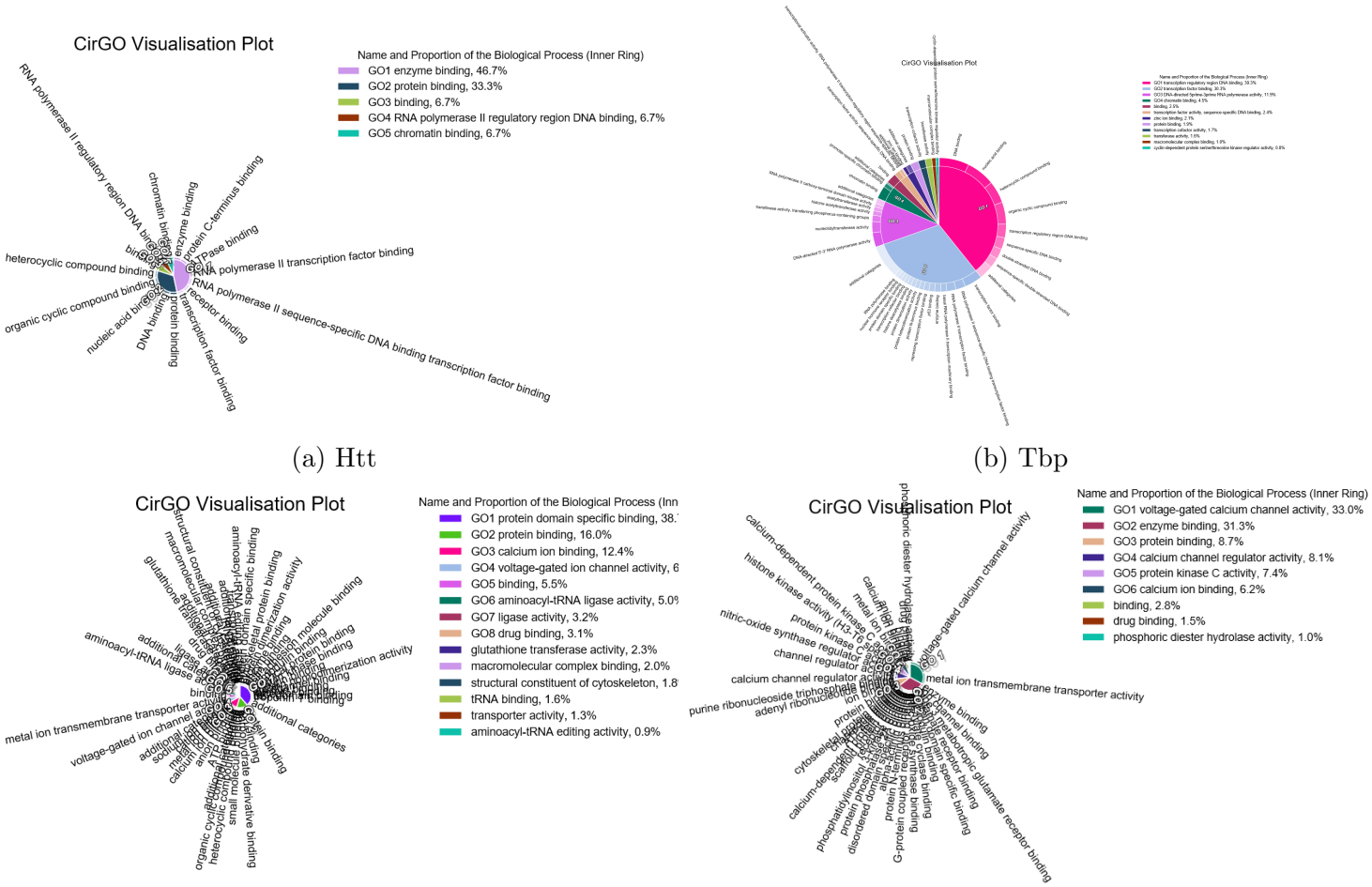


(c) Atxn3



(d) Atxn7

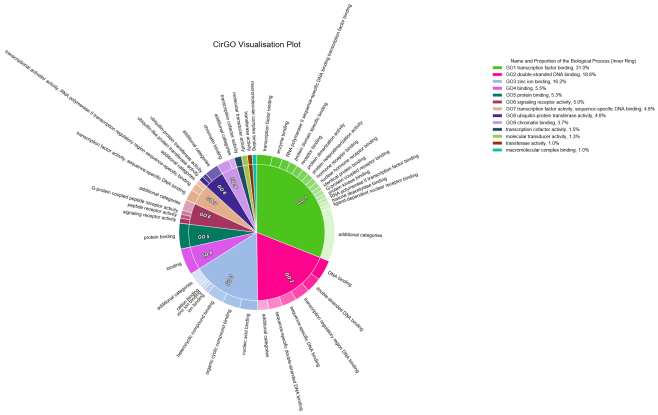
Figure 15: The CirGO GO_function enrichment of 4 of the different disease protein's first neighbour networks.



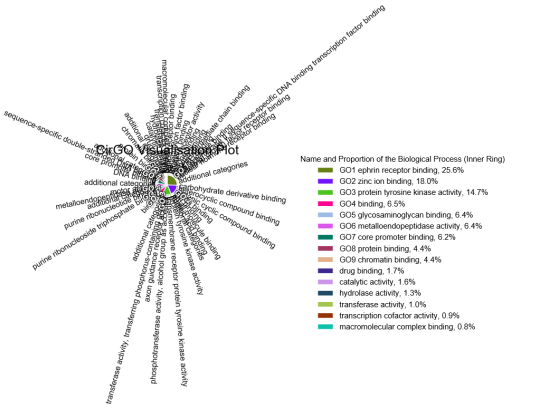
(a) Htt (b) Tbp

(c) Ar (d) Cacna1a

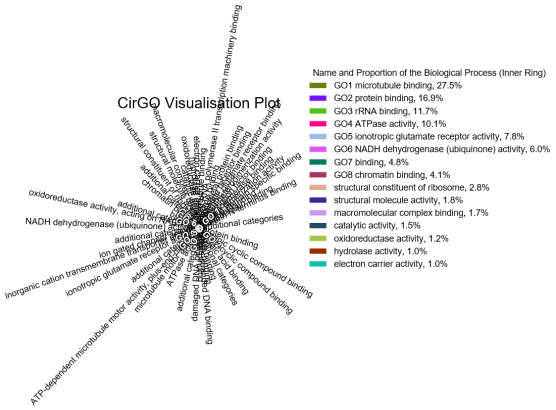
Figure 16: Thhe GO_function enrichment of 4 of the different disease protein's first neighbour networks. Atn1 is not shown in this enrichment, because the FN network of Atn1 did not contain any GO_function.



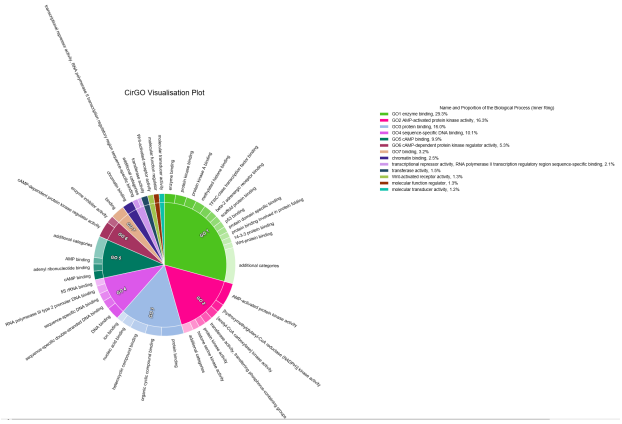
(a) Cluster 4



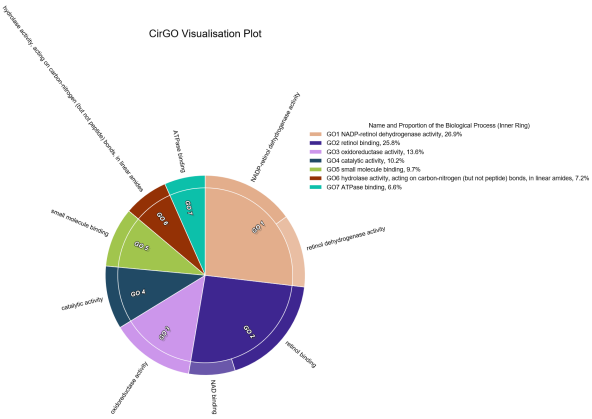
(b) Cluster 7



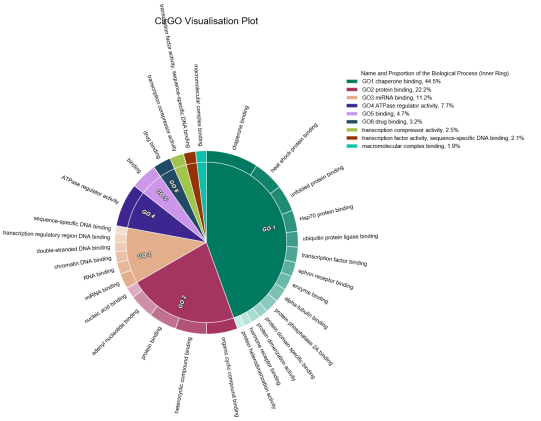
(c) Cluster 8



(d) Cluster 22



(e) Cluster 64



(f) Cluster 73

Figure 17: The CirGO GO_function enrichment of the clusters